Audio-Visual Database for Spanish-Based Speech Recognition Systems

Diana-Margarita Córdova-Esparza[1(✉)], Juan Terven[2], Alejandro Romero[1], and Ana Marcela Herrera-Navarro[1]

[1] Facultad de Informática, UAQ, Universidad Autónoma de Querétaro, Av. de las Ciencias S/N, Campus Juriquilla, 76230 Querétaro, Mexico
diana_mce@hotmail.com, jarg.romero25@gmail.com, anaherreranavarro@gmail.com
[2] AiFi Inc., 2388 Walsh Av., Santa Clara, CA 95051, USA
j.r.terven@ieee.org

Abstract. Automatic speech recognition involves an understanding of what is being said. It can be audio-based, visual-based, or audio/visual-based according to the type of inputs. Modern speech recognition systems are based on machine learning techniques, such as deep learning. Deep learning systems improve their performance when more data are used to train them. Therefore, data has become one of the most valuable assets in the field of Artificial Intelligence. In this work, we present a methodology to create a database for audio/visual speech recognition. Due to the lack of Spanish datasets, we created a comprehensive Spanish-based speech recognition dataset. For this, we selected hundreds of YouTube videos, found the facial features, and aligned the voice beside text with millisecond accuracy using IBM speech-to-text technology. We split the data into three speaker face angles, where the frontal angle represents the simple case, and right-left angles represent harder cases. As a result, we obtained a dataset of more than 100 thousand samples consisting of a small video with its respective annotation. Our approach can be used to generate datasets on any language by merely selecting videos in the desired language. The database and the source code to create it are open-source.

Keywords: Database · Speech recognition · Machine learning

1 Introduction

Visual speech recognition is the ability to recognize speech from visual information only [3]. Audio-visual speech recognition combines acoustic and visual information to recognize what is being said. Applications of this technology include dictation systems, transcription of silent films, assistive technologies for people with hearing disabilities, and solving multi-speaker simultaneous speech [3,4].

In this work, we propose a methodology to semi-automatically create databases for training speech recognition systems based on either visual or audio-only inputs or audio-visual inputs. The database can be created on any language

L. Martínez-Villaseñor et al. (Eds.): MICAI 2019, LNAI 11835, pp. 452–460, 2019.
https://doi.org/10.1007/978-3-030-33749-0_36

by changing the input data and the language of the text-to-speech engine. To test the system usability and contribute to the field, we created an extensive Audio-visual Spanish-based dataset with more than 100,000 samples containing the video with annotated spoken text and 3D mouth features. We also split the data into three face views: frontal view, angle view, and side view.

The source of the videos is YouTube, from where we select videoblogs, TED talks, news, and others. Each sample lasts from one to six seconds and has one to five spoken words with their respective annotation. For storage and copyright issues, we only provide the annotation of the videos with their respective timestamp (start and end of the sample) and the corresponding video link.

1.1 Previous Work

In the literature, we can find several databases that contain annotated images, which are useful for training and testing machine learning algorithms for Computer Vision. The main objective of Computer Vision is the analysis and understanding of a real-world scene captured in an image using a camera. Typical tasks include detection and recognition of objects present in an image, determine what attributes these objects have, and provide a semantic description of the scene. With the advent of public standardize datasets, researchers were able to compare their results on a common benchmark. In the following, we will describe commonly used datasets for computer vision tasks, and we will end with audiovisual speech recognition datasets.

Object recognition is one of the most common tasks in Computer Vision, and it was the one that started the deep learning paradigm -now known as *modern computer vision*- with the requirements of large datasets. Caltech 101 is based on Google images from 101 categories. With approximately 50 images per category [7]. Caltech 256, contains 256 object categories with a total of 30607 images [8], this database represented an improvement for Caltech 101, due to the increase in the number of categories and images for each of them. ImageNet was the stepping stone in large scale benchmark databases used by academia and industry. ImageNet in its current state consists of 12 subtrees with 5247 sets of synonyms (synsets) and 14 million images in total, its main applications are recognition, classification and automatic grouping of objects [5]. Another database focused on recognition is PASCAL visual object classes (VOC), used as a reference in the recognition and detection of categories of visual objects. It provides the user with a set of standard image and annotation data used for machine learning [6].

The Scene Understanding (SUN) database is also used for the recognition of real-world scenes. It contains 899 categories and 130,519 images that show different human activities such as: eating in a restaurant, reading in a library, sleeping, etc. The scenes and their associated functions are closely related to the visual characteristics that structure the space [10].

Microsoft Common Objects in COntext (MS COCO) is a popular database used for detection, recognition, segmentation and human pose estimation, contains 91 categories of everyday objects, 82 of which have more than 5,000

instances labeled. Unlike ImageNet, COCO has fewer categories but more instances in each category, which helps in learning detailed object models [9].

Regarding speech recognition, the work in [4] describes a method to generate a large-scale database from television transmissions automatically. With this approach, they generated the Lip Reading in the Wild (LRW) database with more than one million instances of words, spoken by more than a thousand different people. Lip Reading Sentences in the Wild (LRS2) [3] and Lip Reading in Profile (LRS3) [1] are audiovisual databases also used for speech recognition. LRS2 consists of thousands of sentences obtained from the British Broadcasting Corporation (BCC) and each sentence is up to 100 characters long. LRS3 is a multimodal database for the recognition of audiovisual speech. It includes facial tracks of more than 400 h of TED and TEDx videos, along with their corresponding subtitles.

Inspired by their work, we developed a method to create audiovisual speech recognition datasets for any language automatically. Given the lack of these type of datasets for the Spanish language, we created the first large scale audio-visual speech recognition dataset for Spanish-speaking applications.

The contributions of this work are the following:

1. An extensive database that can be used to train automatic speech recognition systems.
2. The methodology and the free access source code that can be used to generate databases in any other language.

2 Method

This section describes our audio-vision speech recognition dataset generation system. Using this method, we collected thousands of samples covering a vocabulary of more than 10,000 different Spanish words. The procedure is shown in Fig. 1 and consists of the following steps: (1) Select videos, download them, and extract audio; (2) for each audio extract the speech along with timestamps of start and end for each word; (3) crop the speech in samples of no more than five words; (4) determine the video frames involved on each sample and detect the 3D mouth features and face orientation; (5) save the results in file.

In the following subsections, we describe each of these steps with more detail.

2.1 Select Videos and Extract Audio

In this step, we selected Spanish spoken videos from YouTube covering several categories such as TED talks, video blogs, and news. We downloaded the videos an extracted the audio using the following *ffmpeg* command:

$$ffmpeg - ivideo.mp4 - acodecpcm_s16le - ac1 - araudio.wav \quad (1)$$

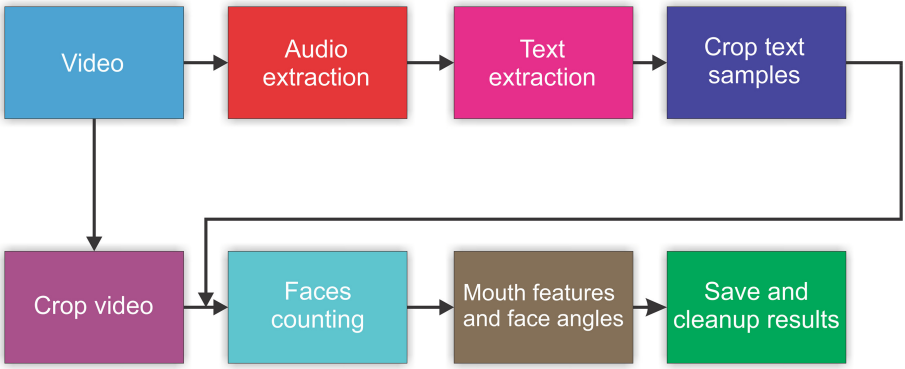


Fig. 1. Procedure to generate the audio-visual speech recognition dataset.

2.2 Timestamped Speech-to-Text

Using IBM's speech-to-text engine, we extracted each word from the audio files including its recognition confidence level and start-end timestamps with millisecond accuracy. To accomplish this, we use *SpeechToTextV1* class from *watson developer cloud* Python library.

2.3 Crop Samples

From the speech-to-text results, we cropped the data in samples containing sentences of one to five words to provide context for homophones words in the training data. Homophones words are identical visually, but its meaning is different, for example, the Spanish words *as*, *has*, *haz*. By providing sentences of multiple words, the machine learning models can learn to disambiguate between these words. We omit samples with single words whose length is less than four letters due to ambiguous recognition.

2.4 Selecting Video Frames and Detecting 3D Mouth Features with Face Orientation

Using the speech-to-text timestamps, we calculate the starting video frame and the ending video frame for each data sample using Eqs. (2) and (3):

$$frame1 = s_{min} \times fps \times 60 + (s_{sec} \times fps) \tag{2}$$

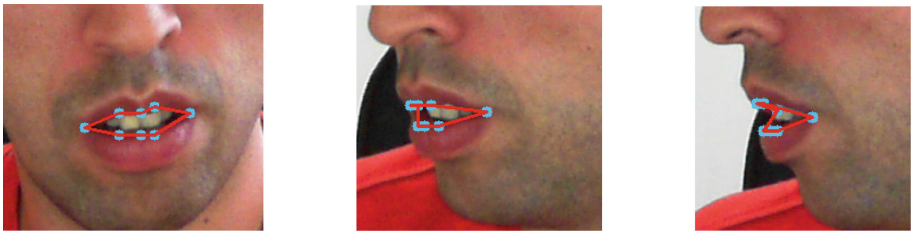
$$frame2 = e_{min} \times fps \times 60 + (e_{sec} \times fps), \tag{3}$$

where s_{min}, s_{sec}, e_{min}, and e_{sec} represent the initial minute, initial second, final minute and final second respectively.

Using these frames numbers, we select that portion of the video and run the 3D face alignment algorithm from [2] to determine if there is a face in the frame and obtain the mouth 3D landmarks. If a frame in the video does not contain a single face, we skip to the next sample. Once we determine that there is a face in the frame, we calculate the face orientation (yaw angle). To do this, we generate a unit 3D vector d using only two facial features: the tip of the nose and the middle point between the eyes. The yaw face angle is calculated with Eq. 4.

$$yaw = atan2(d.x, d.z) \tag{4}$$

We split the samples in easy, medium, and hard samples, according to the face angle. *Easy* samples consists of faces with angles in the range of 0 to 35°. *Medium* samples consist of faces with angles in the range of 36 to 65°. *Hard* samples include faces with angles higher than 65°. Figure 2 shows an image with the mouth landmarks and face angle.



(a) Easy (b) Medium (c) Hard

Fig. 2. Face angles. (a) Shows an example of an *Easy* sample with frontal face. (b) Shows an example of a *Medium* sample with face angle in the range of 36 to 65°. (c) Shows a *Hard* sample with a side view (faces with angles greater than 65°).

After filtering the samples without faces, we save the results in a CSV file with the fields *Link*, *Text*, *Confidence*, *Start*, *End*, *Mouth features*, and *Difficulty*. Table 1 describes the content of each of these fields.

2.5 Cleanup Results

The generation of the dataset is automatic, however, the results are not perfect. We performed a manual cleanup step where we check each sample and determine if we leave it or remove it from the final results.

3 Results

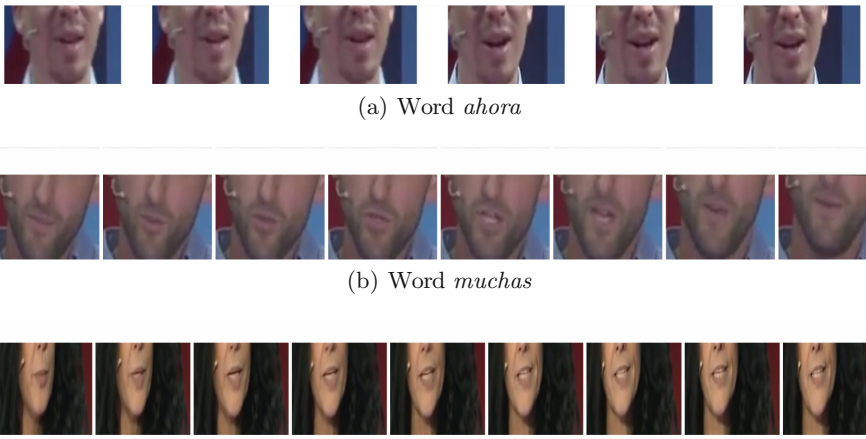
In total, we annotated hundreds of YouTube videos, creating a corpus of more than 100,000 samples with 32342 different words. Our dataset contains the video link, the text, starting and ending timestamp in seconds with respect to the start

Table 1. Data annotations.

Annotation	Description
Link	YouTube video link without www.youtube.com/
Text	Spoken text in the video sample
Confidence	Average speech recognition score in the range $[0, 1]$
Start, end	Sample start and end timestamp in seconds
Mouth features	Eight 3D mouth features covering the inner lips
Difficulty	Easy, medium or hard, according to the faces angles

of the video, the 3D mouth features with eight features surrounding the inner lip and the sample difficulty. The current version of the dataset contains 34061 easy samples, 20574 medium samples, and 12896 hard samples.

Figure 3 shows the visual sequences for the words *ahora*, *muchas*, and *parece*.



(a) Word *ahora*

(b) Word *muchas*

(c) Word *parece*

Fig. 3. Sample sequences. (a) Shows the visual sequence for the word *ahora*. (b) Shows the visual sequence for the word *muchas*, (c) Shows the visual sequence for the word *parece*

Figure 4 shows the ten most frequent words in the dataset. As described in Sect. 2.3, each sample contains from one to five words. Figure 5 shows the number of samples for each of these cases.

The dataset can be downloaded from the following address http://www.lipreadingdata.com. To ensure reproducible results, we open source the code and it is available at https://github.com/jrterven/audio-visual-dataset.

Fig. 4. Word count distribution for the ten most frequent words.

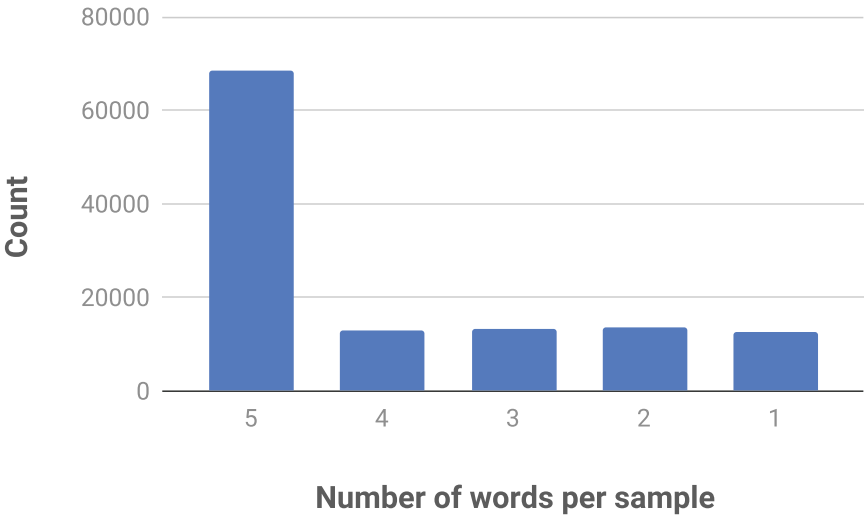


Fig. 5. Number of words distribution per data sample.

4 Discussion

Deep learning systems demand large amounts of data to satisfy their learning capabilities and be able to generalize without over-fitting. In this work, we collected more than 100 thousand training samples where each sample consists of a video with its respective annotation. Each video contains a person quoting from one to five words. The set of video and annotated text can be used to

train speech recognition systems based on vision (lip reading), synthetic generation of talking faces for animation purposes or video games. However, this is not new, and there are multiple databases of this type solely for the English language. In this work, we generated an audio-visual speech recognition dataset in Spanish to support the Spanish-speaking scientific community. Furthermore, the proposed methodology can be used to automatically generate databases to train models of audio-visual speech recognition in any language by merely providing the videos in the desired language. One shortcoming of this method is that it makes use of the IBM Audio-to-text engine that runs in the cloud and has a cost per minute. The advantage is that it allows the alignment between video and text with milliseconds accuracy, which is not offered by other free services such as Google Cloud. Another constraint of the proposed methodology is that it only extracts data from videos that contain a single person speaking to ensure that the voice corresponds to the face in the video. This limitation excludes videos that can provide a high diversity of words, such as those available in discussions, interviews, etc. To address this limitation, as future work, we are working on a mouth movement detection system used to extract information from multi-person videos and extend the capabilities of the automatic systems.

5 Conclusions

This paper describes the method of semi-automatic database creation for audio-visual speech recognition systems. To test the technique, we created an extensive database for training computer vision based automatic lip reading systems in the Spanish language. The created database contains more than 100 thousand samples annotated with accurate timestamps and confidence level. The source code used to generate the database is freely available and can be used to create similar databases for other languages.

Acknowledgements. The authors wish to acknowledge the support for this work to Universidad Autónoma de Querétaro (UAQ) through project FIF-2018-06.

References

1. Afouras, T., Chung, J.S., Zisserman, A.: LRS3-TED: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
2. Bulat, A., Tzimiropoulos, G.: 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, pp. 1021–1030 (2017)
3. Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Lip reading sentences in the wild. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3444–3453. IEEE (2017)
4. Chung, J.S., Zisserman, A.: Lip reading in the wild. In: Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y. (eds.) ACCV 2016. LNCS, vol. 10112, pp. 87–103. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54184-6_6

5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)
6. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. **106**(1), 59–70 (2007)
8. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)
9. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
10. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: SUN database: large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3485–3492. IEEE (2010)