Universidade de Aveiro
Departamento de
Eletrónica, Telecomunicações e Informática
2016

Nuno Miguel
Carreira Vitor

Silent Speech Interface for an AAL scenario.

Interfaces de Fala Silenciosa para um cenário AAL.

Universidade de Aveiro
Departamento de
Eletrónica, Telecomunicações e Informática
2016

Nuno Miguel
Carreira Vitor

Silent Speech Interface for an AAL scenario.

Interfaces de Fala Silenciosa para um cenário AAL.

Nuno Miguel Carreira Vitor

Silent Speech Interface for an AAL scenario.

Interfaces de Fala Silenciosa para um cenrio AAL.

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Eletrónica e Telecomunicações, realizada sob a orientação do Doutor António Joaquim da Silva Teixeira, Professor Associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro.

o júri / the jury

presidente / president

Doutora Beatriz Sousa Santos
Professora Associada com Agregação da Universidade de Aveiro

vogais / examiners committee

Doutor António Joaquim da Silva Teixeira
Professor Associado da Universidade de Aveiro (orientador)

Doutor João Dinis Colaço de Freitas
CTO, DefinedCrowd

Acknowledgements

First of all, I want to thank to my father and my mother who unfortunately died 2 years ago, but helped me to become the person I am today. I also want to thank to my grandmother. Without their support and love, I would not achieve this. Thank you so much.

I want to thank Prof. António Teixeira for his guidance and help through the whole work and all months.

I also want thank to Daniel and Ana who helped me getting more results for this thesis. Thank you all.

Resumo

Desde a década de 80 que começaram a surgir estudos relacionados com o reconhecimento audiovisual da fala. Contudo, chegou-se à conclusão que, em certas circunstâncias, o uso da informação audio não poderia ser considerada devido a ambientes ruidosos ou outro tipo de condicionantes. Desde então, começaram a realizar-se estudos tendo em conta o reconhecimento visual da fala.

Com o lançamento da Kinect por parte da Microsoft, que inclui câmera RGB, sensor de profundidade e microfone por um custo relativamente baixo comparativamente a outras câmeras do mesmo segmento, abriu novas portas e trouxe novas possibilidades no âmbito do reconhecimento da fala. Com o lançamento da Kinect One em 2014, uma câmera com maior resolução e um sensor de profundidade com tecnologia de "tempo de voo", mais precisa, permite ainda obter melhores resultados e abrir ainda mais portas no que toca ao reconhecimento visual da fala.

Esta dissertação foi desenvolvida com base na Kinect One da Microsoft e tem como objectivo o reconhecimento visual da fala, mais especificamente de comandos, em Português, ditos pela pessoa que se encontra de frente para a câmara, com o intuito de controlar o VLC, uma aplicação relevante para um cenário AAL, um player de conteúdos multimédia, o mais utilizado em todo o mundo.

O sistema desenvolvido encontra-se assim projetado para uma realidade de ambiente assistido, para pessoas com dificuldades motoras ou apenas como uma ferramenta de auxílio para uma melhor experiência cinematográfica em casa sem a necessidade do uso de um controlo remoto.

O protótipo segue a abordagem clássica em reconhecimento de padrões, integrando extração de features e classificação. As features adotadas no protótipo realizado foram a posição dos lábios e a posição do queixo. Em termos dos classificadores foram experimentados os algoritmos Support Vector Machine (SVM), Random Forest, Sequential Minimal Optimization (SMO), AdaBoost e Naive Bayes.

O protótipo no decorrer desta dissertação demonstrou conseguir atingir taxas de reconhecimento na ordem dos 80 por cento num mundo de 8 comandos escolhidos de forma a serem o mais intuitivos possível tendo em conta o objectivo desta tese, controlar o reprodutor VLC usando reconhecimento visual da fala.

Abstract

Since the 80's started to emerge studies regarding the audio-visual recognition of speech. However, in certain circumstances, the use of the audio information can not be considered due to noisy environments or other types of conditioning. Since than, studies started to emerge regarding visual speech recognition.

With the launch of Kinect by Microsoft, which includes a RGB, depth sensor and microphone for a relatively low price compared to other cameras in its segment, permited new possibilities in the speech recognition field. The launch of Kinect One in 2014 brought a new RGB-D camera with bigger resolution and a depth sensor with "Time of Flight" technology, more precise, witch allows to get better results and better accuracy in Visual Recognition Systems.

This dissertation was developed with the Kinect One from Microsoft and has the objective of Visual Speech Recognition, especially commands, in Portuguese, said by the person that is standing in front of the camera, with the intention of controlling the VLC player, a relevant application VLC for an Ambient Assisted Living (AAL) scenario, a multimedia player, the most used in the world.

The system developed in this dissertation is projected for an AAL scenario, for people with speech incapacity, noisy environments or only to improve and create a better home cinema experience, without the need for a remote control.

The prototype follows a classic approach in pattern recognition, integrating features and classifiers. The adopted features were the position of the lips and chin. In therms of classifiers the Support Vector Machine (SVM), Random Forest, Sequential Minimal Optimization (SMO), AdaBoost and Naive Bayes algorithms were tested.

The prototype developed in this dissertation achieved an accuracy of around 80 percent in a universe of 8 commands chosen to be the most intuitive as possible regarding the objective of this dissertation, to create a working prototype (VLC as chosen) using visual speech recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In a Human-Computer Interaction based in spoken language sometimes the ambient noise or the need for privacy is a reality, which results in a limitation to the interaction process. Regarding this necessities and taking in consideration speech interaction between humans and machines, a concept of Silent Speech is needed.

Silent Speech Interfaces (SSI) do not use acoustic signals to recognize a word or a sequence of words from the speaker. In fact, the audible acoustic noise is just the result and the end of a complex process of speech production [Freitas et al., 2016]. Silent Speech Interfaces uses a set of other different information to tackle the decision process. These types of information can be, for example, Depth, Surface, Electromyography, Ultrasonic Doppler [Freitas et al., 2013] or information taken from the lip movements [Abreu, 2014].

Silent Speech can be very helpful in some particular cases. Cases that the speaker has a speech handicap or as part of a communications system operating in silence-required or high-background noise environments [Denby et al., 2010]. It is also suitable for situations where privacy and confidentiality is required [Freitas et al., 2013]. A laryngectomyzed patient could use this kind of system (SSI) as an alternative to oesophageal speech, tracheo-oesophageal speech or the electrolarynx [Hueber et al., 2008].

1.1 Motivation

In some cases speech recognition can not be achieved using the acoustic signal of the speaker due to high-background noise environments or if the speaker is deaf for example. To solve these problems, many people started working in Silent Speech solutions. Nowadays, we live in an era where technology is in a great progress and new technology devices are constantly hitting the market.

Back in 1952, in the Bell labs, was created the world's first approach to speech recognition. However, this system was based only in the speaker voice and it was made to recognize spoken digits. Despite being the first system in speech recognition ever created, it had a good accuracy between 97 and 99 percent [Davis et al., 1952], however the capacity in terms of what the system could recognize were limited. Till late 70s speech recognition systems were only based in the speaker voice. Then, psychologist Harry McGurk and his assistant, John MacDonald, discovered by accident an audio-visual illusion, known since then as the McGurk effect or McGurk illusion, leading to bimodal speech recognition systems [Abreu, 2014].

Despite the good results in bimodal or multimodal speech recognition systems, sometimes

the use of the speaker voice is not possible, due to the noisy environments, privacy and confidentiality. For example, a noisy environments where the Silent Speech Interface can be helpful is by reading the lips of a speaker that is watching television. The noise coming from the TV added to the voice of the speaker turn out to be impossible to take the voice of the speaker as input to decide what he is saying. In this case, with a SSI, the speaker can, for example, turn up the volume or switch channel by the movement of his lips.

Silent Speech interfaces are needed in a Human-Machine interaction to control or communicate with the machines using speech. The need of an interaction that can be done without using acoustic signals from the speech production takes into account its usability in noisy environments, cases where privacy and confidentiality are required or in cases that the speaker has some kind of speech handicap.

In the Department of Electronics, Telecommunication and Informatics of the University of Aveiro and other research departments the interest of research for the interaction with high complex devices like smart phones for Ambient Assisted Living (AAL) proposes or domotics is increasing. Speech is the easy way for humans to communicate but sometimes the constant noisy environments due to sound of televisions or music made those systems not able to use audible acoustic signal in the recognizing process. Than Silent Speech systems where created. Recently experiences were done regarding Silent Speech systems for European Portuguese [Abreu, 2014] , [Freitas et al., 2013] to analyze the capability of recognition by machines using SSI. To evaluate the potential practical usability of a SSI in European Portuguese, originated the work reported in this dissertation.

1.2 Problems / Challenges

It is not easy to create a real time application for silent speech recognition due to the complexity of the code and algorithms that are need to tackle the recognition process. If the complexity is reduced, the real time experience is increased but the accuracy is reduced.

One of the problems in SSI and Visual Speech Recognition (VSR) systems is that the use of the sound does not enter in the equation, and of course in a multimodal approach the sound of the speaker saying a word is a good clue in the aim to discover what he/she is saying. However, it's possible to tackle the discovery of the words using other type of information like the lips, tongue, the chin, the ears etc.

One big problem about VSR systems over Automatic Speech Recognition (ASR) systems is that ASR can have an accuracy far more superior than the VSR systems and in VSR systems the accuracy ranges can oscillate quite a bit, between 42 percent and 66.9 percent [Abreu, 2014].

Another problem tends to be the creation of a system totally automatic on the task of recognizing words. For example, in the direct case of this dissertation, it is complicated to create a system that allows people, to stand in front of the television, talking to other people and the system being capable of recognizing the moment when the speaker announces a command to be sent to VLC.

The motion of the head of the speaker is another problem as well. If the speaker moves the head joint too much during a word, the evaluation of the word pronounced would be harder.

1.3 Objectives

The main objective of this dissertation is to create a working prototype capable of recognizing a small set of words/commands, in real time.

In terms of scenario the aim is to address one in the Ambient Assisted Living (AAL) area. The chosen scenario was VLC controlling using SSI.

It is not an objective to create a speaker dependent system only, if possible, to assess the speaker dependency of the developed system.

1.4 Dissertation structure

Chapter 2 presents a review of the speech production process in human beings, speech perception, SSI basics, used types of cameras and technologies in VSR systems and a review of the state of the art in Silent Speech recognition systems.

Chapter 3 includes all the information needed to understand the system produced in this dissertation like the way the features were extracted, automatic activity detection and classification process.

Chapter 4 presents results from several evaluations made to the system and its components developed in this dissertation. Some tables and confusion matrices were developed to better understand the performance of the system. The Speaker Dependency, Distance Dependency, Live performance and best classification algorithms are also investigated.

Last, in chapter 5, conclusions about this dissertation are available to discuss the all performance of the system developed. Some comments are made and some future work suggestions are also made.

Chapter 2

Background and Related work

2.1 Background

Due to the complexity of the speech production and perception process, in this section will be presented some topics needed to understand the SSI basics, how the SSI are usually made, different possibilities in SSI systems, how different effects of the speech production can be obtained for SSI purposes, how and what features are extracted (using cameras like RGB, depth or both) and different algorithms used during the classification process.

2.1.1 Speech production

Speech production requires a complex series of events and is considered the most complex motor task performed by humans [Seikel et al., 2009]. In a fluent conversation, we are able to produce two or three words per second. This is possible thanks to the mental lexicon, which contains at least 50-100 thousand words in a normal, literate adult person. However, the mental lexicon does not have problems to deal with the complexity and speed of the word production, since it does not fails more than once or twice in 1000 words [Levelt, 1999]. The speech production model (see Figure 2.1) can be divided into several stages [Freitas et al., 2011]. In order, these stages are: Conceptual preparation (thought), Formulation (linguistic representation) and finally Articulation (facial movements + acoustics).

Conceptual preparation consists in the activation of a lexical concept for which you have a word in your lexicon [Levelt, 1999].

Formulation is the process corresponding to the linguistic representation required for the expansion of the desired message [De Smedt, 1996].

Articulation is a series of neuromuscular commands to cause the vocal cords to vibrate, producing an acoustic signal as the final output. These commands must simultaneously control all the aspects of the articulatory motion, including the lips, jaw, tongue and velum [Rabiner and Juang, 1993].

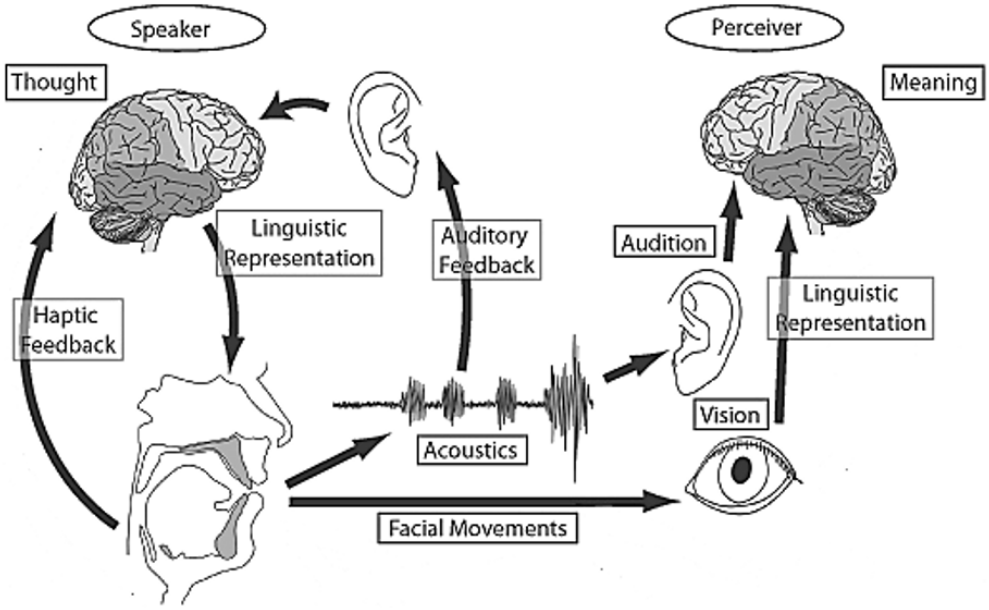

Figure 2.1: Multimodal speech chain representation with feedback loops [Gick et al., 2013].

After an intent to express has been developed, our brain maps it into muscular movements. The nervous system takes charge of activating the motor unit and the associated activation rate. Nerve impulses go from anterior horn of spinal cord to the spinal column and then to the end of the nerve via motor neurons. These neurons send the signals from the brain to the exterior body parts through axons and then the neuromuscular junction. When a nerve impulse reaches this junction causes sodium and potassium cation (positive charge) channels in the muscle fiber. This generates an electromagnetic field in the area surrounding the muscle fibers.

In speech production, the articulatory muscles like tongue represent a vital role because they can shape the air stream into a recognizable speech. Mandibular movement also represents an important role in this process.

Although the importance of cavities, surfaces and organs in speech production, the articulators are the most important in the pronunciation of different sounds regarding the language that the person is talking. Its position defines the articulatory and resonant characteristics of the vocal tract.

Articulators can be active or passive. The active articulators include the lips, tongue, lower jaw and velum, being the tongue the most important of them all, participating in the majority of the sounds. The passive include the teeth, alveolar ridge and hard palate. In Figure 2.2 is represented a sagittal view of several articulators.

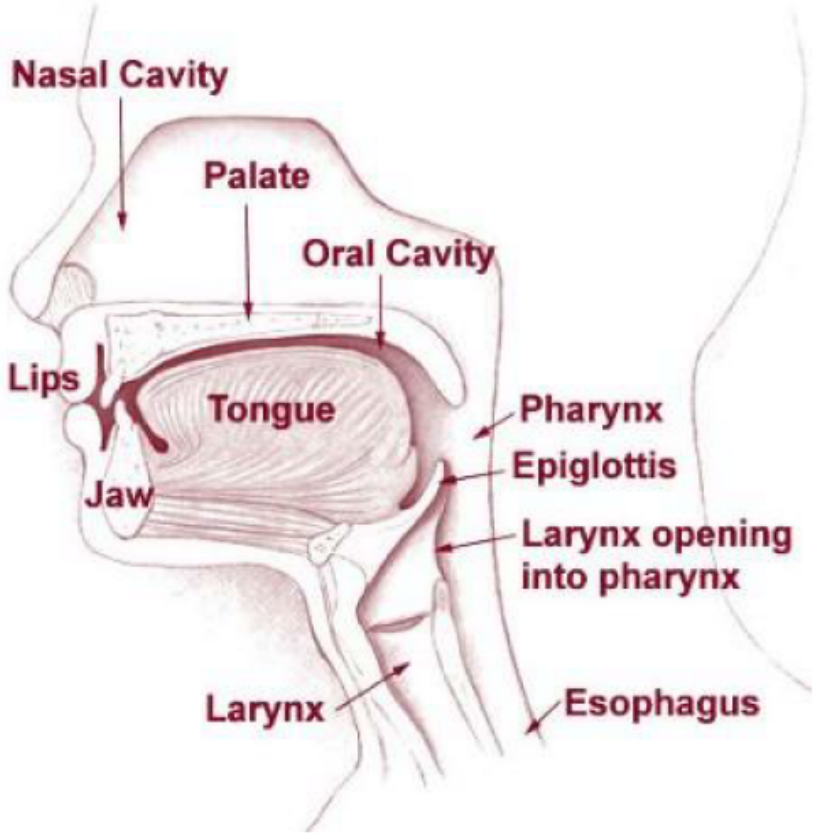

Figure 2.2: Sagittal view of the vocal tract depicting its main regions and several articulators [Freitas et al., 2016].

The most visible effects of the speech production chain are the movement of the lips, tongue, lower jaw and indirectly the chin.

2.1.2 Speech perception

As humans, we can understand and process a large amount of words in a small amount of time. Just with the speech information, we can take conclusions about the genre of the speaker, the age or the regional dialect. Humans are capable of processing words from another person even in noisy environments.

Speech perception is one of the main issues regarding the building of systems concerning the processing of speech. The human system for speech recognition and understanding is the one known example of a robust speech recognizer, since it is insensitive to variability over the range of non-linguistic factors in the speech signal [Gold et al., 2011].

Unlikely what people may think, the speech perception process is not only based in something we hear, it is a multimodal process. There is now overwhelming evidence that the brain treats speech as something we hear, see, and even feel [Rosenblum, 2013]. Humans speech perception rely on sound and visual information like visible movements of the lips, tongue and gestures.

2.1.3 SSI basics

A Silent Speech interface comprises a system that interprets human signals other than the audible acoustic signal enabling speech communication [Denby et al., 2010].

A SSI system is commonly characterized by the acquisition of information from the human speech production process such as articulations, facial muscle movement or brain activity. These systems represents a potential solution for communications in noisy environments, a help for people with disabilities at the vocal tract level such as the elderly.

It is possible to say that the SSI systems extends the human speech production process by exploring biometric signals other that voice, using sensors, cameras, ultrasonic waves and so on [Freitas et al., 2016].

Nowadays there are multiple works in SSI done in every stage of the speech production stage. For example in the first stage (Conceptualization) works were done on the interpolation of the signals from implants in the speech-motor cortex [Brumberg et al., 2010] and from Electroencephalography (EEG) sensors [Porbadnigk et al., 2009]. In Articulation stage works were made regarding the movement of the lips [Wand et al., 2016] , [Abreu, 2014] or the movements of a speaker's face through Ultrasonic Doppler sensing for example [Freitas et al., 2012]. The usual architecture of a SSI system is depicted in Figure 2.3. It is composed by some Facial Image Processing, Features Extraction, Classification process and Fusion.

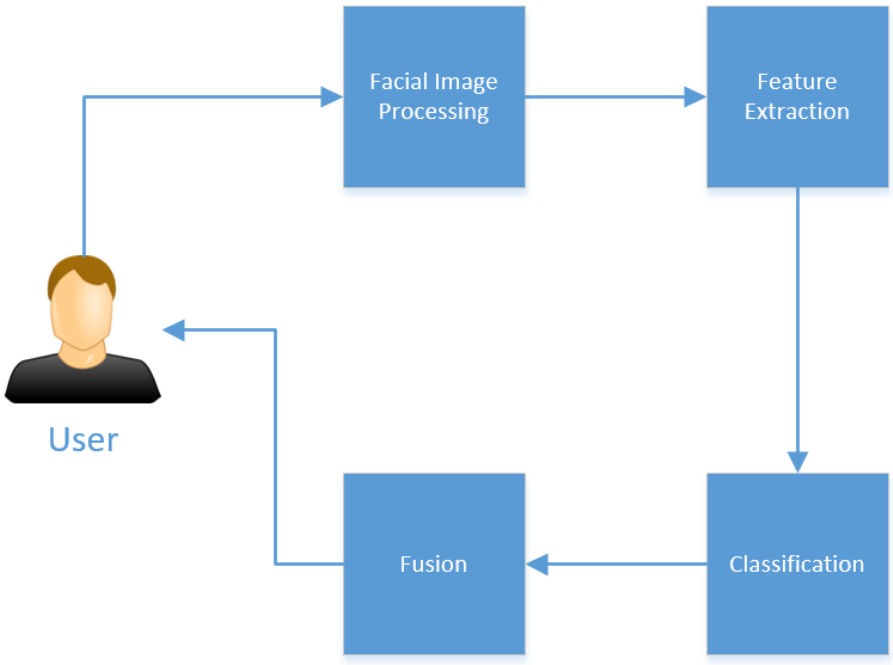

Figure 2.3: SSI system usual architecture.

The signals from the selected stage of the human speech production process can be done using an invasive or obtrusive modality or non of them. The invasive modality needs a medical attention to be used or requires the use of sensors. The obtrusive modality requires wearing some type of equipment like gloves. Choosing the best between these implementations is not an easy task because they have different advantages and disadvantages such as the price, the

usability, the accuracy, the speaker's independence.

An example of a SSI system that is invasiveness and obtrusiveness is the one used by Helder Abreu in his dissertation [Abreu, 2014]. For the system to be able to understand what a speaker is saying (in a speech perception scenario but taking into account a machine approach), it is common to extract features like the width and height of the mouth and then a classification process is done to classify the words. These kind of characteristics are stored in a suitable model (commonly known as feature vector) to be used next in the classification process of the word. This process can be compared with the meaning attribution on the human perception [Moore and Cutler, 2001].

2.1.4 Methods to collect visual information in SSI systems

The most used way to collect visual information from the speech production process that is non-invasive and non-obtrusive is through cameras. There are some different cameras on the market like RGB cameras that collect information on the color space, depth cameras that collect depth information through the stereo vision approach, infrared or time of flight technology. Since 2012, the Kinect for Windows camera is available, and it uses both types of technologies at an affordable price. In this subsection these cameras will be discussed.

2.1.4.1 RGB cameras

RGB cameras are used to collect information pixel per pixel in a RGB color space. Today these kind of cameras uses CMOS or charge-coupled device (CCD) image sensor and operate in a Bayer filter arrangement (see Figure 2.4) where green gives twice as many detectors as red and blue (red-green-blue-green (RGBG) color filter array (CFA)) in order to give better luminance resolution than chrominance resolution.

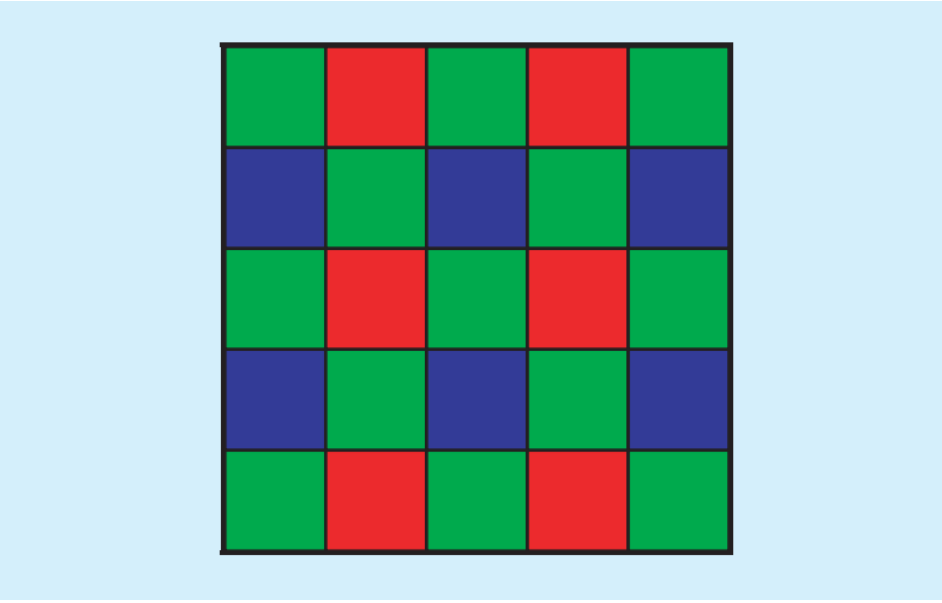

Figure 2.4: The Bayer color filter array arrangement [Schafer and Mersereau, 2005].

2.1.4.2 Depth cameras

Recognizing human actions can have many potential applications including video surveillance, human computer interfaces, sports video analysis and video retrieval. Detecting human motion taken from video sensors can be a very difficult task. However, with the recent introduction of the cost-effective depth cameras may change the picture by providing 3D depth data of the scene [Wang et al., 2012]. To get the information of depth of the various pixels in a image, depth cameras can perform that in two different ways [Henry et al., 2012]. The first one is by active stereo or stereo vision (Figure 2.5). In stereo vision, features in two (or more) images taken at the same time from separate cameras are matched with the corresponding features in the other images, and the differences are analyzed to yield depth information [Bradski and Kaehler, 2008].

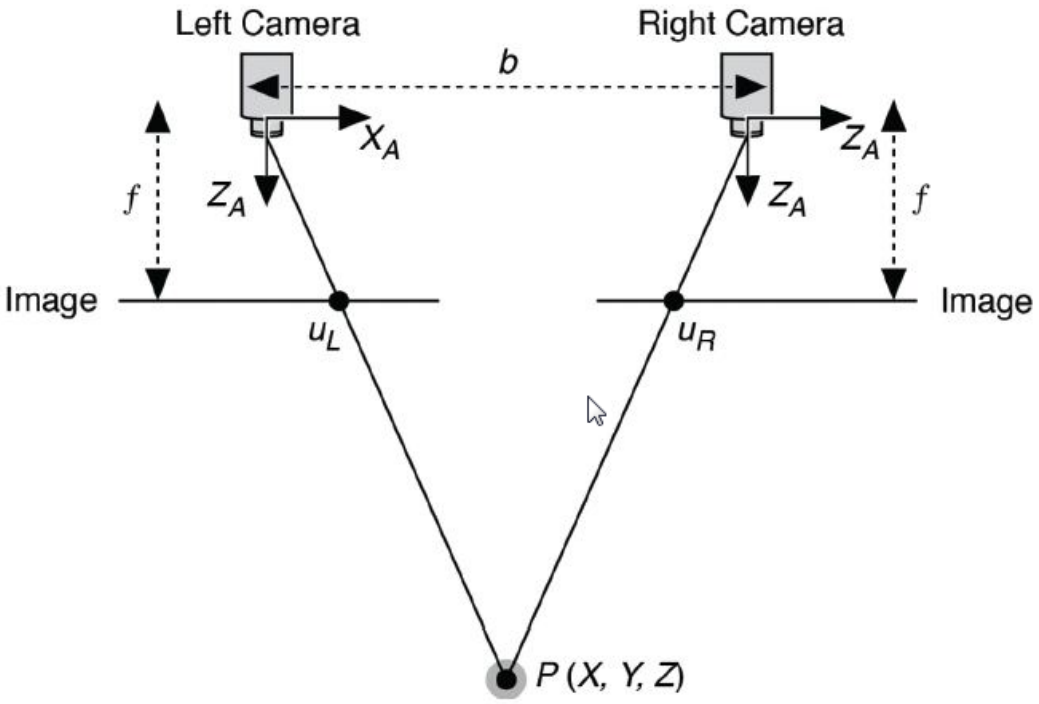

Figure 2.5: Stereo Vision System [Instruments, 2013].

The second one is by Time of Flight (TOF) technology. TOF cameras produces intensity modulated infrared light, which is not visible to humans. Then, a Photon Mixing Device sensor captures the reflected light and evaluates the distance information in every pixel to get depth information, allowing to obtain the distance to objects in the scene [Gokturk et al., 2004].

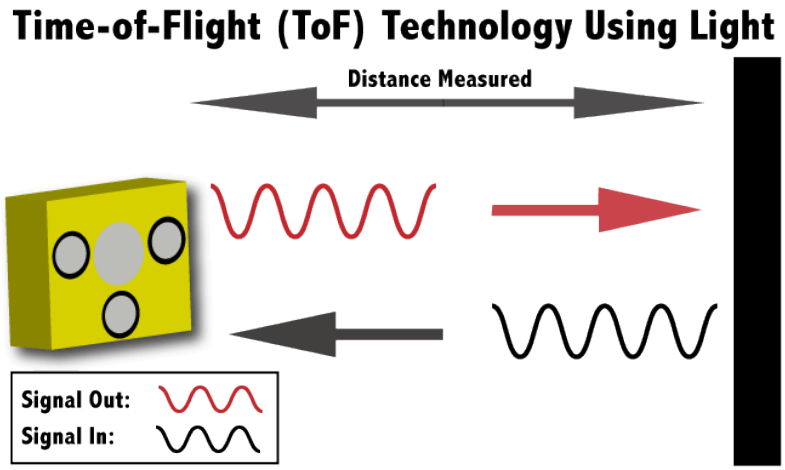

Figure 2.6: Time of Flight [TeraRanger, 2016].

In Visual Recognition Systems, the camera is one of the key issues. The resolution of the camera is extremely important since it will define the detail of each image representing the data collected. Frame rate is another important specification regarding the amount of information that the camera could record in a second. This becomes a key factor in terms of speech recognition systems considering the movement of the lips.

2.1.4.3 Kinect

Back in 2010 Microsoft and Prime Sense released the Kinect (for Xbox 360 and Xbox One) and Kinect for Windows in 2012. With its capability of tracking 48 points from the human skeleton, this camera brought a complete new approach in fields like human gesture recognition, face tracking, Audio-Visual Speech Recognition and so on. It features a RGB camera, a depth sensor (infrared) and a microphone array (see Figure 2.7).

Figure 2.7: Kinect released in 2010 [Ron, 2013]

Many systems were born using this kinect camera since its release [Yargic and Dogan, 2013], [Galatas et al., 2012], [Oikonomidis et al., 2011]. However, this camera was far from

being perfect because of its low resolution (640x480) and its depth information technique known as structured light which is problematic in contours regions.

2.1.4.4 Kinect One

In July 2014 Microsoft released the second version of the Kinect camera, called the Kinect One (see Figure 2.8), along with the public preview of the Kinect for Windows SDK 2.0.

Figure 2.8: Kinect One for Windows

This new version brought several improvements such as a better resolution (Full HD, 1920x1080 in RGB images), way better depth images thanks to the Time of Flight (TOF) technology, has greater accuracy over its predecessor, processing 2 gigabits of data per second, the new version can track up to 6 skeletons at once and a wider field of view.

This new version soon became an important piece in visual speech recognition systems because of its relation in performance over price.

2.2 Related work and State-of-the-art

2.2.1 Recent developments in SSI

On this section recent works in SSI will be presented. These works do not use just visual information taken from a camera, which means that the works depicted in this section can be either invasive or obtrusive. On the next section works will be presented based only on visual, which is a more important section for the purpose of this dissertation.

In 2014 João Freitas et. al. created a SSI system, for European Portuguese language, following sensing technologies such as Video and Depth input, Ultrasonic Doppler sensing (UDS) and Surface Electromyography. These independent streams of information are synchronously acquired with the aim of supporting research and development of a multimodal SSI [Freitas et al., 2014a]. His work is non-invasive, however it is obtrusive, as Electromyography (EMG) sensors were needed for the Surface Electromyography signals (see Figure 2.9).

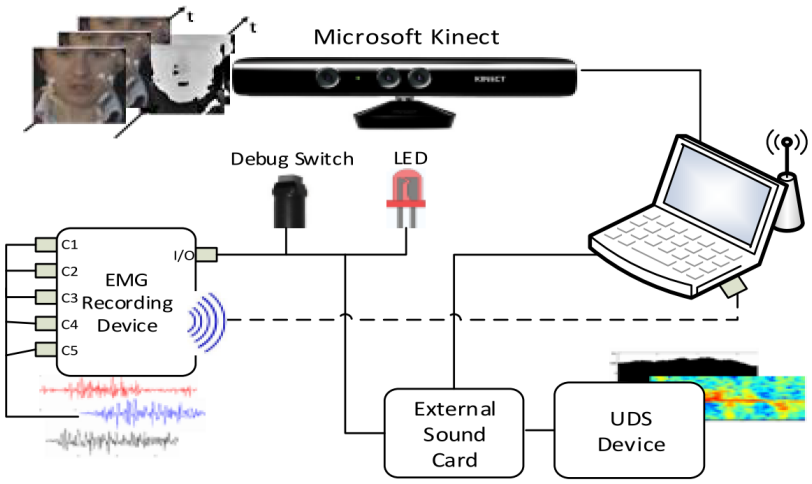

Figure 2.9: Diagram of the alignment scheme of João Freitas, António J. S. Teixeira and Miguel Sales Dias work [Freitas et al., 2014a].

A vocabulary of 32 words in European Portuguese were used regarding an AAL context, divided in sets of digits, pairs of common words and AAL words. For classification, Dynamic Time Warping (DTW) and k-Nearest Neighbor (KNN) classifiers were used. His results points towards performance advantages using a multimodal solution to implement an SSI, especially for Ultrasonic Doppler sensing and Surface Electromyography. However, a final conclusion can not be taken regarding which approach represents a higher gain. The best results had nearly 94 percent accuracy (for ALL words with features from Video+Depth+UDS+EMG with DTW classification) and the worst were nearly 65 percent (for a Vocabulary Mix using features from Video+Depth with DTW+k-NN classification).

A study from Japan (2014) regarding Silent Speech (in Japanese), with EEG signals from 63 channels (obtrusive) shown that previous SSI without adaptative collection could have better classification accuracies if they have adaptative collection, an increase from 56-72 percent to 73-92 percent [Matsumoto, 2014]. The only classification algorithm used in this study was Support Vector Machine (SVM) with Gaussian Kernel.

Still in 2014, from Georgia Institute of Technology, USA, a wearable system (obtrusive and intrusive) was created [Sahni et al., 2014] to capture tongue and jaw movements during silent speech (in English), during pronunciation of 11 distinct phrases (Figure 2.10). To achieve that, a two system part was created: one part with a Tongue Magnet Interface, which utilizes the 3-axis magnetometer aboard Google Glass to measure the movement of a small magnet glued to the user's tongue, and the second part a Outer Ear Interface which measures the deformation in the ear canal caused by jaw movements using proximity sensors embedded in a set of earmolds. The classification was done using hidden Markov model-based techniques to select one of the 11 phrases. The average user dependent recognition accuracy was 90.5 percent using both parts of the system. Using just the part of the Outer Ear Interface (non-intrusive but still obtrusive) the system performs with an accuracy of 85.45 percent.

Figure 2.10: Tongue magnetometer and Outer Ear Interface [Sahni et al., 2014].

2.2.2 Silent Speech based on visual - Visual Speech Recognition

One of the first studies in VSR was in 1994. This study was based on a word recognition system with a lip modeling approach for the recognition task [Rao and Mersereau, 1994]. This system had a 85 percent accuracy using the height and width of the lips, but only 2 words were tested.

In 2007, Werda created an Automatic Lip Feature Extraction prototype (named as ALiFE) that could automatically localize lip feature points in a speaker's face and carry out a spatial-temporal tracking of these points [Werda et al., 2007]. The points of interest in Werda work were the top center of the upper lip, bottom center of the lower lip and corners (Figure 2.11).

By using these points it was possible to extract features like the width (distance between the corners points), the height (distance between the top and bottom points) and also the area consisting of the inside of the mouth. They used multiple speakers in their tests (females and males), French was the used language in the tests and the accuracy obtained was 72.73 percent.

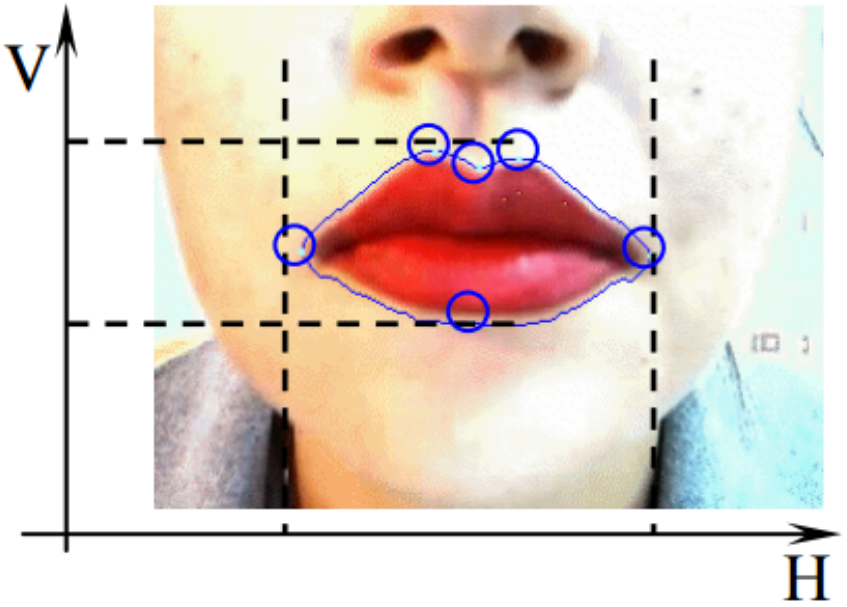

Figure 2.11: Points of interest detection by the projection of final contour on horizontal and vertical axis (H and V) [Werda et al., 2007]

In a more recent scenario and regarding the Kinect camera, using the RGB camera and depth information, without the information of sound (VSR), using 18 points of the lips (Figure 2.15) and extracting the angles between all these points, [Yargic and Dogan, 2013] created a system with a Turkish vocabulary of 15 words obtaining an accuracy rate of 78.22 percent. One more recent system based on visual lip movement recognition was proposed [Frisky et al., 2015] by applying video content analysis technique. Using spatiotemporal features descriptors, features were extracted from video containing visual lip information. A preprocessing step is employed by removing the noise and enhancing the contrast of images in every frames of video. This system had an accuracy beetwen 25.9 percent and 89.02 percent.

2.2.3 Silent Speech for Portuguese

Regarding SSI for European Portuguese (EP), in 2010 João Freitas started working, during his PhD, on a multimodal solution that addressed the issues raised by adapting existing work on SSIs to a new language. His SSI was aimed as a HCI modality to interact with computing systems and smartphones, respectively, in indoor home scenarios and mobility environments. His paper was focused on the approaches of VSR and Acoustic Doppler Sensors (ADS) for speech recognition, evaluating novel methodologies in order to cope with EP language characteristics. João Freitas, in his work, used a classification scheme based in Dynamic Time Warping (DTW) technique, achieved a Word Error Rate (WER) of 8.63 percent (STD 4.01 percent) with the best figure of 2.5 percent WER for the best run [Freitas et al., 2011].

Back in 2013, João Freitas has done another work about SSI for EP. This time, he selected 4 non-invasive modalities Visual data from Video and Depth, Surface Electromyography and Ultrasonic Doppler - and created a system that explores the synchronous combination of all 4, or of a subset of them, into a multimodal SSI. For classification, an example based

recognition approach based on Dynamic Time Warping followed by a weighted k-Nearest Neighbor classifier. His results showed that a significant difference in recognition rates can be found between unimodal and multimodal approaches in favor or the latter, and that benefits can be obtained by aligning several modalities, especially when registering Video, Depth and UDS, or Video and Depth. Results also indicate a slight better performance when using a decision fusion approach with DTW followed by a k-NN classifier [Freitas et al., 2013].

In 2014, another work from João Freitas was done, about the SSI for EP as well [Freitas et al., 2014b]. This time, Freitas et al. proposed a non-invasive method surface Electromyography (EMG) electrodes - positioned in the face and neck regions (see Figure 2.12) to explore the existence of useful information about the velum movement. The applied procedure takes advantage of Real-Time Magnetic Resonance Imaging (RT-MRI) data, collected from the same speakers, to interpret and validate EMG data. The results of his study showed that, in a real use situation, error rates as low as 23.4 percent can be achieved.

Figure 2.12: EMG electrodes [Freitas et al., 2014b].

One of the most recent works in a Silent Speech for Portuguese (and also Visual Silent Recognition) is the Master's Dissertation by Hélder Abreu from University of Minho, Portugal. On his dissertation Abreu used the last version of Kinect (Kinect One) to extract geometric and articulatory features from the lips. To obtain the lips region, Region of Interest (ROI), Abreu made some pre-processing using the collected depth images from the speaker that is talking in front of the camera. The Kinect Software Development Kit (SDK) gives information abount the head joint position and, knowing that the speaker has his face turned frontally to the Kinect camera, the coordinates of the tip of the nose can be obtained by searching the point with the closest depth value in the center of the frame. The lips are located under the nose, therefore the region of interest (lips region) are located under these coordinates (see Figure 2.13).

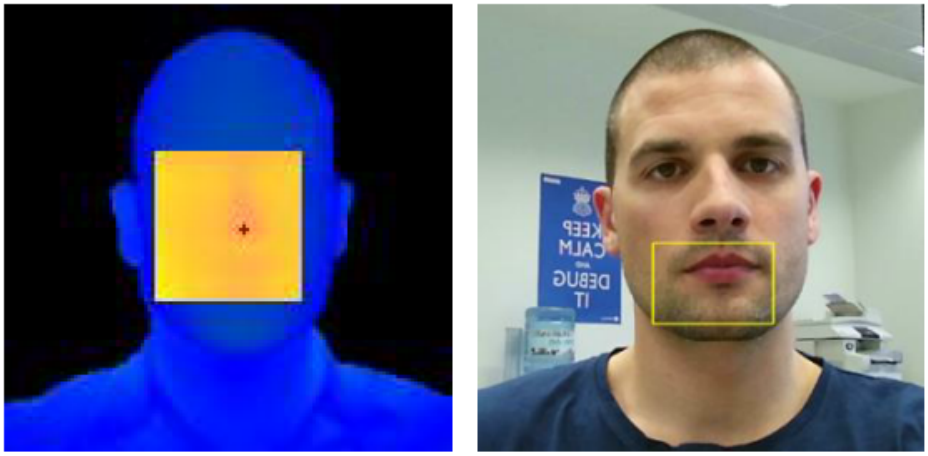

Figure 2.13: Tip of the nose detected (left) and the region of interest including the lips (right) [Abreu, 2014].

For the lip segmentation, Abreu considered two color spaces: RGB and YCbCr. From the RGB frames he used the green channel in order to extract the external points of the lips and from the YCbCr color space the Cr channel was used to obtain the internal points of the lips. These color spaces have shown a higher contrast in the ROI with emphasis on the lips. Then, a Convex Hull technique was applied in order to have a set of points representing the lips (see Figure 2.14) which were used for feature extraction purposes.

Figure 2.14: Set of points (in red) obtained with the Convex Hull technique [Abreu, 2014]

After the features taken Abreu made some normalizations like length normalizations to the feature vectors to be sent to the classifiers and some distance normalization.

The selected vocabulary consisted of 25 European Portuguese words, which were divided into 2 sets: one with a widely used set of words used in speech recognition literature, digits from zero to nine and the other taken from a Ambient Assisted Living context.

The classification process was done using SVM classifiers and the best accuracy of his system (ViKi - Visual Speech Recgnition for Kinect) was 68 percent based on geometric

features using Inverse Multiquadric kernel and 34 percent of recognition accuracy based on articulatory features using Gaussian kernel. A hybrid solution using both geometric and articulatory was also tested achieving an accuracy of 49 percent using Inverse Multiquadratic kernel.

One of the main features that is extracted in SSI based in visual information is the lips and its position/movement over time. Studies are being developed to find them as accurate and quick as possible [Dalka et al., 2014].

About the study from [Yargic and Dogan, 2013] regarding the lips using the kinect camera, the features were extracted using 18 point of the lips as depicted in (Figure 2.15).

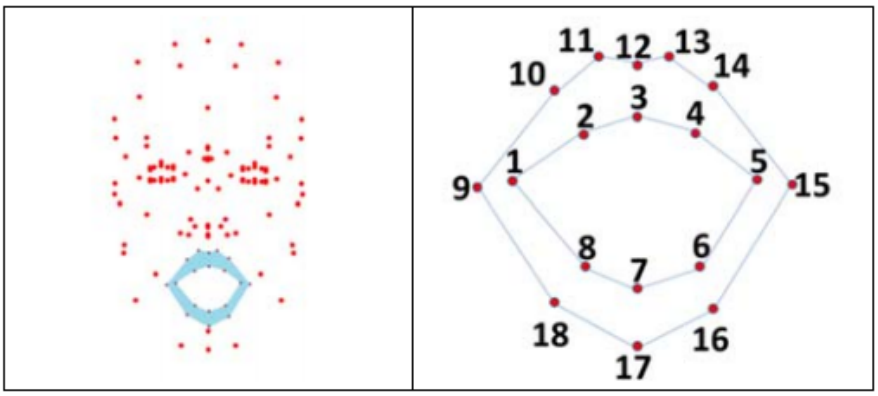

Figure 2.15: 18 lip feature points and their assigned ID values [Yargic and Dogan, 2013].

Next, a set of isolated words about Turkish color names are constructed and they are classified by the KNN classifier. The features are selected from the angles of some points on the lip and reasonable results are obtained by using angle features.

2.2.4 Classifiers

To classify the data from the features that are extracted, some machine learning is mandatory. To achieve that, learning algorithms that analyze the feature data are needed.

The most common used algorithm in Visual Speech Recognition Systems is SVM. SVM is a regression analysis model, which given a set of training data decides each class the examples belong to, separated by an hyperplane (i.e. decision boundary) linearly separating our classes (see Figure 2.16).

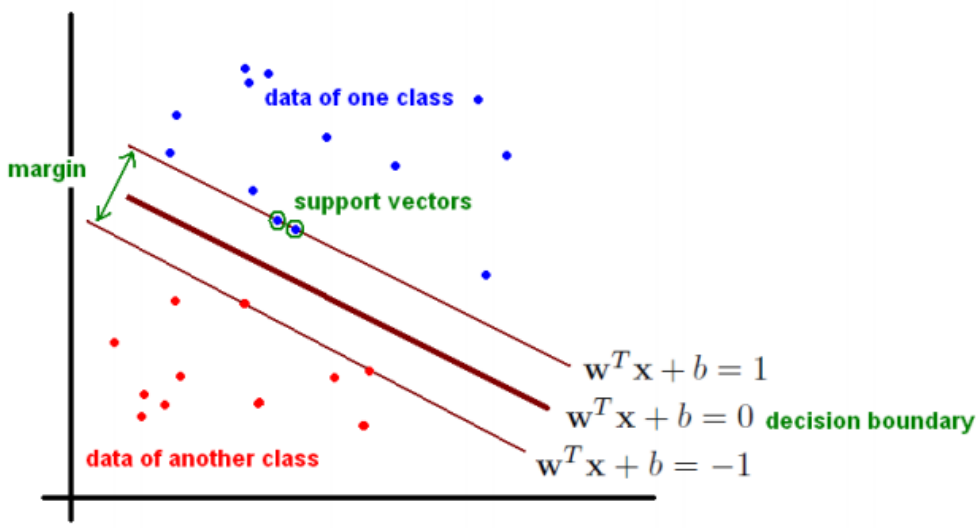

Figure 2.16: SVM model example with decision boundary separating the two different classes [Gavrilov, 2012].

.

This decision is done by finding the hyperplane (decision boundary) linearly separating the two classes. It's intuitive that the decision boundary must have a maximized distance between both classes in order to not have classification mistakes [Gavrilov, 2012].

Anything above the decision boundary should have label 1 and, similarly, anything below the decision boundary should have label -1.

Another common used learning algorithm is the Random Forest tree based algorithm. Tree based methods, unlike linear models, map non-linear relationships quite well and are adaptable at solving any kind of problem at hand (classification or regression). In Random Forest learning algorithm the feature vectors are given as inputs in each trees created for classification task and classification is done for the tree that has more "votes" (see Figure 2.17).

A decision tree splits the nodes with respect to the target variable and in pursuit to a most homogeneous sub-nodes [Saraswat, 2016].

Other types of classifiers can also be used in the classification task of visual speech recognition systems. Classifiers like AdaBoost (Adaptative Boosting, used in conjunction with other learning algorithms to improve performance), Naive Bayes (is a probabilistic classifier that uses Bayes theorem with indepencence assumptions between features) and Sequential Minimal Optimization (SMO, an algorithm made to deal with the quadratic programming problem that arises during training SVM's).

Figure 2.17: Decision tree example [Saraswat, 2016].

2.3 Summary

Speech perception is a task that is easy for human beings, with one or two fails in 1000 words. However, the speech perception using machines represents a huge challenge.

Speech production by humans is divided in 3 different steps: Conceptual preparation, Formulation and Articulation. Many SSI systems were made regarding one of these 3 phases or, in some cases, more than one phase in the same system.

The first studies in speech recognition were done in the beginning of the 50's but rely only on audio information. Visual speech systems are needed in cases like noisy environments or people with speech issues, so studies started to be done regarding this task.

Regarding the field of features extraction, SSI systems can be invasive, obtrusive or none of them. A system is invasive if it requires medical expertise or permanent attachment of sensors and obtrusive if it requires wearing or using sensors. A system can be non-invasive and non-obtrusive, like visual speech recognition systems using Kinect (RGB-D camera). All systems have advantages and disadvantages such as the price, accuracy or practicality.

Many types of classification algorithms are available for the classification process, like Dinamic Time Warping, k-Nearest Neighbor, hidden Markov, SVM, etc. The most commonly used of all is the SVM classifier.

Chapter 3

SSI prototype

In this chapter the created prototype is presented. The first section presents the requirements needed for the prototype and explains the AAL scenario chosen. The system overview is presented and next a detailed description of all steps presented in it. It is described how the activity detection was done, as well as the features extraction, classification and data bases.

3.1 Requirements

As requirements the system has to permit some real daily life experiences, for example the controlling of a television from the sofa at a certain distance. However, with a turned on television a noisy ambient is created and the system has to work without using audio information. A SSI with visual speech recognition must be implemented.

Another requirement is that the prototype must detect the speaker's face automatically and start and stop recording features also automatically (push to talk implementations cannot be used). So, a more user friendly solution is made for people with some kind of motor limitations for deaf people.

3.2 AAL Scenario

The most important areas in AAL are related in allowing the user with some kind of limitations to control entertainment systems, access to social networks, etc.

The AAL scenario chosen was the controlling of a media player, increasingly important to access memories and information from friends and family.

Regarding these interests, the VLC player was chosen, because it's the world's most used open multimedia player [Lanaria, 2016], it is simply to use, can load up content that other players simply can't, it's open source and can aim directly at AAL scenario, which is the objective of this dissertation.

In VLC is possible to load a set of videos in the interest of the user. The speaker can control the sound, the video that he wants to watch, the speed of the video and stop and play functions, all of these controls are used with Silent Speech detection by the movements of the lips and the chin of the speaker. This allows the user to control the system either in a noisy ambient, in case the user wants some privacy or if there is some speech production limitation from the speaker.

The words that were considered in this work are in European Portuguese and were considered to be the most intuitive and short as possible. Several iterations were done to reach the set of words that better affords the interests of this dissertation. These words were chosen regarding the AAL scenario of controlling the VLC player. The Table 3.1 has, in the first column, the set of words chosen in Portuguese and in the second column their translation in English. The commands with two words were also chosen to be the most different as possible one from the other.

Portuguese	**English**
Ver Filme	Watch a Movie
Parar	Stop
Continuar	Continue
Aumentar Volume	Increase Volume
Baixar Volume	Decrease Volume
Mais Rápido	Faster
Mais Lento	Slower
Próximo Filme	Next Movie

Table 3.1: Set of words chosen regarding the AAL context of using the VLC

3.3 System Overview and Global architecture

The global architecture of the work is presented in Figure 3.1. The system created in this dissertation follows the architecture of traditional VSR systems [Abreu, 2014];[Saenko et al., 2004];[Galatas et al., 2012] and takes the advantages of the Kinect One Camera to extract the features from the lips and chin of the speaker.

The system is divided into 4 main blocks (Activity Detection, Feature Extraction, Classification and Data Base Creator). Two paths can be previously chosen before the start of the usage of the system: Train or Test. All these paths and blocks will be explained in detail in the next section.

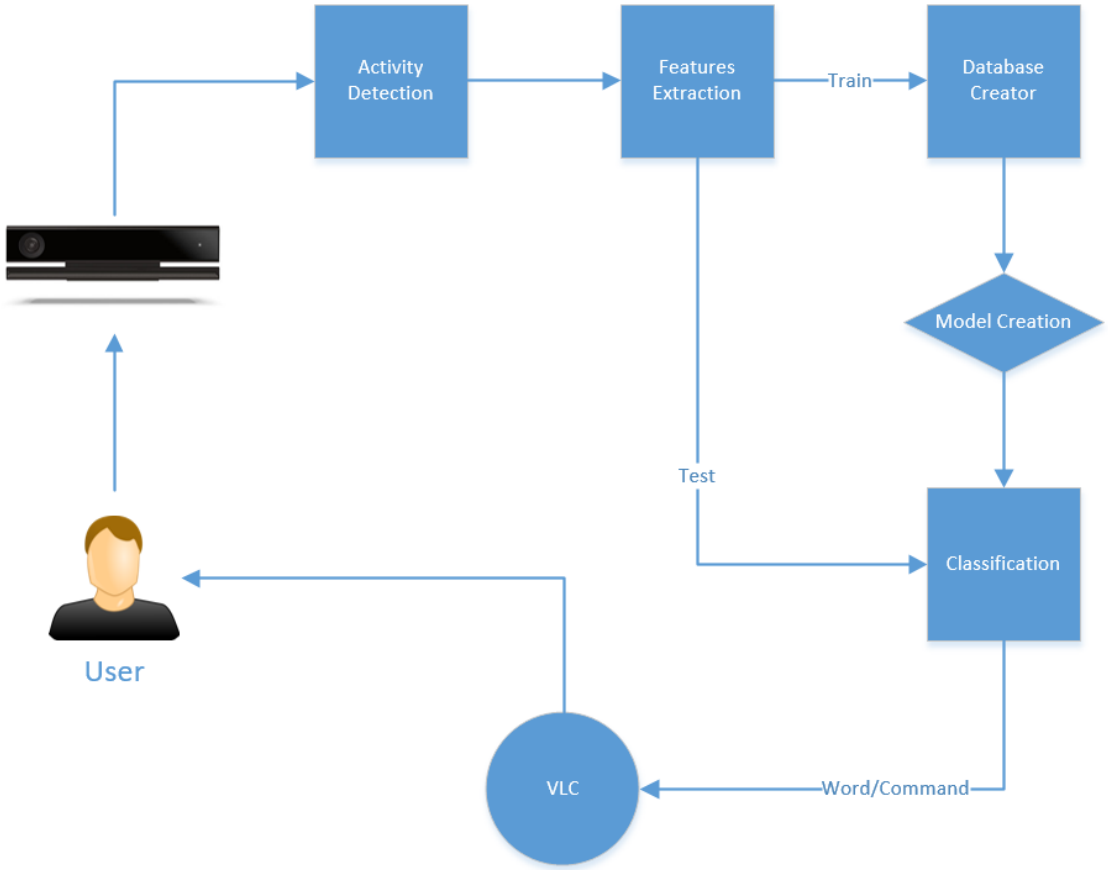

Figure 3.1: System Architecture

3.4 Activity detection

The first step is the Activity Detection. In this step the system searches for the face of the speaker, from every frame that arrives from the Kinect Camera, and when it finds it (with Microsoft Kinect SDK), a rectangular box is drawn surrounding the face of the speaker. Points are drawn in the speaker's eyes, nose, lips and chin as well. Other kind of information are also available like if the speaker is happy, wearing glasses, if the right or left eye is closed etc. All these informations are presented in Figure 3.2.

As we are working in a prototype that can be used by people with some kind of movement limitations, an automatic recording of the movement of the lips and chin in the beginning of the word is mandatory. So, the speaker can give the commands to VLC only with the lips and chin detection, the system does not need for example a click of the mouse to inform that a word will be pronounced.

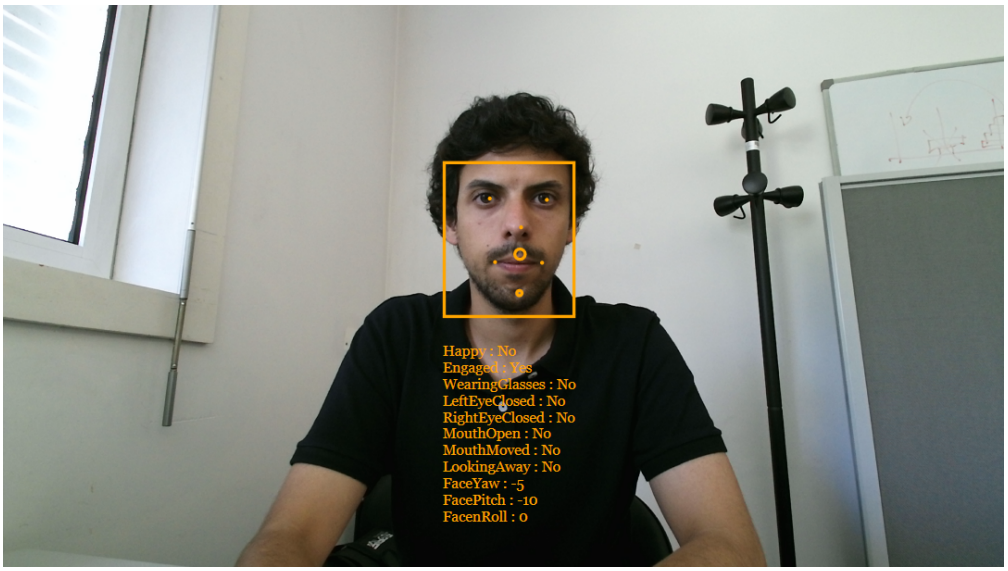

Figure 3.2: Face detection by Kinect and other speaker information shown.

To achieve this, first the speaker has to have the face and lips stable for around one second. Then, the system informs the speaker that it is ready to record a word by displaying in the window program a text message showing this and the window background turns green to be easier for the speaker to see this state change. In the state of Ready to Record, the system starts recording as soon as the speaker opens his mouth (information obtained with Kinect for Windows SDK 2.0, using "MouthOpen" and "MouthMoved" properties [Microsoft, 2016b]). This is used as a sign that the speaker is starting to say a word.

With the same approach, the system stops recording when the speaker has the face and lips stable for around one second.

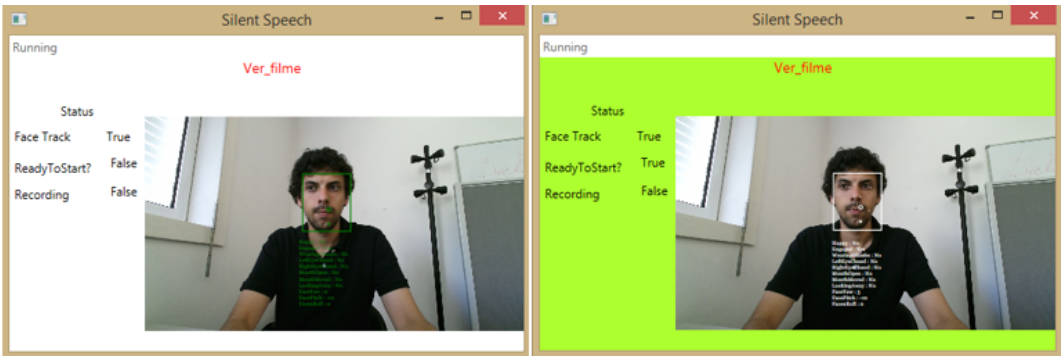

Figure 3.3: SSI system not ready (left) and ready (right) to record the command "Ver Filme".

3.5 Feature extraction

The features to be extracted to send to the classifier will be the position of the lips (width and height), the protrusion of the lips (upper lip and bottom lip) and the chin position (X and Y coordinates), see Figure 3.4. The position of the lips was chosen because it has proven

to give good results in previous works. The Chin position was chosen because the lower jaw takes part of the speech production process by humans, so the chin position could be interesting to have as a feature as well.

A total of 6 features were taken in consideration in this work for the recognition task of the words. To obtain these points, Microsoft Kinect packages were used, precisely the HighDetailFacePoints in Kinect20.face.lib [Microsoft, 2016a].

Lip width is the distance between the left and rigth corners of the lips. Lip height is the distance between the top and bottom corners of the lips. Lip Protrusion represents how close or how far the top and bottom lips are to the Kinect sensor.

To track the Chin position in each frame, the HighDetailFacePoints.ChinCenter in Kinect20.face.lib was used to store the X and Y position of the chin in each frame.

Figure 3.4: Points tracked in mouth and chin for feature extraction proposes.

In the previous section was presented the method of automatic recording when the speaker can start to say a word. In this moment, the features regarding the lips and chin are starting to be stored in different queues, each of them representing one different feature of the word itself. These queues are filled up to the time that the recording stops. However, the numbers representing the features must be normalized to deal with the different distances that the speaker could be to the Kinect and some normalization to better focus the major changes of the movements of the lips and chin are also considered. To solve this distance issue, a distance normalization is applied to a fixed distance (see Equation 3.1).

$$Nd = \frac{Cd * Fd}{dFN} \tag{3.1}$$

where Nd is the normalized distance, Cd is the current distance to the Kinect (before normalization), Fd is the depth value of the forehead which every frame is normalized (also obtained with HighDetailFacePoints from Microsoft Kinect library) and dFN is the depth value for which every frame is normalized.

To achieve a better focus on the major changes of the movements, a z-score normalization is implemented (see Equation 3.2).

$$z = \frac{X - \mu}{\sigma} \tag{3.2}$$

where z is the result of the z-score normalization, X is the current feature value, μ is the mean obtained in one database and σ represents the variance.

So at this point we have a set of queues with features already normalized but with different sizes (because of the time duration of the different commands). Due to a fixed size needed

in the Classification part which is going to be done later in Weka (will be explained in more detail in next section) a length normalization of the features matrix is needed.

The fixed length was chosen regarding the 30fps recording rate of the Kinect One and since 2 seconds was enough to record the total of the words pronounced, the fixed length decided for each array was 60.

The normalization method consists in reading each queue of features (6 queues, each one representing a different feature) and removing the first 10 numbers and, after this, the removal of the numbers beyond the position number 60 (see Figure 3.5).

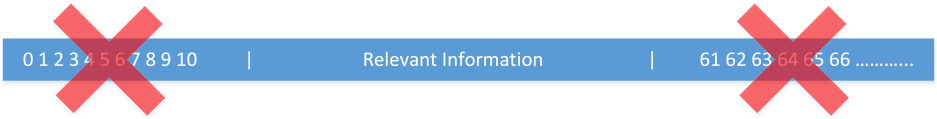

Figure 3.5: Process of length normalization of each array of features regarding the fixed size needed in the classification part.

Now, after a word being said and features captured and recorded, we have 6 arrays already normalized and with a fixed length. Next, two different paths can be chosen: Train or Test. The Train path consists in creating a database regarding the speaker that will use the SSI system. The test path is the path where the system is used to control the VLC. The Test path cannot be used without creating previously a database regarding the speaker in the Train path first, otherwise the system will not have data to compare in order to decide which word the speaker is saying. A database from a different speaker can be used, however the results are expected to be much worst (this part will be shown in more detail later in this dissertation).

3.6 Classifiers / Classification

In the previous section, feature vectors (or feature arrays) with fixed size and needed normalizations are created. Now, it is time to analyze each one of them in order to give a correct hint and return a correct command to VLC regarding the command that the speaker just said, in a Live usage scenario application.

To classify the words, Weka software [Witten et al., 2011] was used for the training, validation and classification processes of this dissertation. The algorithms that are used in this work are: Random Forest, SMO, Naive Bayes, AdaBoost and SVM. The Kernel used for the SVM classifier was the linear one and the classification algorithm used with the AdaBoost was the SMO algorithm.

Classification experiments were made to the recorded databases in the Train part was done. All five classification algorithms were used and a Cross-Validation technique was used with a different number of folds (5, 10, 15, 20, 25 and 30 folds).

The classification in the Test part is the most important one this dissertation, because it's in the Test part that VLC is controlled by the commands recorded regarding the created Silent Speech system. Since this is a real time implementation, the usage of the five classification algorithms turned out to not be the best solution, due to the amount of time taken by the AdaBoost and SVM classifiers. So, the other three classification algorithms were taken to work

in a "Winner Takes All" perspective (the most common hint between the three classifiers is the hint that is selected).

3.7 Databases

Beyond the prototype created, several databases for training and evaluation were also made.

The databases created in this dissertation are created regarding the "Arff" file format, making it possible to use them directly in Weka [of Waikato, 2008]. Every row recorded represents all features arrays, each of them already normalized and with size equals to 60, as depicted in Feature Extracted section. Two types of data bases are created in this dissertation: data bases for Train part and data bases for Test part. Both of them have the same structure, composed of the same Sources, Relation, Attributes and Classes. The difference between the two data bases resides in their length, due to the different repetitions of every word in the different usage of the two data bases.

3.7.1 Databases - Train part

In the Train part, the database is recorded in sets of 10 repetitions of every command, which gives a total of 80 commands spoken per set and a time taken for every set of around 5 minutes to record.

It was decided to record databases of 30 repetitions of every command in the Train part. So, the data bases in the Train part have a total of 86400 numbers representing every feature in every frame of the recordings (30 repetitions * 8 commands * 60 frames * 6 different features).

3.7.2 Databases - Test part

In the Test part, the databases created are a set of 6 arff files, each of them represents one feature and has only one row of information, because in the test part we want a system that works in real time for the recognition of one command at the time the speaker announces the word. So, when the speaker ends the announcement of one command, every queue of features is normalized and forced to a fixed size (60). Then, all the arrays are stored in the same row in the database. This gives a total of just 360 numbers representing every feature in every frame recorded.

3.8 Summary

This chapter introduced the developed SSI prototype for VLC controlling created in this dissertation: the global architecture, how the activity detection works, how the features are extracted and normalized, which classifiers were used and how and how the data bases were created.

When the system finds a face in front of the Kinect One camera, an output is given to the speaker. After that, the speaker needs to have his head stable and his mouth closed for around one second so the system starts recording a word. When the system is recording, the system window turns green for better announcement to the speaker of this state change. To

stop recording, the speaker just needs to have his head stable and lips closed again for around one second again.

Six different features are extracted during the recording (4 regarding the lips and 2 regarding the chin). All these features recorded are z-score normalized and size fixed after the recording.

Two different paths can be used in the system: Train or Test. In the Train path is possible to train the system in order to have a better performance later in the Test path (live usability) to control the VLC player. In both paths data bases are created, however with different purposes and lengths.

Chapter 4

Results

This chapter presents several results from the evaluations that were made. Databases with three different persons were evaluated and, to test these databases and to classify the words, Weka software and libraries were used [Witten et al., 2011].

4.1 Databases for evaluation

Before going for the Test part, which is using the system to control the VLC using Silent Speech, data bases must be recorded in order to train the classifiers.

Three different persons participated in the evaluation of the system: Speaker1 (author of the dissertation, finalist of the Integrated Masters Degree in Electronics and Telecommunication Engineering, 23 year old male), Speaker2 (22 year old male, student of the Integrated Masters Degree in Electronics and Telecommunication Engineering, University of Aveiro) and Speaker3 (29 year old female (Master in Gerontology and PhD in science and health technologies, University of Aveiro), from the island of Madeira, Portugal, where the pronunciation is different when compared to other regions of the country.

In this dissertation, 5 different databases were created: 3 databases regarding the Speaker1 (one recorded at 0.6 meters, 1 meter and 2 meters away from the Kinect Camera), 1 database regarding the Speaker2 and 1 database regarding the Speaker3, both recorded at 1 meter away from the Kinect Camera.

Every database took about half an hour to record. For the other two speakers besides the author of this dissertation, the time taken was way bigger (twice), perhaps they were not comfortable to be sat talking in front of the camera. With Speaker2 and Speaker3 the first databases recorded by each were not well recorded and the tests in Weka showed an accuracy that was too low when compared to the Speaker1. The second database recorded on both cases proved to be way better than the first one and closer to the results from Speaker1.

The databases were recorded in a IEETA lab (Institute of Electronics and Informatics Engineering of Aveiro), in conditions of low noise, during July of 2016. The Speaker1, author of the dissertation, recorded all the databases in Silent Speech and Speakers 2 and 3 recorded the databases pronouncing the words.

4.2 Evaluation - Effect of classifiers

In this section, all databases are evaluated in Weka with a set of five different classifiers that were chosen to evaluate all these databases. After a experiment with all the classifiers available in Weka software, the five that proved to be the best in the classification task were chosen: Random Forest, SMO, Naive Bayes, AdaBoost and SVM. The evaluation of the databases was done for each speaker and the results presented in this section are showed separately.

4.2.1 Evaluation with Speaker1

Speaker1 was the only one between the 3 Speakers that recorded at different distances from the Kinect camera. The results of the databases evaluation are shown in Table 4.1.

Fold	Distance from Kinect (m)	Random Forest (%)	SMO (%)	Naive Bayes (%)	AdaBoost (%)	SVM (%)
5	0.6	79.6	76.7	72.9	76.3	79.6
	1	70.4	66.7	65.8	67.1	69.2
	2	77.1	78.8	67.5	78.3	78.3
10	0.6	78.8	78.8	75.8	77.1	79.1
	1	72.5	65.8	64.6	67.1	68.8
	2	80	79.6	72.5	78.8	79.2
15	0.6	78.3	80	74.2	77.9	79.6
	1	72.5	65.8	65	66.7	67.1
	2	79.6	79.1	72.9	78.8	79.2
20	0.6	79.6	78.3	74.2	77.1	78.6
	1	72.5	66.3	64.2	65.8	69.2
	2	**82.5**	78.8	73.8	79.6	78.3
25	0.6	80	80	75.8	78.3	78.3
	1	71.3	63.8	62.1	67.1	68.3
	2	79.6	80	73.8	79.2	**80**
30	0.6	77.5	80.4	**76.3**	77.9	77.9
	1	70.4	65	62.9	64.6	66.3
	2	80.8	**80.8**	73.3	**80**	**80**

Table 4.1: Results of cross-validation using five different classifiers regarding the Speaker1 at 0.6, 1 and 2 meters away from the Kinect.

By analyzing Table 4.1 we can conclude that the database recorded at 1 meter did not have results as good as the ones recorded at 0.6 and 2 meters away from the Kinect. This data base may have not be well recorded, with the words recorded less articulated compared with the other two data bases. The best result achieved by the system with the Speaker1 was 82.5 percent for the Random Forest classification and cross-validation of 20 folds at 2 meters from the Kinect. The average accuracy of the system (for all distances and for all classifiers) was 74.3 percent with a standard deviation of 5.7. The best classification algorithm, in this case, proved to be the Random Forest (76.8 percent average accuracy, for all Random Forest

values) and the second was the most common SVM (75.4 percent average accuracy, for all SVM values). The SMO, Naive Bayes and AdaBoost reached an average accuracy of 74.7, 70.4 and 74.3 percent, respectively.

To give a better understanding of the system accuracy word by word, a confusion matrix is presentend in Figure 4.1. This confusion matrix was obtained the best result presentend in Table 4.1 (Random Forest Algorithm at 2 meters with 20 folds).

```
 a  b  c  d  e  f  g  h   <-- classified as
26  0  1  0  0  0  0  3 |  a = Ver_filme
 0 28  0  1  0  1  0  0 |  b = Parar
 1  0 28  1  0  0  0  0 |  c = Continuar
 0  1  1 28  0  0  0  0 |  d = Aumentar_Volume
 1  1  0  0 20  2  4  2 |  e = Baixar_Volume
 0  3  0  0  3 19  5  0 |  f = Mais_Rapido
 1  0  0  0  5  3 21  0 |  g = Mais_Lento
 1  0  1  0  0  0  0 28 |  h = Proximo_Filme
```

Figure 4.1: Confusion Matrix obtained in Weka software with Speaker1 database recorded at 2 meters, classified with Random Forest algorithm and fold of 20.

The words/commands "Parar", "Continuar", "Aumentar Volume" and "Próximo Filme" showed to be the ones with less errors (just 2 mistaken hints by the classifier in 30 repetitions). The word/command that had the worst result was "Mais Rápido", with only 19 correct hints in a total of 30 repetitions and in 5 hints the classifier classified this command as "Mais Lento".

4.2.2 Evaluation with Speaker2

Regarding the Speaker2 data base evaluation, the same process of classification was done with the same algorithms. However, was depicted in the beginning of this chapter, Speaker2 only recorded one database at 1 meter distance. The results are presented in Table 4.2.

Fold	Random Forest (%)	SMO (%)	Naive Bayes (%)	AdaBoost (%)	SVM (%)
5	56.3	55	42.5	54.6	55.4
10	**63.3**	57.1	46.3	55.8	53.3
15	61.7	59.6	44.2	56.3	61.3
20	57.5	57.5	46.3	**59.6**	58.3
25	58.8	**61.3**	47.5	58.8	**62.1**
30	60.8	58.8	**48.3**	58.8	60

Table 4.2: Results of cross-validation using five different classifiers regarding the Speaker2 at 1 meter away from the Kinect.

Once again, the best result achieved by the system was obtained by the Random Forest classification algorithm, 63.3 percent (with cross-validation of 10 folds). The average accuracy of the system was 55.9 percent with a standard deviation of 5.7. The best classification

algorithm was the Random Forest once again, with 59.7 percent accuracy and the second was the most common SVM (58.4 percent accuracy). The SMO, Naive Bayes and AdaBoost reached an accuracy of 58.2, 45.8 and 57.3 percent, respectively.

These results from the Speaker2 compared to the previous results obtained with the Speaker1 are considerably worst. There are at least three possible reasons that can explain this lack of accuracy: First is that the Speaker1 is the author of this dissertation and is the person that better knows how the system works. Second the Speaker2 tended to Hyper-Articulate too much and in the cases that the commands recorded had two words (cases like "Ver Filme", "Aumentar Volume", "Baixar Volume", "Mais Rápido", "Mais Lento", "Próximo Filme"), Speaker2 made a unnatural pause among the two words. In those pauses between two words the system was recording information that did not help in the task of classifying the different words. Third the Speaker2 tended to be too slow and took too much time saying the words and as the system only records each command for 2 seconds, Speaker2 exceeded the time limit several times.

4.2.3 Evaluation with Speaker3

Regarding the Speaker3 data base evaluation, the same process of classification was done with the same algorithms. Again, as depicted in the beginning of this chapter, Speaker3 (as well as Speaker2) only recorded one database at 1 meter distance. The results are presented in Table 4.3.

Fold	Random Forest (%)	SMO (%)	Naive Bayes (%)	AdaBoost (%)	SVM (%)
5	46.3	43.8	31.6	42.5	**43.8**
10	52.5	43.8	39.2	**45.8**	**43.8**
15	49.2	**45.4**	37.1	42.1	**43.8**
20	**53.8**	44.2	**40.8**	45	**43.8**
25	50	44.6	39.2	43.8	42.9
30	53.3	43.3	40	43.8	42.1

Table 4.3: Results of cross-validation using five different classifiers regarding the Speaker3 at 1 meter away from the Kinect.

Once again, like in both Speaker1 and Speaker2 evaluations, the best result achieved by the system was obtained regarding the Random Forest classification algorithm and it was 53.8 percent (with cross-validation of 20 folds). The average accuracy of the system was 44 percent with a standard deviation of 4.6. The best classification algorithm was the Random Forest one more time (50.9 percent accuracy) and the second was the SMO (44.2 percent accuracy). The SVM, Naive Bayes and AdaBoost reached an accuracy of 43.4, 37.9 and 43.8 percent, respectively.

These results from the Speaker3 compared to the previous results depicted for the Speaker1 and Speaker2 are worst. In this case, the worst results can be explained by the native region of the speaker, which is a Portuguese Island, Madeira, and her pronunciation is a lot different compared to the Speaker1 and Speaker2 (Portugal Continental). Her pronunciation tended to be faster and a lot less articulated than the other two speakers. Other factor can be some nervousness to be talking in front of a camera.

4.3 Live evaluation

In this section, the results of live evaluation are presented. These evaluations consist in classifying a word in real time for VLC controlling proposes.

In a Live system performance, the speed of the evaluation algorithm is mandatory. The best case scenario would be to have all the 5 algorithms tested in the previous section working together at the same time to give the best hint in a work recognition task, in a perspective of "Winner Takes All". However, this scenario proved to be quite slow for a real time purpose, sometimes taking more than 5 seconds to select a word. So, instead of heading for 5 classifiers, only 3 were chosen. The first classifier chosen (and the obvious choice) was the Random Forest Algorithm, regarding its results (it outperforms the other four) and it is quite fast to decide a word as well. Regarding the other 2 classifiers, the choice turned out to be the SMO and the Naive Bayes Algorithms. Despite the good results on the SVM classifier (that proved to be in much cases the second best classifier), the SVM turned out to be too slow for the real time scenario that we want, as well as the AdaBoost classifier.

In conclusion, the Live evaluation of the words, regarding the control of the VLC player, are in charge of tree different types of classifiers: Random Forest, SMO and Naive Bayes. The classification decision of the word is made adopting a "Winner Takes All" strategy. The mean time of the all tree classifiers regarding the word recognition task was 1.3 seconds. The tests for the Live performance were made by the same three speakers already presented. In all of these tests, 10 repetitions were made, each repetition has the words already presented in the Table 3.1, which gives a total of 80 words spoken for each test that was made.

The first tests to be presented correspond to a match of a Speaker talking with the system evaluating the corresponding Speaker's database already recorded. Additionally, some Speaker Dependency tests were made to figure out if the system can perform well in an evaluation of the data bases of the other speakers. In addiction, distance tests were made to verify the Equation 3.1's performance.

4.3.1 Live evaluation - Train and test with same speaker

In this subsection, Live performances were made and the evaluation task considered the appropriate data base for the Speaker that was talking. It is important to say that regarding the special case of Speaker1, for the 3 different distances tested the training models were recorded at the same distance the Live evaluation took place. In the Table 4.4 the results achieved by the system can be seen.

Speaker	Distance (m)	Hit	Miss	Hit (%)
Speaker1	0.6	52	28	65
	1	44	36	55
	2	56	24	70
Speaker2	1	46	34	57.5
Speaker3	1	25	55	31.3

Table 4.4: Live performance of the system using the corresponding data bases of the speaker talking.

With this live performance we can conclude that the results turned out to be similar to the ones obtained with the evaluation of the data bases with Weka. The best result was

achieved with the Speaker1 at a distance of 2 meters away from the Kinect, with 70 percent hits. The Speaker3 had the worst results, which was expected as well regarding the results from the data base recordings.

4.3.2 Live evaluation - Effect of distance of the speaker

To test the distance dependency, 4 live recordings were made, all by Speaker1 but tested at different distances from the Databases that were recorded in the Train part, contrary to what was done in the previous subsection (4.3.1): the speaker at 0.6 meters away from the Kinect and the classification with the data base at 1 meter and 2 meters and the speaker at 1 meter away from the Kinect and the classification with the data base at 0.6 meters and 2 meters. The results can be seen in the Table 4.5 down below.

Speaker Distance (m)	Classifiers trained with (m)	Hit	Miss	Hit (%)
0.6	1	42	38	52.5
	2	44	36	55
1	0.6	49	31	61.3
	2	65	15	**81.3**

Table 4.5: Live recordings comparative with data bases recorded at different distances for the same speaker (Speaker1)

The results proved that the distance is not an issue (the hits are similar to the ones obtained in Table 4.4) and proved that the Equation 3.1 is useful regarding the different distances that the Speaker can be from the Kinect camera in a AAl scenario. In Table 4.5 the best live performance of this work was obtained (81.3 percent) with the Speaker at 1 meter away from the Kinect and the data base recorded at 2 meters.

4.3.3 Live evaluation - Speaker Dependency

To finish the evaluation, the Speaker Dependency of the system was tested. The objective was to understand if the system is capable of being used by a Speaker that has no training data. In other words, if the system can perform with Speaker X in the Test part against the Speaker Y data from Speaker Y Train part.

To test the Speaker Dependency of the system, 3 tests are made: Speaker1 at 1 meter away from the Kinect performing against the databases of the Speaker2 and Speaker3 and another test with the Speaker2 at 1 meter away from the Kinect performing against the data base of the Speaker1 also at 1 meter. The results can be seen at Table 4.6 down below.

Speaker	Data Base to be Compared	Hit	Miss	Hit (%)
Speaker1	Speaker2	14	66	17.5
	Speaker3	12	68	15
Speaker2	Speaker1	17	63	21.3

Table 4.6: Live recordings results regarding the Speaker Dependency tests.

The results have shown that the system accuracy decreases dramatically in comparison to the results obtained if the data base corresponded to the Speaker that is talking. Analyzing the results presented in Table 4.6 we can conclude that the system is clearly speaker dependent.

4.4 Summary

In this chapter the accuracy of the system obtained from 3 different Speakers with different ages, genres and pronunciations was tested. First the data bases were recorded and their evaluation was made with the classification being done by 5 different algorithms (Random Forest, SMO, Naive Bayes, AdaBoost and SVM). The best global performance was achieved by the Random Forest which gave also the best result of all tests (82.5 percent hits in cross-validation with 20 folds).

The results for the 3 speakers turned out to be quite different, especially with Speaker3, a 29 year old female from the Portuguese Island Madeira. With her low articulation of the words and fast pronunciation, the accuracy of the system dropped to 50.9 percent.

The system proved not to have much dependence on distance, with similar results at any distance of recording. In the field of Speaker Dependency the results are worst. The best result obtained in a Speaker Dependency test was 21.3 percent and the worst was 15 percent, which can be conclusive about the non independence of the system to classify a word with a different data base that the speaker that is talking.

Considering the AAL scenario of controlling the VLC, the system have shown to perform reasonably well even in real time. The best classification result achieved was 81.3 percent to a distance of 1 meter compared to a data base recorded at 2 meters away from the kinect, with a classification time of around 1.3 seconds.

Chapter 5

Conclusions

To conclude this dissertation in this chapter is presented the summary of work, main results and ideas/suggestions for future work.

5.1 Summary of work

In this dissertation the first working SSI prototype for Portuguese was created, regarding a AAL scenario of controlling the VLC player, with interesting results for at least a speaker. The requirements for the prototype were defined based only on visual information.

Different parts of the prototype were developed: activity detection (automatic recording started by movements of the lips), features extraction, train of classifiers and integration with VLC player. The Microsoft Kinect for Windows was used to read the lips of the speaker.

Three different adults, with different ages, genres and pronunciations, tested the system. Databases for each of them were recorded and were evaluated with different distances from the Kinect and also the Speaker Dependency of the system.

5.2 Main results

Using the Kinect One, the system created obtained an accuracy of 82.5 percent in a pre-recorded database and 81.3 percent during the live recordings. The first step to use the system stands to create a data base with 30 repetitions of every 8 words regarding the AAL scenario of controlling the VLC player. After the data base is created, some tests can be done with classifiers using the Weka classifier, to see if the results were good. The results can vary a lot from Speaker to Speaker. Some people pronounce the words too slow with hyper articulation of the lips, others talk too fast and almost do not move the lips at all. The system performs better if the words are correctly articulated during all the repetitions and if the words are correctly recorded during the 2 seconds available to extract the features from the lips and chin.

The system revealed to perform well at the job of real time VLC controlling with recognition of the commands in SS using the Kinect Camera, with an accuracy of 81.3 percent and a time of around 1.3 seconds taken in classification. The effect of distance of the speaker was also tested, proving to not be an issue in terms of the system accuracy.

5.3 Future Work

There are multiple possibilities of future work in this dissertation. Some future work can be done in controlling other applications regarding other AAL scenarios, for example Youtube, Facebook or Spotify.

In therms of the Live usage of the system, the liberty of speaker to start saying a command needs to be improved. The best case scenario would be a system in which the speaker could, for example, be sited on the couch and just say the command whenever he wants, like turn up or down the volume, change movie etc. The implementation of this dissertation does not give this liberty for the speaker. It is hard for a visual speech recognition system to know the exact moment that the speaker starts to say the command that he wants, while the speaker is talking to his family, for example.

A non limited time to record the features extracted from the lips and chin is also needed due to the different ways that people talk, some people talk faster, someones do not. For people that talk slower, recording a large command like "Aumentar Volume" (Increase Volume) can take more that 2 seconds and the implementation of this dissertation will not record all the needed information to take the decision.

The system that was created is a Visual Speech Recognition system, non-invasive and non-obtrusive. However, it would be interesting to create and evaluate the results from a multimodal system using the features extracted with the created prototype in this dissertation and using information from other phases of the humans speech production, for example from the Conceptual Preparation, using signals from electroencephalographic sensors.

The speaker independency needs to be improved, since the results from this dissertation have shown that the prototype created is without doubt speaker dependent, which means that every user needs to record a database for himself before using the system.

Bibliography

Hélder Abreu. Visual speech recognition for european portuguese. Master thesis, Universidade do Minho, 2014.

Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library.* " O'Reilly Media, Inc.", 2008.

Jonathan S. Brumberg, Alfonso Nieto-Castanon, Philip R. Kennedy, and Frank H. Guenther. Brain-computer interfaces for speech communication. *Speech Commun.*, 52(4):367–379, April 2010. ISSN 0167-6393. doi: 10.1016/j.specom.2010.01.001. URL http://dx.doi.org/10.1016/j.specom.2010.01.001.

P. Dalka, P. Bratoszewski, and A. Czyzewski. Visual lip contour detection for the purpose of speech recognition. In *Proc. Int Signals and Electronic Systems (ICSES) Conf*, pages 1–4, September 2014. doi: 10.1109/ICSES.2014.6948716.

KH Davis, R Biddulph, and Stephen Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

Koenraad De Smedt. 11 computational models of incremental grammatical encoding. *Computational psycholinguistics: AI and connectionist models of human language processing*, pages 279–307, 1996.

Bruce Denby, Thomas Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. Silent speech interfaces. *Speech Communication*, 52(4):270–287, 2010.

João Freitas, António Teixeira, Carlos Bastos, and Miguel Dias. Towards a multimodal silent speech interface for european portuguese. In *Speech technologies*, pages 125–149. InTech, 2011. doi: 10.5772/16935.

João Freitas, António Teixeira, and Miguel Sales Dias. Towards a silent speech interface for portuguese. *Proc. Biosignals*, pages 91–100, 2012.

João Freitas, António Teixeira, and Miguel Sales Dias. Multimodal silent speech interface based on video, depth, surface electromyography and ultrasonic doppler: Data collection and first recognition results. In *Int. Workshop on Speech Production in Automatic Speech Recognition*, 2013.

João Freitas, António JS Teixeira, and Miguel Sales Dias. Multimodal corpora for silent speech interaction. In *LREC*, pages 4507–4511, 2014a.

João Freitas, António JS Teixeira, Samuel S Silva, Catarina Oliveira, and Miguel Sales Dias. Velum movement detection based on surface electromyography for speech interface. In *BIOSIGNALS*, pages 13–20, 2014b.

João Freitas, António Teixeira, Miguel Dias, and Samuel Silva. *An Introduction to Silent Speech Interfaces*. Springer, 2016. ISBN 978-3-319-40173-7. doi: 10.1007/978-3-319-40174-4.

A. Z. K. Frisky, C. Y. Wang, A. Santoso, and J. C. Wang. Lip-based visual speech recognition system. In *Proc. Int Security Technology (ICCST) Carnahan Conf*, pages 315–319, September 2015. doi: 10.1109/CCST.2015.7389703.

Georgios Galatas, Gerasimos Potamianos, and Fillia Makedon. Audio-visual speech recognition incorporating facial depth information captured by the kinect. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 2714–2717. IEEE, 2012.

Zoya Gavrilov. Svm tutorial, 2012. URL http://web.mit.edu/zoya/www/SVM.pdf.

B. Gick, I. Wilson, and D. Derrick. *Articulatory Phonetics*. Oxford: Backwell Publishing, 1st edition, 2013.

S Burak Gokturk, Hakan Yalcin, and Cyrus Bamji. A time-of-flight depth sensor-system description, issues and solutions. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 35–35. IEEE, 2004.

Ben Gold, Nelson Morgan, and Dan Ellis. *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Oxford, New York, NY, USA, 2nd edition, 2011. ISBN 0470195363, 9780470195369.

Peter Henry, Michael Krainin, Evan Herbst, Xiaofeng Ren, and Dieter Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. *The International Journal of Robotics Research*, 31(5):647–663, 2012.

Thomas Hueber, Gérard Chollet, Bruce Denby, and Maureen Stone. Acquisition of ultrasound, video and acoustic speech data for a silent-speech interface application. *Proc. of ISSP*, pages 365–369, 2008.

National Instruments. 3d imaging with ni labview, 2013. URL http://www.ni.com/white-paper/14103/en/.

Vincent Lanaria. Vlc, the worlds most popular media player, turns 15 years old: Heres why you should download it now, 2016. URL http://www.techtimes.com/articles/129950/20160203/vlc-the-world-s-most-popular-media-player-turns-15-years-old-here-s-why-you-should-download-it-now.htm.

Willem JM Levelt. Models of word production. *Trends in cognitive sciences*, 3(6):223–232, 1999.

Mariko Matsumoto. Silent speech decoder using adaptive collection. In *Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces*, IUI Companion '14, pages 73–76, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2729-9. doi: 10.1145/2559184.2559190. URL http://doi.acm.org/10.1145/2559184.2559190.

Microsoft. High detail face points, 2016a. URL https://msdn.microsoft.com/en-us/library/microsoft.kinect.face.highdetailfacepoints.

Microsoft. Face tracking, 2016b. URL https://msdn.microsoft.com/pt-pt/library/dn782034.aspx.

Roger K Moore and Anne Cutler. Constraints on theories of human vs. machine recognition of speech. In *Workshop on Speech Recognition as Pattern Classification (SPRAAC)*, pages 145–150. Max Planck Institute for Psycholinguistics, 2001.

WEKA University of Waikato. Attribute-relation file format (arff), 2008. URL http://www.cs.waikato.ac.nz/ml/weka/arff.html.

Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011.

Anne Porbadnigk, Marek Wester, Jan p Calliess, and Tanja Schultz. Eeg-based speech recognition impact of temporal effects. In *2nd International Conference on Bio-inspired Systems and Signal Processing (Biosignals 2009)*, 2009.

Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice hall, 1993.

R. A. Rao and R. M. Mersereau. Lip modeling for visual speech recognition. In *Proc. Conf Signals, Systems and Computers Record of the Twenty-Eighth Asilomar Conf*, volume 1, pages 587–590 vol.1, October 1994. doi: 10.1109/ACSSC.1994.471520.

WinBeta Ron. Company behind microsofts kinect sensor sold to apple for 345 million dollars, 2013. URL www.winbeta.org/news/company-behind-kinect-sold-apple-345-million.

L. D. Rosenblum. Speech perception as a multimodal phenomenon. *National Institutes of Health*, 17:405–409, 2013.

Kate Saenko, Trevor Darrell, and James R. Glass. Articulatory features for robust visual speech recognition. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 152–158, New York, NY, USA, 2004. ACM. ISBN 1-58113-995-0. doi: 10.1145/1027933.1027960. URL http://doi.acm.org/10.1145/1027933.1027960.

Himanshu Sahni, Abdelkareem Bedri, Gabriel Reyes, Pavleen Thukral, Zehua Guo, Thad Starner, and Maysam Ghovanloo. The tongue and ear interface: A wearable system for silent speech recognition. In *Proceedings of the 2014 ACM International Symposium on Wearable Computers*, ISWC '14, pages 47–54, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2969-9. doi: 10.1145/2634317.2634322. URL http://doi.acm.org/10.1145/2634317.2634322.

Manish Saraswat. A complete tutorial on tree based modeling from scratch (in r & python), 2016. URL https://www.analyticsvidhya.com/blog/2016/04/complete-tutorial-tree-based-modeling-scratch-in-python/.

Ronald W Schafer and Russel M Mersereau. Demosaicking: color filter array interpolation. *IEEE Signal Process*, 22:44–54, 2005.

J. A. Seikel, D. W. King, and D. G. Drumright. *Anatomy and physiology for speech, Language, and hearing*. Delmar Learning, 4th edition, 2009.

TeraRanger. Time-of-flight principle, 2016. URL http://www.teraranger.com/technology/time-of-flight-principle/.

Michael Wand, Jan Koutn, et al. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, 2016.

Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

Salah Werda, Walid Mahdi, and Abdelmajid Ben Hamadou. Lip localization and viseme classification for visual speech recognition. *arXiv preprint arXiv:1301.4558*, 2007.

Ian H. Witten, Eibe Frank, and Mark A. Hall. *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, third edition, 2011.

Alper Yargic and Muzaffer Dogan. A lip reading application on ms kinect camera. In *Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on*, pages 1–5. IEEE, 2013.