Polytechnic University of Leiria

School of Technology and Management

Department of Electrical Engineering

Master's in Electrical and Electronic Engineering

Deep Learning applied to Visual Speech Recognition

Carlos Manuel Simões dos Santos

Number: 2180284

Leiria, 2023, September

Polytechnic University of Leiria
School of Technology and Management
Department of Electrical Engineering
Master's in Electrical and Electronic Engineering

Deep Learning applied to Visual Speech Recognition

Carlos Manuel Simões dos Santos

Number: 2180284

Dissertation realized under supervision of
Doctor Paulo Jorge Simões Coelho (paulo.coelho@ipleiria.pt) and
Doctor António Manuel Trigueiros da Silva Cunha (acunha@utad.pt).

Leiria, September de 2023

Acknowledgements

First of all, I would like to thank Professor Paulo Coelho for his technical and non-technical support, for believing from the beginning in the possibility of realizing the idea that served as the basis for this work. Plus, he made this task fun to take on. Many times, when I wasn't understanding what was to be done, I would start the task because Professor Paulo Coelho was there.

Secondly to Professor António Cunha, for accepting this challenge of providing technical support remotely, to a student he had never met, and for the pragmatic approach he imposed from meeting to meeting.

Thirdly, I want to thank João Santos Silva (PhD), for the greatest technical help I could have. In addition to showing how to overcome this and that code error, he served as an example of how to approach a problem and thus increased the likelihood of solving problems. It was a pleasure and an inspiration to watch him reason and evolve.

I want to thank the Cenfim Centre of Marinha Grande, in the person of Eng. Carlos Silva. Through the granted permission to use the Robotics Laboratory (RoboLab), it was possible for students and colleagues to pass through and give their contribution, that is the essence of the Database. I now thank my colleagues and students who enriched the Database with variety, clarity and voluntarism, in addition to providing ideas that made the process faster and easier.

Lastly, I would like to thank my wife Elisabete. She carried out all my tasks and assumed most of my family responsibilities, so I could work. I would also like to thank her for her patience, when I was talking about nothing other than the work that follows.

Dedicatory

I dedicate my efforts to my children Beatriz, Guilherme and Violeta, for whom those are potential examples. I also dedicate them to my parents Dulce and Fernando, from whom I inherited and continue to inherit the best examples.

I dedicate my perseverance to my grandfather Alberto, who, as a blind man, made me see that disability is a difficulty to overcome.

I dedicate the results to the hearing-impaired students who, despite such difficulties, strive to learn.

Abstract

Visual Speech Recognition (VSR) or Automatic Lip-Reading (ALR), the artificial process used to infer visemes, words, or sentences from video inputs, is an efficient yet far from being a day-to-day tool. With the evolution of deep learning models and the proliferation of databases (DB), vocabularies increase in quality and quantity. Large DB feed end-to-end deep learning (DL) models that extract speech, solely on the visual recognition of the speaker's lips movements. However, large DB production requires large resources, unavailable to the majority of ALR researchers, impairing a larger scale evolution.

This dissertation contributes to the development of ALR by diversifying training data, on which the DL depends upon. This includes producing a new DB, in Portuguese language, capable of state-of-the-art (SOTA) performance. As DL only shows a SOTA performance if trained on a large DB, whose resources are not on the scope of this dissertation, a knowledge leveraging method emerges, as a necessary subsequent objective.

A large DB and a SOTA model are selected and used as templates, from which a smaller DB (LusaPt) is created, comprising 100 phrases by 10 speakers, uttering 50 typical Portuguese digits and words, recorded and processed by day-to-day equipment. After having pre-trained on the SOTA DB, the new model is then fine-tuned on the new DB. For LusaPt's validation, the performance of new and the SOTA's are compared.

Results reveal that, if the same video is recurrently subject to the same model, the same prediction is obtained. Tests also show a clear increase on the word recognition rate (WRR), from the 0% when inferring with the SOTA model with no further training on the new DB, to an over 95% when inferring with the new model.

Besides showing a "powerful belief" of the SOTA model in its predictions, this work also validates the new DB and its creation methodology. It reenforces that the transfer learning process is efficient in learning a new language, therefore new words. Another contribution is to demonstrate that, with a day-to-day equipment and limited human resources, it is possible to enrich the DB corpora and, ultimately, to positively impact the performance and future of Automatic Lip-Reading.

Keywords: Automatic Lip-Reading, Visual Speech Recognition, Deep Learning, Database, Transfer Learning.

Resumo

A leitura automática de lábios (LAL), o processo artificial para inferir visemas, palavras, orações ou frases a partir do movimento dos lábios de um orador, é uma ferramenta eficiente, mas ainda longe de fazer parte do dia-a-dia. Com o desenvolvimento de modelos de aprendizagem profunda (*deep learning* - DL) e a proliferação de bases de dados (*databases* - DB), os vocabulários aumentam em qualidade e quantidade. Bases de dados de grandes dimensões alimentam modelos de aprendizagem profunda de ponta a ponta, que por sua vez reconhecem o conteúdo do que é dito. No entanto, a produção destas grandes DB requer igualmente grandes recursos, indisponíveis para a maioria dos pesquisadores da LAL, representando um obstáculo para um desenvolvimento em maior escala.

Esta dissertação contribui para o desenvolvimento da LAL, ao aumentar a diversidade de dados para treino, dos quais depende a aprendizagem profunda. Este objetivo subdivide-se na produção de uma nova BD, em língua portuguesa e validação, utilizando-a num modelo de última geração (SOTA). Como a DL só apresenta bons desempenhos com treinos em grandes DB, cujos recursos necessários não estão no âmbito desta dissertação, um método de alavancagem de conhecimento surge como objetivo subsequente e necessário.

Uma grande DB e um modelo SOTA são selecionados e usados como bitola, a partir da qual uma base de dados menor é criada e aprendizagem por transferência é aplicada. A nova DB (LusaPt) é composta por 100 orações, pronunciadas por 10 oradores, proferindo 50 palavras típicas portuguesas, gravadas e processadas por equipamentos de uso diário. Pré-treinado na grande DB, um novo modelo é então obtido por afinação na LusaPt. Para a validação da nova DB, os desempenhos do novo modelo e do SOTA são comparados.

Os resultados mostram que, se o mesmo vídeo for sujeito repetidamente ao mesmo modelo, a mesma previsão é obtida. Mostram também uma clara evolução na taxa de reconhecimento de palavras da LusaPt, de 0% ao inferir com o modelo SOTA sem treino adicional, para mais de 95% ao inferir com o novo modelo.

Além de expor uma forte crença do modelo SOTA nas suas próprias previsões, este trabalho também valida a nova DB e correspondente metodologia de criação, reforça a exequibilidade da aprendizagem por transferência nesta área e que é eficiente na aprendizagem de uma nova língua, logo de novas palavras. Outra conclusão é que, com um equipamento quotidiano e

recursos humanos limitados, é possível que um leque mais alargado de investigadores enriqueça o corpora das DB, que treine e enrobusteça modelos, logo contribua para o desempenho e para o futuro da Leitura Automática de Lábios.

Palavras-chave: Leitura Automática de Lábios, Aprendizagem Profunda, Base de Dados, Aprendizagem por Transferência.

Index

List of Figures

List of Tables

List of Abbreviations and Acronyms

AI	Artificial Intelligence
ALR	Automatic Lip-Reading
ASR	Automatic Speech Recognition
AV-ASR	Audio-Visual Automatic Speech Recognition
BLSTM	Bidirectional Long Short-Term Memory
BGRU	Bidirectional Gated Recurrent Unit
CTC	Connectionist Temporal Classification
CNN	Convolution Neural Networks
DB	Database
DL	Deep learning
DNN	Deep Neural Networks
FC	Fully connected layers
FCN	Full convolution network
Fps	Frames per second
GRU	Gate Control Unit
HMM	Hidden Markov Models
LBP	Local Binary Pattern
LSTM	Long Short-Term Memory

Px	Pixel
RoI	Region of Interest
RQ	Research Questions
RBM	Restricted Boltzmann Machines
RNN	Recurrent Neural Networks
SOTA	State-of-the-art
TCN	Temporal convolution network
VSR	Video Speech Recognition
WRR	Word recognition rate

1. Introduction

As a student of the Master's in Electrical and Electronic Engineering at the Polytechnic of Leiria, the author learnt about Computer Vision and Machine Learning. As a teacher in Cenfim[1], the author taught and teaches deaf people. These were the ingredients for the idea to use those technologies to help develop a solution to the problem. This dissertation is the result of this idea.

Lip-Reading is the process to extract and recognize speech content, based solely on the visual recognition of the speaker's lip movements. Besides hearing-impaired people, regular hearing people also resort to visual cues for word disambiguation, every time one is in a noisy environment or wishes discretion. Automatic Lip-Reading, also known as Visual Speech Recognition (VSR), is the technological process to do the same. Nowadays, VSR contributes to a plethora of applications, from forensic study to face liveness detection, and may include inputs such as audio, environment classification, micro-expressions detecting, or combinations. This dissertation's scope is restricted to visual inputs alone.

1.1 Objectives

Within the realm of DL applied to ALR, four research questions (RQ) emerged: RQ1 - Which methods are more suitable for visual clues only automatic lip-reading?; RQ2 - Which methods are mainly used to support the analysis of lip-reading data? RQ3 - Which methods are specifically studied with the available datasets?; RQ4 - What challenges are still open for lip-reading solutions?

Defining the contribution to the development of ALR as the main objective, the following literature review lead to a definition of the SOTA models and DBs. An open challenge emerged: video speech recognition in Portuguese, a language spoken by 280 million people, in nine countries, and in four continents. This unfolded on the specific objectives of creating a DB in the Portuguese language, and its validation on a SOTA model.

[1] www.cenfim.pt – *Vocational Training Centre for Metallurgical and Metalworking Industry.*

1.2 Dissertation structure

This dissertation composes of six chapters. The first gives an overview of the work specifying the area of intervention, the motivations, and the problem to approach. Presents a description of the goals and the provided contributions by this dissertation.

The second chapter presents the background on: human vocal apparatus; lip-reading history; phonemes, visemes, homovisemes and disambiguation; and audio-visual speech recognition.

The third chapter is dedicated to the technical fundamental concepts and the SOTA. In it is presented the history and analyses of visemes, words, sentences and applications, available databases. It is also presented and compared the traditional and DL methods, and presented a literature review.

The fourth chapter presents the methodology followed to reach a new DB, and the testing results. Here is described the pipeline for the DB(LusaPt) construction, final composition, and validation setting.

The fifth chapter presents and discusses the results of the process. Sequentially, are the performance results of: the SOTA model on SOTA DB; the SOTA model on LusaPt; and the new model on LusaPt.

The sixth and final chapter describes the achievements and presents hypothesis for future improvements and developments.

2. Background

The current chapter provides a succinct introduction of the biological and sound formation-related information, which are crucial topics of this work. The first section presents the components and a brief explanation of the human vocal apparatus and how sounds are produced. The second section concisely presents how the referred sounds make up speech bits and, therefore, speech. Finally, the third section depicts pre-technology lip reading, including a brief historical introduction, the basis, early difficulties, and limitations used as the research drivers and the potential focus of this dissertation.

2.1 The human vocal apparatus

Human speech is one of the most important forms of communication through which we convey information, concepts, and ideas; therefore, it is a key driver of Human Evolution. It also enhances the performance of teaching/learning and, as we are social beings, gossip, which makes up a large portion of our human-human interactions.

Figure 1 presents the components of the human vocal tract or articulators. Articulators, both active (like a tongue) and passive (like a hard palate), interact to produce the wide range of sounds necessary for verbal communication.

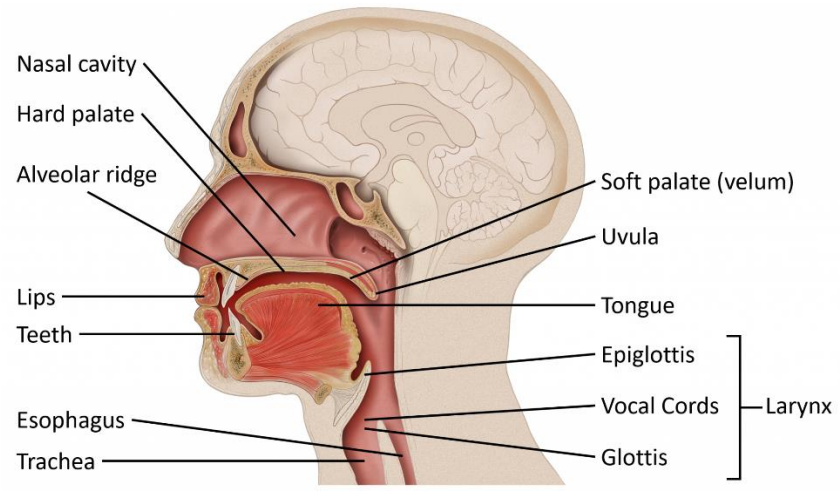

Figure 1 - Parts of the human vocal tract – source [1].

When comparing the physiognomy with other animals or even with human babies, the adult man's larynx position is relatively low, as shown in Figure 2. Contrary to human babies or chimpanzees, who can breathe through their noses while continuing to eat, for adult humans, the pathways to the stomach and the lungs intersect, thus increasing the risks of choking. For Evolution, the descent of the larynx and the resulting ability to speak outweighed the potential for choking in early hominids. It enables the creation of more complex sounds, therefore utterances, words, and ultimately speech and communication.

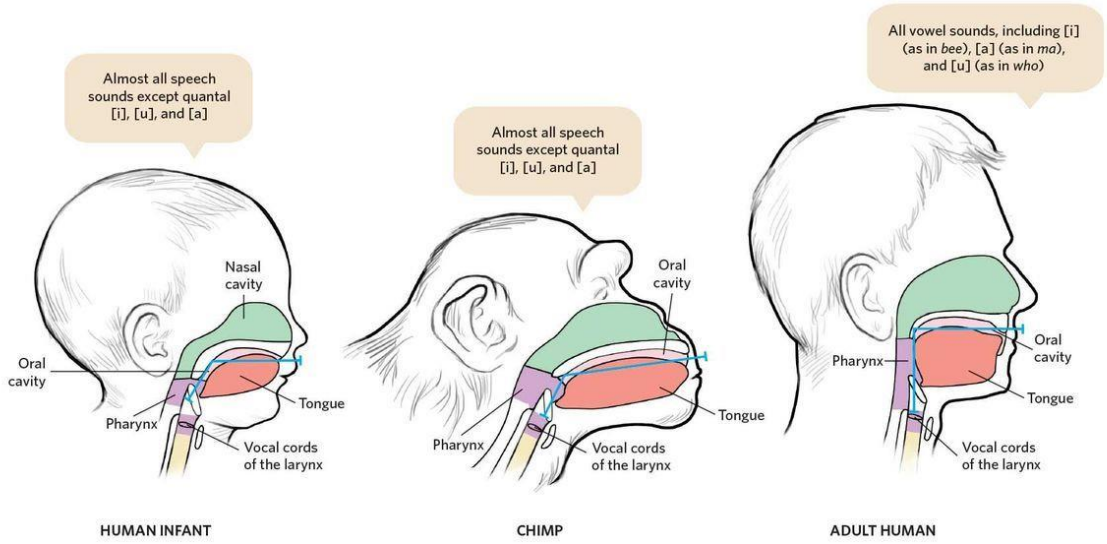

Figure 2 - Larynx relative location – source [2].

The human vocal apparatus may be understood as a conjugation of a wind instrument (lung and corresponding muscles), a string instrument (vocal cords), and a series of chambers that resonate (the pharynx, the mouth, and the nasal cavities), like depicted in Figure 3.

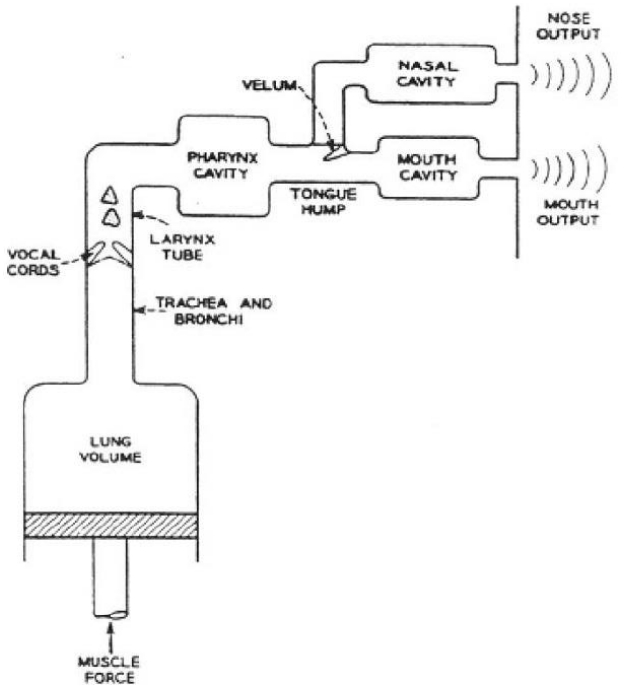

Figure 3 - Simplified model of the human vocal tract – source [3].

Acting as a flow generator, from a fluid dynamics point of view, as the respiratory muscles (diaphragm, rib cage's and abdominals') relax, a small overpressure is produced in the lungs, forcing the air to flow outwards. Further downstream and functioning as a bifurcation control device, the velum (or soft palate) either blocks the passage to the nasal cavity or leaves it open so that the airstream can flow through.

The cavities may produce different sounds. The jaw may open or close, the tongue may change shape or position, and the lips may alter shape by opening, closing, pursing, or stretching. Depending on how and how much of its components are stimulated, these cavities take different geometries, turn into different resonating chambers, and produce different sounds, as one does by blowing different-sized whistles, as depicted in Figure 4.

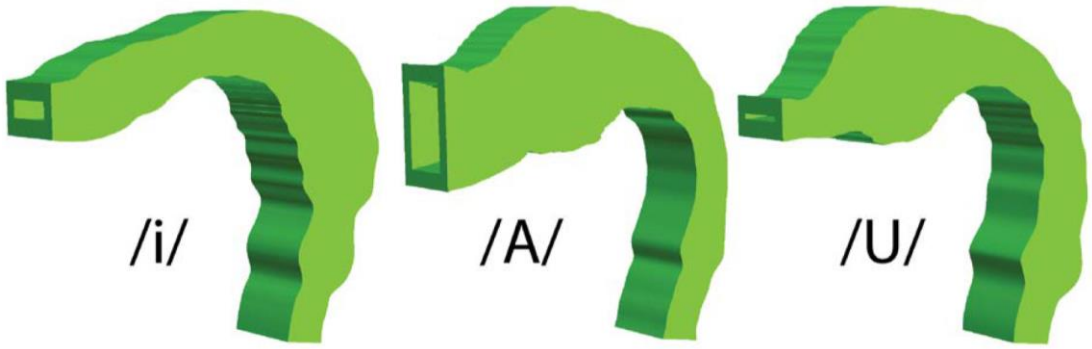

Figure 4 - Mouth cavity shapes for vowel production – source [4].

The position and tension of the vocal cords can also be controlled. As they are tightened, they open and close in superior frequencies, provoking closer (smaller) compression/ depression airwaves, therefore producing higher pitches, as it happens when shortening or tightening the strings from a guitar.

Figure 5 presents artificial human-like vocal cords. Mechanically, tension can be increased or loosened by pulling the cords, generating higher pitches or no sound at all, as the vibrations stop.

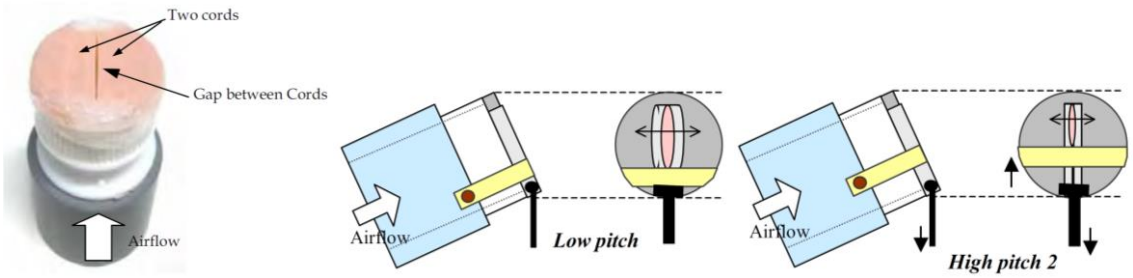

Figure 5 - Artificial vocal cords – source [5].

Hideyuki Sawada also produced an example of a mechanical mouth. Although it lacks some of the human vocal tract components responsible for some kinds of sounds, it serves as proof of concept.

Figure 6 shows the mechanical voice system. It is observed that the flow generator (Air Pump), the vocal cords (as shown above), the Phoneme-Motor and its five pistons (which enables the Resonance Tube's different configurations), and piston #6 (which acts as the velum).

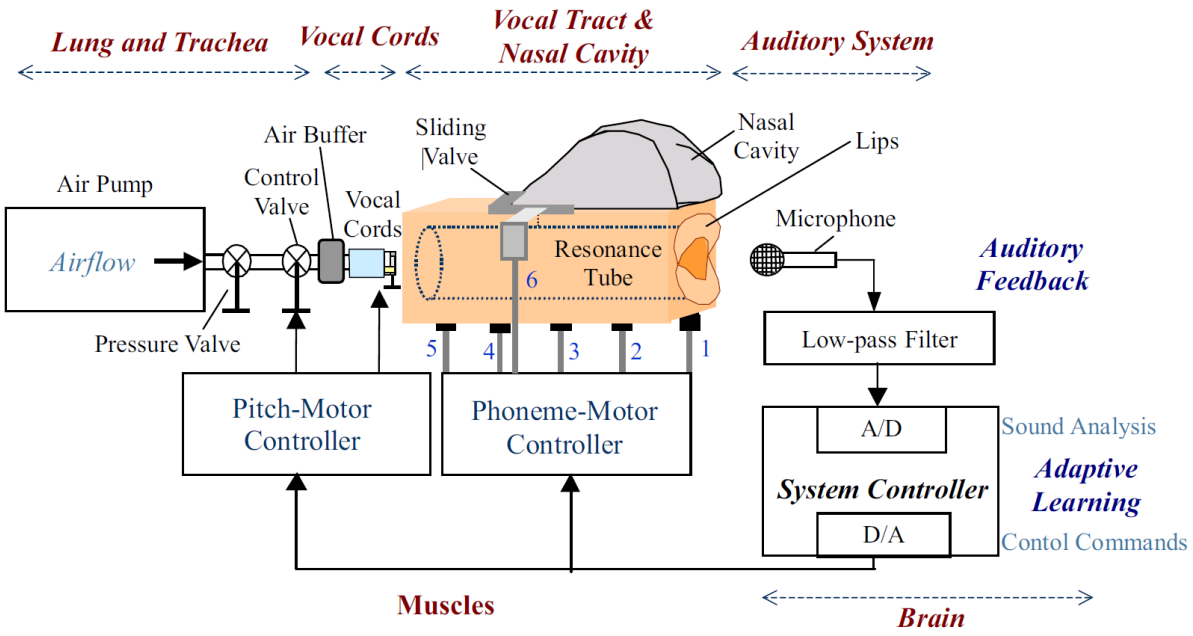

Figure 6 - Configuration of the talking robot – source [5].

2.2 The sounds of language – Phonemes

Phonetics refers to the study of speech sounds and their production. This subsection illustrates how humans, by combining the techniques described in the previous section, produce the phonemes of spoken language. Phonemes are basic speech structures (or minimal units of speech) that enable to distinguish one word from another. For example, the phoneme /p/ distinguishes the word *pat* from *bat* [6].

First, just considering whether the way the air flows through the vocal tract is obstructed or not, there are two basic categories of sound: Vowels and Consonants. Vowels are sonorant phonemes produced without any obstruction. Obstruent phonemes are produced with modification to the airflow and are called Consonants. This obstruction results from the physical contact between articulators and parts of the vocal tract, called places of articulation, as illustrated in Figure 7.

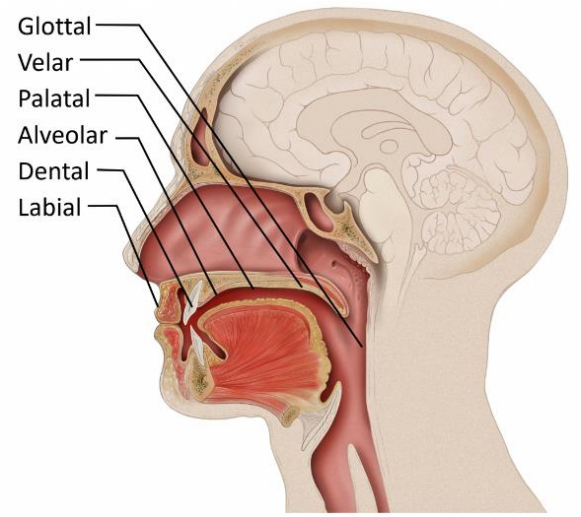

Figure 7 - Places of articulation – source [1].

Consonants are further subdivided by the manner of articulation. A consonant produced with a strong burst of breath is called Aspirate, with the airflow blocked before release is called Stop, with the nasal passage open along with the oral tract is called Nasal, with the forcing of air through a narrow gap between two articulators is called Fricative and is called Affricative if it begins with a Stop and releases a Fricative. Liquid consonants are produced by lateral approximation of articulators but no touching. Figure 8 demonstrates the most commonly used hierarchization for phoneme classification in the English language.

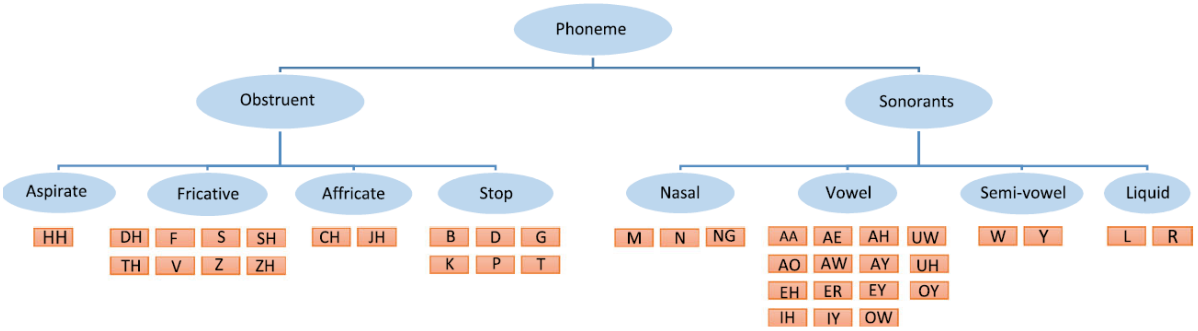

Figure 8 - Classification of phonemes in the English language – source [7].

The combination of articulators and places of articulation gives each phoneme its characteristic sound and, therefore, the variety of consonants. Figure 9 shows most of these, which occur in the human mouth and corresponding phonemes.

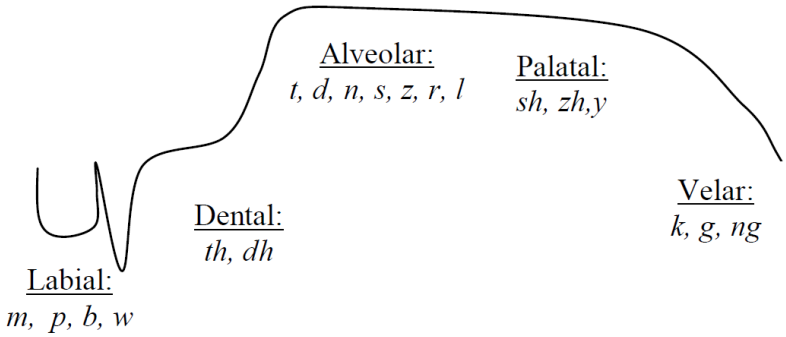

Figure 9 - Major places of consonant articulation – source [6].

Table 1 presents the 44 phonemes (20 vowels, 24 consonants) in English, as this is the language of this written document.

Table 1 - The 44 phonemes of the English Language – source [8].

	monophthongs				diphthongs			Phonemic Chart voiced unvoiced
VOWELS	iː sheep	ɪ ship	ʊ good	uː shoot	ɪə here	eɪ wait		
	e bed	ə teacher	ɜː bird	ɔː door	ʊə tourist	ɔɪ boy	əʊ show	
	æ cat	ʌ up	ɑː far	ɒ on	eə hair	aɪ my	aʊ cow	
CONSONANTS	p pea	b boat	t tea	d dog	ʧ cheese	ʤ June	k car	g go
	f fly	v video	θ think	ð this	s see	z zoo	ʃ shall	ʒ television
	m man	n now	ŋ sing	h hat	l love	r red	w wet	j yes

Different languages may present different phoneme numbers. For example, Turkish presents 31 phonemes (8 vowels, 23 consonants), and the Portuguese Language is 37 phonemes rich (14 vowels [9], 23 consonants [10]).

2.3 Lip Reading

Lip reading or visual speech recognition is a technique used to understand or interpret speech by analysing the movement of lips. It is used to complete relayed information by people with hearing difficulties, either by being congenitally deaf or just by a significant decrease in speech-to-noise ratio, i.e., by augmenting the noise level and maintaining the speech level or by maintaining the noise level but diminishing the speech level [11]. Bauman's study reported that hearing-impaired people understood 21% of speech just using residual hearing, 64% if they

combined residual hearing with either a hearing aid or with speechreading, and 90% if they used their residual hearing, hearing aids, and speechreading [12].

2.3.1 History

Lip reading is not a newly sought skill. In the author's opinion, already in pre-history, humans started having the necessity to understand information transmitted by whispering or in complete silence, for secrecy intents. The first successful lip-reading teacher was Pietro Ponce, a Spanish Monk of the 16th Century. Lip reading teaching spread to other European countries, and Samuel Heinecke opened the first lip-reading school in Leipzig, in 1787. On the other hand, Charles-Michel de l'Épée, an 18th Century French Abbe, introduced an oral method and sign language, still in use today [13].

Different methods such as the Muller-Walle have been described in the literature for human lip reading. Muller-Walle's method is based on movement rather than positions. It focuses on the classification of Direction, Time, Measure, Duration, Rhythm, and Classification. The method also emphasizes the importance of Position, Light, The practice in general conversation, and Expression as hints to pay attention to improve intelligibility. Despite all the progress, one of the reported most significant difficulties is following a conversation when the subject is unknown, a common challenge to today's approach [14].

With the consistent development of skills, pupils leave such schools equipped with the aid of the oral method and with the art of reading the speech from the lips, when at the beginning, it seemed out of the question for children who were born deaf or for adults who became deaf. Through continuous practice, students may fine-tune to see smaller and less significant movements in the mouth with greater rapidity and accuracy.

2.3.2 Audio-visual Speech Reading

Audio-visual speech perception is the composite of perceptual and intellectual processes by which people understand speech when relying on visual and hearing senses. As mentioned above, humans rely on vision to complete information.

McGurk and MacDonald (1976) experimented with this bi-modality by giving test-subjects audio and visual incongruent information [15]. The study demonstrates some degree of influence of vision upon speech perception. As an example, given [*ba*] as audio stimuli and [*ga*] as visual stimuli, test subjects fused onto [*da*]. Then, by closing the eyes, [*da*] became [*ba*],

only to revert to [*da*] when the eyes were opened again. One other example is shown in Table 2. The combination of acoustical "*map*" with optical "*tap*" was mainly identified as "*nap*", while the reverse assignment, acoustical "*tap*" with optical "*map*", yielded "*pap*".

Table 2 - Classification on incongruent synchronized information

	Vision	**Audition**	**Reported**
	t	*m*	*n*
Place	Alveolar	Bilabial	Alveolar
Voicing	Voiceless	Voiced	Voiced
Nasality	Non-Nasal	Nasal	Nasal
	m	*t*	*p*
Place	Bilabial	Alveolar	Bilabial
Voicing	Voiced	Voiceless	Voiceless
Nasality	Nasal	Non-Nasal	Non-Nasal

According to Figure 9, Table 2 also shows the mismatch (in red) in articulation places, for when the visual and audio information are not in par.

2.3.3 Visemes

Although humans can visually distinguish consonants drawn from different groups (e.g., /p/ from /w/), we cannot do the same to those drawn from the same group (e.g., /p/ from /b/). Consonants such as /p/, /b/, and /m/ generally cannot be visually distinguished, are called ”homovisemes” or ”homophemes” and constitute a single viseme. Viseme is a shortened version of the phrase visual phoneme and refers to any individual and contrastive visually perceived unit [17]. It can also be understood as the mouth shape (or appearance) or sequence of mouth dynamics required to generate a phoneme in the visual domain [14] or as the shortest visually recognizable part of speech.

Figure 10 presents only 14 visemes against the previously mentioned 44 phonemes for the English language. For example, for the /b/, /m/, and /p/ phonemes, there is a single viseme for processing the words bop, mop, and pop [17].

Viseme	Phoneme	Output	Viseme	Phoneme	Output
Ah	ɑ, ɔ, a		LNTD	l, n, t, d, ʃ, ʟ, ɾ	
Aa	æ		GK	g, k, ŋ, q, ɢ	
Eh	e, ɛ		MBP	b, m, p	
Ee	i		R	ɹ	
Ih	ɪ		WA_PEDAL	w, ʊ, ʍ	
Oh	o, ɒ		JY	j, dʒ, c, ɟ	
Uh	ʊ, ʌ, ɞ, ɐ, œ, ö, or, ɨ		S	s, z, ɣ	
U	u		ShChZh	ʃ, tʃ, ʒ, ɬ, ɮ,	
Eu	œ, y, ɯ, ø, ɵ		Th	θ, ð	
Schwa	ə, ɘ		FV	f, v, ɱ	

Figure 10 - Audio Driven Animator – source [17].

These visemes cover a small subspace of the mouth's motions, which contributes to the poor performance of the visemic approach when lip-reading is based on visemes. By analogy, it is as digitalizing the signal of the spoken word, with an insufficient resolution, leads to information loss [14]. An alternative is the holistic approach, such as the "visual words", which considers the signature of the whole word and presents better performances [18].

3. Fundamental Concepts and State-of-the-art

The state-of-the-art, to which a major part of this chapter is dedicated, is understood as the definition of the starting point of the author's dissertation. The objective is to present the technical developments of lip-reading systems, where Automatic Lip-Reading technology stands nowadays, what is the author's view on all the literature to be read.

The chapter is divided into six parts, worked and structured to provide the most complete, accurate, coherent, and simple-to-understand overview of the Automatic Lip-Reading possible. The first part is a historical introduction, presenting the initial works that proved that technology-based Lip Reading Systems were possible, and set the standards for what laid ahead. The second part presents the basic understanding of the technology presented forward. The third part is dedicated to traditional-based ALR systems, and the fourth is dedicated to DL based ALR systems. The fifth part is dedicated to the author's literature review, which updates and deepens the study presented in the two previous parts. The sixth part is dedicated to presenting the conclusion drawn from those mentioned above and the research questions to be answered in the coming research on which the next chapters are built.

3.1 – History

Nassimbene [19] signs the first technological trial towards Automatic Lip-Reading. The patented invention consists of a device for determining the position of the facial parts during speech as an adjunct to voice recognition. It measures the reflectivity from the surface of the oral cavity. This invention may unveil mouth attitude indications, which, combined with the acoustic output of a microphone voice reader, may be understood as a rudimentary ALR system.

Figure 11 shows the invention as ideally used by an operator, where 11 and 12 represent source lights, and 13 is the microphone case and the emitters and corresponding receivers. For example, sensor 12' receives light only if the mouth is open, but sensor 11' only does so if the mouth is closed.

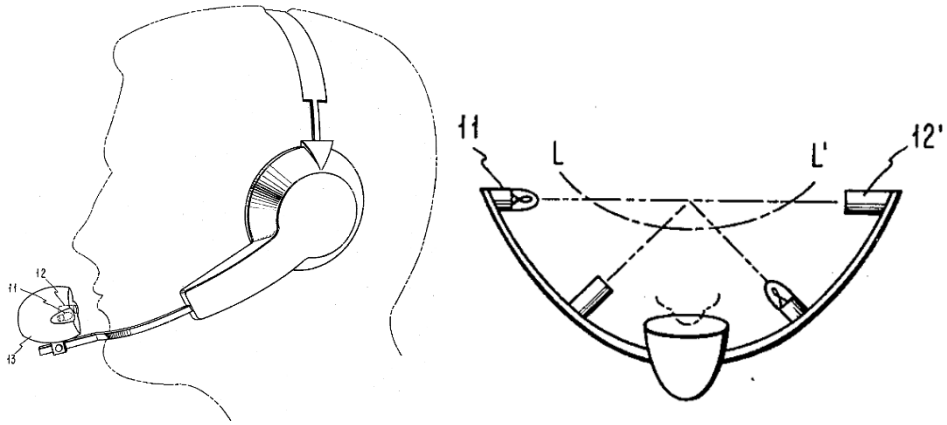

Figure 11 - Pictorial representation of the invention – source [19].

Automatic Lip-Reading or Video Speech Recognition are terms used when inferring speech (utterances, digits, words, phrases, or sentences) using Artificial Intelligence (AI). Petajan [20] describes the first attempt at lip-reading automatically, though to enhance Automatic Speech Recognition. The ALR system was developed for speaker-dependent isolated utterance recognition (as the minimum recognition unit). In the training phase, the method acquired video data samples and reduced them to a template of visual speech parameter time sequences. In the testing phase, it carries out the Nearest Neighbour search between the incoming template and all the trained templates to achieve a recognition candidate (prediction).

The accuracy of the combination of audio and visual recognition predictions was demonstrated clearly to exceed the accuracy of the predictions of audio input alone. Figure 12, adapted from the paper, shows how raw data was 36 years ago. Nevertheless, ALR was making its way and proving its point [20].

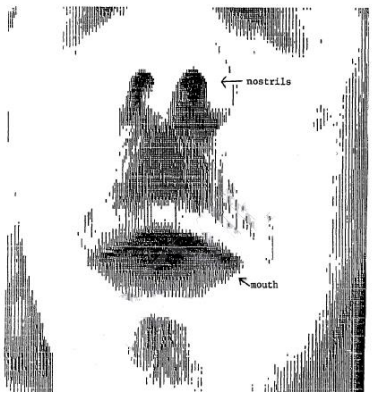

Figure 12 - Unsmoothed facial image in 4 layers of grey – source [20].

3.2 – Development

Automatic lip-reading or visual speech recognition systems may often be divided into:

i) Visual Input - Videos of people uttering the speech to be decoded [21]. The videos are divided into frames, susceptible to selection to reduce redundancy;

ii) Audio Input – Voice recordings of the utterances described above. This input is included parallel to visual input in Audio-Visual Automatic Speech Recognition (AV-ASR), is exclusive in Automatic Speech Recognition , and is excluded from VSR;

iii) Pre-processing [22] – Processing of the raw image data to locate and extract the Region of Interest (RoI), generally comprising the lips. Transformations such as cropping can be performed to reduce the number of operations for the subsequent operations;

iv) Feature Extraction (Frontend) – Process to extract meaningful features, transforming high-dimensional image data into a lower-dimensional representation;

v) Classification (Backend) - Process to establish correlations between the extracted features as observations to infer speech. Speech is decoded in classes or units, eventually encoded as words or sentences.

Thanks to the availability of larger audio-visual datasets with continuous speech, later lip-reading systems have focused on classifying entire sentences employing a wider range of vocabulary. Figure 13 depicts the interpretations of lip movements and the classification schema for lip-reading.

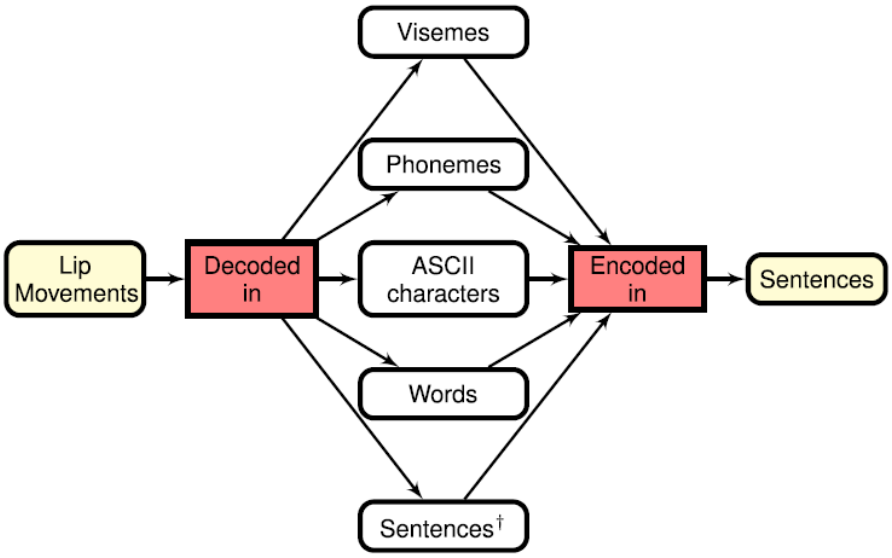

Figure 13 - Different classification schema – source [23].

3.2.1 – Lip-reading Visemes

For ALR, using visemes as classes has advantages as [24]:

i) Fewer classes are needed, which can reduce computational bottleneck when compared to ASCII characters and words;
ii) Pre-trained lexicons are not required;
iii) Because many different words, including from different languages, share the same visemes. Theoretically, a viseme-based lipreading system can classify unknown words, including those from different languages.

However, the one-to-many relationship between sets of visemes and words turns the classification much dependent on viseme-to-word accuracy.

The general classification performance for individually segmented visemes has been less satisfactory than the classification of words since visemes tend to have a shorter duration than words. This results in there being less temporal information available to distinguish between different classes, as well as there being more visual ambiguity when it comes to class recognition. One possible way to address this problem is to significantly increase the training data available to enhance the system's ability to distinguish between classes, and this is why a high volume of training videos has been utilized.

3.2.2 – Lip-reading Words

Lip reading systems designed to classify words often use individual words as the classification schema, where every word is treated as a class. Contrastingly, however, lip-reading sentences have not succeeded in attaining accuracy and word-based approaches [25].

It remains an ongoing challenging task to automatically lip-reading people uttering sentences that cover a wide range of vocabulary and contain words that may not have appeared in the training phase while using the fewest classes possible.

3.2.3 – Lip-reading Sentences

The main obstacles to lip-reading sentences are [25]:

i) Systems that use words or ASCII characters as classes can only predict trained words;
ii) A significant vocabulary size requires a significant number of parameters in the models to be optimized and a significant volume of training data to be used;

iii) They often require curriculum learning-based strategies that involve further pre-processing.

3.2.4 – Lip-reading Applications

Lip-reading, nowadays, has various applications, such as:

i) Speech synthesizer for people able to move lips but unable to utter;
ii) Multi-view mouth rendering as assistance to people with hearing disabilities;
iii) Lip motion silent passwords;
iv) Audio-less video transcriber and re-dubber;
v) Speech recognition under noisy conditions;
vi) Isolation of individual speakers from multi-talker simultaneous speech;
vii) Extracting speech from surveillance videos for forensic study;
viii) Face liveness detection for security (e.g., the system may present a word/phrase that one must reproduce, which prevents hacking with recorded video).

3.2.5 – Databases

During this study, limited time was dedicated to deeply addressing the available databases. There are databases of major and minor sizes, labeled and unlabelled, more complex and simpler. Thus, the author considers it relatively easy to comply with whatever requirements emerge, and this will not be a limitation to the scope of this study. Surely, a more powerful dataset allows for a more robust deep learning model. However, historically all the initial models were trained in modest datasets, and what works with a small dataset will work better in a powerful dataset.

3.3 – Traditional lip-reading methods

The first ALR methods are called Traditional. They are non-deep learning models based on handcrafted features, and are unreliable under unconstrained conditions. Although becoming overcome by DL-based ALR systems, traditional methods are briefly presented for a complete picture understanding and as a DL's methods prelude.

Traditional lipreading methods milestones are [26]:

i) In 1954, Sumby and Pollack [11] first proposed that lip motion features could be used to identify the speaker's speech content;

ii) In 1984, Petajan [20]extracted features from lip movement and combined them with speech recognition to form the first AV-ASR system. Results showed more robustness than ordinary speech recognition systems;

iii) Goldschen et al. in 1994, realized lipreading by using the extracted motion features as the input of the Hidden Markov Models (HMMs)[27], and in 1997, used HMM to model features and achieved good recognition results[28];

iv) In 1998, Potamianos et al. [29] studied a visual front-end of automatic lip-reading based on the HMM. They proposed two methods for extracting lip features: the feature method based on the lip contour and the method based on image change;

v) In 2007, Zhao et al. [30] used a new feature representation method based on the spatiotemporal Local Binary Pattern (LBP) to solve the problem of isolated phrase recognition. They used a Support Vector Machine to recognize phrases.

The main traditional lip-reading steps are as follows: lip detection extraction, feature extraction transformation, and classification, as shown in Figure 14.

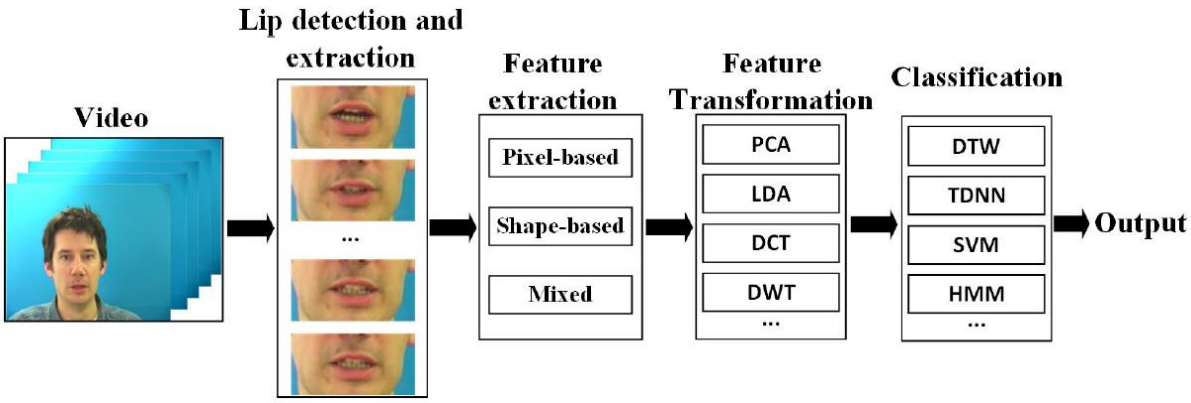

Figure 14 - Traditional lip-reading process – source [26].

3.4 – Deep Learning based lip-reading methods

As the datasets became more complex, the number of speakers increased, the posture became diverse, and the lighting conditions changed. The traditional, manual feature extraction methods became overwhelmed and inefficient. The researchers found that the deep learning method can learn deeper features from the experimental data, which shows good robustness in the case of big data [31]. Instead of manually designing feature extraction methods, researchers are turning

to the deep network's powerful representation learning ability to automatically learn good features according to the task objectives [26].

This work uses the word recognition rate (WRR) for accuracy metrics. This is calculated by Equation (1):

$$WRR = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \times 100\% \quad (1)$$

Deep Learning's milestones are:

i) In 2011, Ngiam et al. [31] proposed an AV-ASR system, building the front-end based on depth Autoencoder and Restricted Boltzmann Machines (RBMs). The visual feature extraction method based on deep learning is introduced into multimodal speech recognition for the first time, superseding traditional feature extraction techniques like PCA.

ii) In 2014, Noda et al. [26] used Convolution Neural Networks (CNN) for feature extraction. The experimental results significantly improvement from traditional methods, including PCA.

iii) In 2016, Wand et al. [26] Long Short-Term Memory (LSTM) for lipreading, achieving a 79.6% WRR on the GRID corpus [32]. In the same year, Chung and Zisserman established LRW, the first large-scale lipreading DB under natural conditions, according to the BBC program.

iv) In 2017, Assael et al. [33] proposed LipNet. This first end-to-end sentence-level lipreading model simultaneously learns spatiotemporal visual features and a sequence model on the GRID corpus, achieving a 95.2% sentence-level accuracy. In the same year, Chung et al. [34] proposed the Watch, Listen, Attend and Spell (WLAS) network, composed of CNN and Recurrent Neural Networks (RNN), which obtained a 46.8% sentence accuracy rate on the LRS database [34], [34], [35] with 104 sample sentences.

v) In 2019, Yang et al. [36] established LRW-1000, the largest Chinese lipreading database under natural conditions according to the China CCTV program.

Petridis et al. [37] categorize deep models into three generations:

1st Generation - Deep bottleneck architectures reduce the dimensionality of audio-visual features extracted from the RoI and the audio signal. These features are then fed to a classifier

like an SVM (ignoring the temporal dynamics of speech) or an HMM (considering the temporal dynamics of speech) ;

2nd Generation - Deep bottleneck architectures extract bottleneck features directly from the pixels. As examples, bottleneck features are extracted from raw mouth RoIs (via deep feedforward network) and fed to the LSTM network for classification, or bottleneck features are extracted from dynamic representations of images (via CNN) and fed to HMM or HMM together with audio features for utterances classification ;

3rd Generation –The main approaches can be divided into: fully connected layers (FC) to extract features and LSTM layers to model the temporal dynamics of the sequence; 3D CNNs or 3DCNNs followed by residual networks (ResNet) and then combined with LSTMs or GRU.

The lipreading flow framework based on the deep learning method is shown in Figure 15.

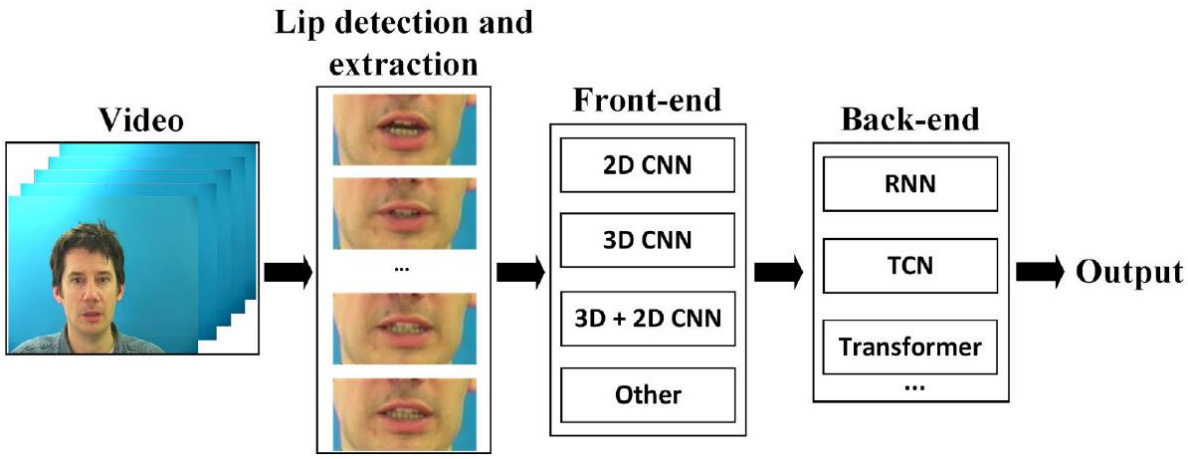

Figure 15 - Deep learning-based lipreading process – source [38].

3.4.1 – Lip detection and extraction

Deep learning-based ALR benefits from technology evolution to extract the RoI with pre-trained models applied for face detection. For example, the Dlib library is used to detect 68 landmarks of the face, 20 of those from lips alone, to serve as the front-end input [26].

3.4.2 – Front-end

The front-end works as a filter to capture local correlations along the spatial dimensions. It can also be extended from 2D to 3D convolutions, capturing the temporal correlations among successive images on top of the spatial correlations in an image [37].

The most common Deep Neural Networks (DNN) applied to the front-end network are:

i) Feed-forward neural network - the most basic neural network that can be used for feature extraction, it simply compresses image data without being able to learn the spatial and temporal features needed for processing sequential inputs [21]. Figure 16 is an example of this type of network in two different styles. On the left, every node is drawn, and on the right, a node is drawn for each entire vector. Matrix W describes the mapping from *x* to *h*, and vector w describes the mapping from *h* to *y*.

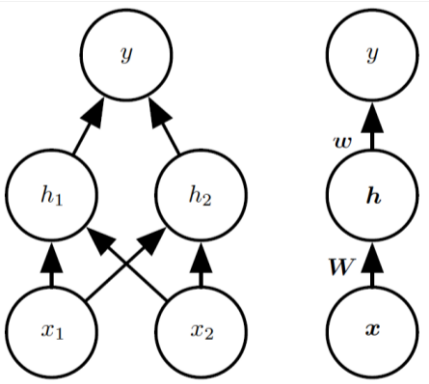

Figure 16 - An example of a feedforward network – source [39].

ii) Autoencoders - networks used for learning compressed distributions of data, consisting of an encoder (which converts data in higher-dimensional space to lower-dimensional space) and a decoder (which converts it back to the original format). They are trained to attempt to copy its input to its output to preserve as much information as possible when an image is run through the encoder. Then, the decoder, but also to make the new representation have various properties. Different autoencoders aim for different kinds of properties [39].

Figure 17 represents the generic autoencoder structure, where f(x) is an encoder function and g(h) is the decoder that produces a reconstruction r.

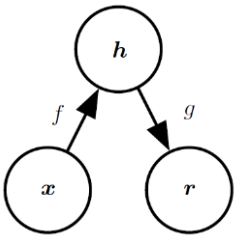

Figure 17 - The generic autoencoder structure – source [39].

If the autoencoder simply learns to set g(f(x)) = x everywhere, it is useless. Autoencoders are designed to restrict perfect copy learning and only copy input that resembles the training data. As it prioritizes which input aspects should be copied, it often learns useful properties of the data [39].

Boltzmann Machines [21] have a similar structure to Autoencoders; however, they use stochastic units that make random decisions with a particular distribution (mainly Binary of Gaussian) instead of a deterministic distribution.

iii) Convolutional Neural Networks - Convolutional Networks or ConvNets, stand out as an example of neuroscientific principles influencing deep learning. These neural networks use convolution instead of general matrix multiplication in at least one of their layers [39] and have been the most common and effective network architecture for feature extraction [31]. An example of 2-D convolution is shown in Figure 18. Drawn are the input matrix, the 2x2 kernel filter (convoluted along with the input elements), and the output tensor.

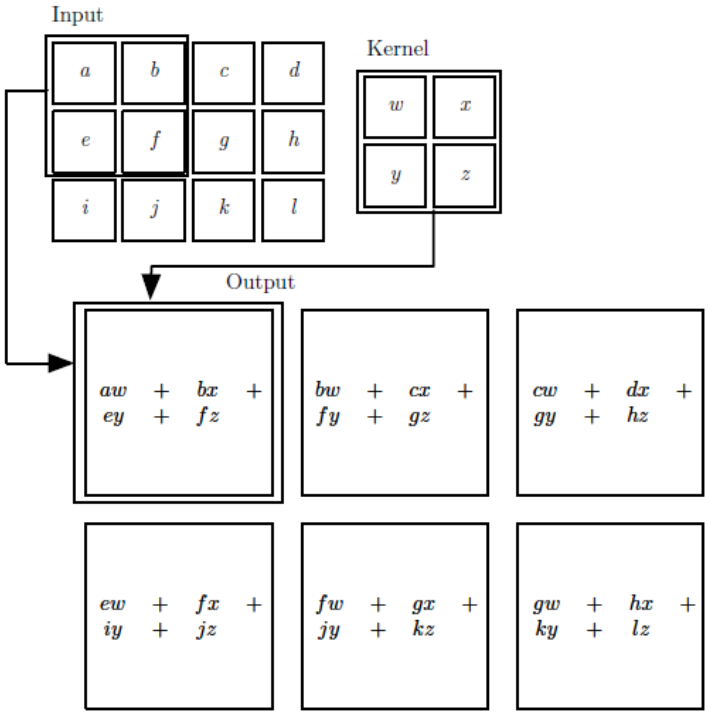

Figure 18 - 2D Convolution example – source [39].

The main examples of 2D CNN structures include ImageNet[2] winners. ImageNet is a large-scale ontology of images built upon the backbone of the WordNet structure [40] and a Large Scale Visual Recognition Challenge. ImageNet challenge started in 2010 with image classification, however, the scope was broadened to video classification since 2015. A time dimension was added to process the temporal information in video, giving way to 3D CNN, and showing the versatility of CNNs.

3.4.3 – Back-end

The back-end network mainly models the features extracted from the front-end network in time, decodes the long-term dependence, and then ascribes speech from facial movements that have been transformed into a lower-dimensional feature vector.

The most common back-end solutions are:

i) Recurrent neural networks - CNNs have no memory, and each input is processed independently, with no state kept between inputs, so to process a sequence or a temporal series of data points, the entire sequence must be shown to the network at once, and turn it into a single data point. In contrast, while you're reading the present

[2] https://www.image-net.org/

sentence, you are keeping memories, thus having an evolving representation of its meaning [41].

"Biological intelligence processes information incrementally while maintaining an internal model of what it's processing, built from past information and constantly updated as new information comes in" [41]. An RNN adopts the same principle, as shown in Figure 19.

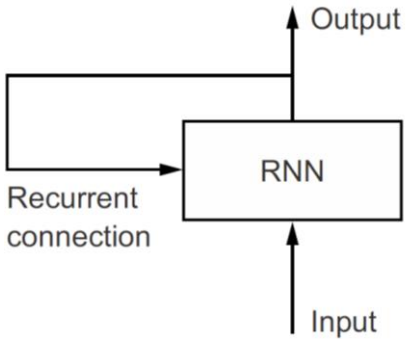

Figure 19 - A Recurrent Neural Network – source [41].

The output of a given step is considered with the input of the next step, which allows the network to remember previous steps in a sequence.

RNN architecture, by learning conditional dependencies, gains discriminative power when distinguishing between classes [21] and is mainly used to deal with the timing problem. However, the common RNN only has a short-term memory, hence has difficulties in learning the long-term dependences [31].

ii) Long Short-Term Memory – an improved RNN, which establishes a state of information and gates to control the state of information and output at different times to solve the short-term memory problem. It uses the forget gate, which determines if the new state information may be forgotten because it is irrelevant, the input gate which determines what new information should be added or updated to the internal state, and the output gate which determines which information part should be output [41].

Figure 20 illustrates the working principle of an LSTM cell. The input feature is commonly computed, and its value will be considered in the state if the input gate determines it (input gate tends towards 1) or not (input gate tends towards 0). By determination of the forget gate, the self-loop may be disregarded disabling the update, and by determination of the output gate, the output of the cell can be shut off.

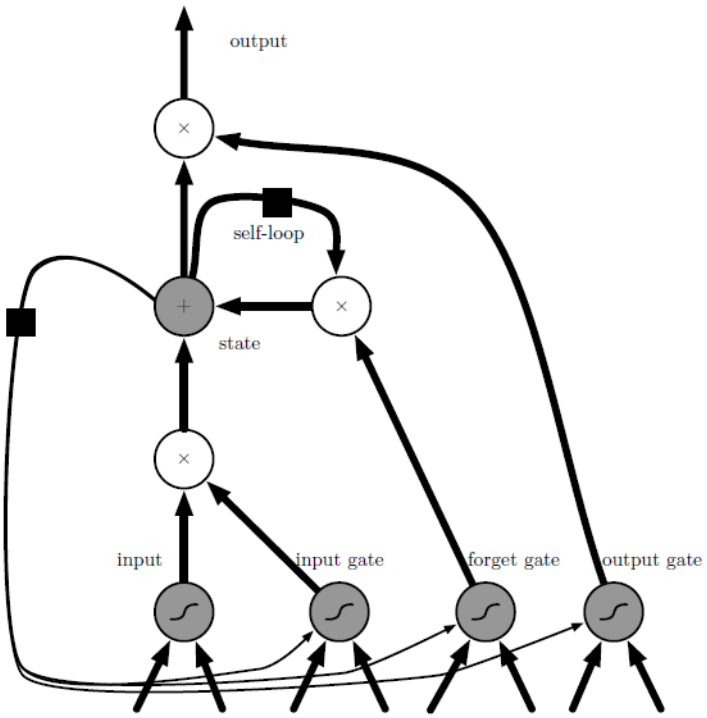

Figure 20 - Block diagram of an LSTM – source [39].

LSTM uses gate structure to combine long-term and short-term memory, which addresses the problem of vanishing gradient, which happens whenever many layers are deep. As layers keep adding up, the network eventually becomes untrainable.

An improvement of LSTM is the Gate Control Unit (GRU), which simplifies the gates into two: update gate and reset gate [26], [31].

In classical RNN, the state's transmission is one-directional only, typically from front to back. However, the output may depend on both previous and posterior factors, hence the evolution towards bi-directional recurrent neural networks (Bi-RNN), such as Bi-LSTM and Bi-GRU [31];

iii) Attention Mechanisms – The human eye is mostly low resolution, except for a tiny patch that only observes an area about the size of a thumbnail held at arm's length. The human brain makes several eye movements to glimpse the most task-relevant parts of a scene [39]. Incorporating similar attention mechanisms into deep learning models allows the limitation of the processing space to the region of importance, drastically reducing computational complexity and noise by excluding irrelevant parts of the visual scene from processing, allowing a contextual representation of the scene without 'clutter' [42].

In ALR, the length of the image series may differ due to the different speech speeds of each person, so to learn how to align predictions of an input sequence temporally, an attention mechanism may be used [31];

iv) Transformers – A new trend in the use of Transformers has emerged in some of the most recent approaches to lip-reading classification, appearing to replace RNNs in many lip-reading systems. They are designed to take advantage of Graphics Processing Unit, allowing parallel computation by processing entire inputs simultaneously, rather than sequentially, as RNNs do. Transformers require less time to train, by bypassing recursion, are better at capturing long-term dependencies and modelling long-range global context. However, they are less capable of extracting fine-grained local feature patterns due to not considering local information [42];

v) Temporal convolution network (TCN) – TCNs have emerged as a promising alternative to LSTMs, taking a time-indexed sequence of feature vectors as input and maps it into another such sequence (i.e., the length of the sequence is not altered) through the use of a 1D temporal convolution [43].
As well as Transformers, TCNs have the advantage of parallel processing over every timestep sequentially RNNs' processing. Another advantage is the flexibility in changing receptive field size, as TCNs are not exposed to exploding or vanishing gradients problems [21];

vi) A full convolution network (FCN) or deep filter – computes a nonlinear filter, while a general deep net computes a general nonlinear function. Each layer of data in a CNN is a 3D array of size $w \times l \times d$. The first layer is the image, with pixel size $w \times l$, and d colour channels. Locations in higher layers are called receptive fields. When these overlap significantly, feedforward computation and backpropagation are much more efficient when computed layer-by-layer than independently patch-by-patch [44].
Long et al. [44] built an FCN, whose input has arbitrary size, trained the end-to-end network for pixel-wise prediction and from supervised pre-training, and produced correspondingly sized output with efficient inference and learning. Figure 21 shows Semantic Segmentation, a classification of the object class for each pixel within an image, a per-pixel labelling.

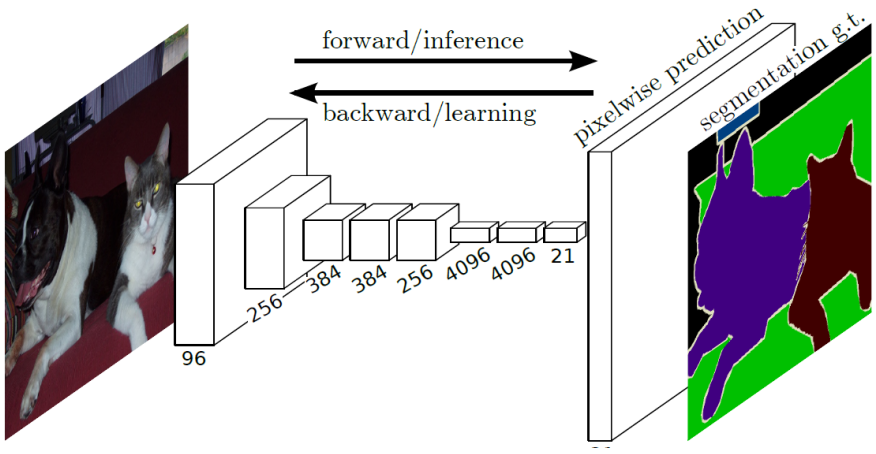

Figure 21 - Semantic Segmentation by FCN – source [44].

Results show advantages over the other two systems (transformer-based and Bi-LSTM-based): they use fewer parameters and are quicker to train [21].

3.4.4 – Connectionist Temporal Classification

RNNs can be discriminatively trained, and their internal state provides a general mechanism for modelling time series, thus becoming powerful sequence learners well suited for ALR. However, they require pre-segmented training data and post-processing to transform their outputs into label sequences. To overcome these requirements, Graves et al. [22] present a method for training RNNs to label unsegmented sequences directly, modelling all aspects of the sequence within a single network architecture.

On a given input sequence, Connectionist Temporal Classification considers the output as a probability distribution over all other possible and equivalent labelling sequences, and an objective function can then be derived for the maximization of correct labelling [22].

The connectionist temporal classification (CTC) loss is widely used in speech recognition, as it eliminates the need for training data that aligns inputs to target outputs [33].

3.5 – Literature review

Before the “official” start of the work of this Dissertation, discussions and “vagabond” reading were done. Although this kind of work easily ends up nowhere and may not be considered accountable, it is fun, permits a general view of the subject of study, and lets the reader have a feeling on whether or not will enjoy the process.

Beyond reading the referenced surveys, necessary for the completion and framing of the literature review, a previous literature selection was performed employing keywords and a time interval as filters in the most relevant academic search engines. The criteria used were:

- Search engines: Science Direct, and Scopus;
- Search dates: 27th to 30th of May, 2022;
- Keywords: ALR, VSR, Lip-reading, Deep-learning;
- Publication years: 2018 -2022.

The mentioned search engines allowed reading the Abstract of the hit results. This reading was the next filter applied to the first search results. The articles whose Abstract demonstrated not being related to ALR, VSR, or AV-ASR were excluded. Few exceptions were made to include either historical articles, eminent authors, or articles referenced by previous or posterior readings that did not appear in the first search results.

The choice of 4 years was due to the number of papers dedicated to ALR published each year. Just with the keyword lip-reading, one can expect about 6 papers-per-day, nowadays.

The following reading summarizes the selected articles, including each's major contributions and conclusions, according to the structure presented in subsections 3.2, 3.3, and 3.4. The author also highlights and explains sparse techniques, a selection solely driven by intuition and consideration of relevancy to his knowledge, a subjective criterion that may unveil objective research questions.

3.5.1 – Normal vs Impaired vs VSR

Lopez *et al.* [23] aimed to study the upper bound of visual-only speech recognition in controlled conditions. Since the literature is unclear on whether hearing-impaired people are better lip-readers than normal-hearing people, the authors compared 9 and 15 subjects, respectively. A database was constructed, and the speakers were instructed to facilitate lip-reading. Another study compared human and VSR systems' performances under optimal and directly comparable conditions. In the authors' tests, hearing-impaired participants just nearly outperformed normal-hearing participants. A 44% to 20 % spoken message decoding decrease gap was observed when comparing humans' performance to visual-only automatic systems. However, similar performances were obtained in phonemes, suggesting that the gap between human and

automatic speech-reading might be more related to the use of context than the ability to interpret mouth appearance.

3.5.2 – Homophemes

Jeon *et al.* [38] address the homophemes as word ambiguity enablers and words under 0,02s as "a", "an", "eight", and "bin", as they do not provide sufficient visual information to learn from. A novel lipreading architecture is presented, combining three different CNNs: 3D CNN, to extract features from consecutive frames efficiently; densely connected 3D CNN – to fully utilize the features; and multi-layer feature fusion 3D CNN with a pixel dropout layer and spatial dropout layer – to avoid overfitting and to extract shapes with strong spatial correlations with fine movements, while exploring the context information both in temporal and spatial domains. Then follows a two-layer bi-directional gated recurrent unit. The network was trained using CTC. The results of the proposed architecture show character (5,681%) and word (11,282%) error rates reductions for the unseen-speaker dataset, even when visual ambiguity arises.

Wang [45] also addresses the homophemes question, accumulating diverse lip appearances and motion patterns among the speakers by capturing the nuances between words and different speakers' different styles, respectively. As for the front-end, the method utilizes 2D (spatial only) and 3D (spatial-temporal) ConvNets to extract both frame-wise spatial fine-grained and short-term medium-grained spatio-temporal features to capture both grained patterns of each word and various conditions in speaker identity, lighting conditions, and so on. Then fuses the different granularity features with an adaptive mask (bidirectional ConvLSTM, augmented with temporal attention, which aggregates spatio-temporal information in the entire input sequence), to obtain discriminative representations for words with similar phonemes, as a multi-grained spatio-temporal novel modelling of the speaking process. The proposed model demonstrates state-of-the-art performance on two challenging lip-reading datasets. As future work, the authors propose simplifying the front-end and extracting multi-grained features with a more lightweight structure.

3.5.3 – Viseme-based lipreading systems

Fenghour *et al.* [24] focus on viseme-based lipreading systems that have been well suited to decoding videos of people uttering entire sentences. As the paper points out, the high classification accuracy of visemes (e.g., over 90%) contrasts with a comparatively very low

classification accuracy of words (e.g., only just over 60%) due to the homovisemes phenomenon which leads to a one-to-many problem (e.g., “I Love You” = “Olive Juice” = “Elephant Shoes”). Aiming for a more efficient viseme-to-word conversion method to tackle this accuracy decline, the authors developed a DNN model with an Attention-based Gated Recurrent Unit. They compared it against three other approaches (Perplexity-Iterator, Feed-Forward Neural Network, and Hidden Markov Model) through the LRS2 and LRS3 corpora. Results show that the proposed model effectively discriminates between words sharing visemes that are either semantically or syntactically different and at modelling long and short-term dependencies, therefore being robust to incorrectly classified visemes.

Viseme-based lip-reading systems do not require pre-trained lexicons and can be used to classify both unknown words and different languages. Fenghour *et al.* [25] explores this fact to classify visemes in continuous speech, uses visemes as a classification schema for reading sentences, and uses perplexity analysis for visemes to word conversion, stating that all contributions improve sentence-level lip reading. The proposed method uses visemes as a very limited number of classes, a unique deep learning model for classification, and perplexity analysis for recognized visemes to possible word conversion, resorting to purely visual cues from LRS2 and being robust to varying lighting levels. Results demonstrate a significant improvement in the classification accuracy of words compared to state-of-the-art works. For future research, the authors hint towards a more suitable architecture to further enhance the generalization capability and a higher training/test number of samples ratio.

VSR is highly influenced by the selection of visual features, which can be categorized into static (geometrically based) and dynamic (motion-based). Radha *et al.* [46] propose a three viseme models study, one as the control group and two considering both categories, one fused at the features level and the other fused at the model level. Motion History Image (MHI) is calculated from all visemes, from which discrete cosine transform (DCT), Wavelet, and Zernike coefficients are extracted for dynamic-motion feature extraction. An Active Shape Model (ASM) is used for static-geometric feature extraction. Fusion models are individually built by the Gaussian Mixture Model Left-to-Right Hidden Markov Model (GMM L-R HMM). The results show an improvement in performance due to the fusion and the presence of complementary cues in the motion-based and geometric-based features and that geometric cues provide better discrimination of visemes.

3.5.4 – RoI definition and Lip Refinement

Rethinking the RoI for ALR is the aim of Zhang *et al.* [47]. This paper questions the standard RoI for ALR papers, as human lip-readers do not just look at the lips during a conversation. Fernandez-Lopez et al. [23] state that facial expressions help to decode the spoken message and context framing the speech (e.g., a sad expression augments the probability of sad-related words/sentences). Resorting to state-of-the-art word-level and sentence-level VSR models, a comprehensive study is performed to evaluate the effects of extraoral information, including the mouth, the whole face, the upper face, and the cheeks. The Zhang et al. [47] proposed model was trained on large-scale "in-the-wild" VSR datasets, depicting many real-world variations, such as pose, lighting, scale, background clutter, makeup, expression, different speaking manners, etc. According to the study, using cut-out augmentation (similar to dropout pixels) with aligned face inputs can yield stronger features, improving recognition by forcing the model to learn the less obvious extraoral cues from data. Using mouth inputs alone makes VSR an isolated problem, as the process is not distracted by other parts of the face. Therefore, there has been no consensus on choosing RoIs, and intuition still plays a large role in RoI cropping.

Das *et al.* [48] shows a refinement in automatic lip contour extraction using pixel-based segmentation. This embodies an alternative to pixels classification of different colour planes, a potential difficulty to lip contours detection in adverse conditions like variations in illumination and clothing. The mouth region is extracted by k-means clustering binary classification based on R/G ratio thresholding. A big connected region around the cropped image's centre is considered the RoI to avoid false detection,. Next, the upper and lower lip areas are detected by the k-means clustering algorithm binary classification of the Green plane and the weighted RGB plane, respectively. The combined lip area is further processed to detect the centrally located big connected region. By finding the centrally located big connected region, rather than the biggest connected region in the whole binary classified image, variations in illumination and clothing effects are overcome, and RoI is restricted around the mouth region. For smooth edges, piece-wise polynomial fitting is employed with a higher degree for the upper lip and a lower degree for the lower lip. The proposed method works well, even for images with varying illumination and clothing effects. A future aim is to use this algorithm to derive the best possible lip contour related.

Lip segmentation accuracy plays an important role in automatic lip-reading, and can directly affect the recognition rate. Lu and Liu [49] propose a localized active contour model-based

method, using two initial contours in a combined colour space: a rhombus as the initial contour of a closed mouth; a combined semi-ellipse as the initial contours of both outer and inner lip boundaries for an open mouth. The method first applies illumination equalization to RGB images to reduce interference of uneven illumination, then adopts a combined colour space, which involves the U component in the CIE-LUV colour space and the sum components of DHT. Finally, the shape of the initial contours is determined by the positions of four key points in the combined colour space. The method improves segmentation results and gets more similar to the true lip boundary, compared with using a circle as the initial contour to segment grey images and images in combined colour space.

Lu *et al.* [50] also propose a lip segmentation implementing method, but now in the framework of the maximum a posteriori Markov random field (MAP-MRF), a statistical segmentation method that considers the interactions between spatial pixels of an image. The proposed method sets up a multi-layer hierarchical model in which each pixel of each layer corresponds to the four nodes in a quad-tree structure (QTS). The probability of a branch node can be derived from the probability of the previous one throughout the tree structure. Then a Markov random field derived from the model is obtained, so the unsupervised segmentation is formulated as a labelling optimization problem. The method also proposes a variable weight segmentation approach to improve over-segmentation robustness. Results show that the proposed method has better performance than the related methods. However, it runs between 3 and 4 s, therefore, it is not suitable for real-time applications.

Ma *et al.* [51] focus specifically on lip feature extraction under variant lighting conditions, since research has been mainly conducted for ideal conditions, therefore ideal lighting. The method consists of a pre-processing chain of illumination normalization and improved LBP features. The first is applied to remove the influence of external illumination noise before the lip feature extraction in four steps: median filtering, gamma correction, multi-scale Retinex filtering, and contrast equalization. LBP is an illumination invariant descriptor of edges, improving lip-reading recognition rate under variant lighting conditions. Experiments show that the proposed algorithm has a lower recognition rate in natural than traditional pixel-based feature extraction method but higher under variant lighting conditions.

3.5.5 – Datasets limitations overcoming

"No Data, no Deep Learning", is a common hearing among AI researchers. Petridis *et al.* [37] focus on lip-reading for isolated word recognition training on small-scale datasets. The

proposed method consists of two streams (each consisting of an encoder and a Bidirectional Long Short-Term Memory (BLSTM)): one stream encodes static information, using raw mouth RoIs as input; the other stream encodes local temporal dynamics, taking as input the difference between two consecutive frames. Each stream's temporal dynamics are modelled by a BLSTM, and stream fusion is done by another BLSTM. Four different datasets are used before very large lip-reading datasets are introduced. The proposed method learns simultaneously to extract features and perform classification using LSTM networks. Results demonstrate that the proposed model achieves state-of-the-art performance, outperforming all other approaches reported in the literature.

Afouras *et al.* [52] aim to boost lip reading performance by training strong models learning from ASR strong models, and not requiring human-annotated ground truth data. The proposed method, depicted in Figure 22, distils (transfers knowledge/weights from a large model to a smaller one) from an ASR model, trained on a large-scale audio-only unlabelled corpus, with a teacher-student approach (the teacher's prediction is used to train the student). The cross-modal distillation combines CTC with a frame-wise cross-entropy loss, minimizing the KL-divergence between the student and teacher posterior distributions. The method and paper's contributions show that: ground truth transcriptions are not essential to train a lip-reading system; arbitrary amounts of unlabelled video data can be leveraged to improve performance; distillation significantly speeds up training; state-of-the-art results on (publicly available) LRS2 and LRS3 datasets can be obtained.

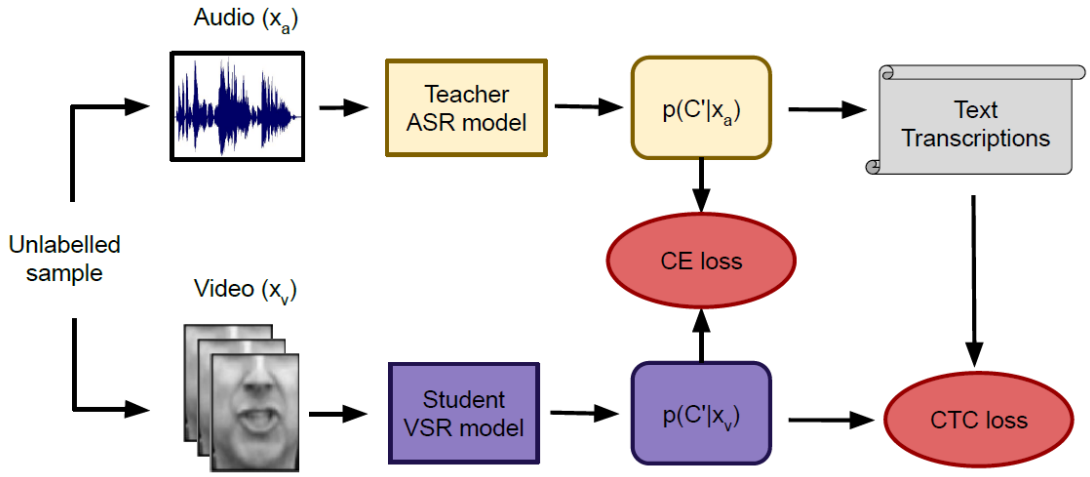

Figure 22 - Proposed Architecture for Afouras et al. work – source [52].

Results demonstrate effectiveness in training strong models for VSR by distilling knowledge from a pre-trained ASR model and, more generally, from any available video of talking heads, e.g., from YouTube, therefore, from any arbitrarily large amount of data.

3.5.6 – Front-end technique update

Weng and Katani [53] conducted experiments on word-level visual lipreading from video input with no audio by replacing shallow 3DCNN + a deep 2DCNN with deep 3DCNN two-stream Inflated Convolution Networks (I3D) and evaluating different combinations of front-end and back-end modules, with the greyscale video and optical flow inputs on the LRW dataset. 3D convolution networks can capture the short-term dynamics and be advantageous in visual lipreading, even when RNNs are deployed for the back-end. However, due to the huge number of parameters introduced by the 3D kernels, state-of-the-art methods in lipreading have only explored the shallow (under 3 layers) 3DCNNs. The authors present the first word-level lipreading pipeline using deep (over 3 layers) 3DCNNs to explore these networks to their maximum. The experiments show that: compared to the shallow 3D CNNs + deep 2D CNNs front-end, the deep 3D CNNs front-end with two-round pre-training on the large-scale image and video datasets can improve the classification accuracy; using the optical flow input alone can achieve comparable performance as using the greyscale video as input; the two-stream network using both the greyscale video and optical flow inputs can further improve the performance.

Lu and Yan [54] propose a CNN and BLSTM that uses a hybrid neural network architecture for an automatic lip-reading system. The method first extracts key frames from each isolated

video clip uses five key points to locate the mouth region, extracts features from raw mouth images using an eight-layered CNN, and uses BLSTM to capture the correlation of sequential information among frame features in both directions in time, and uses the softmax layer to predict final recognition result. The limited number of key points reduces redundant information in consecutive frames, therefore, the complexity of computation and processing. The CNN copes with image deformation by translation, rotation, and distortion, hence strengthening robustness and fault-tolerant capability, and a fully connected layer is used to get static features of a single mouth image. BLSTM improves both finding and exploiting long-time dependencies from sequential data, so the relationship of the features among frames is built and strengthened. The results show that the proposed DNN can effectively predict words from the mouth area on own established database (6 speakers, 9 digits), compared to traditional algorithms that combine handcrafted features with a classification model.

Mesbah et al. [55] developed a visual-only speech recognition system, proposing Hahn Convolutional Neural Network (HCNN), seizing their ability to represent images with less redundancy and being parameterized to retain the global or local characteristics of the image in the lowest orders. The proposed architecture consists of Hahn moments as a filter in the first layer, with its ability to hold and extract the most useful information in images effectively and the performance of the CNNs in learning patterns and image classification. The results show reduced processing time, normal to spatiotemporal modelling features, and visual feature extraction with 3D CNN, ResNet, and Bidirectional LSTM.

3.5.7 – Back-end techniques update

Martinez *et al.* [43] address the Bidirectional Gated Recurrent Unit (BGRU) limitations and propose corresponding improvement proposals. First, the mouth region was extracted, DCT was used to feature transform, and then fed to HMM to model the temporal dynamics. The authors proposed to address the limitations of the model: BGRU layers are replaced with TCN to improve the overall performance; a cosine scheduler was adopted to reduce training time (from 3 weeks to 1 week GPU-time) and avoid relying on a cumbersome 3-stage sequential training; cosine variable-length augmentation was proposed to improve the generalization capabilities. The authors propose a multi-scale TCN, as shown in Figure 23. As each TCN receptive field is defined by the kernel and stride sizes, several temporal convolutional blocks are achieved and stacked sequentially to act as a deep feature sequence encoder. Next, a dense layer is applied to each time-indexed feature vector, and a simple averaging consensus function

is used. With different-sized kernels and multiple temporal scales, long and short-term information can be mixed up during the feature encoding.

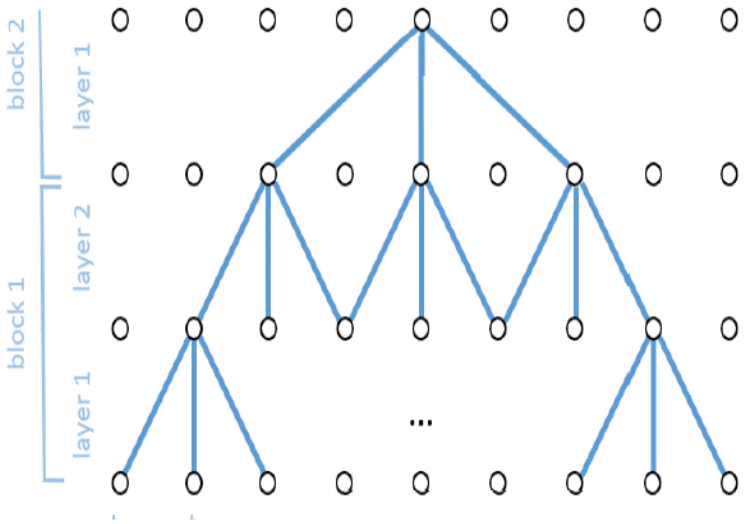

Figure 23 - Multiscale TCN – source [43].

Results on the largest publicly available datasets for isolated word recognition in English and Mandarin, LRW and LRW1000, show that a new state-of-the-art performance was achieved.

Huang *et al.* [29] propose a novel lip reading model using a transformer network, to achieve higher accuracy. The method uses the pre-trained neural network VGG16 to extract the lip features from the GRID corpus, adopts dimensionality reduction towards the originally high dimensions extracted features, and processes the features through the author's Transformer network for training. The transformer adopts a self-attention mechanism instead of CNN and RNN, as is commonly used in deep learning. RNN tends to be slow in some sequential processing tasks. On the other hand, transformers' parallel processing greatly improves training speed, by parallel processing. The experiment shows a significant reduction in training costs without compromising the enhancement of the lip-reading accuracy of the model.

In Lu and Li [56], the authors tackle the difficulty of meeting the requirements of practical applications for ALR due to the complexity of image processing, hard-to-train classification, and long-term recognition processes in three steps. Firstly, they extract keyframes from their own established independent database. Secondly, they use the Visual Geometry Group of Oxford University and the Google DeepMind (VGG) network to extract the lip image features. Then, as an attention-based RNN, they compare two lip-reading models: a fusion model with an attention mechanism; and a fusion model of two networks. The results of the proposed hybrid neural network architecture of CNN and attention-based LSTM, show an increase of 3.3% to

the general CNN-RNN. The authors manifested the future intention to train the model on datasets of real-time broadcast videos.

3.5.8 – Outliers

Intending to learn strong models that recognize speech in silent videos, Prajwal *et al.* [57] focus on challenges in lip reading and propose tailored solutions, contributing to lip movement representations aggregation, and to robustness improvement to ambiguity by sub-word units-based modellin. The paper proposes an end-to-end trainable attention-based pooling mechanism that learns to track and aggregate the lip movement representations, a sub-word (word-pieces) tokenization that not only matches with multiple adjacent frames but also with those that are semantically meaningful for learning a language easily, therefore greatly reducing the run-time and memory requirements, and a model for VSD trained on top of the lip-reading network since there is no automated procedure for cropping out the clips where the person is speaking. The results show a state-of-the-art Word Error Rate (WER), outperforming work trained on public data, even industrial models trained on orders of magnitude more data. Also, the designed Visual Speech Detection obtains state-of-the-art results on this task and even outperforms audio-visual baselines.

Deep Learning methods have been used for developing ALR systems. As DL is vulnerable to adversarial attacks, so will ALR DL-based systems. Historically, adversarial attacks towards video classification have been less explored than towards image classification. They add a well-crafted minimal and imperceptible perturbation to the input, so its classification is incorrect. Gupta *et al.* [58] proposed Fooling AuTomAtic Lip Reading (FATALRead), a method to perform adversarial attacks on state-of-the-art word-level ALR systems, conducted on the publicly available dataset, because of making models design more robust and resilient against engineered attacks. The proposed model aims to replace the target output for another by adding perturbations that alter the classification prediction. FATALRead attack successfully fools state-of-the-art ALR systems based on sequential and temporal convolutional architectures. The results show the vulnerability of the sequential and TCN architectures to an adversarial attack in the domain of ALR.

3.6 – Summary

In the last 4 decades, ALR evolved from a template recognition model (1984, with only 4 layers of grey) to end-to-end deep neural network models, where linear considerations try to answer a non-linear question.

The number of researchers and research teams paying attention to ALR has been steadily growing, and the approaches come from multiple geographic locations and multicultural teams, which adds to the technical diversity of other areas that use artificial vision and intelligence. Equally vast are the non-mentioned areas in which artificial vision and intelligence evolve parallelly to ALR, resulting in later and indirect contributions to its progress.

In many of the referenced papers, statements reinforce the lack of consensus, which only acts as a catalyst for more research and work. The evolution has been so swift that decade-old methods are most likely to be obsolete, and derivations are so diversified that what will happen in the next decade is as unpredictable as the movement of a simple magnetic pendulum subjected just to 3 magnetic fields besides the gravitational one.

Although uncertain, in the author's point of view, the future will be based on:

i) The continuation of the application of newer and newer methods, followed by the simplification of the same;
ii) The simplification of parts of the process, as attention narrows inputs' processing target;
iii) The complication of parts of the process, such as weighing other data to narrow the outputs' scope (if you are with your spouse, you are more likely to say 'I Love You', than to say 'Olive Juice');
iv) Incorporating other inputs, such as emotional states, for context.

Answering the Research Questions risen in the beginning of this chapter:

RQ 1 – End-to-end deep learning, resorting to Attention-based LSTM or Transformers appear to be more suitable for visual clues for automatic lip-reading;

RQ 2 – The same answer as in RQ1;

RQ 3 – No specific methods are studied with the available datasets; Although some datasets are more commonly explored in the presented papers, namely GRID;

RQ 4 – Despite all technology and evolution, papers keep referring to the same difficulties sources as in the pre-AI era, such as visibility conditions or facial hair. These still challenge the hearing impaired and researchers alike, catalysing and emphasizing the relevance of mutual attention in coming years, with mutual yielding. Nevertheless, as there is no database for ALR in the Portuguese language, a resource gap became apparent and an objective for this dissertation.

4. Methodology

This work aims to contribute to the evolution of Automatic Lip-Reading. Accordingly to the previous chapter's the literature review, a research gap is a database for ALR in the Portuguese language. This chapter is dedicated to the production methodology, and validation of this new DB on a SOTA model.

A selected SOTA DB (GRID) was used as a template to verify the video composition in specific details like the speaker poses, background colours, etc.. As there was no available documentation on the making of this DB, to the best of the author's knowledge, an educated guess was performed to attain a valid methodology, and final product. The DB production methodology is depicted in Figure 24.

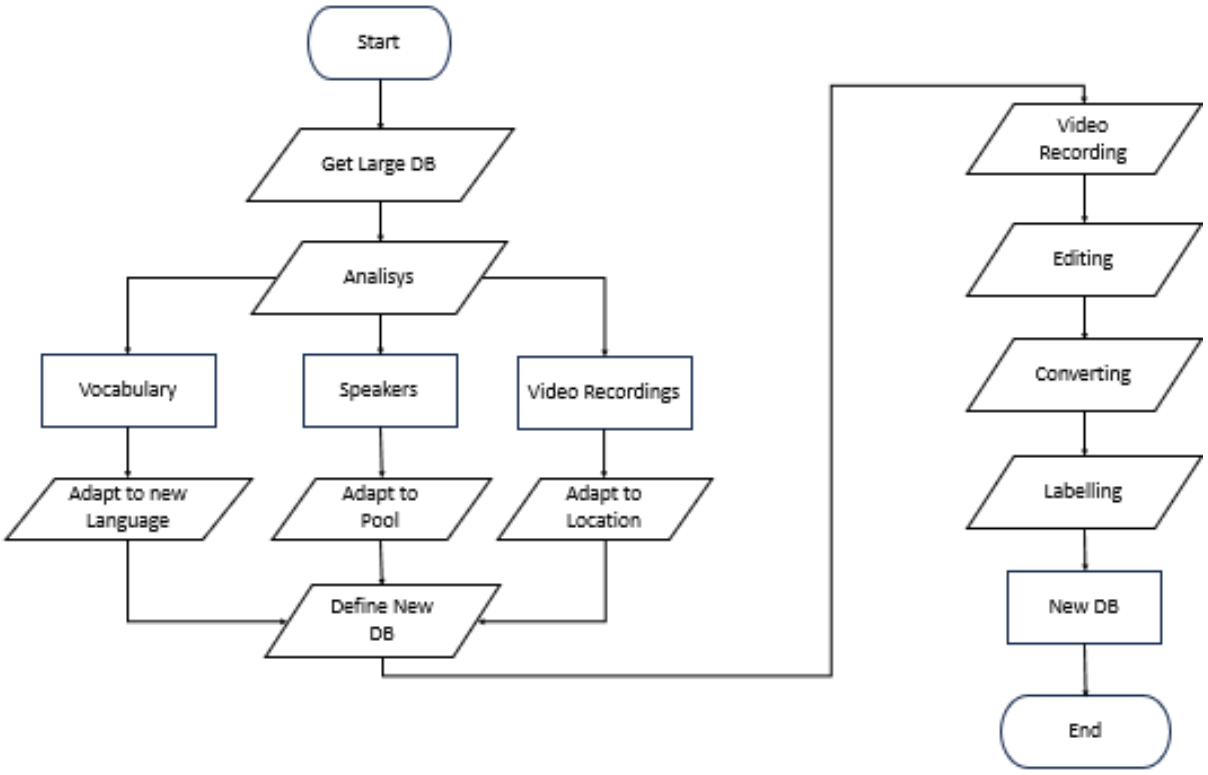

Figure 24 - Pipeline for new database creation.

The large-scale DB was obtained; its Vocabulary, Speakers, and Video Recordings were analysed; the adaptations for the new language, speakers pool and location were performed; and the new DB was structured. The videos were then recorded, edited, converted and labelled. LusaPt was ready to be validated.

The following objective was to validate LusaPt. A SOTA model (LipNet) was selected and obtained to perform the tests required to validate the inference and training procedures. As depicted in Figure 25, a series of inference and training tests were performed, with the original and new DBs.

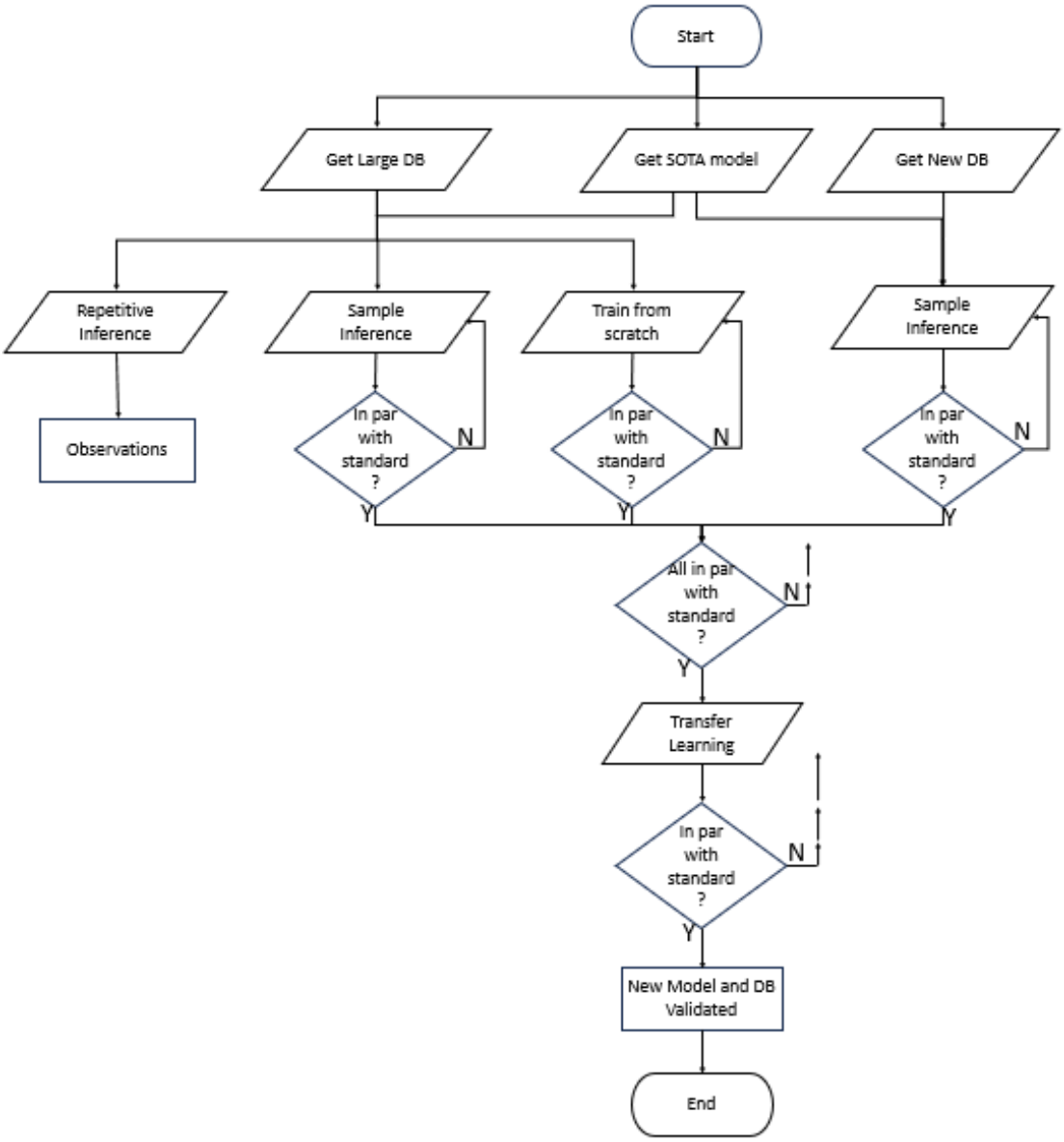

Figure 25 - Database and model validation pipeline.

First the SOTA model, the large and the new DBs were obtained. From the large DB and SOTA model, repetitive and sample inferences were performed and also a train from scratch. Inference tests were also performed by the SOTA model on the new DB. All these test values were then compared to the SOTA's. When in par the procedures were considered valid, otherwise debugging was in order. Finally, a transfer learning method was applied, pre-training with the available model and fine-tunning on LusaPt.

4.1 - Database construction

With the decision to produce a database in Portuguese, the next step was to determine what to say, in what manner, by whom, where, and when. Following the philosophy of "Transfer Learning", in the sense that the author wanted to extract the main characteristics of GRID database, the plan set was to learn the most from the GRID Corpus, reproduce some of them to a new DB, and make the necessary adjustments.

4.1.1 - GRID corpus

The corpus presents a sentence-like structure, as shown in Table 3. It is composed by 34 speakers, uttering 1000 different sentences each, summing to almost 33.000 available videos (one speaker is disabled and some videos are corrupted, therefore not usable) within 51 tokens vocabulary.

Table 3 - GRID corpus sentence structure – source [54].

command	color*	preposition	letter*	digit*	adverb
bin	blue	at	A–Z	1–9, zero	again
lay	green	by	excluding W		now
place	red	in			please
set	white	with			soon

The cast is composed by an almost equal amount of both male and female talkers, staff and students of the Departments of Computer Science and Human Communication Science at the University of Sheffield, paid participants, all English speaking as their first language, encompassing a range of English accents, and ranging from 18 to 49 years. The job was done by a team of 7 [59].

Analysing the videos, the author understood that the lighting was diffuse (to avoid shadows), there was a homogeneous blue background (to avoid visual noise), and the speakers were facing

the camera, as shown in Figure 26. Furthermore, the speakers were at a consistent angle and distance, and there was no audible noise.

Figure 26 - GRID corpus example (Speaker 1) – source [57].

All the videos were 75 frames long, as they have 3 seconds at the rate of 25 frames per second(fps), with 288x360[px].

4.1.2 - LusaPt corpus

The first adjustment had to be the size of the corpus, for the work was to be solely done by the author of this dissertation, there was no funding, and the timeframe had to fit this dissertation's schedule. Therefore, the corpus size was limited to 200 at first and ultimately reduced to 100, considered enough for proof of concept.

4.1.2.1 - Token selection

The possibilities range from visemes to words and digits for the used model's classification objectives. Due to LipNet being words and digits oriented, and the implementation accomplished an over 90% WRR performance [33], the same path was followed.

In the author's understanding, the "round" number of 10 digits is appropriate as the number of lines, at the vocabulary matrix depicted in Table 4. The rest of the words were selected for being mostly between 2 and 4 syllables (for size normalization). They were selected from typical day-to-day words to benefit natural understanding, acceptance and uttering. They were also arranged into four more categories: Countries (that speak Portuguese). As they are one less than the number of Digits, *Cabo Verde* was divided into *Cabo* and *Verde*; Food; Accompaniment; and

Companions. At last, the option for pursuing phrases and not sentences derives from the focus on not widening the scope of this dissertation, keeping it simple.

The 50 vocabulary tokens created, shown in Table 4, differs from the GRID's 51tokens, inducing a vocabulary size incompatibility for the model. A dummy word ("*Fantoche*" – the Portuguese word for dummy) was then introduced in the newly created file *VocabPT.txt*, which substitutes the original *Vocab.txt*.

Table 4 - Vocabulary and Categories.

Digits	Countries	Foods	Accompaniments	Companions
Zero	Portugal	Pão	Tinto	Família
Um	Angola	Peixe	Cerveja	Amigos
Dois	Moçambique	Chouriço	Água	Colegas
Três	Guiné-Bissau	Queijo	Aguardente	Vizinhos
Quatro	Brazil	Sopa	Poesia	Professores
Cinco	Cabo	Tomate	Fado	Alunos
Seis	Verde	Alface	Sumo	Portugueses
Sete	Macau	Bacalhau	Saudade	Estrangeiros
Oito	São-Tomé	Nata	Porto	Sózinho
Nove	Timor	Salada	Ginja	Todos

4.1.2.2 - Speakers selection

At the time of this dissertation submission, the LusaPt corpus comprises 10 speakers (5 males and 5 females), uttering 10 phrases[3] each, all correct and available. Of the 10 selected speakers, 8 speak Portuguese as their first language, and 2 speak Crioulo between peers, with Portuguese as the official language in their home country. Yet, one of the 8 selected speakers lives in Portugal for so long that he can speak with and without the Brazilian accent. In order to make the DB as diversified as possible, within the finite number of speakers and videos per speaker, the former spoke with the Brazilian accent, and the selection of speakers with different skin colour, ages, and facial hair was privileged.

This resulted in the cast presented on Figure 27.

[3] Phrase differ from sentence, as the latter collection of words must have meaning and be grammatically correct.

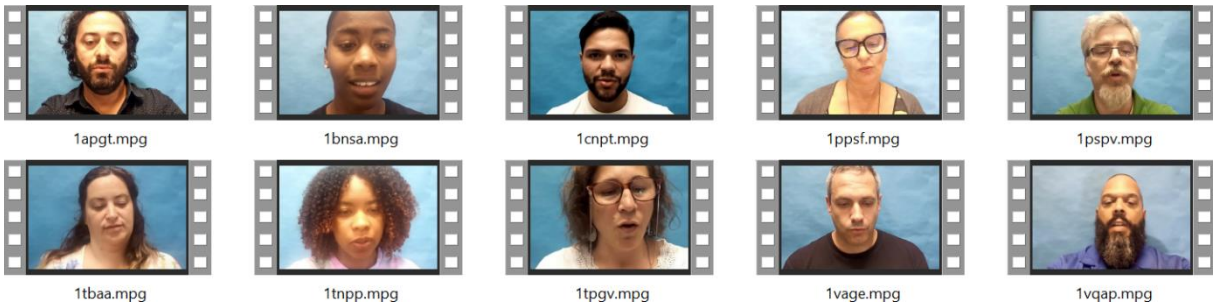

Figure 27 - Data Base Speakers.

All the recordings were made in the RoboLab, at the Marinha Grande's CENFIM Training Centre. As the centre did not stop during the video recordings, there were naturally occurring metalworking noises amongst the recorded utterances.

4.1.2.3 - Videos recording

In order to obtain similar conditions to the GRID corpus, which proved successful, various layouts were tried, as presented in Figure 28, a) and b).

a) b)

Figure 28 - Layout trials.

In both pictures, one can see the speaker sitting position, the cameras at face level, the blue homogeneous background, and the lighting distribution. The left picture was of the same layout on trial day when a microphone (on the chair in front of the potential speaker) was still considered to be used. The right picture was already taken in RoboLab, still considering the recording to be made by a webcam into a computer. One important, invisible feature was one 20W LED light projector covered with two layers of kitchen paper (for light diffusion). It was

beside the recorder, near ground level, towards the speaker as a spotlight to attenuate the upwards and downwards lighting differences.

In practice, the cameras were ultimately substituted by the author's smartphone, with the speaker holding his smartphone and reading from the picture of the scrambled words, as shown in Figure 29.

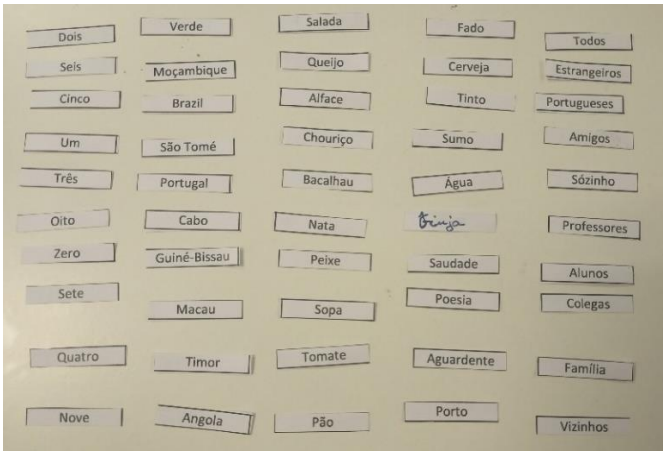

Figure 29 - Scrambled phrases.

The author believes that scrambling the words allows the model to learn words coming from and to different possibilities. If the same word would always precede one word, the classification of the first would influence the latter's classification, diminishing the possibility of generalization and, therefore robustness. As an example, whenever the model may infer the word *Guiné*, it will expect the word *Bissau*. On the other hand, if the model infers *Água* it should not expect to be *Aguardente*.

At the end of the third and final recording day, the whole process would take around 5 minutes per speaker. This time was enough for each one to read and sign the informed consent, scramble the words, re-arrange the phrases, take it's picture, sit and get the lighting adjusted, read the phrases, perform any correction the author considered necessary at the time and leave. The recordings became a light and fluid task, ultimately convincing enough volunteers to give their best, surpassing the initial 10 speakers goal.

4.1.2.4 - Videos processing and editing

Following the videos recording, the task was to process the videos into a format suited to inference and training by the models, as the original mp4 format is not recognizable by the model.

To edit, he first step was too look inside each speaker's videos and understand if: they were correctly recorded from start to finish; they were not blurred, and the correct and entire words were pronounced.

As the instructions given to the speakers were to say the same phrase twice, separated by an pause of 1-2 seconds, the next step was to select between the best of the two registers. This operation, as well as the frame selection and cutting, was done by a video editor.

To select the right video edition software, which can process the input extension (MP4), run at a slower speed if necessary, track frame by frame showing the frame number at all times, cut, divide, and convert into a lower resolution format, the author consulted three experienced video editors. They all referenced a variety of available software, but ultimately, the capabilities of the Microsoft Video Editor[4] revealed to be enough for the job and was selected.

Figure 30 shows the video editor environment. One can see: the Project library towards where one must add the raw video; the Graphic Script, where to draw the same video and where to view the end result; and the screen with a timeline and commands for play, advance, and return 1 frame, and frame time.

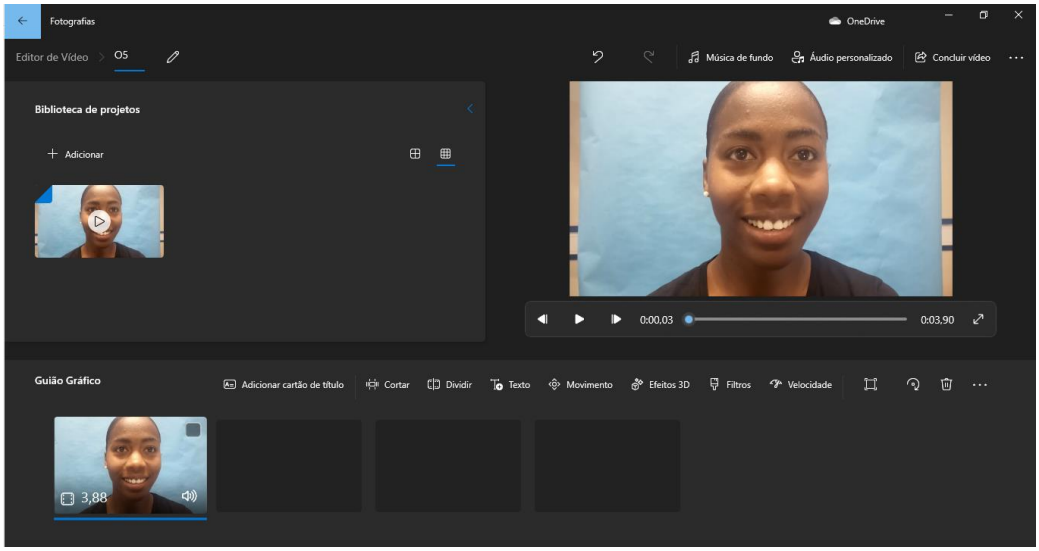

Figure 30 - Video editor environment.

[4] *Microsoft 2022.30100.19004.0 © 2020 Microsoft Corporation*

4.1.2.6 - Video formatting

A suitable conversion was performed with the objective of processing into a model processable format. The original full-HD format was revealed to be too much for the *Ffmpeg* converter command's dedicated cache memory, despite the attempt to augment its upper limit. A downgrade to an intermedium resolution was performed, recurring to the video editor; the cache memory limit was revealed to be enough, and the conversion in bulk could follow.

The conversion of the hundred videos was performed by a python script, named *convertermp4mpg.py*. Basically, the script takes every file type *.mp4 (e.g.: *nameX.mp4*); holds the value of the name from the split of the file (e.g.: *nameX)*; converts the file into another with the same name, but of type *.mpg (e.g.: *nameX*.mpg); and proceeds to the next file.

4.2 – Repository composition

The author's objective is that LusaPt corpus will be available to the academic community, adding to the existing corpora from which GRID was the example used in this dissertation. LusaPt is composed of videos and corresponding alignments.

4.2.1 - Videos

The main processing task resulted in a repository of 10 videos for each of the 10 speakers in three different formats:

i) MP4 (original), Full HD resolution (1920x1080 [px]), 30 fps;
ii) MP4 (downgraded), Intermedium resolution (1280x720 [px]), 30 fps;
iii)MPG (converted), Final resolution (1280x720 [px]), 25 fps.

4.2.2 - Word alignments

Following the video recording and processing, the task was to perform the word alignment. Aligning words, as shown in Figure 31, is to state the start and finish frames of each token. It allows for temporal referencing of the start and finish of each token, essential to the model to "know" it when training.

As patent in Figure 31 from GRID, the 3 seconds at 25fps rate (summing 75 frames) was subdivided as if there were 4 partitions of 250/1000 of a frame. This way, one can understand

that the initial silence perdures from 0 to ¾ of the 24th frame. As the author understands, the 25fps was further divided into 4 parts, similarly as if there were actually 100fps.

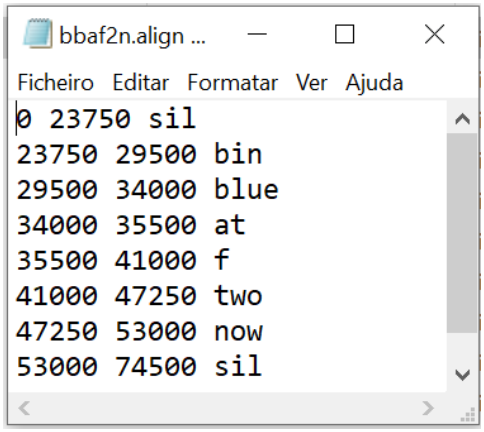

Figure 31 – GRID word alignment example – source [59].

For the word alignment process, the video editor was used again, in mp4 full-HD format, to attain the best quality source of visual information.

The first intuition was to label the beginning of the word with the frame when the sound was emitted. However, along the task, it became clear that the reading of the lips begins at the start of its movement. Therefore, the initial frame of the token was considered to be when the rest position of silence breaks or the passage from the last movement of the preceding token to the starting position of the next is clear. The focus on the starting frame of each token was deliberate, as the ending can be reduced or cut, when one speaks faster. The assumption is that, if deemed necessary, a typical speaker prefers loosening on the final quality of a word (or even sentence), rather than on the beginning.

The frame of the ending of the final token was labelled according to when the movement to make the final phoneme stopped. Again, the sound was not a defining reference due to the fact that, on various occasions, the last frames were simply a maintenance of the lips position to prolong the sound.

The final product of the alignment is visible in the sample shown in Figure 32. The 30 fps rate is shown in the 3 hundredths of a second framework timestamps.

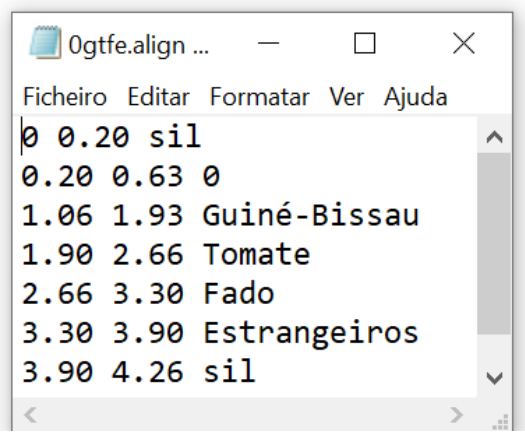

Figure 32 - LusaPt's word alignment (sample).

The contrast between the apparent 100fps, or even 25fps, and the author's 30 fps was also subjected to analysis. There emerged the possibility that the model was prone to a specific type of word alignment time rate dependency.

To the depth of code analysis resorted by the author for this contrast, no time rate dependency was found. Such led to the preliminary assumption that, no specific temporal measurement dependency may exist. Therefore, the timestamps acquired from the video editor were adopted.

4.3 – Database validation setting

Intending to validate LusaPt, the author tested it on a state-of-the-art ALR model. For the model selection, end-to-end deep learning, namely Transformers, appears to be more suitable for visual clues only automatic lip-reading [60].

4.3.1 – Model Selection

The choice of the particular model to use fell on LipNet, for it is an end-to-end sentence-level ALR state-of-the-art model on the GRID corpus, as referred on section 3.4. Acting accordingly, the author then searched GitHub for available implementation repositories to select from. The 131 hits were then filtered for Python implementation only, which was reduced to 56 hits, a small enough number to have an overview of which was performed.

Sorted by "best match" by GitHub, the 56 repositories were overviewed. It became apparent to the author that, as the repositories went down the list, fewer jobs were made upon LipNet. As

examples are: the first on the list[5], which presents fully available code with comprehensive comments within a comprehensive file structure, 55/63 closed Issues (used for debugging and updating) and work instructions; and the last that presents an unstructured file system, almost uncommented files, and a single still open Issue.

The following step was to study and implement the selected LipNet application on visually user-friendly Visual Studio Code, as the author's computer runs on Microsoft Windows. The model was cloned, and the referenced versions met the package requirements. However, the model did not run as announced. After updating the 6 years gap, and compatibility errors, Ubuntu[6] became the environment for code running, and Visual Studio Code was exclusively used for code editing, and another repository was adopted [7,8], also an end-to-end sentence-level model, running on the GRID corpus.

4.3.2 – Model customization

Having a functional model, the next task was to understand and adjust it to the incoming new database. Furthermore, because of the size of the new database, training from scratch would result in poorer performance, making any results comparison to the reference impracticable, impairing the capability to the database validation. Relatively small database training leads to overfitting, low robustness and low generalization capability. Therefore, the necessity for a solution to this question emerged.

4.3.2.1 – Model description

In its original form, the Lips Don't Lie repository comes with the files and folders shown in Figure 33.

[5] https://github.com/rizkiarm/LipNet
[6] Ubuntu 22.04.2 LTS (GNU/Linux 5.10.16.3-microsoft-standard-WSL2 x86_64)
[7] https://github.com/MitchellKT/LipsDontLie
[8] https://github.com/TomBekor/LipsDontLie

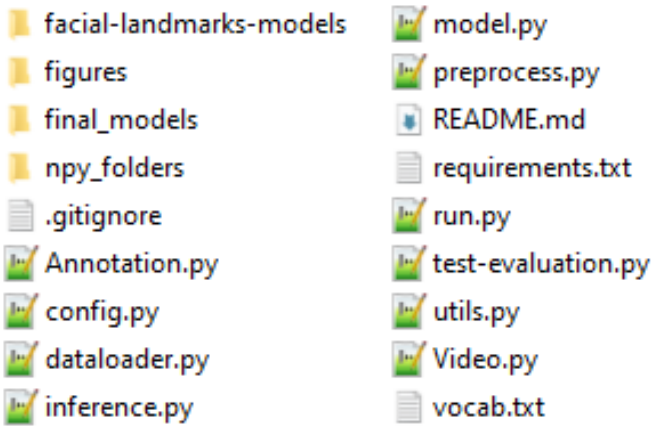

Figure 33 – Lips Don't Lie repository – source [5,6].

This repository will be briefly explained, for this understanding will be necessary when understanding the changes made, and helpful to understand the results and coming conclusions.

i) **facial-landmarks-models** – folder holding a single file (shape_predictor_68_face _landmarks.dat), which is just a data file containing a trained model for 68 facial landmarks location, by Dlib[9] and as mentioned in section 3.4.1. The process, illustrated in Figure 34, passes by face detection (left), face landmarks detection (middle) and mouth landmarks definition (right);

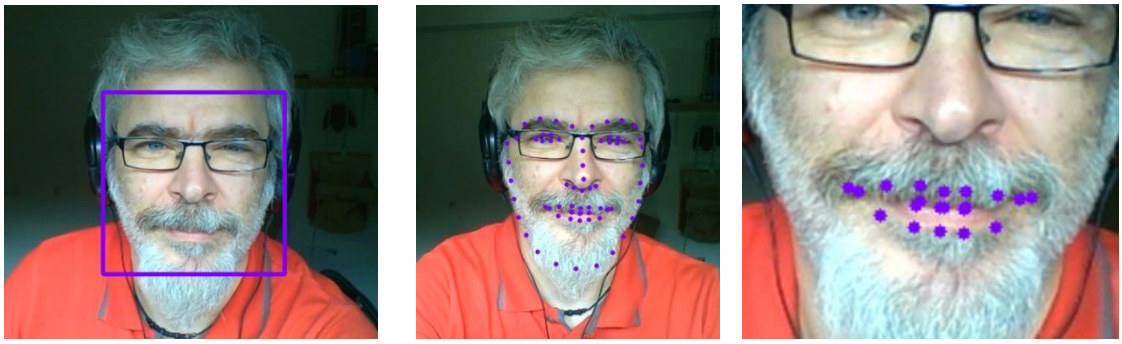

Figure 34 – Face detection and landmarking.

ii) **figures** – folder holding figures of previous training results;

iii) **final-models** – folder holding the trained models' weights and biases. From this files, one can infer with the best training model. Despite this author's search, not a definitive document was found stating the conditions to arrive at this model;

[9] http://dlib.net/face_landmark_detection.py.html

iv) **npy-folders** – a folder holding two folders: npy_alignements, and npy_landmarks. Each holds the alignments and landmarks processed files, split in test, train, and validation folders, for each speaker, and according to the specifications set on *utils.py*;

v) **Annotation.py** – a Python file that reads and compresses alignments files. Instead of the time reference being a timestamp, the words are in a single line, with the temporal reference indicated in the spacing between the tokens;

vi) **config.py** – a Python file which contains constants and hyper-parameters used throughout the project;

vii) **dataloader.py** – a Python file that loads, in batches, raw data from sources and prepares it to be used by the model. It also performs the tokenization, i.e., a method to separate a list into tokens (digits or words);

viii) **inference.py** – a Python file used to infer tokens from a file or files in a dedicated folder;

ix) **model.py –** a Python file that dictates the architecture of the model. One can now update Figure 15 from section 3.4, and Input and Front-end's updates can be seen in Figure 35.

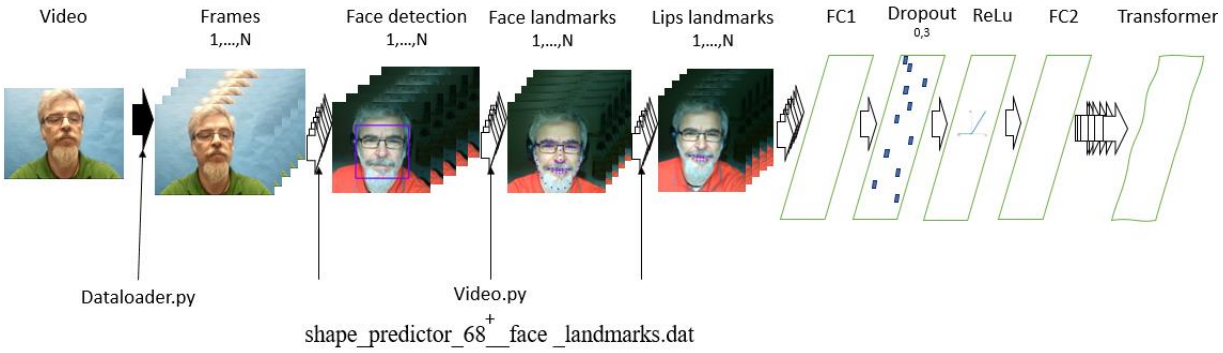

Figure 35 - Input and Front-end architecture update.

In simple terms, each video file is loaded in frames. The model recurs to Dlib's face detector to detect where the face is. To Dlib's shape predictor to find the face points (or landmarks), then focus only on the mouth, feed its landmarks to a FC, drops 30% of the learning neurons off, performs ReLu, then passes by the second fc layer and goes to the Back-end.

The back-end and Output update can be seen in Figure 36

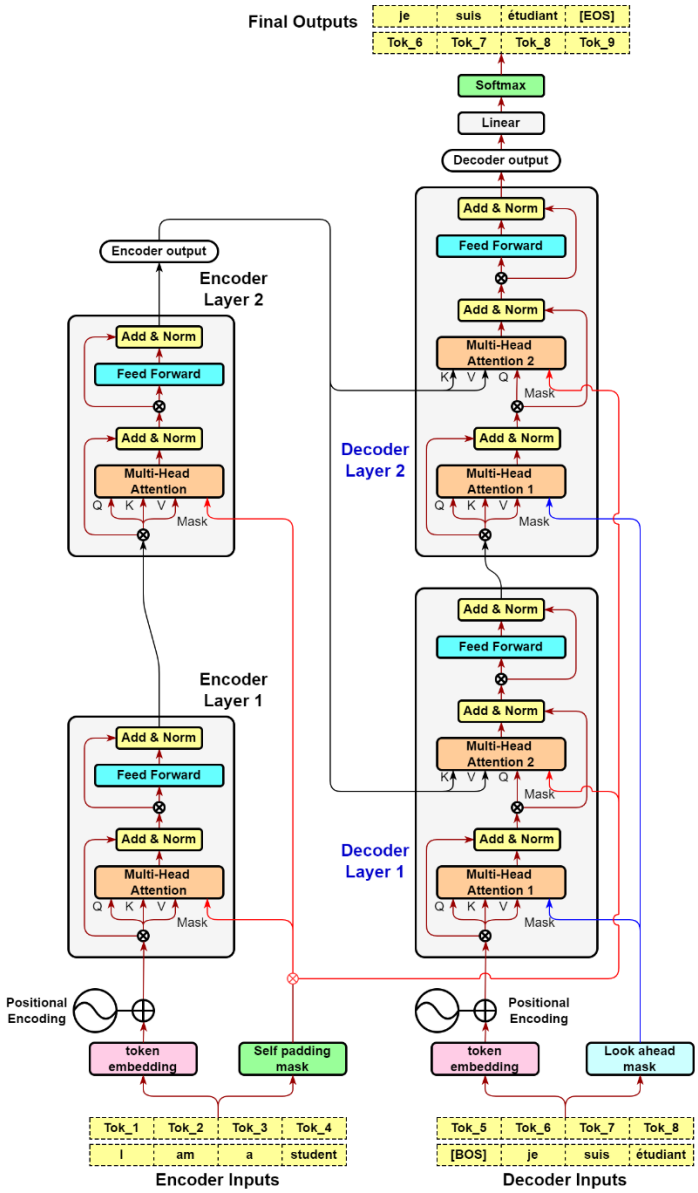

Figure 36 - Back-end and Output architecture update – source [10].

In simple terms, the encoder inputs are the Back-end's inputs and also the Front-end's outputs. It is fed to a 2-layer Encoder and 2-layer Decoder with 4 attention heads, then its output passes by another fc, which serves as the linear classifier, is "softmaxed", and out comes the prediction. To understand less and more complex features, one may understand the double layers as a mean for two hierarchical abstraction level learning. The 4 attention heads work as if there was a team of 4 different persons looking at the same data, and in the end, they would share the 4 different ways to look at the same data, enhancing the representative power and, ultimately, the transformer performance.

[10] https://wikidocs.net/167210

x) **preprocess.py –** a Python file that which splits landmarks and alignments to train validation and test;
xi) **run.py –** a Python file to train the model and to generate predictions on unseen test samples;
xii) **test-evaluation.py –** a Python file that loads pre-trained models, calculates and prints accuracy;
xiii) **utils.py –** a Python file to divide the videos into the train, validation, and test subdirectories, and to define metrics;
xiv) **Video.py –** a Python file to read the video files, find the face(s), then the mouth(s);
xv) **vocabPT.txt –** a text file that states the vocabulary to choose from.

4.3.2.2 – Model adjustment for LusaPt

Besides the necessary size adjustments, as the necessity to create a dummy word mentioned in section 4.1.2.1, the training trial was performed.

Assuming the size of the LusaPt corpus was relatively small, a training-from-scratch approach was dismissed, and alternative solutions were considered. Subsection 3.5.5, describes state-of-the-art methods to overcome the limitation of training on small datasets. In the referred subsection, Petridis *et al.* [37] apply an encoder and a Bi-directional LSTMs whose performance was surpassed by Transformers, and Afouras *et al.* [52] focus on ASR and knowledge distillation, more fitted for training smaller models (Students) learning from larger trained models (Teacher).

Transfer Learning is an alternative method for leveraging knowledge transfer, published in 2017 before my state-of-the-art time scope. This method enables learning from one or more source tasks, to improve learning in a related target task [61]. For example, a model trained to play chess from a large database, can take the advantages of the learnt abilities to learn chequers from a relatively smaller database.

Figure 37 illustrates the learning performance leveraging method. A higher starting level, a higher learning rate (slope), and a higher plateau (asymptote) mean that it already starts with some degree of performance, learns faster and with less data, and achieves a higher final performance, respectively.

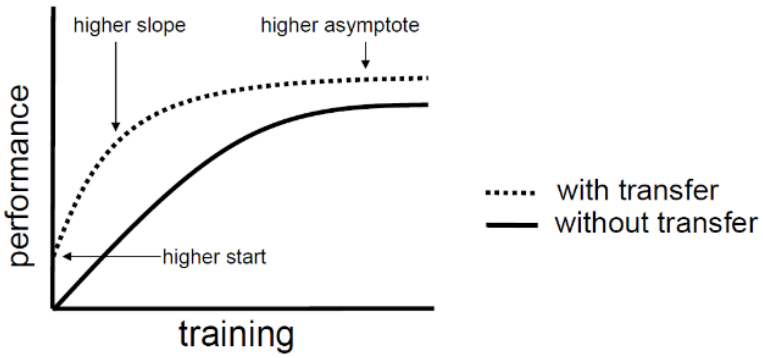

Figure 37 - Learning leveraging – source [53].

It is then possible to take advantage of the knowledge acquired when training the neural network for Automatic Lip-Reading in English with a big dataset as pre-train, fine-tune it, and learn ALR in Portuguese with a small dataset, when in comparison to the first. The pre-train enables the model to learn how to identify mouth landmarks, edges, lines, corners, and shapes, and how to follow them. The fine-tunning will benefit from this learning, expose the model to the new data set and train it to classify in Portuguese, instead.

In practice, transfer learning may be accomplished by adding a new, dimensionally equal, FC at the end of the existing network, allowing the training for this new layer. It can also be achieved by freezing the training until the model's linear classification layer (FC), and train from then onwards. As this last option is simpler, it was the first to be attempted.

Figure 38 shows the part of the code that demonstrates the transfer learning's coding. It shows the building up of the *Embedding* and *Transformer* layers of the model. The only change performed by the author (lines 53-64) was to add the commands for ignoring the changing of the parameters of such layers. Finally, it shows the creation of *Generator* (the classifier layer) which learns by not disabling the parameters changes during training.

```
49      # Transformer architecture and feedforward layer
50      self.transformer = nn.Transformer(d_model=d_model, batch_first=True, nhead=num_heads,
51       num_encoder_layers=num_encoder_layers, num_decoder_layers=num_decoder_layers,
52       dim_feedforward=dim_feedforward, dropout=dropout)
53      # # # # # # # # # # # # # # # # # # # # # # # #
54      #         Freeze to transfer-learning         #
55      # # # # # # # # # # # # # # # # # # # # # # # #
56      #                                             #
57      for param in self.embedding.parameters():   #
58          param.requires_grad = False             #
59      for param in self.transformer.parameters(): #
60          param.requires_grad = False             #
61      #                                             #
62      # # # # # # # # # # # # # # # # # # # # # # # #
63      #                  UnFreeze                   #
64      # # # # # # # # # # # # # # # # # # # # # # # #
65
66
67      self.generator = nn.Linear(d_model, target_size)
```

Figure 38 - Freezing layers learning.

The effectiveness of the transfer method will depend on a strong relationship between the source and the target tasks. If the source task is not sufficiently related or the relationship is not well leveraged, the performance will decrease, i.e., the transfer will negatively affect the performance [61].

4.4 Large language model

This dissertation was worked on for more than one year, in fact, since the first classes of Machine Learning and Computer Vision, when the author was a curious ignorant of such technologies. In the beginning, for every question or doubt, forums, sites, videos, and papers were the only solution, whenever teachers, supervisors and colleagues were unavailable. After the first few months of this year, this AI tool was increasingly used to understand basic concepts, techniques and code, and also to help debug, build, and develop code scripts. The use of this tool was restricted to this tasks alone.

5. Results

This chapter presents the results obtained by following the methodology presented on the previous chapter. It is divided into two sections, one to present the results of the tests made to the SOTA model and the other to present the results of the tests made to the new model.

5.1 Original Model trained on GRID

Having surpassed the natural updating hurdles, a logical, comprehensive, progressive, and conclusive set of tests was performed. These tests obtained and explored the original model's results, to have a standard for the results of the new database and new model.

The first task was to perform a simple inference trial of a 4-video sample size, with results shown in Figure 39.

```
  return torch._native_multi_head_attention(
Video: bgwr4n.mpg | Prediction: ['bin' 'green' 'with' 'r' 'four' 'now'] | runtime: 0:00:00.975452
Video: lban2n.mpg | Prediction: ['lay' 'blue' 'at' 'x' 'two' 'now'] | runtime: 0:00:00.927214
Video: lrbi1a.mpg | Prediction: ['lay' 'red' 'by' 'b' 'one' 'again'] | runtime: 0:00:00.923489
Video: lwwv6n.mpg | Prediction: ['lay' 'white' 'with' 'z' 'six' 'now'] | runtime: 0:00:00.889459
cmssantos@Carlos-Santos:~/LipsDontLie-landmark-based$
```

Figure 39 - Inference run - original model.

By visual analysis, one may read a 1 second per inference in a 3 second videos, and extract the data presented in Table 5.

Table 5 – Results from inference run on original model.

Sentences	Words						Accumulated words	
							Correct	Spoken
Prediction	bin	green	with	r	four	now	6	
Original	bin	green	with	r	four	now		6
Prediction	lay	blue	at	x	two	now	11	
Original	lay	blue	at	n	two	now		12
Prediction	lay	red	by	b	one	again	16	
Original	lay	red	by	i	one	again		18
Prediction	lay	white	with	z	six	now	21	
Original	lay	white	with	v	six	now		24

According to Equation (1), the accuracy was 21/24=87,5%. An observation was that the errors occurred on the small-sized token section, reinforcing the intuition that longer words are less prone to errors, as homovisemes' disambiguation can benefit from redundancy (e.g., "*x*" vs "*Supercalifragilisticexpialidocious*").

The following task was to infer from the same file repeatedly, and check if would vary the predictions and performance. These tests were performed varying videos and speakers, only targeting samples that showed inference errors, and with up to 200 repetitions. One of these trials is shown in Figure 40.

```
cmssantos@Carlos-Santos:~/LipsDontLie-landmark-based$ python3 inference.py
Loading models ...
Done!
/home/cmssantos/.local/lib/python3.10/site-packages/torch/nn/functional.py:4999: UserWarning: Support
ame type for both instead.
  warnings.warn(
/home/cmssantos/.local/lib/python3.10/site-packages/torch/nn/modules/transformer.py:562: UserWarning:
ively affect performance. Prefer to use a boolean mask directly. (Triggered internally at ../aten/src
  return torch._transformer_encoder_layer_fwd(
/home/cmssantos/.local/lib/python3.10/site-packages/torch/nn/functional.py:4999: UserWarning: Support
same type for both instead.
  warnings.warn(
/home/cmssantos/.local/lib/python3.10/site-packages/torch/nn/modules/activation.py:1160: UserWarning:
ively affect performance. Prefer to use a boolean mask directly. (Triggered internally at ../aten/src
  return torch._native_multi_head_attention(
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.838281
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.913116
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.809493
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.804225
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.830664
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.854747
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.803527
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.813855
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.829158
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.846987
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.820757
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.826813
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.815477
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.861293
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.805503
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.822544
Video: bbbs5s.mpg | Prediction: ['place' 'blue' 'with' 'n' 'five' 'soon'] | runtime: 0:00:00.835700
```

Figure 40 - Video repetitive inference.

By visual analysis, one may again read the same 1 second per inference in a 3 second videos, and extract the data, now, presented in Table 6.

Table 6 – Results from video repetitive inference.

Sentences	#	Words						Accumulated words	
								Correct	Spoken
Prediction	1	place	blue	with	n	five	soon	3	
Original		bin	blue	by	s	five	soon		6
Prediction	2	place	blue	with	n	five	soon	6	
Original		bin	blue	by	s	five	soon		12
Prediction	3	place	blue	with	n	five	soon	9	
Original		bin	blue	by	s	five	soon		18
...				...				...	
Prediction	200	place	blue	with	n	five	soon	600	
Original		bin	blue	by	s	five	soon		1200

Results show that the same prediction is obtained, whenever the same video is fed into the model. According to Equation (1), the accuracy was 3/6=6/12=9/18=…=600/1200=50,0%.

Another task was to train from the scratch with the original DB, to validate the training procedure ability. To the date of the delivery of this dissertation, the author was not able to obtain the training conditions in LipNet. A 200-epoch train was defined and performed. This 10 hour of training, resulted in a 79,72% validation, and 80,58% average epoch accuracies. The difference between these WRR results around 80%, and the 95,2% of LipNet[33] may have its origin in the limited numbers of epochs, or a different training dataset.

Test evaluation with 25 iterations, was then ran, confirming an 80,44% test accuracy, and revealing 14.953.270 as the model's total number of parameters, with nearly 99% dedicated to the transformer and the remaining 1% dedicated to the mouth landmarks detection.

In order to test if the SOTA model would accept LusaPt and GRID videos alike, an inference test was conducted testing on three files: one original GRID file; one LusaPt file named *xxx*.mpg; and a copy named *yyy*.mpg. The output of this test is visible on Figure 41.

```
olean mask directly. (Triggered internally at ../aten/src/ATen/native/transformers/attention.cpp:150.)
  return torch._native_multi_head_attention(
Video: bbaf2n.mpg | Prediction: ['bin' 'blue' 'with' 'p' 'four' 'soon'] | runtime: 0:00:00.813619
/mnt/c/users/user/Desktop/Mestrado/4_Semestre/Fase_3_Bolinha/GitHub/2.0_LipsDontLie/LipsDontLie-landmark-based/examples/video(normal)/xxx.mpg
Video: xxx.mpg | Prediction: ['place' 'white' 'at' 'p' 'nine' 'again'] | runtime: 0:00:03.373774
/mnt/c/users/user/Desktop/Mestrado/4_Semestre/Fase_3_Bolinha/GitHub/2.0_LipsDontLie/LipsDontLie-landmark-based/examples/video(normal)/yyy.mpg
Video: yyy.mpg | Prediction: ['place' 'white' 'at' 'p' 'nine' 'again'] | runtime: 0:00:01.663528
cmssantos@Carlos-Santos:/mnt/c/Users/user/Desktop/Mestrado/4_Semestre/Fase_3_Bolinha/GitHub/2.0_LipsDontLie/LipsDontLie-landmark-based$
```

Figure 41 - Original model inference on GRID and LusaPt.

The results show the new DB sample file acceptance. They also show an WRR of 3/6=50% when inferring from known words, again with errors on smaller tokens, and 0/6=0% when inferring from unknown words, what was expected.

Analysing the runtime, one can see:

i) A four-fold from the first run to the second, probably due to the nine fold [(1280x720)/(288x360)≈8,(8)] in resolution, and less probably due to the near doubling the number of frames. ;

ii) The lowering to half (1,663/3,374≈49,3%) from the second run (exact same content). No further tests were made to explore the evolution of runtime vs the number repeatedly inferring from the same file;

iii) The longest runtime is still under the actual video length.

After changing the vocabulary file to the new *vocabPT.txt*, an inference of a batch of miscellaneous videos was performed. As one can observe in Figure 42, 8 different files were predicted, 4 of LusaPt (different frames sizes and 720x1280 [px]), and 4 of GRID (75 frames and 288x360 [px] each).

```
Loading models ...
Done loading models!
Landmarks and Transformer!
(130, 720, 1280)
Video: 0gtfe.mpg | Prediction: ['3' '7' 'Cabo' 'Salada' 'Aguardente' 'Fantoche'] | runtime: 0:00:01.412401
./examplesPT/videos/0gtfe.mpg
(126, 720, 1280)
Video: 1vqap.mpg | Prediction: ['3' '7' '8' 'Guiné-Bissau' 'Amigos' 'Fantoche'] | runtime: 0:00:01.433014
./examplesPT/videos/1vqap.mpg
(145, 720, 1280)
Video: 2mpcv.mpg | Prediction: ['1' '5' 'Verde' 'Nata' 'Amigos' 'Fantoche'] | runtime: 0:00:01.476074
./examplesPT/videos/2mpcv.mpg
(133, 720, 1280)
Video: 3ccpa.mpg | Prediction: ['0' '5' 'Verde' 'Salada' 'Professores' 'Estrangeiros'] | runtime: 0:00:01.479814
./examplesPT/videos/3ccpa.mpg
(75, 288, 360)
Video: bbaf2n.mpg | Prediction: ['0' '4' 'Cabo' 'Porto' 'Amigos' 'Fantoche'] | runtime: 0:00:00.638091
./examplesPT/videos/bbaf2n.mpg
(75, 288, 360)
Video: bbaf3s.mpg | Prediction: ['1' '4' 'Verde' 'Nata' 'Aguardente' 'Fantoche'] | runtime: 0:00:00.600499
./examplesPT/videos/bbaf3s.mpg
(75, 288, 360)
Video: bbaf4p.mpg | Prediction: ['0' '4' 'Cabo' 'Porto' 'Amigos' 'Todos'] | runtime: 0:00:00.640663
./examplesPT/videos/bbaf4p.mpg
(75, 288, 360)
Video: bbaf5a.mpg | Prediction: ['0' '4' 'Verde' 'Portugal' 'Família' 'Fantoche'] | runtime: 0:00:00.653804
./examplesPT/videos/bbaf5a.mpg
carlossantos@Carlos-Santos:~/LDL_PT$
```

Figure 42 - Miscellaneous batch inference on original model.

Results show a full acceptance of the newly created files, the predictions with the 6 tokens-sized phrase format, the confirmation of longer runtime for higher resolution files, a 0/48=0% WRR, and the prediction of dummy word ("*Fantoche*"). This was the only time "*Fantoche*" was predicted, which may be understood as a good index for the robustness of the new model. It is also possible to read that the order of appearance of the predictions follows the order of appearance on the vocabulary. No further study was made on this matter.

5.2 Adjusted model applying LusaPt

Having fully validated the original model's updating, the training and the inference procedures, the next step was to train and test the model on the new DB. Following the proceedings defined

in the last chapter, the model was pre-trained on GRID, updated, and adjusted. The model was fine-tunned with LusaPt.

Tests were performed, with some minor errors in labelling, and a complete training print is annexed to this work, as an example. From this study, the author draws attention to the following results:

i) Original sentence: 7 Queijo Cerveja Alunos <eos>,
Predicted sentence: 7 Queijo Cerveja Alunos <eos> - According to subsection 4.1.2.1, after the Number's token, the Country's token is due to be inferred and is not, neither in the original nor in the predicted sentences. This missing token had its origins in the wrongful introduction of the two words long country *São Tomé*, written without the separating hyphen as was the case of *Guiné-Bissau*. This was corrected, and the error cleared;

ii) Original sentence: 6 Verde Sopa Fado Estrangeiros <eos>,
Predicted sentence: 6 Verde Verde Fado Estrangeiros <eos> - Although the model never trained with two equal consecutive words, this does not exclude such kind of inference. Other examples, besides this one reinforce this statement;

iii) <eos> - (end of string) is presented. To the date of the delivery of this dissertation, the author was not able to debug this. It is probably due to: a size compatibility issue that must be approached in a way that the model is able to accommodate phrases or sentences of various dimensions; any order to ignore this necessary component of the vocabulary file, that was not performed. Although this does not affect the performance of the model, nor of the DB, it is unwanted therefore will be targeted for future work;

iv) Validation accuracy: nan% - Although the author performed successful test evaluations, when working with the model cloned from GitHub and updated, to the delivery date of this dissertation, the author was not able to obtain a valid value of the running model validation accuracy. In the author's view, this occurred while calculating the mean of an empty matrix and may mean that some data did not arrive at this matrix. Validation is used to adjust hyperparameters, and to evaluate the performance during training. Although the overall results ratify the process, this bug is also unwanted and will be targeted for future work.

v) Average epoch acc: 97.81% - This accuracy results are in par with LipNet's. A higher value, when confirmed by the validation test, may confirm the leveraging of the learning transfer, added to the use of longer words, therefore more visemes to recur to;

vi) The dummy word (*Fantoche*) did not appear in the inferences.

The process results also validates the preliminary assumptions that: there is no specific temporal measurement type dependency; and the transfer learning method has a positive effect on the

performance, meaning a strong relationship between the source and the target tasks, therefore it applies to learn a new language from a model trained on another.

As an outlier, a false positive was detected in face/mouth landmarks. Inside the yellow mark in Figure 43, a false face is detected on the facial hair. This is a distraction to the model, another focus that will be processed by the back-end to get nowhere.

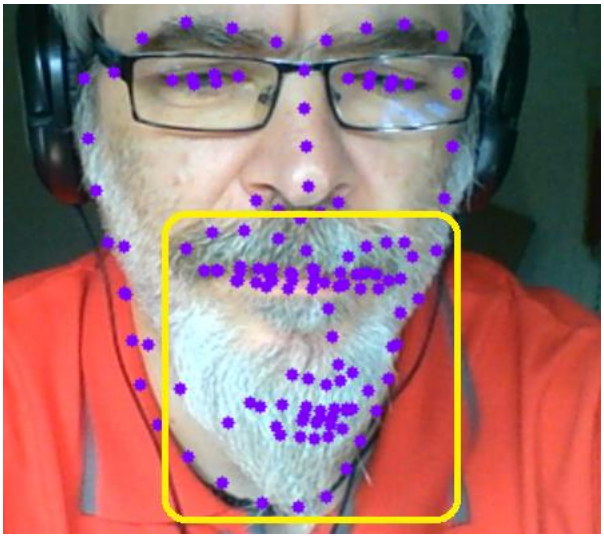

Figure 43 - False landmarks.

One may infer that, by being sporadic, this false input would not go far in the network and be discarded/forgotten.

6. Conclusion and future work

Automatic Lipreading is a tool for hearing impaired people, temporarily or otherwise. Its objective is to gradually diminish the gap of understanding of what is being spoken based on visual clues only, therefore, to improve their ability to communicate with others. This dissertation, review paper [60], and database are the author's contribution for this tool.

The production of large databases to train from scratch end-to-end deep learning models for visual speech recognition, demands considerable human effort, time, and financial resources inaccessible to most researchers. The development of technology brought new alternatives for model training, by transfer learning from large DB and fine-tunning with a small-sized DB, with LusaPt being one example. When compared to original large DBs, LusaPt presents a lower number of speakers and videos, yet is enough to perform model training and word prediction.

This chapter summarizes the contributions on this dissertation and addresses points that require attention and would be interesting to be analysed and developed in the future.

6.1 Conclusion

Audio speech recognition, a RoI not restricted to the mouth, emotions and context's recognition and combinations of these, will play a part in the future of overall artificial speech recognition. Its evolution will lay on top of this and surely others. This multiple and increasing input sources, are the reason for this dissertation's scope restriction, to visual input clues only ALR. This main objective was subdivided into the creation of a DB in the Portuguese language, and its validation on a SOTA model (LipNet), as the specific objectives.

Considering the performed literature review, an open challenge emerged: video speech recognition in Portuguese. A state-of-the-art end-to-end deep learning model was selected, composed by 20 points mouth landmarking to a transformer of double encoder and decoder layers and quadruple headed attention; an original DB was produced, composed by 10 speakers by uttering 10 phrases of 5 categories each; a train was performed, recurring to pre-train on a large available DB (around 32.000 videos in English) , and fine-tunning on the produced DB (100 videos on Portuguese).

The measure and comparation using WRR will always be a relative one, as happens with any other metric. A superior model predicting tokens with less visemes and higher occurrence of homovisemes, may have an inferior performance then a less robust model predicting tokens with more visemes and less prone to homovisemes. Having this in consideration, the values obtained by LusaPt are in par with the SOTA model on large DB (≈95%). It was the intention of the author for this dissertation to allow the increase of visemes per token, to increase the chance of recognition. Future works may focus on less visemes per token, or consider the inclusion of a homovisemes map as another input to the network.

Although more difficult to measure and prove, due to inherent ambiguity, it is the author's understanding that "*The fun theory*" helped on the videos recording attendance and on the appraisal of computer vision, deep learning, and ALR. This understanding was somehow substantiated by a posterior feedback informal query. At first, this may seem a dissertation's unworthy subject, however as the dissertation lays on the construction of a volunteer based database, and its quality lays on diversity (which implies quantity), it all adds up and may not be a trivial matter.

A prediction of Portuguese digits and words was performed for the first time, to this dissertation date, and to the best of the author's knowledge. If otherwise and there is (are) other(s), it was a missed opportunity to include it (them) on this dissertation, and is an opportunity to combine and learn from, in future works. Nevertheless, this work contributes to the development of DB corpora, requiring less human and financial exigencies.

6.2 Future work

In due time the author left unfinished the task to edit and label the remainder videos, besides the debugging mentioned in section 5.2, to further improve the contribution. LusaPt is projected to have over 250 valid, correct, and labelled videos, available to the academic community, in the standard presented in this dissertation. It was constructed with an today's everyday equipment, thus on a minimal budget compared to the firsts and larger DBs, and obtaining a higher resolution than not a long time ago. Maybe in a decade, there will be a system that performs lip reading, integrated on a then everyday equipment, and resorting to online or mini-batch training, possibly by the user himself. This is a goal to aim for, in the author's understanding.

As the inference runtime normally took equal or lesser time than the video itself, this work maintains the open door for the asymptote of simultaneous ALR. However, disambiguation seems to be unavoidable as a necessity, as shows the example on Table 2 or the existence of homovisemes. However, unavoidable also seems to be ALR's evolution. These contributions, when available, will allow and support other researchers to "easily" apply this model and database producing method, increase the overall vocabulary, vary the conditions in which the video is captured, or to use simultaneous speakers, and in that way also contribute to ALR evolution, no matter the language.

In one of this dissertation's trials, there was no prediction change when repeatedly inferring from the same file. The author understands that either this model strongly "believes" in the its own predictions and the standard variation was not significant to vary even once in about 400 inferences, or it is absolutely sure and considers no doubt in both process or conclusions. In case of the last hypothesis, this is a good example of the differences between Artificial and Human Intelligences. As the author views, we consider some degree of doubt, even on our own conclusions, to be as Human as Intelligent.

7. Publications

From work developed and presented in this dissertation, were also produced the following results:

Santos, C., Cunha, A., Coelho, P. (2023). A Review on Deep Learning-Based Automatic Lipreading. In: Cunha, A., M. Garcia, N., Marx Gómez, J., Pereira, S. (eds) Wireless Mobile Communication and Healthcare. MobiHealth 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 484. Springer, Cham. https://doi.org/10.1007/978-3-031-32029-3_17

References

[1] D. Ramoo, M. Bartlett, BCcampus, and BC Open Textbook Project, *Psychology of language*. 2021. Accessed: Apr. 18, 2022. [Online]. Available: https://open.bccampus.ca/browse-our-collection/find-open-textbooks/?uuid=7a70435d-ff9f-4d0b-8bf4-18062806d395&contributor=&keyword=&subject=

[2] P. Lieberman, 'Why human speech is special', *The scientist*, Jul. 2018, Accessed: Apr. 18, 2022. [Online]. Available: https://www.the-scientist.com/features/why-human-speech-is-special--64351

[3] D. Dalva, 'Automatic speech recognition system for Turkish spoken language', 2012.

[4] Ian S Howard, 'Towards a mechanical vocal apparatus for vowel production', presented at the ESSV, Leipzig, Germany, Mar. 2016.

[5] H. Sawada, 'Talking Robot and the Autonomous Acquisition of Vocalization and Singing Skill', in *Robust Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds., I-Tech Education and Publishing, 2007. doi: 10.5772/4761.

[6] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[7] H. Hamooni, A. Mueen, and A. Neel, 'Phoneme sequence recognition via DTW-based classification', *Knowl Inf Syst*, vol. 48, no. 2, pp. 253–275, Aug. 2016, doi: 10.1007/s10115-015-0885-9.

[8] A. S. Geylanioğlu, 'Developing English language learners' pronunciation through conceptualization', MA Thesis). Retrieved from Yükseköğretim Kurulu Başkanlığı.(Accesion No. 448446), 2016.

[9] A. Sucena, S. L. Castro, and P. Seymour, 'Developmental dyslexia in an orthography of intermediate depth: the case of European Portuguese', *Read Writ*, vol. 22, no. 7, Art. no. 7, Aug. 2009, doi: 10.1007/s11145-008-9156-4.

[10] M. M. Azevedo, *Portuguese: a linguistic introduction*. Cambridge, UK ; New York: Cambridge University Press, 2005.

[11] W. H. Sumby and I. Pollack, 'Visual Contribution to Speech Intelligibility in Noise', *The Journal of the Acoustical Society of America*, vol. 26, no. 2, Art. no. 2, Mar. 1954, doi: 10.1121/1.1907309.

[12] N. Bauman, 'Speechreading (Lip-Reading)'. Accessed: Apr. 18, 2022. [Online]. Available: https://hearinglosshelp.com/blog/speechreading-lip-reading/

[13] M. E. Bruhn, *The Müller-Walle Method*, U.S. Department of Health, Education and Wellfare, Public. National Library of Medicine, 2015.

[14] A. B. A. Hassanat, 'Visual Speech Recognition', in *Speech and Language Technologies*, I. Ipsic, Ed., InTech, 2011. doi: 10.5772/19361.

[15] H. Mcgurk and J. Macdonald, 'Hearing lips and seeing voices', *Nature*, vol. 264, no. 5588, Art. no. 5588, Dec. 1976, doi: 10.1038/264746a0.

[16] Q. Summerfield, 'Audio-visual Speech Perception, Lipreading and Artificial Stimulation', in *Hearing Science and Hearing Disorders*, Elsevier, 1983, pp. 131–182. doi: 10.1016/B978-0-12-460440-7.50010-7.

[17] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, 'VisemeNet: Audio-Driven Animator-Centric Speech Animation', 2018, doi: 10.48550/ARXIV.1805.09488.

[18] A. B. Hassanat, 'Visual Words for Automatic Lip-Reading', 2014, doi: 10.48550/ARXIV.1409.6689.

[19] E. G. Nassimbene, 'Electronic Lip Reader', 3192321, Jun. 29, 1965

[20] E. D. Petajan, ‘Automatic lipreading to enhance speech recognition’, degree of Doctor of Philosophy in Electrica l Engineering, University of Illinois, Urbana-Champaign, 1984.
[21] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, ‘Deep Learning-Based Automated Lip-Reading: A Survey’, *IEEE Access*, vol. 9, pp. 121184–121205, 2021, doi: 10.1109/ACCESS.2021.3107946.
[22] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, ‘Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks’, in *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 369–376. doi: 10.1145/1143844.1143891.
[23] A. Fernandez-Lopez, O. Martinez, and F. M. Sukno, ‘Towards Estimating the Upper Bound of Visual-Speech Recognition: The Visual Lip-Reading Feasibility Database’, in *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, Washington, DC, DC, USA: IEEE, May 2017, pp. 208–215. doi: 10.1109/FG.2017.34.
[24] S. Fenghour, D. Chen, K. Guo, B. Li, and P. Xiao, ‘An Effective Conversion of Visemes to Words for High-Performance Automatic Lipreading’, *Sensors*, vol. 21, no. 23, p. 7890, Nov. 2021, doi: 10.3390/s21237890.
[25] S. Fenghour, D. Chen, K. Guo, and P. Xiao, ‘Lip Reading Sentences Using Deep Learning With Only Visual Cues’, *IEEE Access*, vol. 8, pp. 215516–215530, 2020, doi: 10.1109/ACCESS.2020.3040906.
[26] M. Hao, M. Mamut, N. Yadikar, A. Aysa, and K. Ubul, ‘A Survey of Research on Lipreading Technology’, *IEEE Access*, vol. 8, pp. 204518–204544, 2020, doi: 10.1109/ACCESS.2020.3036865.
[27] A. Goldschen, O. Garcia, and E. D. Petajan, ‘Continuous optical automatic speech recognition by lipreading’, *IEEE*, doi: 10.1109/ACSSC.1994.471517.
[28] A. Goldschen, O. Garcia, and E. D. Petajan, ‘Continuous automatic speech recognition by lipreading’, *Springer*, vol. Motion-Based Recognition, 1997.
[29] H. Huang *et al.*, ‘A Novel Machine Lip Reading Model’, *Procedia Computer Science*, vol. 199, pp. 1432–1437, 2022, doi: 10.1016/j.procs.2022.01.181.
[30] G. Zhao, M. Pietikäinen, and A. Hadid, ‘Local Spatiotemporal Descriptors for Visual Recognition of Spoken Phrases’, *Proc. ACM Int. Multimedia Conf. Exhib*, pp. 57–66, 2007.
[31] M. Hao, M. Mamut, and K. Ubul, ‘A Survey of Lipreading Methods Based on Deep Learning’, in *2020 2nd International Conference on Image Processing and Machine Vision*, Bangkok Thailand: ACM, Aug. 2020, pp. 31–39. doi: 10.1145/3421558.3421563.
[32] M. Cooke, J. Barker, S. Cunningham, and X. Shao, ‘Grid AV speech corpus sample’. Mar. 22, 2013.
[33] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, ‘LipNet: End-to-End Sentence-level Lipreading’. arXiv, Dec. 16, 2016. Accessed: Jun. 04, 2022. [Online]. Available: http://arxiv.org/abs/1611.01599
[34] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, ‘Lip Reading Sentences in the Wild’, 2016, doi: 10.48550/ARXIV.1611.05358.
[35] J. S. Chung and A. Zisserman, ‘Lip Reading in Profile’, 2017.
[36] S. Yang, Y. Zhang, D. Feng, and M. Yang, ‘LRW-1000: A Naturally-Distributed Large-Scale Benchmark for Lip Reading in the Wild’, IEEE, 2019. doi: 10.1109/FG.2019.8756582.
[37] S. Petridis, Y. Wang, P. Ma, Z. Li, and M. Pantic, ‘End-to-End Visual Speech Recognition for Small-Scale Datasets’, 2019, doi: 10.48550/ARXIV.1904.01954.

[38] S. Jeon, A. Elsharkawy, and M. S. Kim, 'Lipreading Architecture Based on Multiple Convolutional Neural Networks for Sentence-Level Visual Speech Recognition', *Sensors*, vol. 22, no. 1, p. 72, Dec. 2021, doi: 10.3390/s22010072.
[39] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016.
[40] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, 'ImageNet: A large-scale hierarchical image database', in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, Jun. 2009, pp. 248–255. doi: 10.1109/CVPR.2009.5206848.
[41] F. Chollet, *Deep learning with Python*, Second edition. Shelter Island: Manning Publications, 2021.
[42] M. Alam, M. D. Samad, L. Vidyaratne, A. Glandon, and K. M. Iftekharuddin, 'Survey on Deep Neural Networks in Speech and Vision Systems', *Neurocomputing*, vol. 417, pp. 302–321, Dec. 2020, doi: 10.1016/j.neucom.2020.07.053.
[43] B. Martinez, P. Ma, S. Petridis, and M. Pantic, 'Lipreading using Temporal Convolutional Networks', 2020, doi: 10.48550/ARXIV.2001.08702.
[44] J. Long, E. Shelhamer, and T. Darrell, 'Fully Convolutional Networks for Semantic Segmentation'. arXiv, Mar. 08, 2015. Accessed: Jun. 25, 2022. [Online]. Available: http://arxiv.org/abs/1411.4038
[45] C. Wang, 'Multi-Grained Spatio-temporal Modeling for Lip-reading', 2019, doi: 10.48550/ARXIV.1908.11618.
[46] R. N, S. A, and N. K. A, 'Visual Speech Recognition using Fusion of Motion and Geometric Features', *Procedia Computer Science*, vol. 171, pp. 924–933, 2020, doi: 10.1016/j.procs.2020.04.100.
[47] Y. Zhang, S. Yang, J. Xiao, S. Shan, and X. Chen, 'Can We Read Speech Beyond the Lips? Rethinking RoI Selection for Deep Visual Speech Recognition', 2020, doi: 10.48550/ARXIV.2003.03206.
[48] S. K. Das, S. Nandakishor, and D. Pati, 'Automatic lip contour extraction using pixel-based segmentation and piece-wise polynomial fitting', in *2017 14th IEEE India Council International Conference (INDICON)*, Roorkee: IEEE, Dec. 2017, pp. 1–5. doi: 10.1109/INDICON.2017.8487538.
[49] Y. Lu and Q. Liu, 'Lip segmentation using automatic selected initial contours based on localized active contour model', *J Image Video Proc.*, vol. 2018, no. 1, p. 7, Dec. 2018, doi: 10.1186/s13640-017-0243-9.
[50] Y. Lu, X. Zhu, and K. Xiao, 'Unsupervised lip segmentation based on quad-tree MRF framework in wavelet domain', *Measurement*, vol. 141, pp. 95–101, Jul. 2019, doi: 10.1016/j.measurement.2019.03.009.
[51] X. Ma, H. Zhang, and Y. Li, 'Feature Extraction Method for Lip-reading under Variant Lighting Conditions', in *Proceedings of the 9th International Conference on Machine Learning and Computing*, Singapore Singapore: ACM, Feb. 2017, pp. 320–326. doi: 10.1145/3055635.3056576.
[52] T. Afouras, J. S. Chung, and A. Zisserman, 'ASR is all you need: cross-modal distillation for lip reading'. arXiv, Mar. 31, 2020. Accessed: Jun. 02, 2022. [Online]. Available: http://arxiv.org/abs/1911.12747
[53] X. Weng and K. Kitani, 'Learning Spatio-Temporal Features with Two-Stream Deep 3D CNNs for Lipreading'. arXiv, Jul. 18, 2019. Accessed: Jul. 10, 2022. [Online]. Available: http://arxiv.org/abs/1905.02540
[54] Y. Lu and J. Yan, 'Automatic Lip Reading Using Convolution Neural Network and Bidirectional Long Short-term Memory', *Int. J. Patt. Recogn. Artif. Intell.*, vol. 34, no. 01, p. 2054003, Jan. 2020, doi: 10.1142/S0218001420540038.

[55] A. Mesbah, A. Berrahou, H. Hammouchi, H. Berbia, H. Qjidaa, and M. Daoudi, ‘Lip reading with Hahn Convolutional Neural Networks’, *Image and Vision Computing*, vol. 88, pp. 76–83, Aug. 2019, doi: 10.1016/j.imavis.2019.04.010.

[56] Y. Lu and H. Li, ‘Automatic Lip-Reading System Based on Deep Convolutional Neural Network and Attention-Based Long Short-Term Memory’, *Applied Sciences*, vol. 9, no. 8, p. 1599, Apr. 2019, doi: 10.3390/app9081599.

[57] K. R. Prajwal, T. Afouras, and A. Zisserman, ‘Sub-word Level Lip Reading With Visual Attention’. arXiv, Dec. 03, 2021. Accessed: May 28, 2022. [Online]. Available: http://arxiv.org/abs/2110.07603

[58] A. K. Gupta, P. Gupta, and E. Rahtu, ‘FATALRead - Fooling visual speech recognition models: Put words on Lips’, *Appl Intell*, Nov. 2021, doi: 10.1007/s10489-021-02846-w.

[59] M. Cooke, J. Barker, S. Cunningham, and X. Shao, ‘An audio-visual corpus for speech perception and automatic speech recognition’, *26 June 2006*, pp. 2421–2424, doi: 10.1121/1.2229005.

[60] C. Santos, P. Coelho, and A. Cunha, ‘A Review on Deep Learning-based Automatic Lipreading’, *MobiHealth 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, vol. 484, pp. 180–195, doi: https://doi.org/10.1007/978-3-031-32029-3_17.

[61] E. Olivas, J. Guerrero, M. Sober, J. Benedito, and A. Lopez, *Handbook of Research on Machine Learning Applications,*. Information science reference, 2009.

Appendix