BACHELOR' S THESIS / BIOMEDICAL ENGINEERING 2022

Continuous Lip Reading in Spanish

Yadira Ronquillo

upf. Universitat Pompeu Fabra Barcelona

Continuous Lip Reading in Spanish

Yadira Ronquillo

Bachelor's thesis UPF 2021/2022

Thesis supervisor(s):

Federico Sukno and Adriana Fernández López

Acknowledgments

I would like to express my gratitude to my supervisors Federico Sukno and Adriana Fernández López, for their help and guidance during the whole project.

Also, I would like to acknowledge my reviewers in advance, Gloria Haro and Nicolás Gutiérrez, for acceding to read my thesis.

Summary/Abstract

Lip reading, also known as visual speech recognition, is the task of decoding text from lip movement, which involves analysing the change in the speaker's lip shape. It has a wide application in the fields of security, assisted driving systems, virtual reality, speech transcription for cases where audio is not available, and communication of people who are hearing-impaired. For the last case, it can also be an extremely helpful tool for these people to communicate through video calls or to understand what the other person is speaking. For such reasons, lip-reading has been the subject of a vast research effort over the last few decades. Currently, deep learning is being used to deal with this task. However, the training of the lip-reading model relies on a large amount of data. Therefore, lip reading has limited its applicability to English since this is the only language with large-scale datasets. In this work, we used a new audio-visual dataset in the Spanish language, which has been built from a subset of the RTVE database. This is the largest publicly available sentence-level lip reading dataset to date in the Spanish language and it consists of over 13 hours of video, extracted from Canal 24 horas. We used it to develop an Automatic Lip-Reading (ALR) system for continuous speech recognition in Spanish. For this purpose, we employed Audio-Visual Hidden Unit BERT (AV-HuBERT) model, based on transformer network. The system obtained can differentiate some short sentences. On the other hand, we observed transfer learning works better when the languages are similar, and that there is a relationship between the size of the dataset and the learning transfer method.

Keywords

Deep Learning, Automatic Speech Recognition, Transfer learning, Lip Reading,

Preface or prologue

In the last years, deep learning techniques have experienced exponential growth, being used in many fields like computer vision or automatic speech recognition. Deep learning techniques are also increasingly essential in Automatic Lip-Reading (ALR) systems. Currently, most research uses it, as it has shown that it can achieve better results than traditional methods.

Index

List of Figures

List of Tables

1. Introduction

Every day, speech is used to transmit information, ideas, and thoughts to others, who use speech perception to better understand the message delivered. Thanks to McGurk and MacDonald, it is known that speech perception involves both auditory and visual information [1]. In their research, they discovered a curious phenomenon when they were watching a video that contained people pronouncing different syllables. Due to an error in the editing, the video was not synchronized with the sound, so they saw a recording of a person pronouncing something that did not correspond to the sound heard.

At the time of playing the video, they heard a third phoneme instead of the one that was articulated with the lips and the one that was emitted. A clear example of this phenomenon, known as the McGurk effect (Figure 1), would be to see someone making lip movements that correspond to the syllable /ga/, but while emitting the syllable /ba/ it will be perceived as /da/. This research revealed that visual information plays an important role in perception.

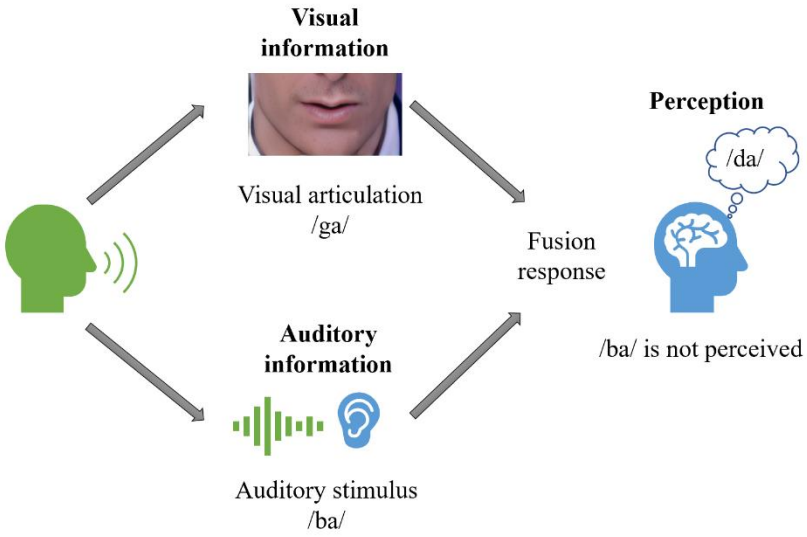

Figure 1: The illusion occurs when the auditory component of one sound is paired with the visual component of another sound, leading to the perception of third sound. If the listener closes his eyes, so that he no longer sees the speaker's lips, he hears /ba/.

Furthermore, it has been observed that the visual signal helps us to better capture the message transmitted in a noisy environment. For example, when two persons try to have a conversation on the terrace of a bar, on a busy street or in a school classroom. When there is loud background noise, we find it difficult to understand what the person in front of us is saying. In order to understand something, we use the instinctive trick of watching the mouth while the person is speaking.

1.1. Applications

In this section, we will mention the applications that have been made in recent years and proposed applications.

ALR can help deaf or hard-of-hearing people understand what the other person is talking about in real-time, such as when making video calls. The system would capture the lip movements and return the transcription of the emitted message in time real, so it would not be necessary for these people to try to read lips because the ALR system would do it automatically. Thus, this could lead to improved social relations.

An ALR system for the Spanish language could help many people considering that there are more than one million deaf people in Spain and only 27.300 use sign language according to the Instituto Nacional de Estadística [2]. Therefore, ALR in Spanish could have a significant impact on the day-to-day life of many people. But this could also have a major role in hospitals. There are cases where patients cannot speak, due to having a tracheotomy, stroke, trauma, or other conditions. This fact doesn't allow these patients to communicate with family members and healthcare staff. For this reason, the Liopa company developed a mobile app called SRAVI (Speech Recognition App for the Voice Impaired) in 2019 [3], which is a communication aid for speech impaired patients. SRAVI has been trialled successfully on ICU patients who have had tracheostomies. It is worth mentioning that SRAVI can improve communications and provide significant improvements in quality of care, speed of addressing patient needs and costs incurred in patient care (via optimization of staff time). To date, it is only available in the UK.

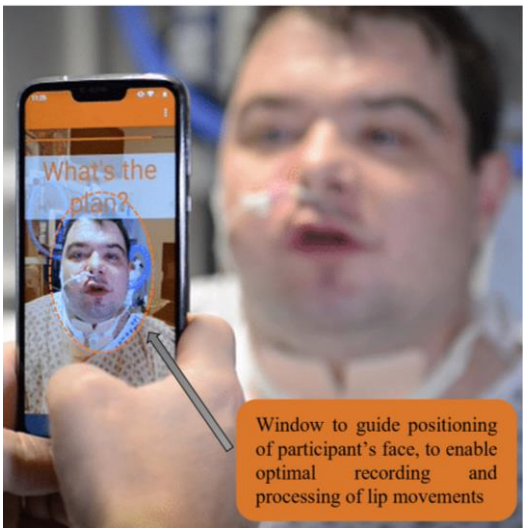

Figure 2: Interface of SRAVI. It is a clear example of how an ALR System can be used in hospitals.

Furthermore, it has numerous real-world applications as well, such as biometric identification [4], visual password [5], Closed-Circuit TeleVision (CCTV) footoge to assist law enforcement, forensic video analysis [6], speech transcription for cases where audio is not available [7], and it can be in implemented in assisted driving systems [8] and virtual reality (VR) systems to enhance the actual immersive VR experience [9], among others.

On the other hand, ALR is considered a new mode of human-computer interaction. A clear example of this can be seen in the research performed by Yuanyao et al. [10], ALR is used to realize the control of the set-top box to turn on and off, change the channel and adjust the volume. Instead, the research of Pandey et al. [11] and Sun et al. [12] use it to interact with a mobile device.

As mentioned, ALR systems can be applied to various tasks. Perhaps, ALR systems have attracted a lot of attention in recent years due to their wide applicability in different fields.

1.2. Automatic Lip Reading Systems

In this section, we first introduce the conventional approach and then focus mainly on deep learning architectures.

Typical ALR systems consist of mainly three blocks [13] :

1. **Lips localization** focuses on face and lip detection.
2. **Extraction of Visual Features** provides numerical values to the visual information observable at a given time instant.
3. **And Classification into sequences** maps the extracted features into speech units while ensuring that the decoded message is consistent, and it helps eliminate ambiguity between visually similar speech units using context.

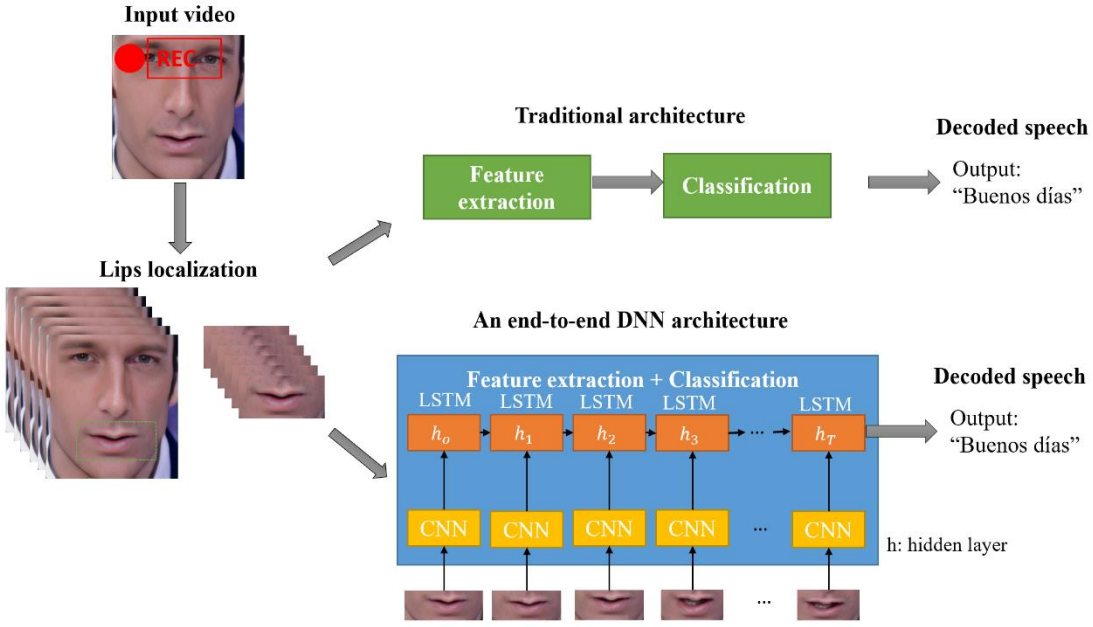

Figure 3:Baseline DNN architecture for lip-reading, consisting of combinations of CNNs and LSTMs.

An example of traditional lip reading systems is to use Principal Component Analysis (PCA) to perform feature extraction and Hidden Markov Models (HMMs) to classify the visual features into speech units. Some reaches that used this architecture were Cappelletta and Harte [14] and Seymour et al. [15]. We refer the readers to [16] and [17] for more information on traditional approaches.

A clear limit of traditional architectures is that each module must be optimized separately according to different criteria. For this reason, a new approach was proposed, which consists in replacing the second and third blocks with a Deep Neural Networks (DNNs) architecture which is trained end-to-end, i.e. whole architecture is trained jointly, rather than step by step. Thus, DNNs enable the use of a unique optimization criterion for system improvement.

Over the past five years, different Convolutional Neural Network (CNN) models have been combined with variants of Recurrent neural networks (RNNs) to create end-to-end DNN architectures. CNN extracts the features from the images, and then RNN classifies them. But traditional RNNs cannot be used because their internal memory has a storage problem for automatic lip-reading recognition, known as the vanishing gradient problem, which makes their memory short-term. Long Short-Term Memory Network (LSTM) is an advanced RNN, a sequential network, that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. Thus, it is well suited for learning from meaningful experiences that have very long lapses of time between them. Therefore, several researchers use LSTMs for classification [18, 19, 20] or other types of RNNs like Bidirectional Long Short-Term Memory networks (BLSTM), which can preserve information from both the past and the future [21, 22, 23].

However, all RNNs have the following problem, the output of hidden layers (the orange block in figure 3) depends on the input and the previous hidden values, which causes a sequential dependency between the hidden layers. For this reason, transformers were introduced with the purpose to avoid recursion, and this way, to allow parallel computation, which reduces training time. Transformers demonstrate decent performance on lip reading [24, 25, 26]. The main characteristics are:

- Non sequential: sentences are processed as a whole rather than word by word.
- Self-Attention: it tries to capture the relations between the words and to calculate a score that shows how much attention should pay to each element in a given sequence.
- Positional embeddings: which means that it modifies the values of each embedding vector to represent its location in the text. Thus, although there is no sequential processing, there is order.

The transformer of [27] is an encoder-decoder structure with self-attention layers, as shown in Figure 4.

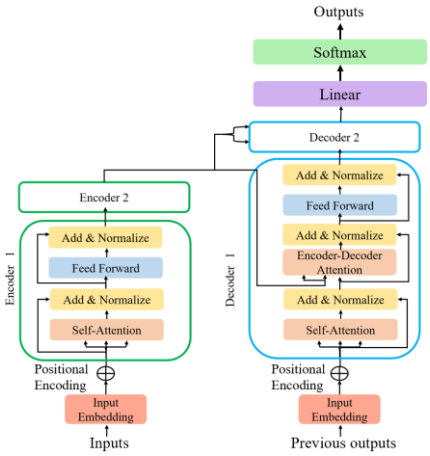

Figure 4: The architecture of the Transformer model proposed by [27].

1.2.1. Limitations

One of the main challenges they must face is the visual ambiguity that exists in the pronunciation process because there are phonemes that when they are pronouncing, they

use the same mouth movements that other phonemes. Phonemes are the minimum distinguishable sound that can change the meaning of a word [28]. An example of visual ambiguity is the consonant phonemes /p/ and /b/ are visually indistinguishable. Both are bilabial, meaning that they are pronounced after closing and opening the lips, and they are distinguishable in that /b/ makes the vocal cords sound and /p/ does not. Homophones also produce a visual ambiguity, an example is the 'hola' word and 'ola', both have different meanings but are pronounced in the same way, so they are difficult to distinguish without context. Furthermore, the accent, speed of speaking, and mumbling are factors that complicate the task [29].

To cope with visual ambiguity several authors have decided to group phonemes that share the same visual appearance. Each grouping is usually referred to as a viseme [30]. However, there is still no consensus on the precise definition of the different visemes or the number of existing visemes, and even their meaning is still under debate. For these reasons, many possible mappings of phonemes to visemes have been proposed [31].

In addition, the rotation of the head in speech can make the mouth area appear incomplete in the image, and therefore part of the mouth information is lost. For example, an image containing a person in profile does not allow to extract all the features of the mouth, as there is only half of the mouth. Another limitation is a dataset with a small number of speakers, which leads to speaker dependency because the extracted features may contain speaker information such as speech habits, which is irrelevant for lip reading.

On the other hand, adversarial imaging conditions, such as poor lighting, strong shadows, motion, resolution, foreshortening and, among others, are factors that make challenges more difficult.

It is worth mentioning that resolution affects feature extraction. Some researchers have seen that degradation in image resolution effect the performance of CNN because it decreases the performance score (accuracy, precision and F1 score) of image classification [32, 33]. In the case of **lip reading,** Jitaru et al. [34] and Schwiebert et al. [35] indicate that **the quality of a data set is of utmost importance for the performance of a model using it**. A higher resolution will provide more details, and thus more robust features can be learned. In addition, the low resolution also affects the landmark prediction performance [36, 37], which is employed for lip detection.

Deep learning models can resolve the low-resolution problem. In our case, the model must restore faces, there are many designed models for this task, but the best is GFP-GAN [38], which consists of GAN pre-trained on the faces (green part) and a U-Net module (blue part) for removing the degradation. Although GFP-GAN was introduced in 2021 it has been used in much research on face image generation [39, 40, 41].

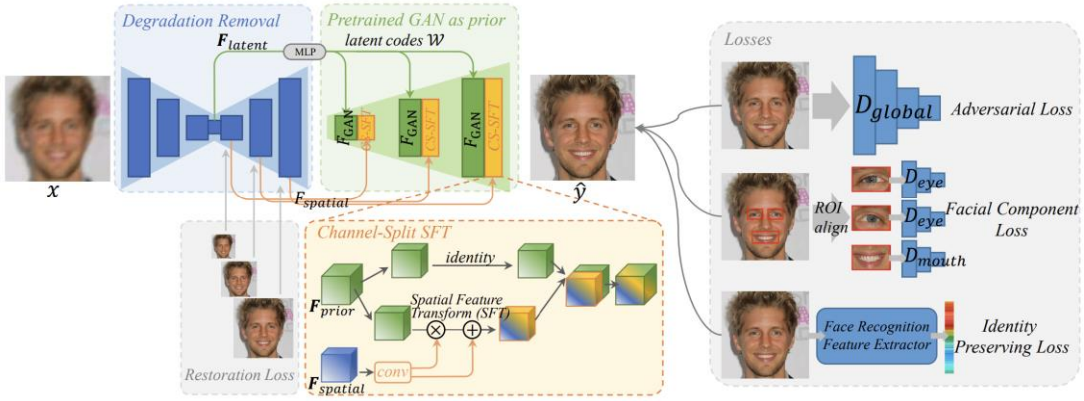

Figure 5: Overview of GFP-GAN framework. It restores well the mouth

1.2.2. Lip Reading Datasets

For twenty years, ALR systems have been used for simple tasks, such as the alphabet or digit recognition, allowing researchers to work in controlled environments, where the position of subjects and the speech content were predefined. The limited vocabulary size, clear pronunciations and recording settings are some essential characteristics of such collections, called controlled datasets. The videos recorded in controlled environments can be used as a benchmark for evaluating VSR pipelines, training models and speeding up the convergence of the training procedure. However, the trained models with those datasets failed when processing videos recorded in real-world conditions due to differences in illumination, the speaker pose, pronunciation distinctiveness, and restricted vocabulary. Therefore, new datasets have been created in the late years. In addition, deep learning needs large amounts of data to obtain high performance. New datasets are composed of words, phrases, and sentences. Table 1 shows a list of datasets which were recorded in a non-supervised scenario or the wild. As can be seen, the dominant language in the available data is English. As a result, this creates a gap in the ALR field as it restricts its applications and benefits for other languages.

Name	Year	Language	Task	Speakers	Hours
LRW	2016	English	Words	1000+	1600 h
LRS2 -BBC	2017	English	Sentence	-	±200 h
LSVSR	2018	English	Sentence	-	±4000 h
LRS3- TED	2018	English	Sentences	9000+	±400 h
LRW-1000	2018	Mandarin	Words	2000+	57 h
VoxCeleb	2019	Multilanguage	Sentences	7000+	2000 h
Wild LRRo	2019	Romanian	Words	35+	21 h
RTVE	**2020**	**Spanish**	**Sentences**	**323**	**13 h**
GLips	2021	German	Words	500	± 80 h

Table 1: Comparison of different datasets in the wild

The new datasets are being used to resolve more complex tasks, such as recognising single words, simple sentences, and continuous speech. In addition, it exists several types of experimental setups that can be done: speaker dependent (SD) or speaker independent (SI), single-view or multi-view. In SD setting, data from all speakers have been seen in both training and evaluation. On the other hand, evaluation is performed on the unseen

speakers in SI setting. In the multi-view setting, we use data recorded from multiple cameras with various angles simultaneously, while data in a single-view setting is recorded from one camera angle.

1.3. Motivation

Most ALR systems developed for continuous lip reading have used English datasets, as these datasets have lots of speakers and hours. These systems can only work with English speakers. Therefore, **we want to take advantage of the recent creation of the RTVE dataset to develop an ALR system in Spanish for continuous lip reading.**

The RTVE dataset, being so recent, has only been used by Gimeno et al. [42]. They investigated what existing methods work best to extract features from the images then these were classified by a traditional method. One could say that their approach was traditional. They conclude the research by indicating that a pure deep learning approach might be more useful.

Deep learning methods are generally avoided for small datasets because they tend to overfit the model, i.e., the model performs well on the training data but does not perform accurately in the evaluation set. Transfer learning can be one approach used to solve this problem since it can train deep neural networks with a small amount of data. Jitaru et al. [34] and Schwiebert et al. [35] recently demonstrated the effectiveness of this approach for lip reading, using word datasets and very different languages between. The models available for transfer learning are limited (models are available at the following links [1,2,3,4]). Moreover, they were trained with word datasets, which are not suitable for continuous lip-reading. Therefore, these models cannot be used by us. But, at the beginning of the year a model trained with sentences in English was introduced, its name is AV-HuBERT [43].

According to Meta AI, which developed **AV-HuBERT**, only a small amount of labeled data is needed to re-train a model for a particular task or a different language. Therefore, this model is a good choice to be **our baseline system**. Nevertheless, we should consider that our dataset has13 hours of videos, and AV-HuBERT used 30 horas to train a model for visual speech recognition. It is relevant to mention that there is no previous work on AV-HuBERT with non-English languages.

On the other hand, English and Spanish use the Roman alphabet. There are even words that are spelt the same in English and Spanish, as well as meaning the same thing, such as 'bar', 'agendas', 'general' or 'doctor'. At least 35% of English words have a Spanish word with which they are closely related, known as cognates [44]. This relationship can be either in sound, meaning or both. These similarities are because the evolution of English and Spanish was not so unrelated to each other. Although English and Spanish come from different roots, they share relatively close periods of development and geographical areas.

Considering that **AV-Hubert has been designed to be used with different languages** and that **English and Spanish have similar aspects**, features of English transferred to

[1] https://github.com/lordmartian/deep_avsr
[2] https://github.com/afourast/deep_lip_reading
[3] https://github.com/VIPL-Audio-Visual-Speech-Understanding/learn-an-effective-lip-reading-model-without-pains
[4] https://github.com/mpc001/Lipreading_using_Temporal_Convolutional_Networks#model-zoo

our model should be very useful in learning the new language. As we mentioned in the first paragraph of this section, the goal is to develop an ALR system in Spanish for continuous lip reading. To do that, we will perform transfer learning from English to Spanish. Our working plan is:

- To read literature for continuous lip-reading
- To get familiar with the RTVE dataset
- To improve the resolution of images
- To organize the videos and text transcriptions to create different subsets or corpus.
- To perform experiments
- To evaluate and analyse the results of experiments

1.4. Related Work

There is a lot of literature on traditional systems and deep learning systems. But in this section, we will mention the literature that has similarities with our study.

Gimeno et al. [42] used the RTVE dataset to see which is the best option for visual speech feature extraction since there is no consensus or agreement in the literature. Therefore, they use both traditional and CNN techniques to extract features and see which one gives better results. Furthermore, they employed a traditional system based on Hidden Markov Models with Gaussian Mixture Models for classification. They found that the combination of eigenlips (obtained after applying PCA) and deep features can be established as the best approach to address automatic lipreading. Nevertheless, the differences were not significant with respect to the exclusive use of deep features. Experiments were performed with 43 different speakers, reaching around 3 hours of data, whereas the test partition comprises 120 samples from 13 speakers, covering 0.13 hours of utterances. Here, the RTVE dataset has small size because they decided to increase the number of seconds that make up this dataset after this one. On the other hand, they decided to relax the task complexity by employing a vocabulary list, which included words from the test partition, because they did not achieve minimally acceptable results (error rates greater than 90%). The results we will obtain will not be compared with this research because there are several differences in the number of speakers and samples.

Fernandez-Lopez and Sukno [45] proposed an alternative learning strategy that allows end-to end training of an ALR system without the need for large-scale data. Their system consists of two modules: the visual module which is a pre-trained CNN and CNN, and the temporal module is an attention-based seq2seq architecture. For training, they divided the whole set of sentences into small speech units. The aim was to control the network learning to ensure that the extracted features were sufficiently representative to help the temporal module predict the character. The dataset used was the VLRF database, which was recorded in a controlled environment. It consists of 600 sentences and 24 speakers. We want to emphasise that this dataset is one of the largest audio-visual datasets in Spanish, although it contains approximately 3 hours of recordings. The results obtained were a 44.77% CER and 72.90% WER on the test data.

Jitaru et al. [34] tried to build a language-independent lip-reading system. But they needed massive amounts of labeled data. For this reason, they explored the effectiveness of transfer learning to address the lack of large datasets, using LRW, LRRo, and LRW-1000 datasets. Those datasets were input data for the D3D (DenseNet 3D) and a visual attention

autoencoder network. On the other hand, the domain dataset is LRW. Therefore, they used it as the source domain for transfer learning on the other two subsets to improve the generalization capabilities of the evaluation models. As a result of that, they found that transfer learning architectures improved predictive capabilities. This research was useful for us to see the effectiveness of transfer learning, but we will not be able to compare their results with ours because the languages used are quite different between them.

Schwiebert et al. [35] used the deep learning approach for word-level lip reading, specifically, the transfer learning method. They did transfer learning from LRW to GLips and vice versa to investigate whether lip reading has language-independent features so that datasets of different languages can be used to improve lip reading models. They showed that transfer learning improved learning speed and performance, particularly for the validation set, compared to learning from scratch. They chose the X3D convolution neural network model to do their experiments.

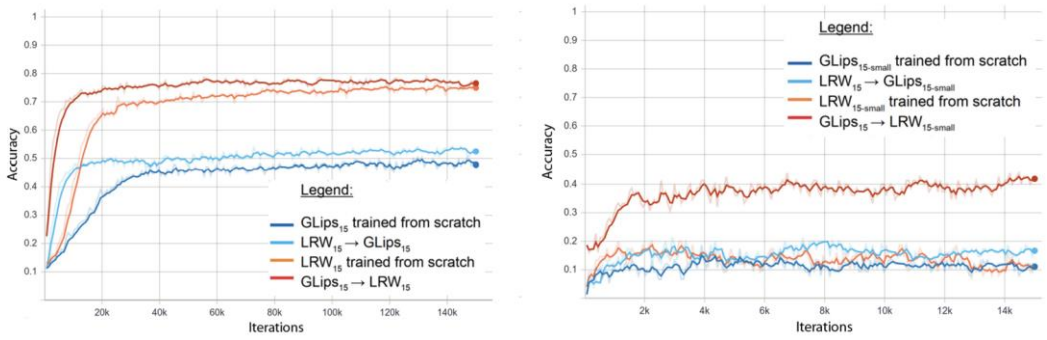

Figure 6: Results of validation (left) accuracy per iteration for transfer learning between same-size datasets and resulting of validation (right) accuracy per iteration for transfer learning from a large data set (~80,56 hours) to a small dataset (~2,5 hours) [35]

On the other hand, Meta AI introduced Audio-Visual Hidden Unit BERT (AV-HuBERT) [43], which is the first system to simultaneously predict speech and lip movements from unlabeled data, in other words, untranscribed video. They used 433 hours and 1759 hours unlabeled data for pre-training, which belong to LRS3 and VoxCeleb datasets, respectively. Once the pre-trained model mastered the structure and correlation, only a minimal quantity of labeled data was required to train a model for visual speech recognition, which reached a 28.6% WER with just 30 hours of labeled data, outperforming the former state-of-the-art approach (33.6 % WER) trained with 31000 hours of transcribed video data [46]. Furthermore, when using 433 hours of labeled data, they achieved a new state of the art at 26.9 % WER. Also, they experimented with labeled data that corresponded to {1, 10, 100} hours (Table 2).

Labeled	WER %
1	92,0
10	63,1
100	48,1

Table 2: WER(%) when it was used different amounts of labeled data for transfer learning

2. RTVE dataset

The subset used in this investigation was pre-processing by Pattern Recognition and Human Language Technologies Research Center of Universitat Politècnica de València in 2020. They compiled it from a subset of the RTVE database [47] which is made up of shows that cover a great variety of scenarios from scripted content to live broadcast, from reading speech to spontaneous speech, different Spanish accents, including Latin-American accents and a great variety of contents. Nevertheless, they carried out the pre-processing considering only the news program 20H broadcast by the Canal 24 horas. The selected scenes have unique conversations of the speakers from different distances to the camera and in diverse scenarios, either inside a recording studio or in outdoor locations. The subset obtained contains approximately 13 hours of conversations by 323 speakers, giving10345 samples with a vocabulary of 9307 words.

2.1. Data preparation

In this section, we explain how we have discarded defective samples using a conventional face detector and CNN.

At first glance, images of objects, backgrounds and two speakers per sample were detected. This fact can hurt feature extraction and lip reading recognition. To avoid faulty samples in our corpus, we designed a verification mechanism using the dlib library for face detection [48]. Dlib is an open-source suite of applications and libraries written in C++. We used Dlib function called *get_frontal_face_detector()*, based on the Histogram of Oriented Gradients (HOG) and Linear SVM face detector. This function only works with grayscale images, so we had to do that first with OpenCV.

The *get_frontal_face_detector()* returned a "detector" which is a function we used to retrieve faces information. Each face is an object that contains the points where the image can be found.

By nature of how the HOG descriptor works, it is not invariant to changes in rotation and viewing angle. For this reason, *cnn_face_detection_model_v1()* was used for non-frontal faces at odd angles. This function loads a pre-trained model, which is a ResNet network (a type of CNN architecture). The network was trained from scratch on a dataset of about 3 million faces.

If no face is detected in the frame after going through both functions. The id of the frame is stored in a file. As none of them is 100% accurate, we manually verified the frames which were annotated in the file to discard the possibility of removing a correct sample. Once done this, 184 samples [5] were eliminated.

[5] ID of deleted speaker:
https://drive.google.com/file/d/1RYDH3t4AwxSkP0PajZ83bF7Tg8YkP2fu/view?usp=sharing

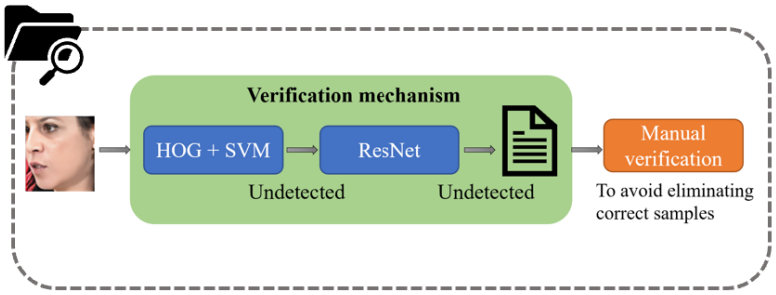

Figure 7: Verification process for each frame

3. Methods

3.1. Preprocessing of the RTVE dataset

Here, each image was restored because some images had the mouth area very blurred due to low resolution. In this way, the quality of the RTVE dataset was improved.

The low quality of the images involves that lips and teeth cannot be distinguished between them. The same happens with lips and beards. Furthermore, it is also very difficult to tell whether the mouth was open or close in some cases. Therefore, the images had to be improved before being used in the following steps, as poor quality can affect lip detection and feature extraction [32, 33, 36, 37]. To achieve high-quality images, we used GFP-GAN [38], which has proven to repair faces much better than other current models. In addition, an interface is available that allows you to insert whole folders of images or just a single image. This has been useful because we have been able to work with folders.

Despite its mighty power to restore, there have been cases where the teeth have not been restored correctly. Because of its limitation when the degradation of real images is severe, the restored facial details by GFP-GAN are twisted with artefacts.

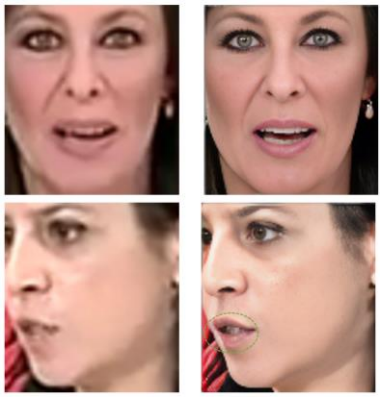

Figure 8: Example of two images from the RTVE dataset restored with GFT-GAN. The green circle shows a poor restoration of teeth.

3.2. Our ALR system

Our baseline system is AV-Hubert which is a variation of the transformer model. The main difference between [27] and AV-Hubert is that it only uses encoder modules.

Furthermore, it has a visual encoder and an audio encoder. To develop our ALR system, we only use the visual encoder.

On the other hand, we will explain lip detection, architecture, the method to create the vocabulary list and how to do transfer learning in this section.

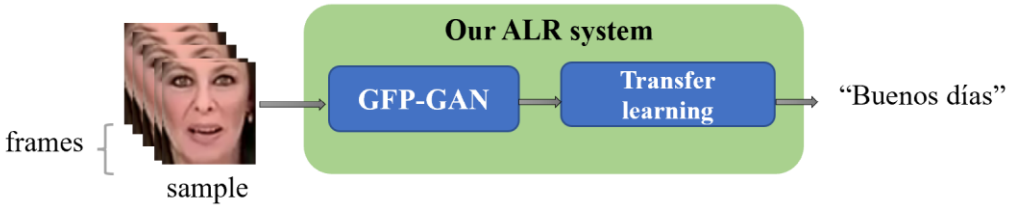

Figure 9: Our ALR system. The input data is non-restored images, then transfer learning gives an output

3.2.1. Lip detection

In each video, the face was localized using *get_frontal_face_detector() and cnn_face_detection_model_v1()*. Given the face region, *shape_predictor* () was applied to estimate the location of 68 coordinates (x,y) that map the facial points on a person's face like the image below. These points are identified from the pre-trained model, where the ibug 300-W dataset was used. The faces are aligned to a reference face frame using affine transformation to remove differences related to rotation and scale. A bounding box of 96x96 was used to crop the mouth Regions-of-Interest (ROIs). The moth region is always roughly centred on the image crop. As data augmentation, ROIs were randomly cropped with a size of 88x88, and a random horizontal flip was also implemented during training. This form of preprocessing has been used in prior works of lip reading [49, 50].

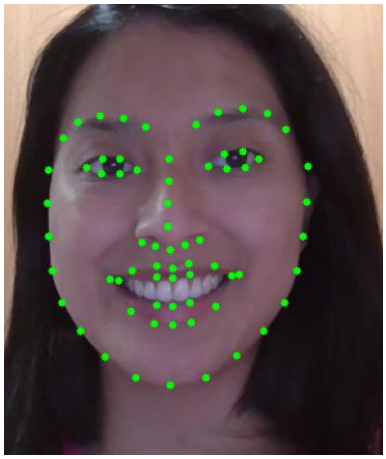

Figure 10: Example of face landmark detection using 68 facial landmark coordinates.

3.2.2. Architecture

The visual feature sequence was obtained by a modified ResNet-18, in which the first convolutional layer is replaced by a 3D convolutional layer with a kernel size 5x7x7. The visual feature is flattened into a single-dimensional vector through a 2D average pooling layer at the end. In contrast, a feed-forward neural network was used to extract the audio

features. This way, the model does not depend strongly on the audio modality because the audio encoder learns simple features. Therefore, the visual encoder is forced to learn meaningful features. The audio-visual features are encoded into a sequence of contextualised features via a transformer encoder followed by a linear projection layer. Two types of transformers were used: BASE with 12 transformer blocks and LARGE with 24 transformer blocks. For BASE and LARGE (Figure 11), the embedding dimension, feed-forward dimension and attention heads in each transformer block are Table4. The number of parameters in BASE and LARGE are 103 M and 325M respectively.

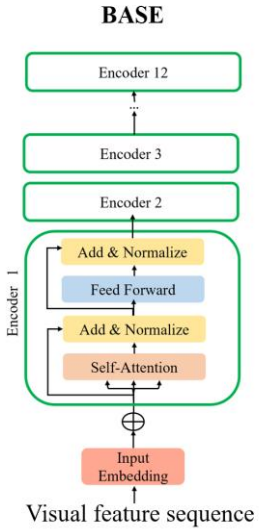

Figure 11:Transformer encoder of BASE

3.2.3. Vocabulary list

Unigram language model [51] is used to create the dictionary which will be used during training. The particularity of this word segmentation algorithm is that it sets a vocabulary size limit for the word segmentation algorithms to use as minimal memory as possible. Larger datasets have a considerable number of words, which involves using a lot of memory to store these words. Nevertheless, these words can be exceedingly rare and only appear once or words which have a complicated meaning. For example, the word ‘abierta’ and ‘abierto’ are the same words but slightly altered to suit the grammatical purpose. Therefore, it is not worth having an additional vector stored in memory (see Figure 12).

The word segmentation algorithm captures sub-words that appear frequently enough to determine the importance of the sub-word but also diverse enough to minimise re-capturing the same information and build up a useful, diverse sub-word vocabulary list.

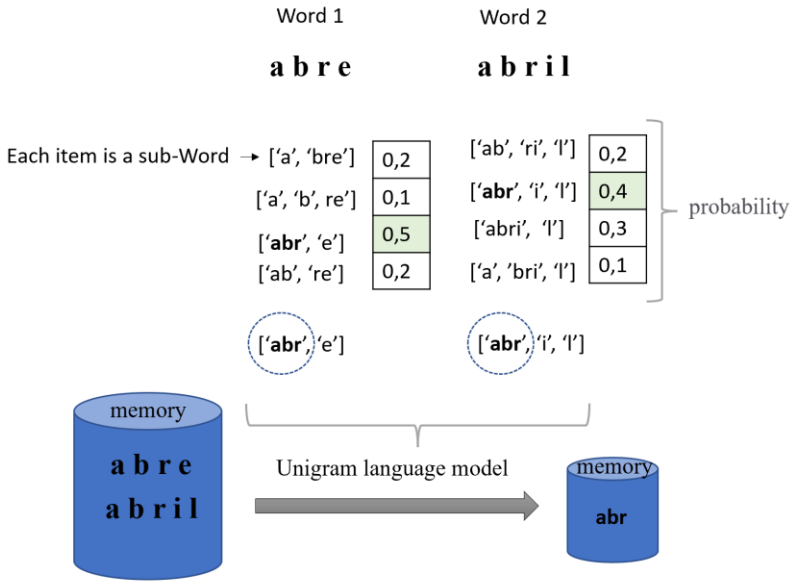

Figure 12: General concept of the unigram language model.

3.2.4. Transfer learning

Fine-tuning is a way of applying or utilizing transfer learning. Specifically, fine-tuning is a process that takes a model that has already been trained for one given task and then tunes the model to make it perform a second similar task, as it has been re-trained with the new data.

Meta AI proposed to remove the audio encoder and replace its outputs with zero-vector in fine-tuning (see Figure 13). The pre-trained encoder is frozen for n steps, and then the rest of the steps are used to fine-tune the whole model. A transformer decoder is appended in the 6-layer of BASE and the 9-layer of LARGE to decode features into unigram-based subword units.

In addition, the models are trained with Adam and a learning rate of 0,001, using labeled data and cross-entropy loss (2).

$$L_{s2s} = -\sum_{t=1}^{s} \log p(w_t | w_{1:t}, e_{1:T}) \quad (1)$$

Where the feature sequence output of our pre-trained model is $e_{1:T}$ and the ground-truth transcription is $w = w_1, w_2, \ldots, w_s$. The transformer decoder decodes the feature sequence $e_{1:T}$ into target probabilities $p(w_t | w_{1:t}, e_{1:T})$

	BASE	BASE (penultimate iter)	LARGE
Max updates	30000	30000	30000
Warmup steps	10000	10000	10000
Decay steps	20000	20000	20000
Freeze finetune updates	24000	24000	30000

Table 3: it shows the hyperparameters used during transfer learning of [43] for visual speech recognition with 30h of data.

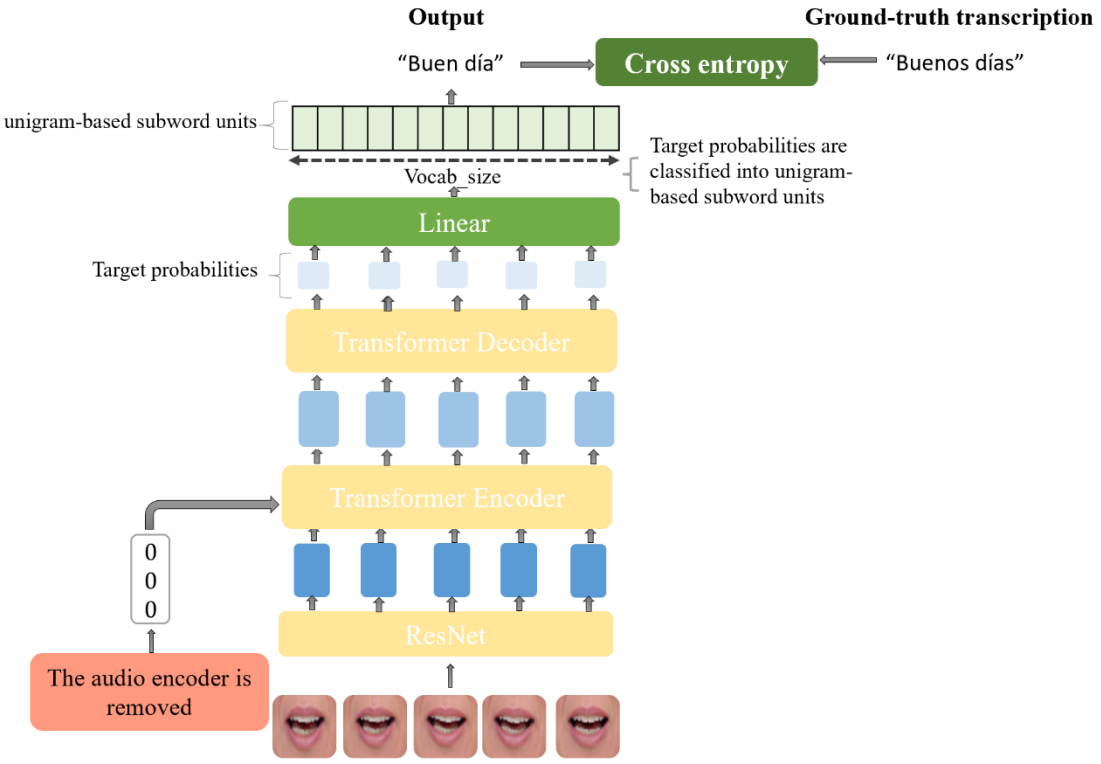

Figure 13: Architecture for transfer learning

4. Results

We will first describe how we split the data, then the metrics used and the experiments.

4.1. Dataset split

After improving the images, we converted images of each sample to videos of 25 frames per second, using the VideoWriter function of OpenCV [52], which is an open-source library for computer vision.

The first corpus created was speaker dependent, and only sentences with a length of 10 words or less were considered. For validation and test set, we randomly selected 10% of sentences in each speaker. The following corpus created contained the same samples as the previous corpus. However, samples were reorganized to create a speaker independent corpus. Therefore, the split was done considering the percentage allocated to each set, that was 80% of the samples for the training set and 10% for the validation and test sets. The last corpus created was also speaker independent. In this case, we excluded the sentences with more than 37 words and used the same split mentioned above. Table 4 shows the number of speakers and samples used in each corpus.

	Corpus 1 (SD): ~ 2 h		Corpus 2 (SI): ~ 2 h		Corpus 3 (SI): ~ 13 h	
	Speakers	Samples	Speakers	Samples	Speakers	Samples
Training	292	3508	105	3410	133	7638
Validation	118*	350	86	418	88	931

Test	118*	394	101	424	107	948

Table 4: Information about the different corpus. * Each speaker has a different number of samples. Although, some speakers have only one or two samples, thus, we could not take 10%.

4.2. Evaluation Criteria

The word error rate (WER) calculation is based on a measurement called the Levenshtein distance which is a number that shows how different two words are. The higher the number, the more different the two words are. Its formula is:

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & otherwise, \end{cases} \quad (2)$$

For example, the Levenshtein distance between “buenos” and ‘buen’ is 2 as two characters were deleted.

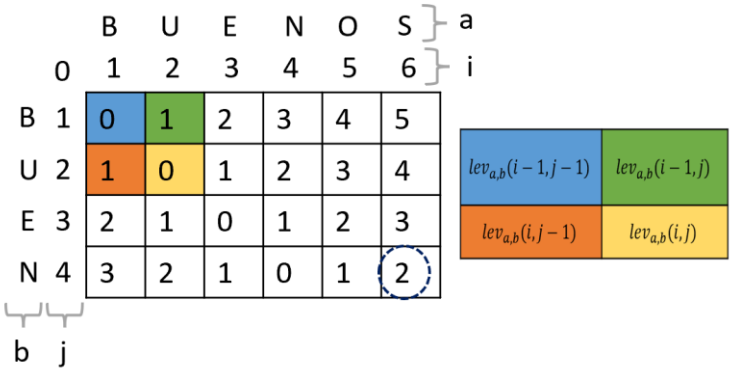

Figure 14: Calculation of the Levenshtein distance for words a and b using a matrix

Word error rate can be computed as:

$$WER = \frac{S + D + I}{N} \quad (3)$$

Where **S** is the number of substitutions, **D** is the number of deletions, **I** is the number of insertions and **N** is the number of words in the reference transcription.

It is also relevant to know the performance of the model at character level, thus character error rate (CER) was implemented.

CER is similar to WER but operates on character instead of word. It can be computes as:

$$CER = \frac{S + D + I}{N} \quad (4)$$

Where **N** is the number of characters in the reference transcription.

Both metrics were used in tests to evaluate the performance of the model.

4.3. Experiments

Six pre-trained models are available. Through the following experiments, we want to see which one is the best for continuous lip reading in Spanish and to obtain the most optimal values for the hyperparameters. We decided to modify the same hyperparameters that Meta IA changed during transfer learning for visual speech recognition. These were:

- **Max updates:** This forces the training to stop at a specified update, i.e. it serves to indicate the maximum steps. The value of the following hyperparameters depends on this. The value of warmup steps is 1/3 of the whole training steps, whereas the freeze finetune updates are 100% or 80%. On the other hand, the decay steps are twice the warmup steps. This configuration was used by [43].
- **Warmup steps:** We linearly increase the learning rate from a low rate to a peak of 0.001 (in our case). This reduces volatility in the early stages of training and helps our network to slowly adapt to the data.
- **Decay steps:** The learning rate decays linearly over n steps, which helps the network to avoid oscillations during the training.
- **Freeze finetune updates:** It freezes the pre-trained encoder for n steps during transfer learning.

We first worked with the speaker dependent dataset to evaluate whether the model had difficulty in learning the new language. The fact that the same speakers were used in all three sets (training, validation, and test) allowed us to see whether the model can use the learned features of one speaker for another speaker. Afterwards, the other datasets were used. On the other hand, the experiments were carried out with 1 GPU Tesla T4, Adam as an optimiser, and the learning rate was 0.001.

Exp 1. **LRS3 vs LRS3 + VoxCeleb 2.** In this first experiment, both the vocabulary size and the hyperparameters match the values implemented by Meta AI in the BASE model. The aim is to see if models trained with more data give a significant advantage.

The model with more data gives better results in validation accuracy. Thus, this model was employed in experiments 2-8

Datasets	epoch	Train loss	Train accuracy	Validation loss	Validation accuracy
LRS3	79	19,852	99,836	89,692	**26,926**
LRS3 + VoxCeleb	79	19,437	99,949	87,963	**29,849**

Exp 2. **Impact of batch size.** Considering that most samples have less than 100 frames and that sample with the most frames has 193 in corpus 1, we used a batch size of 400 and 800 frames. The max updates were changed to max epochs, training stopped when it reached 500 epochs.

Figure 15: Resulting of training and validation with different batch size.

The model with a **800 batch size** has 15,34% of training loss, 99,98% of training accuracy, 75, 22 % of validation loss and **27, 67% of validation accuracy in 100 epochs**. In contrast, the model with a **400 batch size** has 16,12% of training loss, 99,32% of training accuracy, 75,87% of validation loss and **26,92% of validation accuracy in 350 epochs**

Figure 15 shows that the batch size affects the learning speed. When both models reach 350 epochs they have comparable results, the model with a batch size of 800 only needs 100 epochs to achieve those results and the other model requires triple epochs. Therefore, training will be much faster if the batch size is 800. On the other hand, it tends to increase the value of validation loss when the batch size is large, this can be observed in validation loss graph, the orange line after 200 epochs increases its value.

Exp 3. **Different Vocabulary size.** The experiment setup was the same as Experiment 2, but the max epochs is 350 and the batch size is 800.

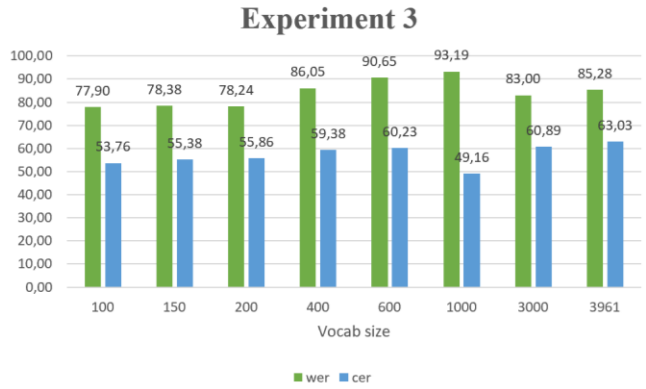

Figure 16: WER and CER with different vocab size.

The first column shows a WER of 77,90% and a CER of 53,76%. These values increase as more items are added to the vocabulary list.

As mentioned in the methods, unigram language model created our vocabulary list according to the desired vocabulary size. This method chooses the most repeated sub-words and then assesses the relevance of the repeated sub-words to select the most appropriate ones, as well as taking into account the white space between words. Setting the vocabulary size to 100 causes our vocabulary list contains only 'syllables' and characters, but if we increase the vocabulary size unigram language model will start to take words into account, so we will stop working with sub-words.

Having words in the vocabulary list is a disadvantage for learning the model because the model will need many samples of those words to learn to classify them. Furthermore, having many items in the list complicates the classification task, because there are many options for classifying the predictions. Therefore, a reduced vocabulary limits the possibilities, which benefit accuracy, WER and CER. This was observed in experiment 3. Each time we added more items to the vocabulary list, validation accuracy dropped, and WER/CER increased.

Exp 4. **Changing the hyperparameters.** Here, we used max updates again, so the other parameters are changed according to the value set in max updates.

Train loss	Train accuracy	Validation loss	Validation accuracy	WER	CER
26,153	99,974	88,273	59,018	**84,246**	**61,865**

These results were obtained using the configuration of [43]. Following experiments prove another configuration.

Exp 5. **Modifying the warmup steps and decay step.** Taking into account the previous experiment, where the warmup steps was 2,88% (in Experiment 2) and 7,48 (in Experiment 3) of the total training steps, we decided to use 10% as the value for warmup steps. Furthermore, this percentage was also used by Meta IA during the experiments of VSR.

Train loss	Train accuracy	Validation loss	Validation accuracy	WER	CER
28,061	99,876	80,073	58,051	**81,225**	**56,72**

CER improved by 5.14% compared to experiment 4. Therefore, this new configuration performs better than the previous one.

Exp 6. **Modifying the freeze finetune updates.** Observing the above experiment where the pre-trained model was frozen the first N% updates, N is 6,92 and 17,96 for experiment 2 and experiment 3, respectively. The freeze finetune updates were set at 15% and 5 % of the total training steps.

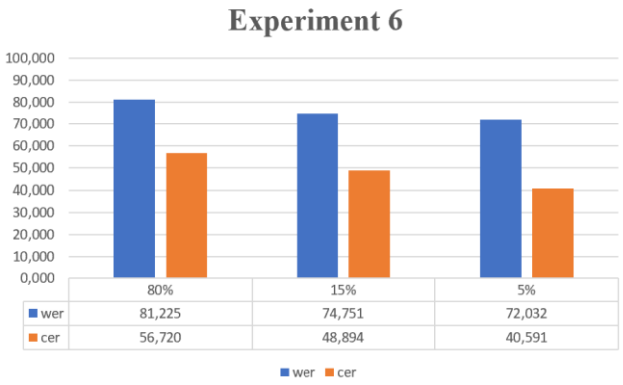

Figure 17: WER and CER with different % of freezing.

Due to the language difference, we cannot freeze the model for 80% of the trained time as proposed by [43] because the remaining 20% of the time is insufficient to learn new features.

Train loss	Train accuracy	Validation loss	Validation accuracy	WER	CER
27,714	99,73	79,788	**59,54**	**72,032**	**40,591**

CER improved by 16.12% (56,72% → 40,59) using a 5% of frozen updates.

Exp 7. **Using a small dataset of speaker independent.** The experiment setup was the same as Experiment 6, but the input was the corpus 2. This way, it can be vitrificated if 5% work for SI dataset.

Frozen updates	Train loss	Train accuracy	Validation loss	Validation accuracy	WER	CER
15%	38,241	99,876	124,243	40,6	95,5	**62,972**
5%	28,032	99,867	105,071	**39,193**	**95,961**	**56,401**

The result obtained concluded that 5% was also beneficial with corpus 2, in concrete for CER. Thus, it can be applied to corpus 3.

Exp 8. **Using a large dataset of speaker independent.** Here, the input was the corpus 3. The max updates were set at 300000, which is equivalent to 200 epochs.

Train loss	Train accuracy	Validation loss	Validation accuracy	WER	CER
79,518	99,979	165,497	**61,597**	**83,53**	**50,974**

This study aimed to develop an ALR system in Spanish using the transfer learning method since the RTVE dataset is small. After performing different experiments, we obtained a model with an accuracy of 39,193% and the other model that reached an accuracy of 61,597%, the difference between the two models is the size of the dataset used. The first one was trained for 2 hours and the second one for 13 hours. Increasing the dataset by 11 hours increased the accuracy by 22,404%.

Possibly, that value could have been higher if some images had not had a poor reconstruction of the teeth, this may influence feature extraction.

4.3.1. Examples of transcriptions

We want to see the performance of the models. So, five random samples were selected for evaluation.

S	Model obtained of Experiment 6: Speaker dependent
1	**GT:** a los tres jugadores de futbol que compartian piso **HYP:** a los luse asugados de futbol de copartia miso
2	**GT:** en el coche segun venia **HYP:** en el coche segun venia
3	**GT:** que hay argumentos tambien para defender **HYP:** que hay argumentos tambien para defender
4	**GT:** dos malas a su juicio puigdemont o un candidato **HYP:** dos vanes de su cuando puigdemontro el hoy
5	**GT:** de cargos de figuras libres de procesos judiciales recordemos **HYP:** de carros de circunas aquien de procestos udien rultimos
	Model obtained of Experiment 8: Speaker independent
6	**GT:** a lo largo de la manana unos cuatrocientos escolares cantabros no han podido **HYP:** a lo largo de la manana unos cuatrocientos sextos tres horas y tenemos su apoyo
7	**GT:** aunque **HYP:** aunte
8	**GT:** su investidura aunque esta no es ni mucho menos **HYP:** todo investidura aunque a esta hoy ser mucho menos
9	**GT:** ofrecer a esta hora desde aqui desde la sede de medicos sin fronteras **HYP:** ofrecer a esta hora desde aqui en esas dias extres piden los fronteras
10	**GT:** dos bajo cero **HYP:** dos bajo cero

Table 5: Transcription from speaker dependent model and speaker independent model. S: sentence. GT: ground-truth. HYP: hypothesis. Blue: new words.

Furthermore, we calculated [6] the number of times a word was said in the training and validation set. Thanks to that, we can find out which words have not been used in the training and validation data set (Table 6). These words are considered new words because they appear for the first time in the test set.

	Training	**Validation**		**Training**	**Validation**
a	517	55	a	2650	247
argumentos	1	1	aqui	206	21
candidato	18	0	aunque	73	2
cargos	3	0	bajo	38	3
coche	4	0	**cantabros**	**0**	**0**
compartian	**0**	**0**	cero	52	3
de	1160	117	cuatrocientos	12	2
defender	6	0	de	6358	540
dos	76	11	desde	224	25
el	582	63	dos	444	51
en	549	65	es	653	139
figuras	1	0	escolares	5	0
futbol	9	0	esta	629	70
hay	50	6	fronteras	3	0
judiciales	4	0	han	530	37
jugadores	6	2	hora	102	10
juicio	1	2	investidura	102	5
libres	1	0	la	3642	337
los	345	38	largo	25	4
malas	**0**	**0**	lo	558	105
o	40	6	manana	217	20
para	174	27	medico	0	1
piso	**0**	**0**	menos	59	5
proceso	**0**	**0**	mucho	32	7
puigdemont	42	4	ni	61	7
que	687	63	no	798	137
recordemos	5	0	**ofrecer**	**0**	**0**
segun	16	0	podido	28	1
su	106	5	sede	27	2
tambien	94	9	sin	136	11
tres	49	4	su	498	31
un	216	25	unos	79	15
venia	2	0			

Table 6: Words that appear in table 5

Speaker-independent and speaker-dependent data were used in this study. Thus, the models obtained with these data were analysed.

[6] Folder with extended information:
https://drive.google.com/drive/folders/1zCIakSa83W3q6UzEM0Bh8T9vc0EA7msy?usp=sharing

Sentences of Experiment 6:

1. The word ‘futbol’ is correctly predicted, although it only appears in the training set, whereas ‘trees y ‘jugadores’ do not. The speaker who says ‘futbol’ has never said this word before, this was checked by looking at speakers who said this word in training. Therefore, the model was able to learn to correctly predict this word from other speakers. That indicates that the features learned of this word are unrelated to the speaker. On the other hand, the word ’compassion’ and ‘peso’ appear for the first time in the test set, they are new words. Both words are not correct because of an error, in the case of ‘compassion’ the letter ‘n’ was deleted and in ‘peso’ the ‘p’ was replaced by ‘m’. This shows that learned features are used to predict new words.

2. The hypothesis of this sentence is absolutely correct. The same thing that happened with the word ‘futbol’ happens with the words ‘coche’, ‘segun’ and ‘venian’.

3. This sentence is totally correct as well. However, it was correctly predicted because this sentence appears in validation and training.

4. The word ‘candidate’ is wrong, although it appears 18 times in training, even ‘puigdemont’ which is repeated 43 times is not correct because the letters ‘r’ and ‘o’ were added, maybe the model added the letter ‘o’ because it is the next word. In addition, the new word ‘malas’ is incorrect as well.

5. In this sentence, ‘de’ is correctly predicted all three times. On the other hand, ‘cargos’ is not correct because the ‘g’ is changed to ‘r’, considering that it only appears 3 times in the training set, it is a good hypothesis. The word ‘procesos’ is a new word, the mistake is to add the letter ‘t’

Sentences of Experiment 8

6. This sentence is correct up to ‘escolares’, all words before this had appeared several times in the training and validation set.

7. It is wrong because the model removed the letter ‘u’ and replaced ‘q’ with ‘t’.

8. In this sentence the word ‘aunque’ is correct, perhaps the context of the sentence helped in the prediction. The model has problems with the set of words ‘ no es ni’ and the word ‘su’.

9. The word ‘ofrecer’ which is a new word and ‘fronteras’ which only appers 3 times are correct.

10. The hypothesis generated by the model is correct. All the words in this sentence were repeated several times in training. The previous sentences and this show that the model learned to predict words that have many repetitions and few repetitions.

5. Discussion

First, we will compare our results with other research that uses the Spanish language. Subsequently, we will discuss whether a model trained on an English dataset can be used to learn a similar language.

Fernandez-Lopez and Sukno [45] obtained the following results CER of 44.77% and WER of 72.9% with a dataset which was acquired in a controlled environment. In contrast, we got a CER of 50.97% and a WER of 83.53%. The results are quite different, but we should consider that the dataset is not the same. Even so, we can see that it is normal for the WER and CER to have values so far apart.

Schwiebert et al. [35] made transfer learning between two languages, English and German. Here, we are going to focus on model LRW_{15} → $GLip_{15\text{-}small}$. It shows a significant improvement of a model trained from scratch. However, the accuracy is very low (corresponds to the light blue line). In contrast, we achieve an accuracy of 39.19% with a similar amount of data (Experiment 7). That may indicate that transfer learning works better when the languages are similar. Because even if English and German have the same origin, they are currently not very similar. Words with the same origin today look and sound completely different, i.e. they differ both orally and in writing, and a high percentage of them have very different meanings. On the other hand, at least 35% of English words have a Spanish word with which they are closely related.

On the other hand, the transfer learning method is more beneficial if the dataset is speaker dependent, with only 2 hours an accuracy of 59.54% was obtained. When the dataset became speaker independent the transfer learning starts to need more data, 2 hours was not enough (accuracy of 39.193%), using 13 hours could be increase 22.404% the accuracy of the model. This indicates that there is a relationship between the size of the dataset and the learning transfer method. More data implies that this method works better when the dataset is speaker independent. But the amount of data needed for this method is not equivalent to the data needed for training from scratch, as it is less.

6. Conclusion

This study demonstrated that AV-HuBERT can be used for continuous lip reading in Spanish by obtaining a model with an accuracy of 61.597%. Furthermore, this model can predict the transcription of some short sentences. Thus, we had to find the most convenient values for the hyperparameters through various experiments.

The transfer learning method is more beneficial if the dataset is speaker dependent, with only 2 hours an accuracy of 59.54% was obtained. When the dataset evolves speaker independent, transfer learning starts to need more data. Two hours was not enough (accuracy of 39.193%). In contrast, 13 hours increases 22.404% the accuracy of the model. That indicates that there is a relationship between the size of the dataset and the learning transfer method. More data implies that this method works better when the dataset is speaker independent. But the amount of data needed for this method is not equivalent to the data needed for training from scratch, as it is less. On the other hand, transfer learning works better when the languages are similar.

The quality of the image plays a significant role in feature extraction. The method used to improve the resolution has some limitations, so on some occasions, it has not been able

to reconstruct the teeth of some speakers. Since we have a limited dataset, we cannot eliminate samples that contain erroneous reconstructions. If we use the whole RTVE dataset, we could replace the wrong samples and increase the training set. Acquiring more videos involves selecting fragments of videos and proceeding to transcribe them, which is a complex and time-consuming process. For this reason, we did not do it in this study because preprocessing the videos would have taken months.

7. Bibliography

[1] MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748. https://doi.org/10.1038/264746a0

[2] Instituto Nacional de Estadística. (2020). Utilización de la lengua de signos por sexo y edad. Población de 6 y más años con discapacidad de audición. INE. https://www.ine.es/jaxi/Datos.htm?tpx=51628#

[3] Flesher, W., Twamley, J., Verlander, M., Nanda, S., Madanapalli Raghunath, S., Williams, A., Laha, S. (2019)."SRAVI - Speech Recognition App for the Voice Impaired. A feasibility study into the use of Visual Speech Recognition software to aid communication in the ICU environment." Intensive Care Society State of the Art Meeting.

[4] McQuillan, L. (2019). Is lip-reading the secret to security? Biometric Technology Today, 2019(6), 5–7. https://doi.org/10.1016/s0969-4765(19)30085-2

[5] Hassanat, A. B. (2014). Visual passwords using automatic lip reading. In *arXiv [cs.CV]*. http://arxiv.org/abs/1409.0924

[6] Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2017). Lip reading sentences in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] Chung, J. S., & Zisserman, A. (2018). Learning to lip read words by watching videos. *Computer Vision and Image Understanding: CVIU*, *173*, 76–85. https://doi.org/10.1016/j.cviu.2018.02.001

[8] Biswas, A., Sahu, P. K., & Chandra, M. (2016). Multiple cameras audio visual speech recognition using active appearance model visual features in car environment. *International Journal of Speech Technology*, *19*(1), 159–171. https://doi.org/10.1007/s10772-016-9332-x

[9] Loomis, J. M., Blascovich, J. J., & Beall, A. C. (1999). Immersive virtual environment technology as a basic research tool in psychology. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *31*(4), 557–564. https://doi.org/10.3758/bf03200735

[10] Lu, Y., & Jiang, H. (2020). Set-top box automated lip-reading controller based on convolutional neural network. In *Advances in Intelligent Systems and Computing* (pp. 422–431). Springer International Publishing.

[11] Pandey, L., Hasan, K., & Arif, A. S. (2021). Acceptability of speech and silent speech input methods in private and public. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

[12] Sun, K., Yu, C., Shi, W., Liu, L., & Shi, Y. (2018). Lip-interact: Improving mobile device interaction with silent speech commands. The 31st Annual ACM Symposium on User Interface Software and Technology - UIST ’18.

[13] Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53–72. https://doi.org/10.1016/j.imavis.2018.07.002

[14] L. Cappelletta, N. Harte Viseme definitions comparison for visual-only speech recognition Proc. European Conference on Signal Processing (2011), pp. 2109-2113

[15] R. Seymour, D. Stewart, J. Ming Comparison of image transform-based features for visual speech recognition in clean and corrupted videos J. Signal Image Video Process. (2008), pp. 14-22

[16] Mingfeng Hao, Mutallip Mamut, Nurbiya Yadikar, Alimjan Aysa, and Kurban Ubul. A survey of research on lipreading technology. IEEE Access, 2020.

[17] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, and Matti Pietikäinen. A compact representation of visual speech data using latent variables. IEEE transactions on pattern analysis and machine intelligence, 36(1):1–1, 2013

[18] Chung, J. S. and Zisserman, A. (2016b). Out of time: automated lip sync in the wild. In Proc. Asian Conference on Computer Vision, pages 251–263.

[19] Petridis, S. and Pantic, M. (2016). Deep complementary bottleneck features for visual speech recognition. In Proc. International Conference on Acoustics, Speech and Signal Processing, pages 2304–2308.

[20] Wand, M., Koutn´ık, J., and Schmidhuber, J. (2016). Lipreading with long short-term memory. In Proc. International Conference on Acoustics, Speech and Signal Processing, pages 6115–6119.

[21] Petridis, S., Wang, Y., Li, Z., and Pantic, M. (2017c). End-to-end Multiview lipreading. In Proc. British Machine Vision Conference.

[22] Stafylakis, T. and Tzimiropoulos, G. (2017). Combining residual networks with LSTMs for lipreading. In Proceedings of Interspeech, pages 3652–3656.

[23] Fung, H. L. and Mak, B. (2018). End-to-end low-resource lip-reading with maxout CNN and LSTM. In Proc. International Conference on Acoustics, Speech and Signal Processing.

[24] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audiovisual speech recognition. IEEE transactions on pattern analysis and machine intelligence, 2018.

[25] Peratham Wiriyathammabhum. Spotfast networks with memory augmented lateral transformers for lipreading. In International Conference on Neural Information Processing, pages 554–561. Springer, 2020.

[26] Souheil Fenghour, Daqing Chen, Kun Guo, and Perry Xiao. Lip reading sentences using deep learning with only visual cues. IEEE Access, 8:215516–215530, 2020.

[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, Aidan N, Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. ArXiv.org. https://arxiv.org/abs/1706.03762

[28] Twaddell, W. F. (1935). On defining the phoneme. Language, 11(1), 5. https://doi.org/10.2307/522070

[29] Chung, J. S., & Zisserman, A. (2017). Lip reading in the wild. In Computer Vision – ACCV 2016 (pp. 87–103). Springer International Publishing.

[30] Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, *11*(4), 796–804. https://doi.org/10.1044/jshr.1104.796

[31] H. L. Bear, R. W. Harvey, B.-J. Theobald, and Y. Lan, ‘‘Which phonemeto-viseme maps best improve visual-only computer lip-reading?’’ in Proc. Int. Symp. Vis. Comput. Cham, Switzerland: Springer, 2014, pp. 230–239.

[32] Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX).

[33] Kannojia, S. P., & Jaiswal, G. (2018). Effects of varying resolution on performance of CNN based image classification an experimental study. International Journal of Computer Sciences and Engineering, 6(9), 451–456. https://doi.org/10.26438/ijcse/v6i9.451456

[34] Jitaru, A.-C., Stefan, L.-D., & Ionescu, B. (2021). Toward language-independent lip reading: A transfer learning approach. 2021 International Symposium on Signals, Circuits and Systems (ISSCS).

[35] Schwiebert, G., Weber, C., Qu, L., Siqueira, H., & Wermter, S. (2022). A multimodal German dataset for automatic Lip Reading systems and transfer learning. In arXiv [cs.CV]. http://arxiv.org/abs/2202.13403

[36] Kumar, A., & Chellappa, R. (2019). Landmark detection in low resolution faces with semi-supervised learning. In arXiv [cs.CV]. http://arxiv.org/abs/1907.13255

[37] Knoche, M., Merget, D., & Rigoll, G. (2017). Improving facial landmark detection via a super-resolution inception network. In Lecture Notes in Computer Science (pp. 239–251). Springer International Publishing.

[38] Wang, X., Li, Y., Zhang, H., & Shan, Y. (2021). Towards real-world blind face restoration with Generative Facial Prior. In arXiv [cs.CV]. http://arxiv.org/abs/2101.04061

[39] Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

[40] Wang, T., Zhang, Y., Fan, Y., Wang, J., & Chen, Q. (2021). High-fidelity GAN inversion for image attribute editing. In arXiv [cs.CV]. http://arxiv.org/abs/2109.06590

[41] Xia, W., Yang, Y., Xue, J.-H., & Wu, B. (2021). Towards open-world text-guided face image generation and manipulation. In arXiv [cs.CV]. http://arxiv.org/abs/2104.08910

[42] Gimeno-Gómez, D., & Martínez-Hinarejos, C.-D. (2021). Analysis of visual features for continuous lipreading in Spanish. IberSPEECH 2021.

[43] Shi, B., Hsu, W.-N., Lakhotia, K., & Mohamed, A. (2022). Learning audio-visual speech representation by masked multimodal cluster prediction. In arXiv [eess.AS]. http://arxiv.org/abs/2201.02184

[44] Lubliner, S., & Hiebert, E. H. (2011). An analysis of English–Spanish cognates as a source of general academic language. Bilingual Research Journal, 34(1), 76–93. https://doi.org/10.1080/15235882.2011.568589

[45] A. Fernandez-Lopez and F. M. Sukno, "End-to-End Lip-Reading Without Large-Scale Data," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, doi: 10.1109/TASLP.2022.3182274.

[46] Makino, T., Liao, H., Assael, Y., Shillingford, B., Garcia, B., Braga, O., & Siohan, O. (2019b). Recurrent neural network transducer for audio-visual speech recognition. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU).

[47] E. Lleida, A. Ortega, A. Miguel, V. Baz´an, C. P´erez, M. Zotano, and A. de Prada, "Rtve2018 database description," Vivolab and Corporaci´on Radiotelevisi´on Espa˜nola, Zaragoza, Spain, 2018, [Online] Available: http://catedrartve.unizar.es/reto2018/RTVE2018DB.pdf.

[48] King, D. E. (n.d.). Dlib-ml: A Machine Learning Toolkit. Mit.Edu. Retrieved June 15, 2022, from https://jmlr.csail.mit.edu/papers/volume10/king09a/king09a.pdf

[49] Martinez, B., Ma, P., Petridis, S., & Pantic, M. (2020). Lipreading using temporal convolutional networks. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[50] Ma, P., Petridis, S., & Pantic, M. (2021). End-to-end audio-visual speech recognition with conformers. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[51] Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[52] Bradski, G. (2000). The OpenCV Library. Dr. Dobb's Journal of Software Tools.