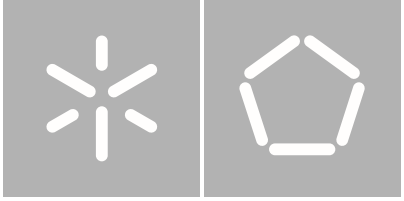
Universidade do Minho
Escola de Engenharia

Hélder Paulo Monteiro Abreu

Visual Speech Recognition for European Portuguese

Outubro 2014



Universidade do Minho
Escola de Engenharia

Hélder Paulo Monteiro Abreu

Visual Speech Recognition for European Portuguese

Dissertação de Mestrado

Mestrado em Engenharia Informática

Trabalho realizado sob a orientação de

Prof. Carlos Silva (Universidade do Minho)

Prof. Miguel Sales Dias (Microsoft Language Development Center, Instituto Universitário de Lisboa, ISCTE-IUL)

Outubro 2014

“All our knowledge has its origin in our perceptions.”

Leonardo da Vinci

ACKNOWLEDGEMENTS

First of all, I want to thank my parents for their endless support in all these years. Without them, I wouldn't achieve this. Thank you.

I also want to thank João Freitas, for his guidance and ability to discuss any issue. To Prof. Miguel Sales Dias for his guidance and for the opportunity to work at the Microsoft Language Development Center (MLDC). It has been an enriching experience.

To Prof. Carlos Silva for the guidance and the helpful insights, since the beginning of this work.

Finally, I want to thank my colleague and friend, Carlos Sotelo, for the frequent and useful exchange of ideas and opinions, and for helping to relativize the stressful moments.

RESUMO

O reconhecimento da fala baseado em características visuais teve início na década de 80, integrado em sistemas de reconhecimento audiovisual da fala. De facto, o objetivo inicial do recurso a características visuais foi o de aumentar a robustez dos sistemas de reconhecimento automático da fala, que perdem precisão rapidamente em ambientes ruidosos. Contudo, o potencial para manter um bom desempenho de reconhecimento de fala em situações em que os dados acústicos estão comprometidos ou em qualquer outra situação em que é necessária uma pessoa capaz de ler os lábios, levou os investigadores e a criar e desenvolver a área de reconhecimento visual da fala.

Os sistemas tradicionais de reconhecimento visual da fala usam apenas informação RGB, seguindo uma abordagem unimodal, uma vez que o recurso a outras modalidades é dispendioso e implica problemas de sincronização entre as mesmas. O lançamento do Microsoft Kinect, que inclui um microfone, uma câmara RGB e um sensor de profundidade, abriu novas portas às áreas de reconhecimento da fala. Para além disso, todas as modalidades podem ser sincronizadas usando as funcionalidades do SDK. Recentemente, a Microsoft lançou o novo Kinect One, que oferece uma melhor câmara e um sensor de profundidade com uma tecnologia diferente e mais precisa. O objetivo principal desta tese consiste em criar um sistema de reconhecimento visual da fala baseado no Kinect e verificar se um sistema multimodal, baseado em RGB e dados de profundidade, é capaz de obter melhores resultados do que um sistema unimodal baseado exclusivamente em RGB.

Considerando o processo de extração de características, uma abordagem recente baseada em características articulatórias tem mostrado resultados promissores, quando comparada com abordagens baseadas em visemas. Esta tese pretende verificar se uma abordagem articulatória obtém melhores resultados que uma abordagem baseada na forma.

O sistema desenvolvido, chamado ViKi (***Vi**sual Speech Recognition for **Ki**nect*), alcançou uma taxa de reconhecimento de 68% num vocabulário de 25 palavras isoladas, com 8 oradores, superando a abordagem unimodal testada. A informação de profundidade provou aumentar a taxa de reconhecimento do sistema, tanto na abordagem articulatória (+8%) como na abordagem baseada na forma (+2%). Num contexto de dependência em relação ao orador, ViKi também alcançou uma

média de ≈70% de taxa de reconhecimento. A abordagem articulatória obteve piores resultados que a abordagem baseada na forma, alcançando 34% de taxa de reconhecimento, contrariando os resultados obtidos em estudos prévios com abordagens baseadas na aparência e a terceira hipótese desta tese.

ABSTRACT

Speech recognition based on visual features began in the early 1980s, embedded on Audio-Visual Speech Recognition systems. In fact, the initial purpose to the use of visual cues was to increase the robustness of Automatic Speech Recognition systems, which rapidly lose accuracy in noisy environments. However, the potential to keep a good accuracy, whenever the use of an acoustic stream is excluded and in any other situations where a human lip reader would be needed, led researchers to create and explore the Visual Speech Recognition (VSR) field.

Traditional VSR systems used only RGB information, following an unimodal approach, since the addition of other visual modalities could be expensive and present synchronization issues. The release of the Microsoft Kinect sensor brought new possibilities for the speech recognition fields. This sensor includes a microphone array, a RGB camera and a depth sensor. Furthermore, all its input modalities can be synchronized using the features of its SDK. Recently, Microsoft released the new Kinect One, offering a better camera and a different and improved depth sensing technology.

This thesis sets the hypothesis that, using the available input HCI modalities of such sensor, such as RGB video and depth, as well as the skeletal tracking features available in the SDK and, by adopting a multimodal VSR articulatory approach, we can improve word recognition rate accuracy of a VSR system, compared to a unimodal approach using only RGB data.

Regarding the feature extraction process, a recent approaches based on articulatory features have been shown promising results, when compared to standard shape-based viseme approaches. In this thesis, we also aim to verify the hypothesis that an articulatory VSR can outperform a shape-based approach, in what concerns word recognition rate.

The VSR system developed in this thesis, named ViKi (**Vi**sual Speech Recognition for **Ki**nect), achieved a 68% word recognition rate on a scenario where 8 speakers, pronounced a vocabulary of 25 isolated words, outperforming our tested unimodal approach. The use of depth information proved to increase the system accuracy, both for the articulatory (+8%) and the shape-based approach (+2%). On a speaker-dependent context, ViKi also achieved an interesting average accuracy of ≈70%. The articulatory approach performed worse than the shape-based, reaching 34% of word

accuracy, contrary to what happens with previous research based on appearance approaches and not confirming our third hypothesis.

INDEX

LIST OF FIGURES

LIST OF TABLES

Abbreviations and Acronyms

AAL	Ambient Assisted Living
AAM	Active Appearance Model
AF	Articulatory Feature
ASR	Automatic Speech Recognition
AVSR	Audio-Visual Speech Recognition
CUAVE	Clemson University Audio Visual Experiments
DARPA	Defense Advanced Research Projects Agency
EP	European Portuguese
FFT	Fast Fourier Transform
FPS	Frames per Second
GUI	Graphical User Interface
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HSR	Human Speech Recognition
IR	Infrared
LPC	Linear Predictive Coding
ROI	Region of Interest
SDK	Software Development Kit
PSL	Portuguese Sign Language
PSLR	Portuguese Sign Language
SVM	Support Vector Machine
TOF	Time-of-Flight
UI	User Interface
ViKi	Visual Speech Recognition for Kinect
VSR	Visual Speech Recognition
WER	Word Error Rate

1. INTRODUCTION

Speech is the most natural form of human communication and researchers have been working for decades to make machines able to understand and pronounce it. Speech recognition, in particular, also known as automatic speech recognition (ASR), can be defined as the process of converting a speech signal to a sequence of words, by means of an algorithm implemented in a computer program. For reasons ranging from technological curiosity to the desire to automate simple tasks, the research in ASR has attracted a great deal of attention for the past sixty years. It has even inspired seventh art characters as *R2D2* and *C3PO* ("Star Wars", 1977) or *Sonny* ("I, Robot", 2004). Nowadays, it still inspires science fiction movies but also TV series like "Person of Interest", where we can find *The Machine* who's not only able to understand human speech but also capable to predict human behavior. All those machines can not only recognize spoken words but they're also able to comprehend its meaning. However, in spite of all the research that have been conducted, we are still far from creating one able to understand spoken discourse on any subject, by any speaker, in any environment. This is due to the complexity of the language itself, but also to a whole range of environmental conditions such as the presence of noise, the speed of speech, the gender or age of the speaker or even the speaker's mood.

As human beings, we use more than just sound to communicate. Facial expressions, as well as lips motion, body language and gestures, give us useful information to interpret words correctly, suggesting the benefits of using more than one source of information (multimodal approach), for the development of a speech recognition system.

1.1. MOTIVATION

The first approach in the speech recognition field was made using only the voice as an input, yielding to a human-computer interaction (HCI) system accepting several commands, as the one created at the Bell labs in 1952. This system may have been the first true word recognizer system, as it could be trained to recognize spoken digits from a single speaker (Davis, et al., 1952).

The visual approach for the speech recognition was only introduced in the late 70's, after the discovery of the psychologist Harry McGurk and his assistant, John MacDonald. They accidentally discovered an audio-visual illusion, known since then as the McGurk effect or McGurk illusion (McGurk & MacDonald, 1976), that demonstrated the combined nature of speech understanding and the importance of the visual features, leading to bimodal speech recognition systems (Petajan, et al., 1988). Thereafter, several multimodal databases, recorded in distinct conditions, emerged to be used in speech recognition research. For example, the CUAVE (*Clemson University Audio-Visual Experiments*) is an English database with a *corpus*[1] containing more than 7000 expressions of a vocabulary[2] comprising isolated and connected digits (Patterson, et al., 2002).

Despite the good results achieved by multimodal speech recognition systems, the audio data captured in unsupervised environments is occasionally unusable. For instance, let's imagine a speech recognition system included in a car radio, allowing the driver to switch the radio station or the CD track. If we consider the engine, the traffic or even the music noise, it becomes obvious that an audio recognition system would be highly unreliable. Visual Speech Recognition (VSR) systems have the advantage of not being influenced by acoustic noise. They can be used by people with speech problems (e.g. laryngectomy[3], muteness or temporary states of aphonia), when, for privacy maters, it is not advisable to make noise, or simply when it is not possible to control the environment's noise. These systems are essentially lip-reading based, which gives them the same weaknesses as the human lip readers. Lip reading weaknesses are essentially related with visibility conditions, but also to the fact that some sounds are visually indistinguishable from others or simply invisible. While phonemes can be defined as units of different sounds perceived of a language or dialect, visemes are based on the visual aspect of the phonemes. Knowing this, it's natural that the number of visemes is smaller than the number of phonemes, since several phonemes are

1 Collection of words or sentences of a database.
2 The body of words used in the research.
3 Removal of the larynx and separation of the airway from the mouth, nose and esophagus.

classified as the same viseme (e.g. in the European Portuguese, the phonemes /f/ and /v/ belong to the same viseme class). To compensate the lack of visemes classes, lip readers often rely on higher-level linguistic information as the pragmatic, semantic, phonetic or syntactic context, associated to the message. Automatic systems based on visual cues don't have access to this kind of information, which hinders the recognition process. On the other hand, VSR systems can collect precise information unavailable to Human lip readers (e.g. mouth features variations, like width and height).

Traditionally, in VSR research, standard video cameras are used to collect the databases (Patterson, et al., 2002; Pera, et al., 2004; Lee, et al., 2004). Therefore, researchers were limited to RGB images only (unimodal approach), to extract the necessary features for the recognition. The release of the Kinect (for Xbox) in 2010 and Kinect for the Windows platform in 2012, brought new possibilities to the ASR, Audio-Visual Speech Recognition (AVSR) and VSR fields. The Kinect sensor enables the users to control and interact with their computer, through a natural user interface (UI) using gestures and spoken commands. It features an RGB camera, a depth sensor and a microphone array, providing full-body 3D motion capture, facial recognition and robust voice recognition capabilities with echo canceling. Both the RGB camera and the depth sensor are able to capture 30 FPS with a 640x480 resolution, while tracking 20 skeletal joints of up to 2 simultaneous skeletons. Since its release, more and more Kinect-based recognition systems emerged, using several combinations of its features (Yargıç & Doğan, 2013; Sun, et al., 2012; Galatas, et al., 2012; Galatas, et al., 2011). One of the greatest advantages of the Kinect over traditional cameras is the ability to track up to 100 face points. Since the face tracking is a key step on any speech recognition system based on visual features, it becomes a very interesting tool for VSR researches.

The Kinect for Windows brought a range of new possibilities for the speech recognition research fields throughout the several released versions (1.5 to 1.8) of its Software Development Kit (SDK) but, as versatile as it looks, it also has some downsides. The VGA resolution of its RGB camera lacks details compared with most of the existing standard cameras. Furthermore, the depth information is collected using a technique known as structured light, which principle is to project a known pattern onto the scene and infer depth from the deformation of that pattern (Zhang, et al., 2002). This technique obtains direct measurements only at scene points illuminated by the light pattern. Depth values at non-illuminated points have to be derived via interpolation or surface fitting (Schaller, 2011), which is problematic in contours regions with abrupt depth variations (e.g. inner contours of the lips when the mouth is open). Moreover, the tracking technique of the 100 face

points is based on Active Appearance Model (AAM). The main problem of using an AAM approach is that it can generate non-realistic shapes/appearance configurations (Gonzalez-Mora, et al., 2007). Since the shape of the mouth is widely used in shape-based approaches of VSR researches, the tracked mouth points must be used cautiously.

The second version of the Kinect, the Kinect One, was released in July 2014, along with the public preview of the Kinect for Windows SDK 2.0. It brings several improvements over its predecessor, like a Full HD color camera able to capture 30 FPS with a 1920 x1080 resolution and a wider field of view. The skeletal tracking can now follow as many as 6 complete skeletons and 25 different joints (including thumbs). The depth sensor has a higher depth fidelity and significantly improved noise floor, using now a Time-of-Flight (TOF) technology. TOF cameras use intensity modulated infrared (IR) light, which is not visible to human eyes to illuminate the scene. Then, a Photon Mixing Device sensor captures the reflected light and evaluates the distance information on the pixel, with no need of values interpolation. These improvements make the new Kinect One even more interesting for VSR purposes than its predecessor, allowing to use of high resolution color images along with precise depth information.

In VSR systems, after capturing and extracting the region of interest (ROI) of each frame, the extraction of visual features normally falls into three main categories: appearance-based approaches (using the color of the pixel), shape-based approaches (e.g. using geometric features, like the width and height of the mouth, or the angle of its corners), or hybrid approaches (using a combination of the previous ones). The extracted features are then sent to some classification system to identify visemes, isolated words or sentences. In the last years, a new approach emerged in the VSR domain, based on the use of articulatory features (AF), proposing a multiple stream of visible linguistic features (e.g. lips closing, lips opening, tongue protrusion and so on) to classify words, as opposed to a single stream of visemes (Saenko, et al., 2004). It was shown that AF-based models can outperform conventional single-stream viseme-based models (using appearance-based approaches), in a medium-vocabulary isolated word task and in small-vocabulary connected word task (Saenko, et al., 2009). However, AF were always extracted manually or by using appearance-based approaches. No references were found of articulatory approaches being compared with shape-based approaches.

Regarding the languages used on VSR, most of the vocabularies are based on the English language. VSR systems based on European Portuguese (EP) vocabularies were not found in the literature. The existing ones were only based on RGB images (Moura, 2005; Moura, et al., 2005).

1.2. PROBLEM

Based on the previously presented information, it is possible to identify the underlying problems that formed the basis for this thesis hypothesis.

- Traditional VSR systems only rely on RGB data to perform feature extraction. Depth data has been collected for a couple of VSR and AVSR systems (Galatas, et al., 2012; Yargıç & Doğan, 2013), but using the first version of the Kinect sensor.
    - Question 1: Does skeletal tracking and improved depth data (multimodal approach) from the new Kinect One sensor, allows us to increase the accuracy of VSR systems when compared with traditional systems using only RGB images (unimodal approach)?
- Few VSR systems follow an articulatory approach (Saenko, et al., 2004; Saenko, et al., 2005; Saenko, et al., 2009) and none of them used depth data for features extraction.
    - Question 2: Does articulatory approaches for VSR based on RGB images and depth data outperform articulatory approaches based only on RGB images?
    - Question 3: Does articulatory approaches for VSR outperform shape-based approaches?

1.3. HYPOTHESIS

From the research questions presented in the section 1.2, we propose the following thesis hypothesis:

- **H1**: A multimodal approach, based on skeletal tracking, RGB imaging and depth data, improves the VSR systems accuracy compared to the use of unimodal approaches using only RGB imaging.
- **H2**: An articulatory VSR approach based on RGB imaging and depth data outperforms RGB-based articulatory approaches.
- **H3**: An articulatory VSR approach outperforms a shape-based approach.

1.4. Objectives

To address the stated hypothesis, we've established the main objective of this thesis, as the development of a multimodal VSR system for the EP language (**O0**).

To address more specifically each of the hypotheses, we have stated the following objectives:

O1: To demonstrate hypothesis **H1**, we have used Computer Vision techniques and the skeleton tracking, RGB imaging and depth data available on the Kinect One sensor, aiming to extract some geometric and articulatory features and assess its influence on the word recognition rate, in a process of recognizing a word out of a 24 words vocabulary, and then comparing the results with the classical RGB-only unimodal approaches.

O2: To demonstrate hypothesis **H2**, we have tested and compared the word recognition rate results of our articulatory approach using RGB and depth data with the same approach using only RGB data, in a word recognition task.

O3: To demonstrate hypothesis **H3**, we have tested and assessed if our articulatory approach (using opening, rounding and the protrusion of the lips as features) outperforms our shape-based approach (using the width, height and depth variation of the lips), for the same word recognition task.

To the best of our knowledge, this is the first VSR system for EP using the new Kinect One sensor.

1.5. Dissertation Overview

This document aims to introduce all the necessary background for the understanding of the VSR field and the developed system.

Chapter 2 presents a review of the state of the art of VSR, starting with a background section to introduce important concepts related to the speech process and speech recognition by humans and machines. The applications of VSR systems on the real world are also discussed, as well as the main issues and challenges on the development of speech recognition systems.

Chapter 3 includes all the key processes on the development of the proposed VSR system – ViKi. The Graphical User Interface (GUI) is introduced, presenting all the features of the developed application. The Data Collection section presents the selected vocabulary, participants and structure of the collected data. All the data processing, including the features extraction and the classification processes are also discussed.

Chapter 4 discusses the analysis and validation of the developed system, using all the corpora and speakers. The depth data influence and the accuracies obtained on each of the words sets separately are discussed. For a better understanding of the intra- and inter-speaker variability, results from an individual training and validation of each speaker is also presented.

Finally, chapter 5 concludes this dissertation with the discussion of the verification of our stated Hypothesis, based on the results of chapter 4. Some comments on the work done and suggestions for future activities, are also made.

2. State of the Art of VSR

Conceptually, the development of speech recognition is closely tied with other developments in speech science and engineering and, as such, it can be viewed as having roots in studies going back to the Greeks. However, the history of speech recognition systems itself started in the 20th century, bonded with the first commercial appearances of the microphone. Thomas Edison (1847–1931) is well known for his work refining the carbon granule microphone[1]. His transmitter was simple and cheap to manufacture, but also very efficient and durable, being employed on the first ever radio broadcast in 1910. It also became the basis for the telephone transmitters, used on millions of telephones around the world for the majority of the last century, as well as an essential tool for the speech recognition field (Robjohns, 2000).

This chapter aims to introduce important concepts related to the speech, providing the necessary background in human speech, for understanding the development of a speech recognition system. The practical applications of speech recognitions systems will be presented, as well as their notable evolution, from a basic toy of the beginning of the 20th century (able to recognize only one word), to the powerful virtual assistants existing on our smartphones today (able to perform complex tasks and learn from our habits). Finally, the recent work in the VSR field and the main issues related with the development of speech recognition systems will also be discussed.

1 Originally created by David Edward Hughes (1831-1900).

2.1. Background

Human Speech Recognition (HSR) has been the subject of many researches for several years, side by side with speech recognition by machines, essentially to assess the evolution of the latter's accuracy. But, considering the complexity of the speech recognition process, it is necessary to have some background from several fields to understand the fundamental aspects of its development. This section will discuss how the human speech is processed and perceived, and the mechanics of speech recognition systems.

2.1.1. Speech Production

In normal fluent conversation we are able to produce two to three words per second, which amounts to about four syllables and ten or twelve phonemes per second. These words are continuously selected from a huge repository, the mental lexicon[1], which contains at least 50–100 thousand words in a normal, literate adult person. Even so, the high speed and complexity of word production does not seem to make it particularly error-prone since we err, on average, no more than once or twice in 1000 words (Levelt, 1999).

The organization of speech processing is grounded in the fact that the brain has to do two kinds of things with the acoustic speech information. This model holds that a ventral stream is involved in processing speech signals for comprehension while a dorsal stream is responsible for translating acoustic speech signals into articulatory representations, which are essential for speech production. A wide range of evidence suggests that the ventral stream is bilaterally organized (although with important computational differences between the two hemispheres). The dorsal stream, on the other hand, is strongly left dominant (Hickok, 2013). In most people, the cerebral area responsible for the speech is in the left cerebral cortex. Two major cortical areas are involved in this process: Wernicke's area (sensory speech area), a portion of the parietal lobe, and Broca's area (motor speech area) in the inferior part of the frontal lobe. Wernicke's area is necessary for understanding and formulating coherent speech. Broca's area initiates the complex series of movements necessary for speech (Seeley, et al., 2008). The two areas are connected by a bundle of neurons known as the arcuate fasciculus (see Figure 1).

1 Or person's vocabulary

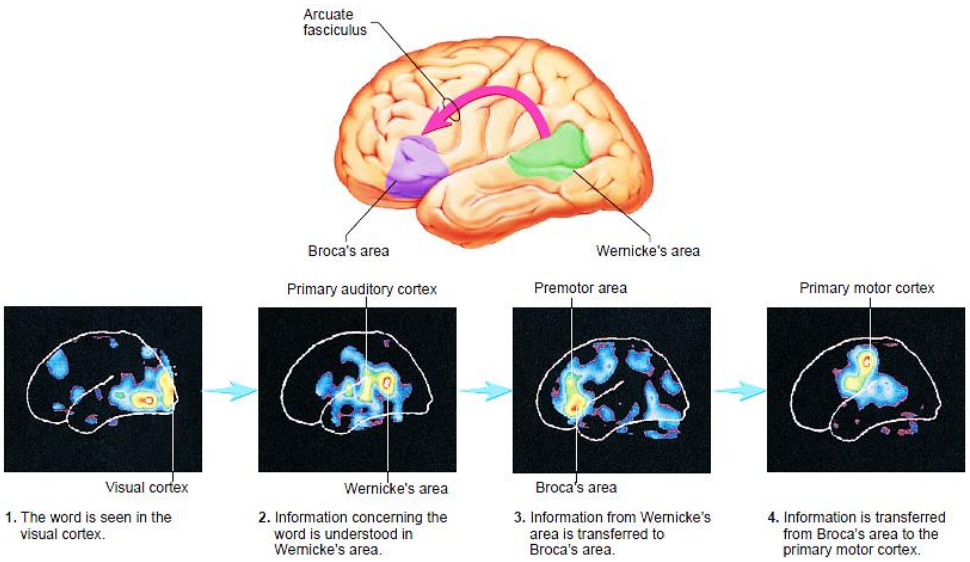


Figure 1 - Description of the cortical activities during speech (Seeley, et al., 2008)

Regarding the steps involved in the speech production, this process can then be divided in 3 main events, as represented in Figure 2 (speaker side):

1. **Conceptual preparation (thought)** – Consists in the activation of a lexical concept for which you have a word in your lexicon. Usually, such a concept is part of a larger message but, even in the simple case of naming a single object, it is not trivial which lexical concept should be activated to refer to that object, depending on the discourse (Levelt, 1999).
2. **Formulation (linguistic representation)** – It is the process corresponding to the linguistic representation required for the expression of the desired message. It comprises both the selection of lexical material and the organization of this material in a syntactic framework (Smedt, 1996).
3. **Articulation (facial movements + acoustics)** - Series of neuromuscular commands to cause the vocal cords to vibrate when appropriate and to shape the vocal tract such that the proper sequence of speech sounds is created and spoken, producing an acoustic signal as the final output. The neuromuscular commands must simultaneously control all the aspects of the articulatory motion, including the control of the lips, jaw, tongue and velum (Rabiner & Biing-Hwang, 1993).

A speech recognition system is based on the data obtained from the articulation step, since it's the only physical process from which we are able to extract information. For VSR purpose, the anatomical structures used as reference are the visible ones (lips, jaw, cheeks, teeth and tongue). Considering this, they will not be presented.

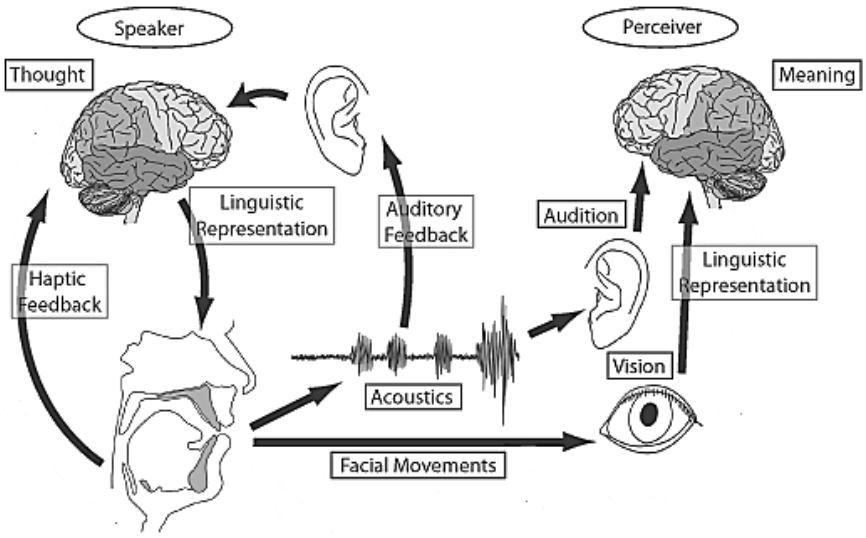


Figure 2 - Multimodal speech chain representation with feedback loops (Gick, et al., 2013)

2.1.2. Speech Perception

Spoken syllables may persist in the world for mere tenths of a second. Yet, as adult listeners, we are able to gather a great deal of information from these fleeting acoustic signals. We may apprehend the physical location of the speaker, the speaker's gender, regional dialect, age, emotional state, or identity. These spatial and indexical factors are conveyed by the acoustic speech signal in parallel with the linguistic message of the speaker. Although these factors are of much interest in their own right, speech perception most commonly refers to the perceptual mapping from acoustic signal to some linguistic representation, such as phonemes, syllables or words (Holt & Lotto, 2010). Speech perception, in addition to be of great scientific interest, is also a main issue to those concerned with building systems for 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 nonlinguistic factors in the speech signal (Gold, et al., 2011).

The ease with which people frequently comprehend speech by audition alone and the methodological rigors associated with controlling several different stimulus modalities are perhaps two reasons why speech perception research has been mostly focused on auditory speech perception (Bernstein & Benoît, 1996). However, speech perception is inherently multimodal. Brain regions once thought sensitive to only auditory speech, are now known to respond to visual speech input. Despite our intuition of speech as something we hear, there is now overwhelming evidence that

the brain treats speech as something we hear, see, and even feel (Rosenblum, 2013). Human speech perception rely then on both acoustic and visual cues to decode speech. These visual cues usually include the position and movement of the visible articulators (e.g. lips, teeth and tongue), and other cues not directly related to the production of speech (e.g. facial expression, head pose and body gestures). As listening conditions degrade, speech perception increasingly relies on these visual cues to assist with decoding spoken words. Indeed, visual speech automatically integrates with auditory speech in a number of different contexts (Hilder, et al., 2009; Rosenblum, 2013).

When we speak, we are also constantly monitoring and adjusting what we're doing. Speech production uses many different sensory mechanisms for feedback. The most commonly known feedback in speech is auditory feedback, though many senses are important in providing feedback in speech. Speech is often thought of largely in terms of sound. Sound is indeed an efficient medium for sharing information: it can be disconnected from its source, can travel a long distance through and around objects, and so on. As such, sound is a powerful modality for communication. Likewise, auditory feedback from sound provides a speaker with a constant flow of feedback about his or her speech (Gick, et al., 2013). But speech can also be perceived visually, by watching movements of the face and body. Lip reading is a god example, since it's used to understand or interpret speech without hearing it, a technique especially mastered by people with hearing difficulties. Lip reading was practiced as early as 1500 AD, and probably before that. The first successful lip reading teacher was the Spanish Benedictine monk, Pietro Ponce, who died in 1588. Lip reading teaching subsequently spread to other countries. The German Samuel Heinecke opened the first lip reading school in Leipzig in 1787. The first speech reading conference was held at Chautauqua, USA in 1894. Several different methods have been described in the literature for human lip reading such as the Muller-Walle, Kinzie, and the Jena methods. Although only 50% or less of speech can be seen, the reader must guesstimate the missed parts. This was the core of the Jena method: training the eye and exercising the mind. However, regardless of the variety of known lip reading methods, all methods still depend on the lip movement that can be perceived by the lip reader (Hassanat, 2011).

As shown in the Speech Production section, speech can be described as a dual route model which includes a ventral stream, involved in processing speech signals for comprehension, and a dorsal stream, responsible for speech production. Speech perception is often conceived of as a complex categorization task accomplished within a highly multidimensional space. One can conceptualize a segment of the speech signal as a point in this space representing values across

multiple acoustic dimensions. In most cases, the dimensions of this space are continuous acoustic variables such as fundamental frequency, formant frequency, formant transition duration, and so forth. That is, speech stimuli are represented by continuous values, as opposed to binary values of the presence or absence of some feature. Speech perception is the process that maps from this space onto representations of phonemes or linguistic features that subsequently define the phoneme (Holt & Lotto, 2010).

Speech processing is not a trivial task. It has been the subject of many researches, leading to several theories. Observations of the automaticity of audiovisual speech have had important influences on these theories (Rosenblum, 2005) but, considering the scope of this dissertation, they will not be described.

2.1.3. Speech Recognition by Machine

The technology has reached a point where large-vocabulary speaker-independent continuous speech recognition is now available online, and where small-vocabulary voice command-and-control is becoming a familiar feature for users of smartphones. But, after years of research and development, the accuracy of speech recognition systems remains one of the key challenges. This leads the interest to the sophisticated mechanisms that give the human speech perception its excellence in comprehension and error correction. Although the simple imitation of human mechanisms may not be a good approach to engineering design, it is still likely to be useful to study the functional characteristics of human speech perception (Gold, et al., 2011).

Modern speech recognition systems, like HSR, have three main processes (see section 2.1.1). Table 1 shows these processes and their correspondence in the human speech perception system. A speech recognition system must always have, at least, one sensor able to capture some speech feature. The most common speech recognition systems are based on the acoustic feature, requiring one or more microphones, but other features can also be captured like the face morphology (using a depth camera) or the electrical activity of the face muscles (using surface electromyography). Similarly, humans use their ears and eyes in this first process to capture audio and visual information. After the capture of the selected characteristic, the data must be processed. While humans convert the captured information into the appropriate linguistic representation, a speech recognition system must extract useful features to be subsequently used in the final process. This

processing depends heavily on the type of speech recognition system but what is normally required is a mathematical and scientific formalism for encoding the data in a computationally useful form: a formalism which can exploit patterns, can generalize from seen data to unseen data, and in response to some efficiency criterion uses the minimum information to achieve its goals. For example, in the VSR field, it is common to extract geometric features like the width and height of the mouth to be used in the classification, since each word can be described in terms of the variation of these features. Finally, the data must be encoded in a suitable model (commonly known as feature vector), to compute the most likely state of the model for a specified input, which is normally achieved through a classification system. This process can be compared with the meaning attribution on the human perception (Moore & Cutler, 2001).

Table 1 - Correspondences between Speech Recognition Systems and Human Speech Perception

Process	Speech Recognition System Feature	Human Speech Perception Feature
Conversion of the audible and/or the visible information into data that can be processed	One or several sensors (e.g. microphone and/or camera)	Ears and eyes
Conversion of the captured information from the sensors in a structured representation	Feature extraction mechanism – feature vector	Linguistic representation
Recognition of the processed information	Classification system	Meaning attribution

2.1.4. Human vs. Machine Results

Since the emergence of ASR, researchers aim to create a system able to perform as a human. Central issues in the study of ASR and HSR are clearly comparable in respect of the aim to understand the process of extracting linguistic information. Nevertheless, the research communities that concern themselves with ASR and HSR are largely distinct. In part this follows from the way the two research tasks are envisaged. In the machine side, the objective has been to define a complete end-to-end process for moving from low-level to high-level descriptions of a speech signal.

ASR models are typically presented in the form of an entire working system that takes acoustic input (or visual input, in case of VSR systems) and produces a lexical or semantic-level description as output. In contrast, HSR research has been subdivided, and its separate components parceled out to different research domains. As a consequence, HSR theories are models of specific aspects of this process, with minimal attention to the feasibility of interfacing to models of other stages in the recognition process (Moore & Cutler, 2001).

The major obstacle for HSR researchers is the impossibility of direct observation. HSR researchers cannot take the system apart, neither can they switch off components one by one to identify their separate contributions, nor can they adjust the system and compare the performance of versions which differ only in a single respect. On the other hand, automatic systems are based on a statistical modelling paradigm, which allows generalization to unseen data. In fact, a small but finite probability can be attached to any acoustic/visual signal - even one unlikely to have been generated by any known physical system. One consequence of this, is that, not only are automatic systems able to ride over a momentary loss of data, but it would appear to hallucinate with data that was well outside of its experience - that is, data on the tails of its probability distributions (Moore & Cutler, 2001).

For these reasons, the direct comparison of human and machine accuracies is a difficult process. Human error rates should ideally be compared only to error rates of the current single best-performing speech recognition system. Such comparisons are impossible because no single recognition system has been tested on all corpora and shown to provide the best performance across all conditions sampled by these corpora. In general, speech recognition systems developed for one corpus degrade on other corpora due to the careful tuning required to provide the best performance. Results obtained with human listeners are also sometimes inflated because experimenters reduced errors caused by inattention by allowing multiple listening passes and by using a committee majority vote across multiple listeners instead of using the average individual error rate. Experimenters also sometimes corrected for spelling errors to eliminate out-of-vocabulary responses. Such out-of-vocabulary responses are not possible for machine recognizers which use limited vocabulary testing when the training and testing vocabularies are known and identical (Lippmann, 1997).

Despite the presented issues, several studies compared the performance of speech recognition systems and humans, in multiple contexts. Lippmann (Lippmann, 1997) presented some results obtained from several studies comparing HSR and ASR. Figure 3 shows some of the key results, ranging from connected digit recognition to the transcription of spontaneous telephone speech. These results clearly indicate that ASR recognition accuracy lags about an order of magnitude behind that of HSR, at the time of writing his paper. Several years later, Meyer et al. (Meyer, et al., 2006) also compared ASR and VSR accuracies obtaining similar results. Across all conditions, humans outperform ASR systems. The differences are clearly visible in the noise-free conditions, where the average human performance is 99.6% and the ASR performance is 74.0 %. They used the German dialect from four different regions (East Frisian, East Phalian, Bavarian and standard German) showing that the better results were obtained with the standard German (HSR: 99.7% accuracy; ASR: 81.1% accuracy). The ASR accuracy was reduced by 2.5 to 10.2% absolute for the dialects while, for the HSR test, an average performance drop of 4% was observed. Other than for the ASR experiment, the differences between dialects were not significant.

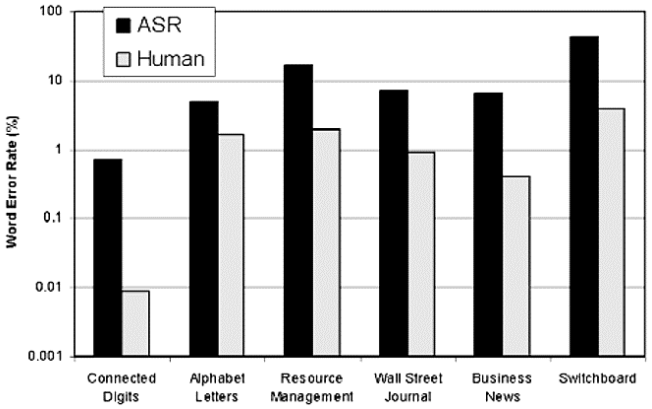


Figure 3 - Comparison of human and automatic speech recognition performance (Moore & Cutler, 2001)

Recently, Juneja compared ASR and HSR but using a null grammar continuous speech vocabulary with varying perplexities, claiming that the results are not biased from a language model while the accuracy tests are done on continuous speech at the same time (Juneja, 2012). In this study, ASR exhibited as much as an order of magnitude more errors than HSR, even if supplied with the same context. That, combined with the observation that the increase in WER is very similar for ASR and HSR for a significant change in perplexity, implies that there were a large difference in the acoustic modeling performance of ASR and HSR. While humans may be adept at context awareness and exception handling, these results shown that the gap in acoustic modeling between humans and machines is still fairly large, even when there are no exceptions in the ordering of words and when the context is the same.

Considering that most of ASR work focuses purely on speech acoustics or on audio-visual speech, it is natural that few studies compared HSR and VSR results. The human ability for lip reading varies from one person to another, and depends mainly on guessing to overcome the lack of visual information. Lip readers need to have a good command of the spoken language, and in some cases the lip reader improvises and uses his/her knowledge of the language and context to pick the nearest word that he/she feels fits into the speech. Moreover, human lip readers benefit from visual cues detected outside the mouth area (e.g. gestures and facial expressions). The complexity and difficulties of modelling these processes present serious challenges to the task of VSR. Nevertheless, Hilder et al. (Hilder, et al., 2009) compared the performance of human and machine lip-reading ability, based on shape-only (using AAM) and shape + appearance information. They found that automatic lip-reading systems outperform human viewers at both the viseme and the phoneme level (80.27% and 91.6% of accuracy compared to 31.6% and 35.4%, respectively) with full shape and appearance information. In the full, isolated, real words recognition task, human recognition outperform the VSR system (18.42% compared to 5%). The authors justified this phenomena by the prior knowledge of English words, which is not provided to the machine-based system. An important fact to notice about this study is that none of the subjects who have participate was a professional lip-reader.

2.2. Real World Applications

Speech recognition has been successfully applied in a range of systems and has the potential to reach multiple fields. The main fields where speech recognition can be found today range from office applications to medical systems, and they can be categorized into several broad classes (Rabiner & Juang, 2000).

- ### Office / Business

The most common office applications supporting speech recognition, normally include dictation, but other features like data entry onto forms, database management and control, and keyboard enhancement can be found. In addition to a large vocabulary speech-to-text capabilities, systems like *Dragon* (Nuance Communications, Inc., 2014) or *VoxSigma Software Suite* (Vocapia

Research SAS, 2014) include adaptive features allowing the transcription of noisy speech, such as speech with background music.

- **TELECOMMUNICATIONS**

The speech recognition technologies have a strong presence in telecommunications. Nowadays, most of mobile phones support voice dialing by name or number and it is common for telecommunications companies to include automation of operator assisted services (voice recognition call processing systems to automate operator service routing according to call types). *VoltDelta OnDemand Solutions* (VoltDelta, 2014) is an example of an automated customer service call centers in the telecommunications market.

- **MEDICAL**

Clinicians at many healthcare delivery networks rely on speech recognition to document patient encounters reporting several benefits including reduced transcription expense, increase of productivity, reduced time spent in documenting care and improved patient care via more detailed documentation and faster results delivery (Nuance Communications, Inc., 2008). Speech recognition has a similar role in the medical field as in the office/business field, delivering free-text dictation systems with medical vocabulary and even regional accent support (1st Provider's Choice, 2014).

- **OTHERS**

Speech recognition has some modest, but growing, applications in several other fields:

› Video Games - The Microsoft Kinect, the XBOX and XBOX One accessory, supports voice commands in several languages;
› Automotive Market - The Android Car, that should be launched in the late 2014, will support several functions like GPS mapping/navigation or music control;
› Biometrics - Biometrics systems based on speech recognition can currently perform authentications through natural voice pattern instead of passwords or pins;
› Digital Personal Assistants – Nowadays, these systems are increasingly present in each one's life, in our smartphones, tablets or laptops (e.g. Google Now, Siri and Cortana, present in the 3 major mobile operating systems – Android, iOS and Windows Phone, respectively), allow to search online or simply add a reminder using the voice.

› Robotics – As shown in the previous chapter, speech recognition has a strong presence in the robotics field, incorporating robots like AIBO or ASIMO.

2.3. The Evolution of Speech Recognition Systems

In the early 1920s machine recognition came into existence. The first machine to recognize speech to any significant degree was a toy manufactured in 1920, commercially named Radio Rex. Rex was a celluloid dog with an iron base and an in-built system sensible to acoustic energy (see Figure 4). The energy contained in the vowel of the word "Rex" was sufficient to trigger the system, making the dog to leave its house.

Figure 4 - Radio Rex toy

Despite its ingenious way of responding when its name was called, the toy responded to many words other than "Rex", and even to many non-speech sounds that had sufficient acoustic energy to trigger it. This inability to reject out-of-vocabulary words/sounds is an issue shared by most recognition systems that followed it. Although much of the work in vocoding and related speech analysis in the 1930s and 1940s was relevant to speech recognition, the next complete system of any significance was developed at Bell Labs in the early 1950s (Moura, 2005; Anusuya & Katti, 2009; Gold, et al., 2011).

The first true ASR system was built in 1952 at Bell Laboratories and it consisted of an isolated digit recognition system. It was able to achieve an accuracy varying between 97% and 99% and it could be trained to recognize digits from a single speaker. This system relied heavily on measuring spectral resonances during the vowel region of each digit (Davis, et al., 1952). In 1956, a similar work aimed to recognize 10 distinct syllables from 10 monosyllabic words in a single speaker system. It also relied in spectral measurements (provided by an analog filter bank) primarily during vowel regions

Figure 5 - Dr. E. A. Quade demonstrating the Shoebox

(Olson & Belar, 1957). In 1962, IBM presented its "Shoebox" machine at the Seattle World's' Fair (see Figure 5). This device recognized and responded to 16 spoken words, including ten digits (from 0 to 9) and six arithmetical operations. It was able to instruct an adding machine when a number and command words such as "plus," "minus" and "total" were spoken and print answers to simple arithmetic problems (IBM, 2003). The particularity of the Shoebox was its size: smaller than a shoe box.

Prior to the 1060s, the primary approach to estimating the short-term spectrum was a filter bank. Then, some breakthroughs occurred that became important for the speech recognition field in the 1970s. Three spectral estimation techniques were developed that were later of great significance for recognition, although their early applications to speech were for vocoding: the Fast Fourier Transform (FFT), Cepstrum analysis, and Linear Predictive Coding (LPC). The FFT is one of the most fundamental numerical algorithms. It computes the Discrete Fourier Transform (DFT) of an n-dimensional signal in $O(n \log n)$ time. The algorithm plays a central role in several application areas, including signal processing and audio / image / video compression. It is also a fundamental subroutine in integer multiplication and encoding / decoding of error-correcting codes (Gold, et al., 2011; Hassanieh, et al., 2012). The Cpestrum analysis was originally developed by Bogert for seismic analysis in 1963 (Bogert, et al., 1963) and consists of a nonlinear signal processing technique. Its significance for speech recognition is primarily as an approach to estimating a smooth spectral envelope. It ultimately became widely used for recognition, particularly in combination with other analysis techniques, being applied to speech and audio signals by Oppenheim, Schafer, and Stockham (Rabiner & Juang, 2000; Gold, et al., 2011). LPC is a mathematical approach to speech modeling and one of the most powerful speech analysis techniques. It provides an efficient way of finding a short-term spectral envelope estimate that has many desirable properties for the representation of speech, in particular the emphasis on the peak spectral values that characterize voiced sounds. It is also a useful method for encoding good quality speech at a low bit rate. A final achievement of note in the 1960s was the pioneering research of Reddy in the field of continuous speech recognition by dynamic tracking of phonemes (Gold, et al., 2011).

In the 1970s, speech recognition research achieved a number of significant milestones. The isolated word or discrete utterance recognition systems became a viable and usable technology, IBM launch a highly successful group effort in large vocabulary speech recognition and researchers at AT&T Bell began a series of experiments aimed at making speech recognition systems that were

truly speaker independent. The Defense Advanced Research Projects Agency (DARPA)[1] also funded a large speech-understanding project which goal was to perform 1000-word ASR system by using a few speakers, connected speech, and constrained grammar with less than a 10% semantic error (Anusuya & Katti, 2009; Gold, et al., 2011). But the 1970s also brought new discoveries in human speech perception that would influence the course of the speech recognition field, like the one made by the psychologist Harry McGurk and his assistant John MacDonald. They were studying how infants perceive speech during different periods of development. For example, they placed a videotape of a mother talking in one location while the sound of her voice played in another. For some reason, they asked their recording technician to create a videotape with the audio syllable “ba” dubbed onto a visual "ga". When they played the tape, McGurk and MacDonald perceived "da". After the initial confusion, they realized that "da" resulted from a quirk in human perception, not an error on the technician’s part. They then reported this phenomenon in their 1976 paper (McGurk & MacDonald, 1976), becoming known as the McGurk effect and proving the multimodal nature of human perception. So, the acoustic speech signal can be seen as the primary cue for recognizing speech, but visual observation of the lips, teeth, tongue, and jaw contribute to perception of phoneme articulation, while the angle of the head and raising of the eyebrows help convey the intonation of speech (Granström & House, 2004).

The findings of McGurk reverberated in the following years, leading to the emergence of AVSR systems in the 1980s. In 1984, Petajan gave the first step in the research of speech recognition based on both acoustic and visual information (Petajan, 1984). He designed a lip-reading system using geometric features, such like the mouth height, width, area and perimeter, to aid his ASR system for the recognition of isolated words from a 100-word vocabulary that included digits and letters. He concluded that the combination of both modalities yield a final recognition accuracy which greatly exceeds the acoustic recognition accuracy alone. This increasing of accuracy is particularly visible in degraded conditions like noisy environments, where traditional ASR systems performs poorly.

The growing number of AVSR systems led to the creation of several audio-visual databases, with different purposes. For example, the Clemson University Audio Visual Experiments or CUAVE (Patterson, et al., 2002), created in 2002, is an English audio-visual database of isolated and connected digits that include an even representation of male and female, for a total of 36 speakers, and several particularities like different skin tones and accents or other visual features such as

1 Agency of the United States Department of Defense responsible for the development of new technologies for use by the military.

glasses, facial hair, and hats. It also includes records with 2 speakers simultaneously and with speakers moving. Another audio-visual database, targeting speech recognition systems in a car environment, is the AVICAR (Lee, et al., 2004). This English database included 100 speakers (50 males and 50 females), 60% of which are native speakers of American English while others have Latin American, European, East Asian and South Asian backgrounds. It includes isolated digits, isolated words, phone numbers and sentences recorded with eight microphones and four cameras, at different driving speeds.

Regarding the EP language, two databases were found: the LPFAV2, an audio-visual database specially designed for an application where the user has a specific motor impairment, and a corpora designed for a SSI that, in addiction to collect audio-visual data, also acquires information from other sensors. The LPFAV2 (Pera, et al., 2004) consists of a single speaker database with 652 sentences from a 68 words vocabulary. It includes numbers from zero to billions ranges, mathematic operators, commands and connectors. The database for the SSI was collected recently (Freitas, et al., 2014) using, among other sensors, the Kinect (for the audio-visual capture of isolated words). The data acquisition included recordings of 9 sessions from 8 native EP speakers (2 female and 6 male) and the vocabulary consisted of 32 words, including 10 digits from 0 to 9.

Figure 6 - AIBO ERS-210

With the enhancement of speech recognition systems, they began to be included in several projects like the Mars Polar Lander. This robotic spacecraft lander, launch by NASA on January 3, 1999 to study the soil and climate of a region near the south pole of Mars, included a microphone similar to those in hearing aids and a microprocessor chip used in speech recognition devices (Baalke & Goodall, 1997). In the same year, Sony released the first robotic pet, the AIBO ERS-110. One year later, the second generation of the AIBO, the ERS-210 (see Figure 6) already included speech recognition capabilities, being able to learn and recognize up to 50 words/instructions (Sony Corporation, 2000). In October 2000, Honda introduced what is now considered the most advanced humanoid robot (see Figure 7). ASIMO, an acronym for Advanced Step in Innovative Mobility, was designed to be a multi-functional mobile assistant and is capable to run, walk on uneven slopes and surfaces, turn smoothly, climb stairs, and reach for and grasp objects. It can also comprehend and respond to simple voice commands, and recognize the face of a select group of individuals. Using its camera eyes, ASIMO can

map its environment and register stationary objects, avoid moving obstacles as it moves through it (Honda Motor Corporation, 2014).

Figure 7 - Asimo humanoid robot

The previous examples show some current applications of ASR and AVSR systems. Most of the focus on constructing speech recognition systems based on visual information has been with respect to AVSR, augmenting an ASR with visual features to improve its robustness to acoustic noise. Pure VSR systems have not been effectively employed in practical applications due to the difficulties in tracking lips robustly, obtaining stable features, and implementing word classifiers. These difficulties result from the inherent dynamic nature of lip shapes and texture, these characteristics changing in accordance with lip motion, illumination changes, pose movement, and alterations in view camera settings. There are also considerable inter-person variations in lip structure. Furthermore, it is also difficult to obtain the appropriate lip reading features in a real-time manner (Shin, et al., 2011). Another major limitation of VSR systems is that not all of the speech articulators are visible, and so the information available is somewhat limited. This means that the differences in articulation for many sounds are where they cannot be seen, and so, many sounds appear visually similar on the lips. For example, the place of articulation for the phonemes /b/, /m/ and /p/ is at the lips (these are bilabial stops and require a closure of the lips). The differences between these sounds are in the voicing and the nasality. Similarly /f/ and /v/ are labiodental fricatives and require the lower lip to come into contact with upper teeth. Again, one of the main differences between these sounds is that /f/ is voiceless whilst /v/ is voiced, a difference that generally cannot be seen at the lips (Newman, et al., 2010). But VSR systems have many potential applications. Speech recognition systems based only on visual features are essential in situations where someone cannot express itself loud and clear due to pathological reasons, where the use of an acoustic stream is precluded and in any situation where a human lip reader would be needed (e.g. security issues or noisy environments). Considering this, recognition based on visual features has received a great deal of attention since the 1990s for its potential use in applications such as HCI, speaker recognition, sign language recognition (SLR) and video surveillance.

2.4. Related Work in VSR

One of the first studies in the VSR field consisted of a word-based system with a lip modeling approach for the recognition task (Rao & Mersereau, 1994). This system achieved an 85% accuracy using the distance between the corners of the lips and the distance between the top of the upper lip and the middle of the lower lip for the recognition task, but only two words were considered.

Saenko et al. (Saenko, et al., 2004) adapted the concept of AF to the visual domain, aiming to improve recognition rates in the presence of image noise. The difference between their method of classifying AF and the conventional method of classifying visemes can be illustrated by the following example. Let's consider the phoneme /m/ in the words, "romantic" and "academic". The Figure 8 shows the closure during the production of /m/ in each word. Both examples belong to a single viseme class (the bilabial viseme) and have the same open/closed feature value (fully closed). However, their appearance is different: in the right image, the distance between the corners of the lips is wider. This suggests the presence of contextual information. In fact, the preceding /ow/ in the word "romantic" causes the /m/ to be rounded, whereas the preceding /eh/ in the word "academic" does not. Therefore, modeling lip rounding and lip opening as two separate AF would recover more information than just modeling the /m/ viseme. From their results, Saenko et al. concluded that, while the viseme classifier's performance degrades with increasing noise levels, the AF-based classifier retains a significantly higher recognition rate.

Figure 8 - Full bilabial closure during the production of the words "romantic" (left) and "academic" (right) (Saenko, et al., 2004)

Werda et al. (Werda, et al., 2007) developed an Automatic Lip Feature Extraction prototype (ALiFE) to automatically localize lip feature points in a speaker face and to carry out a spatial-temporal tracking of these points. Their approach consisted in detecting a lip contour, using the active contour technique known as Snakes, and then use it to identify 4 points of interest (top center of the upper lip, bottom center of the lower lip and corners). Using these points they extract 3 features - the distance between the corners points (width), the distance between the top and bottom points (height) and the area consisting of the inside of the mouth. Then, they classified the vowels (in French) from 42 native speakers, achieving an accuracy of 72.73%.

Since its release for the Microsoft Windows platform in 2012, the Kinect have been used for the creation of corpora (Galatas, et al., 2011; Freitas, et al., 2014) and in several AVSR studies (Galatas, et al., 2011; Galatas, et al., 2013). Not many studies used it for a VSR system development but McKay et al. (McKay, et al., 2013) forecast that the Kinect has the ability to act as an ASR and that, using the RGB camera and the depth information, words can potentially be reconstructed through the observation of lip movement without the presence of sound. Using the 18 points of the lips tracked by the Kinect, Yargiç & Doğan (Yargıç & Doğan, 2013) created a VSR system with a Turkish vocabulary of 15 words (consisting of colors). For the classification task they extracted the angles between all the tracked points of the lips (see Figure 9) obtaining an accuracy rate of 78.22%.

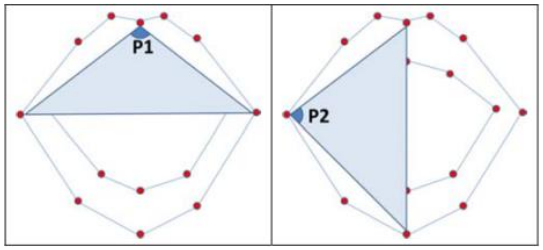


Figure 9 – Example of two extracted angles - p1 and p2 based on the tracked lip points from the Kinect SDK (Yargıç & Doğan, 2013)

The EP language was also used in several speech recognition researches. Pera et al. (Pera, et al., 2004) create an audio-visual dataset from a 68 words vocabulary, that was posteriorly used by Moura (Moura, 2005) in its AVSR system. Moura et al. (Moura, et al., 2005) used this database to create a lip-reading system to be integrated in an ASR system, achieving a WER of 36.88% for the development set and 38.44% for the validation set. They also claimed to have reduced the WER of the ASR in 20% after merging the systems. Freitas et al. (Freitas, et al., 2014) used the Kinect and two other sensors to collect a multimodal corpora for Silent Speech Interaction with a vocabulary of 32 words that included RGB, depth, surface electromyography, ultrasonic doppler sensing and audio data. Considering the variablity of the collected data, this dataset can also be used in ASR, AVSR or VSR systems.

Regarding the classifiers used for the development of speech recognition systems, the Hidden Markov Model (HMM) stands out as the main choice (Moura, et al., 2005; Saenko, et al., 2005; Hilder, et al., 2009; Lan, et al., 2012). Other classifiers found in literature were Support Vector Machine (SVM) (Yau, et al., 2007; Shaikh, et al., 2010), K-Nearest Neighbours (Yargıç & Doğan, 2013) and Fuzzy Set Theory (Baldwin, et al., 1999). Despite the common use of HMM on the speech recognition field, Yau et al. achieved an higher accuracy using SVM, compared with a previous research where they used HMM (Yau, et al., 2007).

2.5. Speech Recognition Challenges

The development of a speech recognition system is not a trivial process, requiring knowledge and expertise from a wide range of disciplines. Therefore it is especially important to have a good understanding of the fundamentals of speech recognition, so that a range of techniques can be applied to solve the problem presented in this dissertation. This section introduces the different types of speech recognition systems, as well as the common issues and possible approaches in their development. VSR will be discussed specifically, addressing some fundamental aspects like the selection of the ROI or the features extraction approaches.

2.5.1. General Speech Recognition Issues

Endow machines with the ability to understand human instructions is an achievement that offers many applications in HCI systems. Therefore, the speech recognition field has been subject of an increasingly number of researches, for the last decades. For a better understanding of its evolution, it is necessary to know first the different existing types of speech recognition systems as well as the issues considered in their development.

- **Modalities**

Considering the variability of data used in the recognition process, a system can be unimodal or multimodal. While a unimodal system is based on just one data type (e.g. audio stream), a multimodal approach includes two or more types on its data collection (e.g. audio + images), merging the extracted features for the classification process or classifying them separately and employing some fusion strategy of both classifications. Multimodal systems normally achieve higher accuracies, especially in noisy environments (Potamianos, et al., 2003; Galatas, et al., 2012) but they can be quite complex depending on the number of modalities included and the features extraction processes. This premise must be taken into account, especially in real time recognition systems.

The most common modalities used in speech recognition are the audio and video streams but, since the release of the Kinect sensor, more and more researches started to include depth information (Galatas, et al., 2012; Galatas, et al., 2013; McKay, et al., 2013).

- **TYPES**

Historically, the first recognition systems were based exclusively on the audio stream. This approach has remained ever since as the standard in the speech recognition field, and the systems based on it known as ASR system. The performance of speech recognition systems is usually specified in terms of accuracy and speed. Accuracy may be measured in terms of performance accuracy which can be rated with Word Error Rate (WER), whereas speed is measured with the real time factor. After years of research and development, the accuracy of ASR remains one of the important research challenges. Although read speech and similar types of speech can be recognized with accuracy higher than 95%, using state-of-the-art speech recognition technology, recognition accuracy drastically decreases for spontaneous speech (Anusuya & Katti, 2009) and in noise environments (Potamianos, et al., 2003).

Both human speech production and perception are bimodal in nature. The visual modality benefit the speech intelligibility, mainly due to three factors: it helps to locate the audio source, it contains speech segmental information that supplements the audio, and it provides complementary information about the place of articulation (due to the partial or full visibility of articulators such as the tongue, teeth and lips). The above facts have motivated significant interest in automatic recognition of visual speech. Work in this field aims at improving ASR by exploiting the visual modality of the speaker mouth region, in addition to the traditional audio modality (Potamianos, et al., 2003), leading to Audio-Visual Automatic Speech Recognition (AV-ASR) systems or AVSR.

In the traditional AVSR paradigm, visual speech information is extracted from planar video of the speaker face and combined with audio in a two-stream classifier fusion framework (Galatas, et al., 2012). Studies showed considerable speech recognition improvements, reaching up to a 27% relative reduction in WER, in noise environments (Potamianos, et al., 2003). However, there are situations where the use of the audio stream is not reliable, since we know in advance that it may be corrupted. In these cases, all the useful speech information will be in the video. That's why some researchers started to focus on visual-only speech recognition as a standalone problem. Systems based on this approach are known as VSR systems and can be applied in pathological contexts of the speech production system or in environments where the noise cannot be controlled (Saenko, et al., 2004; Moura, et al., 2005; Song, et al., 2012).

Despite the benefits of visual information in noise environments, the accuracy of VSR systems is still far from ASR accuracy. Furthermore, the accuracies obtained keep oscillating, which indicates that there is still no visible convergence in this specific field. For example, in isolated word

task, accuracy ranges between 42% and 66.9%. VSR problems represent difficult tasks by nature, as the visual stream provides relevant but little information about speech. Common VSR problems are normally related to the large variations in the way that people speak, errors in face detection or lips localization, visual appearance differences between individuals, light conditions or video quality (Chitu & Rothkrantz, 2012).

Another type of systems allow recognition to take place in the absence of an acoustic signal, which is also adequate for users with some speech impairments or in noise environments. These systems can be based on different sensory types of data, ranging from simple video input to Ultrasonic Doppler Sensing, and are known as Silent Speech Interfaces (SSI). VSR systems can be included in Silent Speech Interfaces since they do not rely on an audio stream.

▪ Speaking Modes

Speech recognition systems can be separated in several different classes based on the type of utterance or speaking mode it is able to recognize. These classes are classified as follows (Rabiner & Juang, 2000; Anusuya & Katti, 2009):

- Isolated word – Systems based on isolated words accept only a single word or a single utterance at a time, drawn from a specified vocabulary. These systems have an "On/Off" states, requiring that the speaker waits between utterances;
- Connected word – This class of systems allows the user to speak fluent speech consisting entirely of words from a specified vocabulary (e.g. telephone numbers). This speaking mode is similar to the isolated word, but allows separate utterances to be 'run-together' with a minimal pause between them;
- Continuous speech – In this mode the user can speak naturally, from a large (often unlimited) vocabulary, while the computer determines the content (e.g. computer dictation). Systems based on continuous speech are difficult to create because they require special methods to determine utterance boundaries.

On the VSR domain, several researches focus on the visemes instead of the whole word recognition.

- **VOCABULARY SIZE**

The vocabulary size is another important issue in a speech recognition system. A larger vocabulary normally means a more difficult recognition task partly because of the greater variability, but also due to the increase number of words that are confusable with one another. However, even small vocabularies can contain extremely confusable words (Gold, et al., 2011). This issue is found in both the acoustic and the visual approaches. Vocabulary's size categories are divided as follow (Rabiner & Juang, 2000):

› Small vocabulary - Up to 100 words;
› Medium vocabulary - From 100 to 1000 words;
› Large vocabulary - Over 1000 words.

- **SPEAKER DEPENDENCE**

Considering the knowledge of the user speech patterns, a recognition system is classified according one of the following classes (Rabiner & Juang, 2000):

› Speaker dependent - Systems which have been custom tailored to each individual talker;
› Speaker independent - Systems which work on broad populations of speakers, most of which the system has never encountered or adapted to;
› Speaker adaptive - Systems able to customize their knowledge to each individual user over time while the system is in use.

2.5.2. VSR ISSUES

The development of a VSR system is notoriously difficult as speech involves more than just the visible articulators. Other face elements were shown to be important during a face to face communication. However, their exact influence is not completely elucidated. During experiments, in which a gaze tracker was used to track the speaker areas of attention during communication, it was found that the human lip-readers focus on four major areas: the mouth, the eyes and the center of the face, depending on the task and the noise level. In normal situations the listener scans the mouth and the other areas relatively equal periods of times. However, when the back-ground noise increases, the center of the face becomes the central point of attention (Chitu &

Rothkrantz, 2012). There is also ambiguity in the VSR process as many sounds are visually indistinguishable. In addition, the visible articulators are free to adopt the position of upcoming sounds if there is no immediate requirement on their position for the current sound (e.g. early lip-rounding during /s/ in anticipation of the rounded vowel /u/ for the word "soon") — a phenomenon known as coarticulation. This means that for the same underlying sound, the visible speech articulators can be in different positions and so the visual features can be very different (Lan, et al., 2012). Therefore, in addition to the General Speech Recognition Issues presented in the previous section, these (and others) specific aspects must be considered in the development of a VSR.

- **CAMERA**

The camera is one of the key issues in VSR. With the proliferation of digital cameras, they became increasingly affordable, smaller and with higher resolutions. The resolution selected for the data collection process is important since it will define the detail of each image. The higher the selected resolution the higher the level of detail/information available, but also higher the processing. For this reason, the selection of the resolution must consider the required post-processing, especially if the aim is to create a multimodal system. The frame rate is another important specification, defining the number of frames per second (FPS) captured by the camera. A higher frame rate allows to capture more images and therefore more dynamic information. It is especially important when we consider the movement of the lips during the speech, since some visemes are only visible during a fraction of a second. If the frame rate is too low, these visemes can be lost.

Traditional digital cameras are commonly used in VSR, but they only allow to capture RGB images, which only gives information about the color of each pixel, and depends heavily on the available illumination. Another useful information that can be used in VSR is the depth information. Depth sensors provide spatial information of each pixel as three-dimensional coordinates. Until recently, these sensors had the downside of being expensive. Besides, to use a digital camera and a depth sensor presents another problem to the researchers: synchronization. Using distinct devices, it becomes difficult to ensure the time alignment between each stream. The release of the Kinect for Windows in 2012 introduced a device that, in addition to a microphone, included a RGB camera and a depth sensor at an affordable price. Furthermore, the synchronization between each stream is ensured by the Kinect SDK. Since its release, the Kinect for Windows has been used for

several AVSR (Galatas, et al., 2012), VSR (McKay, et al., 2013; Yargıç & Doğan, 2013) and even SSI (Freitas, et al., 2014) researches.

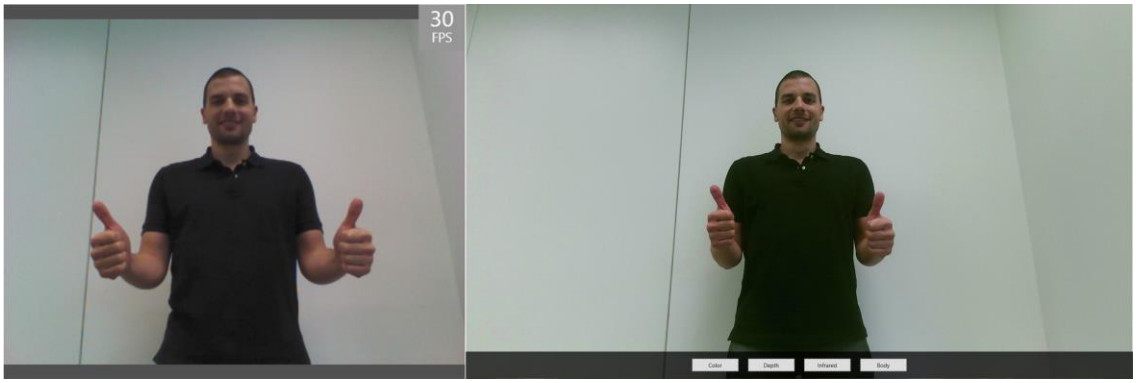


Figure 10 - RGB images from the first version of Kinect (left) and the Kinect One (right)

Recently, Microsoft released the second version of Kinect – the Kinect One. This new device has a RGB camera with a Full HD resolution (see Figure 10) and uses a depth sensor with TOF technology, more precise than the previous one that was based on Structured Light technology[1]. It also includes an IR camera endowing this device with the capacity to see in the dark (i.e. to capture images that do not depend on the light conditions).

- ## Scenario / Light Conditions

VSR systems depend on visual information, therefore the scenario and the light conditions are two key issues to be considered for the data collection process. Considering that every captured image are usually processed, it is important to ensure good conditions during the recordings to prevent additional post-processing load and to reduce segmentation errors. Direct light must be used carefully and should be placed centrally to prevent light peaks or shadows in the speakers face.

Ideally, the captured images should contain only the information required for recognition task. If the scenario contains many objects with different colors, the head segmentation process will be harder. A simple way of facilitating this process it to use a monochromatic background which contrast with the speaker (e.g. a blank wall).

[1] The differences between these technologies were presented in section 1.1.

- **Speaker / Camera Position**

The position of the camera and speaker depends on the aims, but VSR tasks normally require the frontal view of the face, even though some studies considered other positions (Estellers & Thiran, 2012). Therefore, the camera is normally placed right in front of the speaker, while the latter should look frontally to it, to keep the symmetry of the face on the captured images. The camera should also be close to the speaker to prevent the loss of details. When using a depth camera, the distance must be kept in the recommended range to ensure the precision of the depth values.

- **ROI**

In VSR, the ROI is the selected part of the image that is processed to extract the features to be used for the classification process. It always includes the lips but some authors also consider adjacent regions, like the chin or the cheeks (Potamianos & Neti, 2001; Potamianos, et al., 2003).

- **Approach**

VSR systems can follow two distinct approaches depending how the features are extracted from the ROI: Appearance-based and Shape-based.

In the appearance-based approach, all the pixel of the ROI are used as features. The problem of finding the relevant information in the image is left for the classifier to solve. Although it may seem that locating the ROI does not need to be done with very good precision in this case, it has been proven that the normalization of the ROI improves recognition rates. This normalization implies that the mouth is centered, horizontally aligned and scaled. This process turns the recognition invariant to small movements of the speaker and to the distance between the speaker and the camera. The normalization is done by tracking, at least, the two corners of the mouth. Usually the dimensionality of the obtained feature-vector is too large to allow statistical modeling and must be reduced. The most popular dimensionality reduction methods are the image transforms like the Principal Components Analysis and the Discrete Cosine or Wavelet Transforms. These methods are inspired from image compression, reducing the size of the images by eliminating redundancy while keeping the perceived quality constant. However, there is no guarantee that they are appropriate for dimensionality reduction in classification, as they might not preserve the relevant information for speech recognition (Potamianos & Neti, 2001; Gurban, 2009).

The shape-based assumes that the speech information lies in the contours of the lips. However, the extraction of the contours proves to be a difficult task. They can be defined as parametric curves (e.g. B-Spline, Bézier, ellipses, snakes or active contours). They are guided by minimizing a cost function over the area enclosed by the curves. From these contours, a number of high-level features can be extracted: height, width, perimeter of the contour, or the size of the enclosed area. They contain a significant amount of speech information, but the contours themselves are not always correctly estimated. This happens especially in the cases when the color of the lips is close to that of the surrounding skin, or when the lighting is not uniform and shadows appear. Facial hair can also pose problems. A way of improving the quality of the contours is by using models, like lip templates or Active Shape Models (ASMs). However, to obtain such a model, a relatively large number of points has to be marked by hand on the images in the training phase. Lip contours alone may lack the necessary discriminative power, so they are often combined with appearance on hybrid-approaches. For example, it was shown that the addition of appearance to shape significantly improves the lip reading performance of the ASMs, resulting in AAM, which combine shape and appearance parameters into a single feature vector (Saenko, et al., 2004; Gurban, 2009). However, despite the effectiveness of the AAM for representing and modeling non-rigid structures such as the face, it is not appropriate for robustly tracking rapidly changing lips in a real time manner (Shin et al. 2011).

2.6. SUMMARY

Speech recognition commonly refers to the ability to convert human speech to a textual form. As presented in this chapter, humans performs this task without difficulties, failing no more than once or twice in 1000 words. In fact, 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 nonlinguistic factors in the speech signal. For over 60 years, researchers endeavored to endow machines with the same ability, in order to make them perform specific tasks by speaking simple commands or sequences of words. Although the speech recognition researches of the past few decades have resulted in great advances, much more remains to be done to achieve the goal of devices that equal or exceed human performance.

The first ASR research was performed in 1952 yielding to an increasing number of similar researches in the following years. In the late 1970s, the discovery of the McGurk effect proved that human speech perception rely on both acoustic and visual cues to decode speech, giving rise to the AVSR systems. These systems were able to increase the recognition robustness, especially in noisy environments. The study of visual cues on the following 15 years allowed to explore new applications on the speech recognition field, introducing pure VSR systems in the 1990s. Unlike ASR, VSR systems can be used in situations where someone cannot express itself loud and clear due to pathological reasons, where the use of an acoustic stream is precluded and in any situation where a human lip reader would be needed (e.g. security issues or noisy environments).

For the development of a speech recognition system, several issues must be considered. The type of feature(s) used defines the modality (unimodal if only one is used or multimodal otherwise) and the type of the system (e.g. ASR are based on acoustic information, while VSR are based on visual cues). The speaking mode (isolated words, connected words or continuous speech) and the selected vocabulary (by its size and type of words) influence the complexity of the system. Moreover, a speech recognition system can be developed to be used by one person or trained to recognize multiple speakers. Regarding VSR systems, specific factors must also be considered. The selection of the camera must consider the type of features (traditional digital cameras only offer RGB input but the Kinect, for example, allows to capture depth and IR information) and frame rate required for the system. The light conditions and the position of the camera/speaker must be verified if a data collection process is needed. Finally, the selection of the ROI and the approach (appearance or shape-based) define how the collected data will be processed.

3. ViKi - Visual Speech Recognition for Kinect

In this chapter we will present our created system, referred to as ViKi (named after **Vi**sual Speech Recognition for **Ki**nect). The first section presents the system requirements considered for its development. An overview of the system is presented in the second section, introducing its GUI and main features. The methodology used for its development is also discussed in the remaining sections. We describe the data collection process and characteristics of the collected corpora, followed by all the necessary processing for the lip segmentation and features extraction. Finally, we present the word classification system.

3.1. System Requirements

In order to achieve the objectives defined in the section 1.4, we need to set the system requirements of our ViKi application accordingly. Table 2 reminds our thesis objectives, aligned with the stated hypothesis. The system requirements are divided into 5 classes: Data Collection & Processing (DCPR), Training (TR), Classification (CR) and GUI (GR) requirements. These requirements are presented on Table 3.

Table 2 - Thesis objectives aligned with the defined hypothesis.

Objective ID	Thesis Objective Description	Thesis Hypothesis
O0	Development of a multimodal VSR system for the EP language	All
O1	Compare a multimodal approach (using geometric and articulatory features), with a classical unimodal approach, in a word recognition task.	H1
O2	Compare the results of our articulatory approach using RGB and depth data with the same approach using only RGB data, in a word recognition task.	H2
O3	Assess if our articulatory approach (using opening, rounding and the protrusion of the lips as features), outperforms our shape-based approach (using the width, height and depth variation of the lips), for the same word recognition task.	H3

Table 3 - System requirements definition

Requirement ID	Requirement Description	Objective ID	Status
DCPR1	Collect words pronounced by a series of users, using RGB imaging and depth data	O0	Completed
DCPR2	Compute geometric features from RGB imaging and depth data	O1, O2, O3	Completed
DCPR3	Compute articulatory features from geometric features	O1, O2, O3	Completed
TR1	Use a previously collected dataset to train or to test the classification sub-system	O1, O2, O3	Completed
CR1	Develop a classification technique appropriate for the word recognition task	O0	Completed
CR2	Draw plots from the extracted features, allowing to track their progress along a word recognition task	O0	Completed

CR3	Plot the classification results from each approach of the word recognition task	00	Completed
GR1	Display each visual stream (RGB, depth and IR) available	-	Completed
GR2	Start, stop and save a recording	-	Completed
GR3	Define the speaker, the session and the word to be saved with a correct structure	-	Completed
GR4	Load a previously saved recording	-	Completed
GR5	Present the extracted features from the processed frames	-	Completed

3.2. System Overview and GUI

ViKi follows the architecture of traditional VSR systems (Saenko, et al., 2004; Galatas, et al., 2012) and uses a recent sensor (Kinect One) for the data collection process, presenting three main functionalities: data collection (recording), training and recognition. The recording feature can be used to create a dataset or simply to record a word to be recognized in real-time. The speaker can choose any media (RGB, Depth or Infrared - IR) to visualize the recording. When the recording

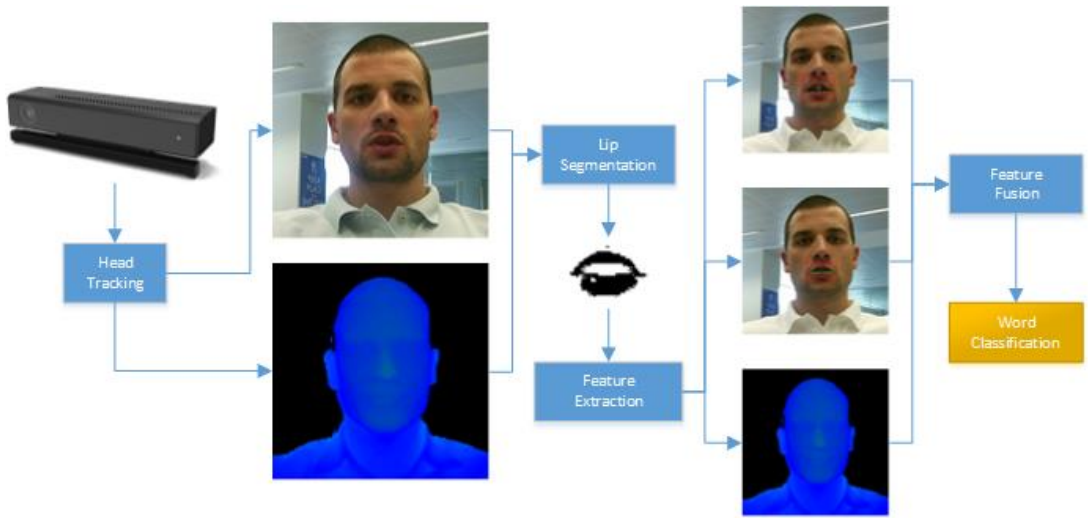


Figure 11 – Logical workflow overview of the Viki system, showing the multimodal input HCI approach. The collected streams are first processed to identify the lips. After the lip segmentation, the height, width and depth variations of the lips are extracted. All these features are then normalized, merged (fused) into a feature vector and sent to the classifier, for word recognition.

stops, the system processes the collected data (already in memory) and initiates the lip segmentation and both shape and articulatory features extraction processes. All the extracted features are then normalized and merged into the feature vector, to be classified. After the recognition, the recorded data can be saved (see Figure 11).

If the word was previously recorded, it must be loaded into the system first to enable its recognition. When an appropriate dataset exists, the system can be trained. The training functionality processes all the words from the training dataset and extracts the information required to create a SVM classifier.

Regarding its GUI, ViKi consists essentially of two different windows. The main window (see Figure 12) can be used to record and save a word, to load previously recorded words or to train the system using a previously collected training dataset. While the word recording requires the Kinect One to be connected, the other tasks can be achieved without its presence.

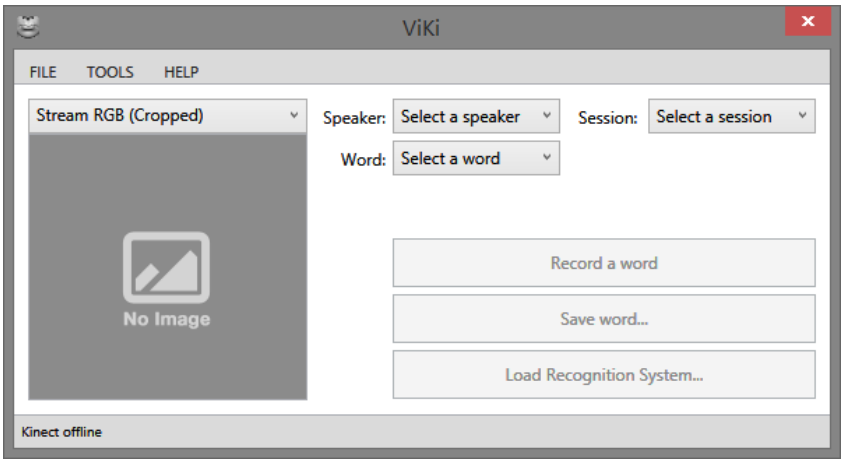


Figure 12 - Main Window of ViKi

To record a word, it is recommended to wait until the speaker face is tracked by ViKi. When the face is tracked, it will be visible on the image area of the main window. The RGB stream is selected by default, showing the RGB image already cropped around the speaker face. The Depth or IR input streams can also be selected. It is advised to check the depth stream before starting to record, to ensure that the speaker face is on the Kinect depth sensor range. To start the recording, the "Record a word" button must be pressed. The same button can be pressed to stop the recording. As soon as it is pressed, a countdown timer also starts, stopping the recording after 6 seconds, if it's not stopped manually (see Figure 13). This mechanism is necessary since all the collected data is saved on memory to be processed after the recording.

Figure 13 – ViKi while recording a word

After a recording is stopped, the collected data is immediately processed and the classification window is showed automatically (see Figure 14). This window includes the words recognized from the different classifiers, as well as information about the number of collected frames, recording length and several plots depicting the extracted features. If this window is shown on behalf of a previously recorded word, since the name of the word is known (has been previously loaded into memory), the "Selected word:" field of the "World info" window, will show details of that word and the classifications from the SVM will be presented in green or red, depending if the classification was correct or not, respectively. The plots are divided in two sections to present both geometric and AF variations for each time frame, along the pronunciation of word. While the plots on the right

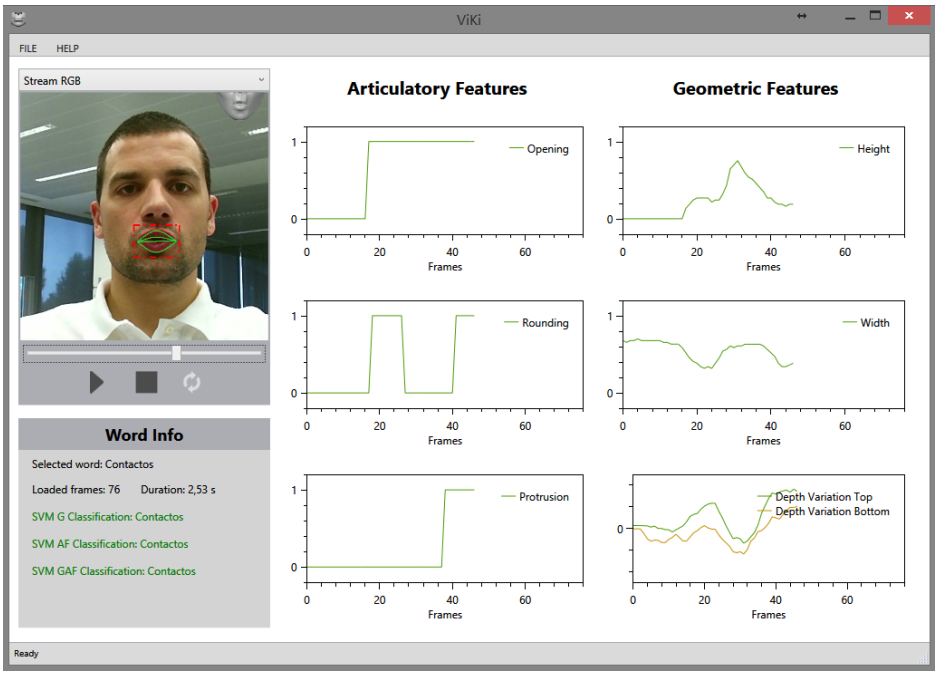


Figure 14 - Classification window from ViKi, showing the classification results from the 3 SVM and the plots of the extracted features.

include continuous values (between 0 and 1) corresponding to the height, width and depth variation of the lips, on the left side, the values of the plots are discrete (0 or 1), indicating the presence or absence of the respective AF (opening, rounding or protrusion).

In addition to the RGB frames, the processed RGB and YCbCr frames can also be visualized allowing to see the images used to the lip segmentation. Furthermore, the ROI and the extracted lip points of the present frame are depicted in the RGB frames. The internal and external contours of the lips are represented using the tracked points, by using the *DrawCurve* method from the *Graphics* library of C#. The recording can be paused and the slider of the player allows to move slower or quicker through the word frames.

To save a recorded word, the classification window must be closed. On the main window, the "Save word..." and "Load Recognition System..." buttons will be available while the recorded word is in memory (i.e. until it is saved, a new word is recorded or the application is closed). To be able to save a word, the Speaker, Session and Word must be selected. This is needed so the created folders and files have the right structure to be later loaded by the application. Speakers and Words can be easily added to ViKi on the "Tools" menu. Previously recorded words can be loaded to the application by clicking on the "Load word..." button from the "File" menu (see Figure 15). After selecting the folder with the recorded data, the classification window will open.

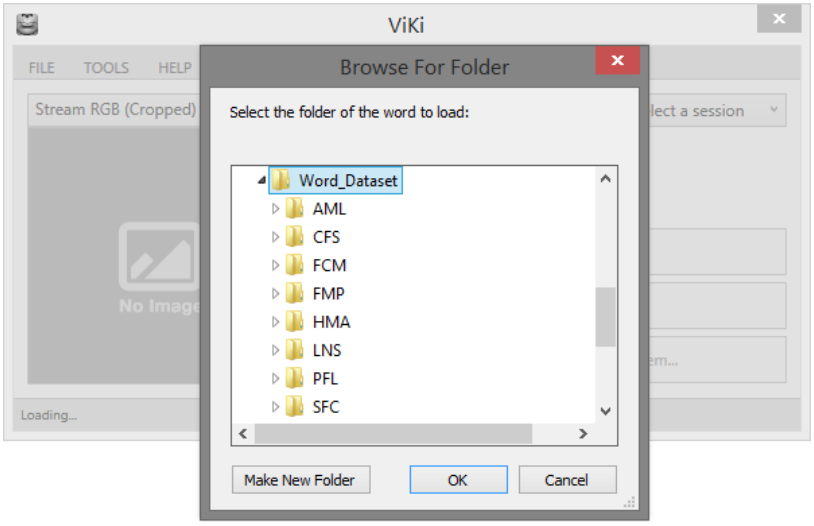


Figure 15 - Browse window to load a word to ViKi

The "Tools" menu also includes the "Train System" submenu. Here, the 3 training options can be found, allowing the creation of a classifier based on geometric features, AF or a fusion of both features. Once a training option is selected, the user must select the dataset folder from the browse window. Depending on the size of the dataset, this process can take minutes to hours.

3.3. DATA COLLECTION

For the data collection process, the Microsoft Kinect One was used to collect three visual modalities: RGB, depth and IR information. Despite the fact that only the RGB and depth data are used in the VSR system development of this thesis, the IR data was also collected to allow this dataset to be use in future VSR studies. The developed data collection tool was based on a preview version of the Kinect SDK. The RGB images were captured with a resolution of 1920 x 1080 pixel, while the depth and IR images had 512x424 pixel. Every stream was captured at 30 FPS and the synchronization of all media was ensure by the Kinect SDK.

The recordings took place mostly in a room with good natural and artificial lighting. However, some sessions were recorded at the end of the afternoon, needing an extra light pointing to the speaker face. Each speaker was briefed about the recording process and to voluntarily sign a consent form which accurately described the experiment and its duration, and what kind of data was going to be collected (see Appendix A)[1]. The Kinect was placed at the height of the neck, between 60 and 100 cm from the speakers. Participants were asked to keep the face turned to the Kinect during the recording and to start and end each word with the lips closed in a rest position. The background scenario was not a key issue, since the SDK allows to keep track of the head of the speaker. Despite the ability of the Kinect to track multiple skeletons simultaneously, our data collection tool was programmed to keep only the information of the closest body. Furthermore, using the tracking feature, the collected images were cropped to include only the head region of the RGB, depth and IR images before saving them.

Figure 16 - Images collected from the RGB (left), depth (center) and IR (right) streams

Eight native EP speakers participated in the recordings - 7 males and 1 female – with an age range from 23 to 32 years old and an average age of 27 years. Each recording session took

[1] A detailed description of the data collections and associated studies/experiments were submitted, evaluated and approved by the ethics committee of the Instituto Universitário de Lisboa (ISCTE-IUL) regulated by the dispatch nº7095/2011.

between 1h to 1h30, generating an average of 15GB of data per speaker, including, for each frame, a text file with the spatial coordinates of the head joint and 3 image files of a square area around the head, respectively, for the RGB, depth and IR streams (see Figure 16). For each recorded word, a file containing the depth and IR values for each pixel of the selected area, is also created. Both the depth image and data file are filtered to include only the depth values closer that the head joint position + a tolerance of 300 mm, ensuring that the depth information of the head is preserved while the background is not.

Table 4 - Set of digits words

Digits Set				
Um (*One*)	Dois (*Two*)	Três (*Three*)	Quatro (*Four*)	Cinco (*Five*)
Seis (*Six*)	Sete (*Seven*)	Oito (*Eight*)	Nove (*Nine*)	Zero (*Zero*)

Table 5 - Set of words extracted from the Ambient Assisted Living context

Ambient Assisted Living Set				
Videos (*Videos*)	Ligar (*Call*)	Contactos (*Contacts*)	Mensagens (*Messages*)	Voltar (*Back*)
Pesquisar (*Search*)	Anterior (*Previous*)	Fotografias (*Photographs*)	Família (*Family*)	Ajuda (*Help*)
Seguinte (*Next*)	Lembretes (*Reminders*)	Calendário (*Calendar*)	E-mail (*E-mail*)	

The selected vocabulary consisted of 25 EP words, which can be divided into 2 distinct sets. The first set, widely used in speech recognition literature works, consists of 10 digits from zero to nine (see Table 4). The second set included 14 common words (see Table 5) taken from context free grammars of an Ambient Assisted Living (AAL) application that supports speech input (Teixeira, et al., 2012). A total of 225 recordings were made per speaker, including 9 repetitions of each word and silence state, totaling 1800 recordings for the 8 speakers.

3.4. PRE-PROCESSING

The data must be pre-processed before the system training and word recognition, to allow the extraction of the necessary features to be sent to the classifier. The aim of this step is to take the collected data (RGB images and depth) and find strategies to locate and keep track of the region of the lips. Since the head tracking is easily achieved using the Kinect SDK, this section will focus on the lips segmentation and tracking.

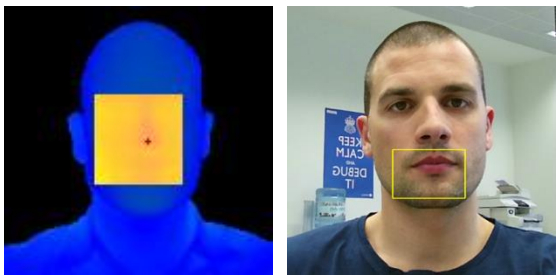


Figure 17 - Lip tracking approach with the depth frame showing the scanned region and its closest point (left) and the selected region to locate the lips based on this point location (right)

The first step of the pre-processing is to achieve the segmentation of the region of the lip. Considering that the coordinates of the head joint (center of the head) is given by the Kinect SDK and that the saved RGB images include only the head region, centered on those same coordinates, it is possible to locate the lips starting from the head joint coordinates. Knowing that the speaker has his face turned frontally to the Kinect (since he was instructed for it), the coordinates of the tip of the nose can be obtained searching by the point with the closest depth in the center of the frame. Therefore, dividing the frame in nine squares, the lower depth value of the square at the center (see Figure 17 - Left) corresponds to the tip of the nose. The lips region (ROI) is then expected to be under this point (see Figure 17 - Right) but, to be able to perform a correct segmentation, the RGB image needs to be processed.

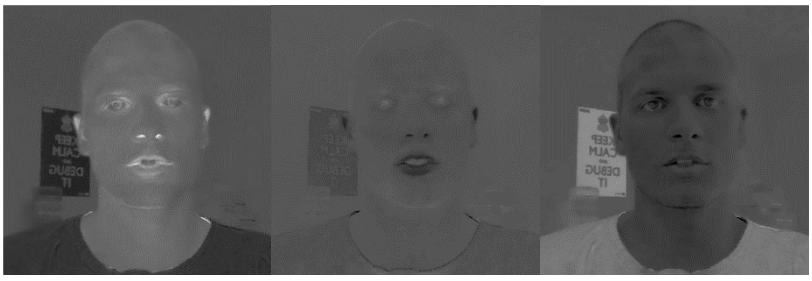


Figure 18 - Red (image on the left), Green (image in the center) and Blue channels (image on the right), from the RGB frame

For the lip segmentation, two color spaces were considered: RGB and YCbCr. From the RGB frame, the green channel was extract to find the external points (top, bottom and corners) of the lips. The Cr channel was extracted from the YCbCr frame to obtain the internal points (top and bottom). Both the green and Cr channels, showed a higher contrast in the ROI with emphasis on the lips (see Figure 18 and Figure 19).

Figure 19 - Y (image on the left), Cb (image in the center) and Cr channels (image on the right) from the YCbCr frame

The extracted channel results in a gray image to which a brightness equalization correction filter is applied. The brightness correction allows the system to be more robust for words recorded on different lighting conditions. After the brightness correction, a median filter is also applied to smooth the image and remove some existing noise, before the final binarization, as can be seen in Figure 20.

Figure 20 - Extracted Green channel from the RGB frame (image on the left), after brightness equalization (image in the center) and after median filter and binarization (image on the right)

The same image processing is applied to the YCbCr frame, including the brightness correction, median and threshold filters, resulting in the right image of Figure 21.

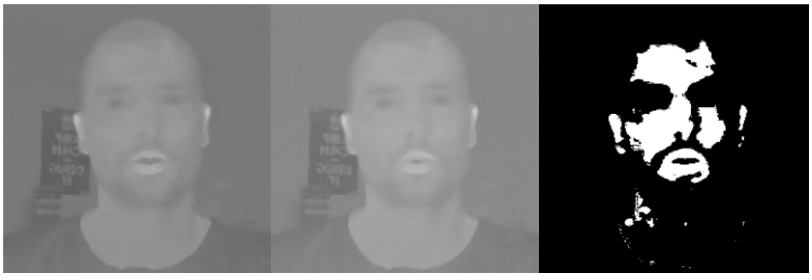


Figure 21 - Extracted Cr channel from the YCbCr frame (1st image), after brightness equalization (2nd image) and after median filter and binarization (3rd image)

As for the applied threshold filter, a customized threshold value is used for each recorded word and for each color space. After applying the previous transformations, the threshold value is readjusted as long as the number of black pixel (for the RGB frame), or white pixel (for the YCbCr frame), are outside a predefined range. This process is repeated for the first 5 frames of each word and the final threshold value, used for binarization, is defined as the average of the 5 results. With the binarized images obtained from the previous processing, the geometric features, of interest for our word recognition task, are then extracted.

3.5. Features Extraction

Two types of features are extracted in the developed system: Geometric and AF. Using a shape-based approach, the internal and external borders of the lips are located to track several key points and estimate the height, width and depth of the lips. The tracked points also allow to determine some articulatory states of the lips (opening, rounding and protrusion). Both types of features are then used for the training and word classification. After a word is recorded and the frames are pre-processed, the next step is to find the contours of the lips. This is achieved using a Convex Hull technique on the processed RGB frames (where the lips are represented as black pixel). The Convex Hull of a set of points (or pixel) can be defined as the smallest convex set containing these points or, more precisely, as the intersection of all convex sets containing the points (Berg, et al., 2000). In this case, the Convex Hull technique was applied to the set of black pixel of the ROI. The result is an array containing the coordinates of the points (see Figure 22).

Figure 22 - Set of points (in red) obtained with the Convex Hull using the processed green channel image

The lips corners points are easily extract as the ones with the lower and higher X coordinate value, but the top and bottom points require some processing. Since the aim is to obtain the top and bottom points from the center of the lips (average value between the X values of the corners points) and the obtained set may not include points with the exact average value, the exact coordinates must be searched. This is achieved by starting at the points with the lower (for the bottom)

and higher (for the top) Y coordinates, from the Convex Hull points and at the average X value between the corners and searching the edge of the lips in the neighbor Y pixel. The search algorithm used for the external top point is presented in Figure 23 (the approach for the external bottom point is similar). Initially, the top point is defined as the point with the average X value between the corners and the maximum Y coordinate from the points of the extracted set. Since we use the maximum Y coordinate value, the search for the lip contour only need to be made downward. Therefore, the contour is found when more than 3 consecutive black pixel are found, assuming the Y coordinate as the first black pixel position.

```
topPoint = pixel(average, maxY);
numberOfBlackPixel = 0;

For(y = maxY; y > maxY – 10; y--) {
        If isBlack(pixel(averageX, y))
                numberOfBlackPixel++;
        Else
                numberOfBlackPixel = 0;

        If numberOfBlackPixel > 3
                topPoint = pixel(averageX, y+4);
                Break;
}
```

Figure 23 - External top lip point search algorithm pseudocode

The same search algorithm is used to find the external bottom point. With the external points found, the next step is to locate the inner top and bottom of the lips. For this task, the YCbCr processed frames are used (where the lips are represented as white pixel). The process starts from the external bottom point and stops when the first black pixel is found (indicating the inside of the mouth) or when a margin before the top external point is reached (closed lips).

The process is repeated for the inner top point but, this time, stopping when a black pixel is found or the inner bottom point is reached. While this approach works well for the majority of the situations, it may fail when the tongue is visible during a word. Since the color of the tongue is similar to the color of the lips, both appear with the same color in the binarized images and, if the tongue approaches the lips, the internal contour is lost (see Figure 24).

Figure 24 - RGB frame (left image), processed green channel (center image) and processed Cr channel (right image) showing the effacement of the inner contour due to the visibility of the tongue during the word "Calendário"

The strategy employed to address this situation is based on the depth information. Despite the similarity of the color between the lips and the tongue, in these situations, the depth values in the tongue region are higher than the ones on the lips (the tongue is further from the Kinect than the lips). Therefore, the inner contour can be detected using this gap in the depth values between the lips and the tongue (see Figure 25). Considering that the tongue is only visible in few words and that this process requires more processing that the one based only on the color of the pixel (involving conversions between color and depth space), it is only used as a correction process when the distance between the external and internal points is higher than a given limit.

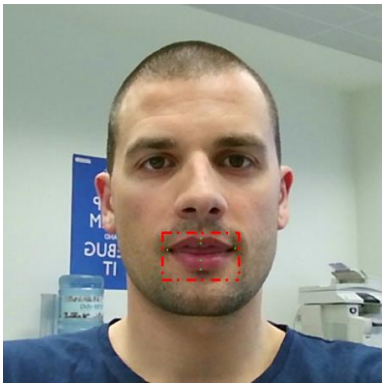


Figure 25 – Same RGB frame as the presented on Figure 24, showing: the ROI (in red) and the six tracked points (in green) after the depth correction of the internal ones.

To stabilize the flickering of the extracted contours points, each one of them assumes the coordinates value of its weighted arithmetic mean of the three last frames, using the Equation 1 (applied to both X and Y coordinates):

$$C_n = 0.5 \times C_n + 0.3 \times C_{n-1} + 0.2 \times C_{n-2} \qquad \textbf{Equation 1}$$

where C_n is the coordinate of a given point in the n^{th} frame. While removing some of the existing flickering, the higher weight of the actual frame helps to keep a good tracking agility on quick movements of the lips.

The ROI for the next frame is based on the external points captured on each frame. A margin is added around these, giving the ROI a dynamic area with adaptive capabilities to the lips position variations. The margin around the external points must be large enough to allow a quick variation of the ROI but it can't be too large so it risks including other regions that could jeopardize the Convex Hull. Furthermore, the margin bellow the bottom point should be larger than the top margin since the lower lip has a greater range of motion (and quicker movements) than the upper lip (see the region limited by the red lines on Figure 25).

After the identification of the six key points from the contours of the lips, the geometric and AF features are extracted and all the process is repeated for each one of the next frames, starting on the Convex Hull step.

3.5.1. Geometric Features

As mentioned in the beginning of this section, three geometric features are extracted from the processed frames: the width, the height and the depth of the lips. The width is obtained using the coordinates of the corners points while the height uses the internal top and bottom points (see Figure 26). The width and height distances are calculated by the Euclidean distance between the pixel where these points are located, using Equation 2:

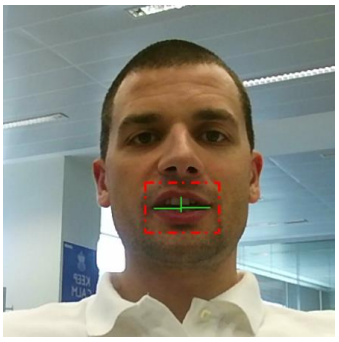


Figure 26 - Width and Height represented on a RGB frame

$$d(p,q) = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \qquad \text{Equation 2}$$

where p and q are pixel, and X and Y their coordinates.

As presented section 3.3, the speakers must be positioned between 60 cm and 1 m from the Kinect. For a given frame, measuring the lips width of a speaker at 60 cm would obviously give a greater value than at 1 m. For this reason, both measured distances must be normalized to a fixed distance (see Equation 3).

$$Nd = Cd \times Hd \div Fd \qquad \text{Equation 3}$$

where Nd is the normalized distance (width or height), Cd is the current distance (width or height before normalization), Hd is the depth value of the head joint on the captured frame (in meters) and Fd is the depth value for which every frame are normalized (in meters). For this system, all width and height distances were normalized to a 0.6 m depth.

Figure 27 - Neutral lip position (left image), lip protrusion (center image) and lip backing (right image).

The depth of the lips is assessed by their position on the sagittal plane[1] (see Figure 27) using the closest depth value of the pixel between the internal and external points, of the top and bottom lips. But considering that the distance of the speaker from the Kinect can vary during the movement of the lips, the depth value of the pixel between those points cannot be used directly. Therefore, the depth of the lips (Ld on Equation 4 and the yellow line on Figure 28) must be obtained by the difference between the head joint distance (Hd on Equation 4 and the green line on Figure 28) and the lips distance from the Kinect (Lk on Equation 4 and the red line on Figure 28).

$$Ld = Hd - Lk \qquad \textbf{Equation 4}$$

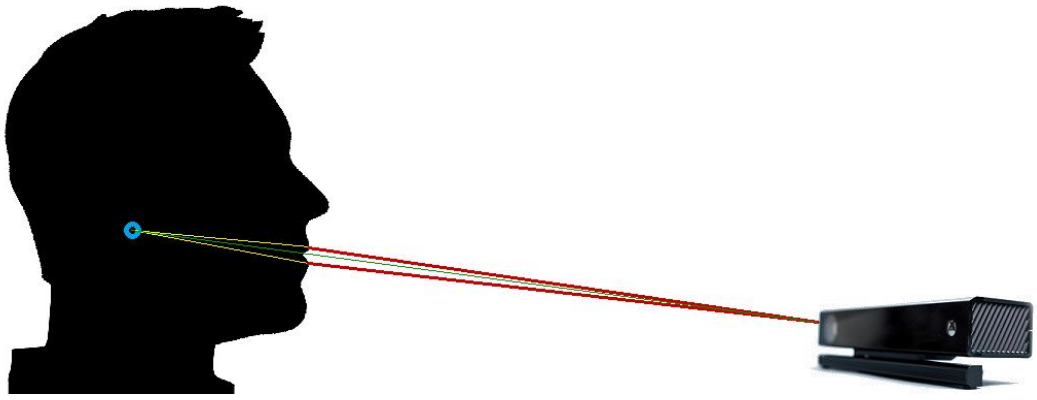


Figure 28 - Scheme showing how the depth of the lips (yellow lines) is obtained using the distance from the Kinect to the head joint (green line) and to the lips (red lines)

Since each speaker is asked to start and to finish each recording with the lips closed in a neutral position, the distance from the lips to the head joint on the first frame (Ld_0) can be used as a reference value to get the lips depth variation in each one of the following frames. Therefore, the depth variation of the lips on the first frame (ΔLd_0) is 0, while on the following frames (ΔLd_n)

[1] Vertical plane which divides the body into right and left halves

it is equal to the difference between the depth value on the present frame (Ld_n) and on the first frame (Ld_0) as expressed on Equation 5.

$$\Delta Ld_n = Ld_n - Ld_0 \qquad \text{Equation 5}$$

Finally, the extracted width, height and depth values are normalized to a range between 0 and 1 and used to create the feature vector for the word classification. Figure 29 shows one of the plots presenting the geometric features of a recorded word on the ViKi output. This plot presents the normalized values of the Width distance for each frame of a recording of the word “Contactos”. We can observe two decreases of this distance, corresponding to the “-on” and “-os” parts of the word.

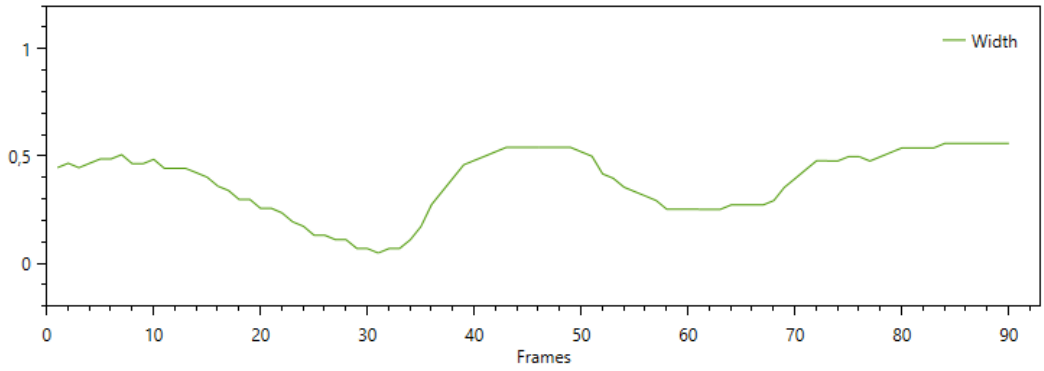


Figure 29 - Plot showing the extracted Width feature from a recording of the word "Contactos"

3.5.2. Articulatory Features

Using the tracked six points of the lips, 3 AFs were extracted: opening, rounding and protrusion. As the name suggests, the opening is present when the mouth opens and the rounding when the corners of the lips approach. The protrusion occurs when the lips are pushed forward (see center image from Figure 27). While the opening and rounding features are defined based on the height and width, respectively, the protrusion depends on the depth values of the lips.

For a given word and for each one of the AFs, an array is created with a size equal to the number of frames. Each position of the arrays will assume the value 1 or 0, depending on whether the corresponding feature is present or not on the respective frame. A feature is considered present when the associated value exceeds a predefined threshold.

Considering that the opening of the mouth during the speech is not related with its size, a fixed threshold value can be used to assess its presence. The first choice for the opening threshold value should naturally be 0.0 since any greater distance between the top and bottom points of the internal contour indicates an opening but, considering that the tracked points are the result of image processing, their position may not be always accurate and a simple tracking deviation would automatically trigger the opening feature. Thus, a greater threshold value was consider and any frame in which the normalized height value is greater or equal than 0.1 is tagged with an opening AF.

For the protrusion case, since the array with the depth variation of the lips included negative values and was normalized to a range between 0 and 1, the resting position should have geometric values around 0.5. Therefore, frames with a normalized depth variation greater than 0.65 were tagged with a protrusion AF.

The threshold for the rounding (Tr) features is harder to select since, unlike the opening feature, it manifests differently from person to person and there is not a specific value from which we can consider that they are present. Therefore, instead of a fix value for all the speakers, a dynamic threshold is defined based on the geometric features of the lips. The average normalized width (Nw) of the first three frames is estimated, setting the feature as present when the normalized width on a frame is less than 75% of this value (see Equation 9). Considering the rounding feature definition, the selected threshold must always be lower than the estimated average value.

$$Tr = 0.75 \times \frac{\sum_{n=0}^{2} Nw_n}{3} \qquad \text{Equation 6}$$

Figure 30 shows one of the plots presenting AF of a recorded word on the ViKi output. This plot presents the values of the rounding feature (1 when it's present and 0 when it's not) for each

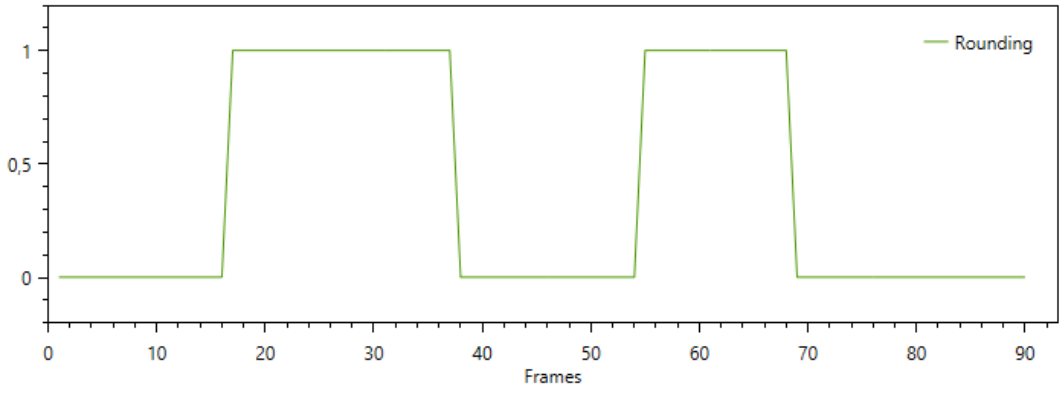


Figure 30 - Plot showing the extracted Rounding feature from a recording of the word "Contactos".

frame of a recording of the word “Contactos”. We can observe that the rounding feature is visible 2 times along the word, corresponding to the lower Width values presented on Figure 29.

Like the Geometric Features, the arrays of AFs created in this step are then used for the feature vector and word classification.

3.6. CLASSIFICATION

Machine learning-based classifiers require specific data structures known as feature vectors. These structures consist of a compact representation of the given input and are obtained through the feature extraction process. For the development of the proposed VSR system, geometric and AF features were extracted and presented in the previous section. The next step is to create the feature vector based on the previously extracted features and use this vector to train the classification system. The Accord.net framework (Souza, 2013) implementations were used for the training, validation and classification processes of this thesis. The developed system uses the SVM - Support Vector Machine technique for the classification process but, for the system to be able to classify a word, it must be previously trained. The collected dataset is used for this task. To assess how the classification results will generalize to an independent dataset, a k-fold Cross-Validation technique is used, that splits the dataset in training and testing sets.

- **FEATURE VECTOR**

For a given word, each one of the selected features is extracted to a single array with the same size as the number of captured frames, in which the n^{th} position corresponds to the n^{th} frame. Since the recorded words rarely have the same number of frames, they also have different sizes of arrays which, in turn, will generate feature vectors with different sizes for each word of the vocabulary. However, classifiers normally require that all feature vectors should have the same size. Therefore, a normalization of the arrays of features is usually required. Furthermore, in speech recognition research, some of the collected

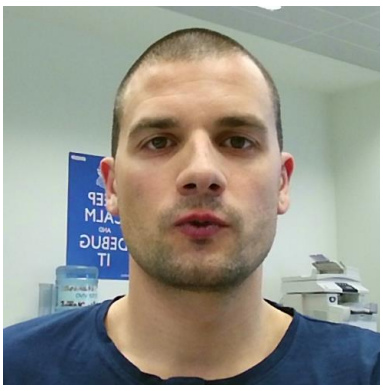


Figure 31 - Frame from the EP word "Fotografia" showing the rounding feature appearing before the opening

data is not useful for the classification and should be removed (e.g. the frames corresponding to the silent states from the beginning and the end of a recorded word, for example). In VSR, the removal of these frames must be processed carefully, since the words are not delimited by the presence of sound but, instead, by visual features like the lip opening. However, important features for the VSR, other than the lip opening, could be present on these frames (e.g. in EP when pronouncing words like "Quatro" or "Fotografia", the speaker can start rounding his lips before opening them, like presented on Figure 31). Considering this, the method applied to remove those unnecessary frames was to search for the first and the last frame with the opening or rounding feature and remove the previous and the following ones from every array of features.

After removing the frames without relevant information for the classification process (silences), the next step was to normalize the size of each array of features. Several normalization sizes were tested and are described in Chapter 4. One of two situations can occur, regarding this normalization. If the original size is greater than the normalized size, then some positions must be removed and the average value of the removed positions is used. The first step to determine the positions to be removed is to calculate the size factor (sf), dividing the original size (os) by the normalization size (ns), as presented on Equation 7.

$$sf = \frac{os}{ns} \qquad \text{Equation 7}$$

The correspondence of the normalized position (np) with the original position (op) is then set by the integer division of the original position by the size factor (sf) (see Equation 9):

$$np = \frac{op}{sf} \qquad \text{Equation 8}$$

If we consider, for example, an original array size of 10 positions and a normalization size of 5, the first position of the normalized array - na_0 - will correspond to the first and the second positions of the original array – oa_0 and oa_1 (since the result from the integer division of 0 and 1 by 2 is 0) . Therefore, the average value (na_0) of these two positions will be used (see Equation 9).

$$na_{np} = \frac{\sum_{i=m}^{n} oa_i}{n - m + 1} \qquad \text{Equation 9}$$

where na_{np} is position of the normalized array obtained from Equation 8, oa_i is the i^{th} position of the original array, and m and n are the first and the last positions of the original array corresponding to np.

For the AF arrays, since they only contain 0 and 1 values, the average of the remove positions must be rounded. When the size of the feature array is smaller than the normalized size, the inserted positions assume the value of the neighboring positions. The inserted or removed positions are defined by the relation between the original and normalized sizes.

After the normalization of the features arrays, the feature vector can be created, joining the previous arrays. Considering that the system includes three classifiers, three feature vectors can be created: one including only geometric features, another including only AF, and the last including both geometric and AF features. Since the considered normalized size of each feature array was 60 (see section 4.1 for a more detailed explanation of the rationale of the selection of this number), the final feature vector for the geometric (width, height, depth) or articulatory (opening, rounding and protrusion) classification has 180 positions, while the one including geometric and AF has 360 positions (width, height, depth variation, opening, rounding and protrusion). Apart from the feature vector, the corresponding label and the word list (vocabulary) are also required for the classifier training.

- **System Validation**

To be able to classify recorded words, the system must first be trained to create a classifier. This training is accomplished using 3 SVMs on the previously collected dataset (one based only on geometric features, another based on AF and a third one using both features). For each word on the dataset, a feature vector is created and sent for training, with its corresponding label. Along with the training, a k-fold cross-validation process is also realized. In k-fold cross-validation, the original sample is randomly partitioned into k equal size subsamples. A k = 9 value was used to ensure a correct division of the dataset, since it includes 72 repetitions of each word (9 repetitions for each of the 8 speakers). Of the 9 subsamples, 1 subsample is retained as the validation (or test) dataset for testing the model, and the remaining 8 subsamples are used as training data. The cross-validation process is then repeated 9 times (the folds), with each of the 9 subsamples used exactly once as validation data. The 9 results from the folds are then averaged to produce a single estimation. The advantage of this method is that all observations are used for both training and validation, and each observation is used for validation exactly once.

Normally, SVM are binary classifiers and, as such, they can only decide between two classes at once. However, several approaches have been suggested to perform multi-class classification

using SVMs. The used framework follows one of these approaches, dividing the multi-class problem into a set of binary problems. Redundant options can be discarded, such as comparing one class with itself (i.e. comparing A vs. A) and one of the two pairs of the same comparison (i.e. comparing A vs. B and B vs. A, in which B vs. A can be ignored). Discarding the redundant options, a typical decision problem can be decomposed in the subset of binary problems obtained from Equation 10:

$$S = \frac{n \times (n-1)}{2} \qquad \text{Equation 10}$$

where S is the number of subsets and n is the number of classes. Therefore, the number of necessary SVM for this approach is S. To decide for a class, a voting scheme is used. The class which receives the most decisions wins the decision process. This approach is known as a one-against-one strategy for multi-class classification. In this thesis, 24 classes were considered (for the 24 words of the selected vocabulary).

Table 6 - Kernels tested for the SVM

Kernels tested			
Linear	Gaussian	Laplacian	Sigmoid
Spline	Log	Histogram Intersection	Quadratic
Multiquadric	Inverse Multiquadric	Power(2)	Power(3)
Polynomial(2)	Polynomial(3)		

A set of 14 different kernels were tested for the SVM (see Table 6), creating the respective classifiers and saving the one that obtained the best accuracy results to be used by the system. The tolerance value used on the sequential minimal optimization was set to 0.01. To be able to identify misrecognition patterns and to assess the accuracy of the classifier for each word of the vocabulary, a confusion matrix was also created, including the training and validation accuracies of each fold.

Although it could be possible to achieve higher accuracies with a better exploration of the tested kernels parameters, considering the available time for realization of this thesis, and that its

aim was not to explore the machine learning field in detail, each kernel was used with the predefined values of the framework implementations.

3.7. SUMMARY

This chapter introduced the methodology used for the development of the created VSR system – ViKi. The data collection process and all necessary data processing for the features extraction were presented. The system classification was also discussed.

ViKi has a simple GUI, allowing recorded words to be immediately recognized or saved. After a word is recorded, the system shows the output window, presenting information about the extracted features. If a dataset is collected, ViKi can use it to train the system and create classifiers based on geometric features, AF, and both features.

The data collection process included RGB, depth and IR images and data arrays, captured respectively, from the RGB, depth and IR streams. The Kinect SDK was used to track the speaker head and crop the collected images. The head joint coordinates is also used to locate the ROI for the first time. After locating the ROI, the lip segmentation is achieved through computer vision techniques, manipulating the RGB frame. Six lip points are identified (corners, external top, external bottom, internal top and internal bottom), allowing the extraction of geometric (height, width and depth) and AF (opening, rounding and protrusion) features. After the normalization of the extracted features, the feature vector is created and sent for classification.

For the word classification task, 14 kernels were tested on the 3 developed SVMs: one based on geometric features, another based on AF and a last one based on both features. For the validation of the system, a k-fold cross-validation technique was employed.

4. Results and Evaluation

This chapter presents the results with several classification approaches for the word recognition task and discuss the cross-validation process. The first section introduces the global evaluation of the system, using the collected dataset, as well as the tests performed to assess the influence of the depth data in the work recognition accuracy results. The following sections present such results, obtained on a speaker-dependent context and for each one of the vocabulary sets.

4.1. Global Analysis, Classification and Validation

As referred in section 3.6, the SVM classifier requires that all the feature vectors to have the same size. A size of 60 positions was used to normalize all arrays of features but several sizes were tested to assess its influence on the accuracy results. The normalization sizes that were tested can be related with the number of frames on a word recording. Considering that a word length ranges approximately between 1 and 2 seconds, non-normalized feature arrays have sizes of 30 to 60 frames (after removing the initial and final silent frames). Therefore, a normalization size ranging between these values, should be used to reduce the bias of the modifications induced by the normalization process, on the data values. Feature arrays of 30, 45 and 60 positions were tested and no significant differences were found but, since the size of 60 obtained a 1% higher accuracy (using the kernel with better results), this one was selected.

As stated previously, ViKi relies on 3 SVMs to classify words based on geometric features (shape-based approach), AF (articulatory approach) and the fusion both features. These classifiers were trained and used on a cross-validation process to assess their capacity of generalization to

an independent data set. The results for each fold of the 3 approaches are presented in Table 7 and in Appendix B. Despite the higher results achieved from the training sets for all the SVMs, they cannot be generalized to independent data sets, since the data used to estimate the accuracy of the system is simultaneously used for its training. However, they can be used to detect some problems. A good example of that are the results presented on Table 7, obtained from the cross-validation process of our approach using a fusion of geometric and AF and features.

Table 7 - Cross-validation results for approaches using geometric and AF features

Fold	Training (%)	Validation (%)
1	99	50
2	99	43
3	99	50
4	99	53
5	99	47
6	99	54
7	99	52
8	99	51
9	99	45
Average (Standard Deviation)	**99.0 (0.0)**	**49.44 (3.71)**

The training sets always achieved high accuracies values, indicating the possibility of over-fitting. Typically, a model is trained by maximizing its performance on the training data. However, its accuracy is determined by its ability to perform well on unseen data, rather than on the training data. Overfitting occurs when the training data is memorized instead of being used to learn and generalize from a trend.

Regarding the validation sets, the average accuracies are presented on Table 8. The table is divided in 4 columns showing the results of the 3 kernels which obtained the best results, for both training and validation sets. The higher results were obtained using the *Inverse Multiquadric* kernel, achieving a 68% average accuracy on the 9 folds, using the shape-based approach. The results obtained with the SVM based on the articulatory approach and from the one based on the combined geometric and AF were considerably lower, achieving precisely half of the geometric features classifier accuracy (34%) for the first and 49% accuracy for the latter. Nevertheless, all the achieved

results are still encouraging since, considering the selected vocabulary, the average accuracy by guessing is around 4%. As referred on section 3.6, predefined values were used for each tested kernel. Therefore, the *Inverse Multiquadric* kernel constant value and the σ value of the *Lapacian* kernel were 1. For the *Gaussian* kernel, an estimation method from the framework was use to define the value of σ, based on the dataset.

Table 8 - Average word accuracies of the training and validation sets from each SVM using all dataset

	Word Accuracy Average (%)					
	Geometric Features		Articulatory Features		Geometric + Articulatory Features	
Set / Kernel	Training	Validation	Training	Validation	Training	Validation
Inverse Multiquadric	97	**68**	71	29	99	**49**
Laplacian	**99**	66	**72**	25	**100**	46
Gaussian	68	57	47	**34**	63	46

The presented results outperformed the ones achieved by Galatas et al. (Galatas, et al., 2012). Using the first version of the Microsoft Kinect, this author achieve a 44.39% of recognition accuracy using an appearance-based approach. However, since they used a dataset consisting of 4500 connected digits from 15 speakers, a direct comparison with our 8 speaker 24 word vocabulary scenario, is hard to make. Yargiç & Doğan (Yargıç & Doğan, 2013) also used the first version of the Kinect to create a VSR system using a dataset with 5 repetitions of a vocabulary of 15 Turkish words and 10 speakers. They followed a shape-based approach, extracting all the angles between the 18 tracked lip points of the Kinect SDK. Their system achieve a 72.44% using the best four angles and 78% accuracy using all the 23 angles, but the recognition process required a significant computation time.

The presented results suggest that the AF have a lower discriminative power than geometric features or even than the fusion of both types of features. Considering that each one of the 3 AF used on the classification have only 2 possible states to represent a frame (0 or 1), the total number

of possible states is 8 which, comparing with geometric features, that can assume any real value between 0 and 1, is significantly lower.

The confusion matrix of the SVM classifier based on geometric features allows us to see the classification errors resulting from the cross-validation process (see Figure 32).

	Vídeos	Ligar	Contactos	Mensagens	Voltar	Pesquisar	Anterior	Fotografias	Família	Ajuda	Seguinte	Lembretes	Calendário	Email	Zero	Um	Dois	Três	Quatro	Cinco	Seis	Sete	Oito	Nove	Word Accuracy
Vídeos	56	0	2	1	0	0	1	0	0	2	1	1	0	0	1	0	0	1	0	5	0	1	0	0	**0,78**
Ligar	0	56	0	1	1	1	1	0	3	0	2	0	0	0	1	0	0	4	0	0	2	0	0	0	**0,78**
Contactos	1	0	59	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	1	2	3	2	**0,82**
Mensagens	1	4	0	43	1	3	1	0	1	0	6	0	0	0	0	0	1	2	0	1	4	4	0	0	**0,6**
Voltar	0	0	0	0	53	5	0	4	3	0	1	1	0	1	0	0	1	0	1	0	1	1	0	0	**0,74**
Pesquisar	0	3	0	1	3	42	0	4	3	0	7	0	1	0	0	0	0	1	0	2	1	3	1	0	**0,58**
Anterior	0	0	0	2	0	1	57	0	0	2	1	3	4	0	0	0	0	1	0	0	1	0	0	0	**0,79**
Fotografias	0	0	4	0	0	0	0	61	1	0	2	1	0	0	0	0	0	0	1	0	0	0	0	2	**0,85**
Família	0	1	0	0	0	0	0	0	58	0	2	2	0	3	2	0	0	0	1	1	0	1	0	1	**0,81**
Ajuda	1	0	0	0	0	1	4	0	0	63	1	0	0	0	0	0	0	0	0	0	0	0	0	2	**0,88**
Seguinte	0	3	0	1	1	3	2	0	0	0	39	1	3	1	0	0	0	5	3	2	4	4	0	0	**0,54**
Lembretes	1	0	1	0	0	0	1	0	2	0	2	48	1	11	0	0	0	0	2	0	0	3	0	0	**0,67**
Calendário	0	3	0	0	1	0	1	0	0	0	8	0	43	2	6	0	0	3	3	0	1	0	0	1	**0,6**
Email	0	0	1	0	0	0	0	1	6	0	3	11	0	43	0	1	0	3	0	1	0	0	0	2	**0,6**
Zero	0	0	0	5	0	1	0	0	1	0	2	2	2	0	43	0	0	2	1	9	2	2	0	0	**0,6**
Um	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	66	0	0	0	2	0	0	1	0	**0,93**
Dois	0	0	1	0	2	0	0	0	1	0	1	0	0	1	0	0	51	1	4	1	1	0	5	3	**0,71**
Três	2	7	0	1	0	0	1	0	0	0	4	0	2	1	1	0	1	41	0	0	8	1	0	2	**0,57**
Quatro	0	0	0	0	2	0	0	0	0	1	3	0	0	0	0	0	4	0	51	0	0	1	8	2	**0,71**
Cinco	4	0	0	0	0	0	0	0	1	0	4	0	0	1	6	0	0	1	2	47	2	2	0	2	**0,65**
Seis	0	1	3	4	0	1	1	0	2	0	5	0	3	0	5	0	0	12	2	2	28	3	0	0	**0,39**
Sete	1	1	0	2	0	3	0	1	2	0	10	2	0	0	4	0	1	5	0	3	3	31	2	1	**0,43**
Oito	0	0	1	0	0	1	0	1	0	0	1	0	0	0	0	2	6	0	10	0	0	0	45	5	**0,63**
Nove	2	0	0	0	0	0	2	0	3	0	5	0	1	1	2	2	2	1	3	1	1	1	3	42	**0,58**

Figure 32 - Confusion matrix obtained from the validation test based on geometric features

The two classification errors with higher prevalence occurred between the words "Três" (tɾˈeʃ)[1], "Seis" (sˈɐjʃ), "Lembretes" (lɐ̃.bɾˈe.tɨʃ)1 and "Email" (ˈɛ.mˈɐj.ɫ)1. The word "Três" was incorrectly classified as "Seis" 12 times, while the opposite occurred 7 times. The words "Lembretes" and "Email" were incorrectly misclassified 22 times (11 times for each one). This classification errors happen when some words have similar features. For instance, if we look at the plots from Figure 33, a quick closure of the lips around the 21st position can be seen on both "Email" (top plot) and "Lembretes" words (bottom plot). This variation is due to the viseme required to pronounce the "-m" and the "-b" on each word. The depth variation of the lips is also similar on

[1] Phonetic transcriptions of the vocabulary words, from the *Instituto de Linguística Teórica e Computacional* (http://www.portaldalinguaportuguesa.org/)

these words, justifying why they are often confused. The three words with higher recognition rates using only geometric features were “Um” (93%), “Ajuda” (88%) and “Fotografias” (85%).

As presented in Table 8, the classification based on AF obtained worst results (34% of accuracy), with 2 words that were never recognized correctly (“Três” and “Seis”). Furthermore, the word “Ligar” obtained a high number of false positive recognitions (see Appendix C). This can be explained by the fact that this word has normally only one AF present – the lip opening. Considering that several other words from the selected vocabulary have a similar behavior, the SVM classified all the recordings with only the lip opening feature as “Ligar”. Other words including other features like lip rounding (e.g. “Um”, “Oito” or “Fotografias”) were never classified as “Ligar”. The words with higher recognitions rates using these features were “Um” (89%), “Ligar” (63%) and “Quatro” (63%). Despite the lower accuracy of the SVM based on AF, an interesting fact is that the word “Um” was also the word with the higher recognition rate and its accuracy was not far from the one obtained by the SVM based on geometric features.

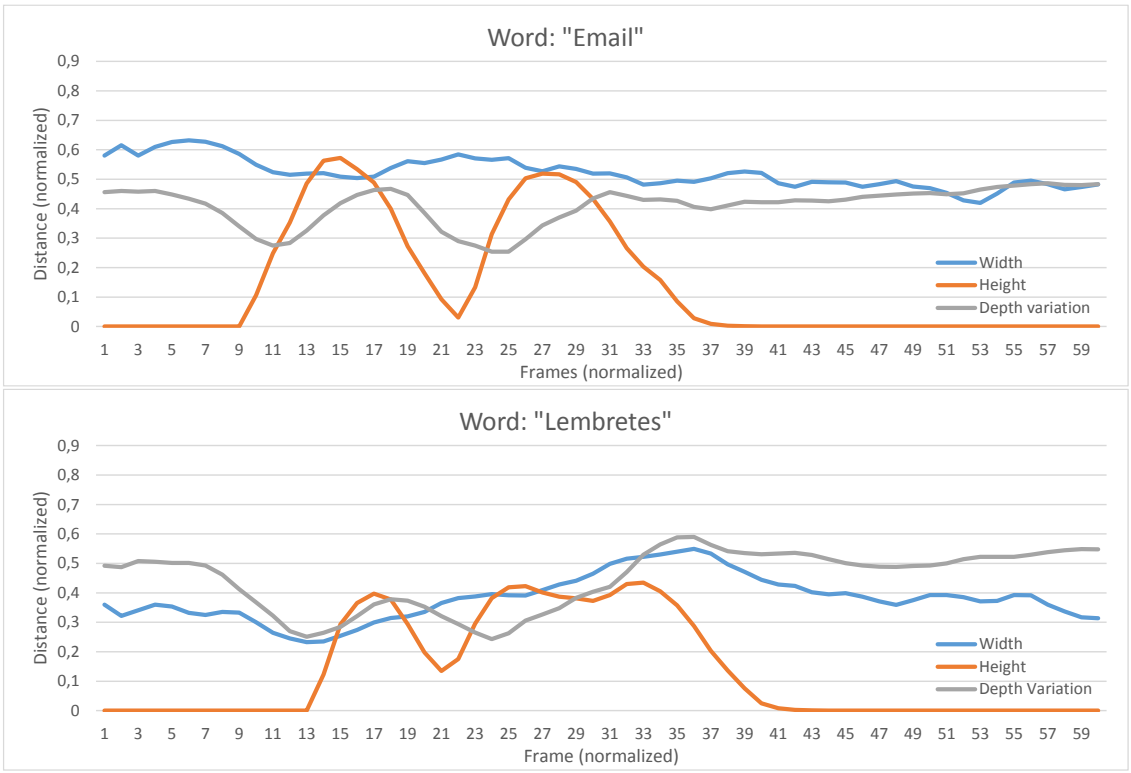


Figure 33 - Plots based on (normalized) geometric features from the words "Email" (top) and "Lembretes" (bottom)

Considering that the SVM based on AF obtained a lower accuracy than the one based on geometric features, it was expected that the results of using both features should be around the average of both. The words with higher recognition rates were “Um”, “Quatro” and “Fotografias”.

While the words “Fotografias” and “Quatro” were some of the most correctly recognized from the SVM based on geometric features and AF, respectively, the word “Um” was the word with the highest accuracy on the three SVMs (93%, 89% and 90%). “Um” is the smallest word of the selected vocabulary (with only one viseme) and it is the only word that can include the 3 extracted AF in all instants of the speech, which can be the reason for this high recognition rates. However, “Um” seems to be an exception on the digits set, since the majority of the other digits obtained accuracies below the average value. Furthermore, the two words with the worst accuracy results on each SVM also belong to the digits set, which can be related with the fact that these words are smaller than the words from the other set and, consequently, have less variability.

One of the hypothesis proposed by this thesis was that the depth information benefits the word classification. To test this, ViKi classifiers were trained with and without depth data. The results obtained from the cross-validation processes are presented on Table 9.

Table 9 - Word accuracies average obtained from each SVM with and without the use of the depth data

	Word Accuracy Average (%)		
Using Depth Data	**Geometric Features**	**Articulatory Features**	**Geometric + Articulatory Features**
Yes	**68**	**34**	49
No	66	26	**54**

The results show a decrease of the accuracy of the SVM for the (geometric) shape-based approach (-2%) but especially of the SVM from the articulatory approach (-8%), when the depth data is not used for classification task. Surprisingly, the accuracy of the SVM based on both features presented an increase of the accuracy (+5%) when removing the depth information. However, these results still suggest that depth information can benefit both shape-based and articulatory approaches. Galatas et al. (Galatas, et al., 2012) and Galatas et al. (Galatas, et al., 2012) also had similar findings on AVSR systems using the first version of Microsoft Kinect and an appearance-based approach.

4.2. Analysis of Speaker Dependency

ViKi was also trained with each one of the speakers separately, with the complete vocabulary, to assess its performance on speaker-dependent scenarios and the variability of results with different speakers.

Table 10 – Word Accuracies average of the training and validation sets from each SVM and each speaker

	Word Accuracy Average (%)		
Speaker	**Geometric Features**	**Articulatory Features**	**Geometric + Articulatory Features**
AML	80	35	71
CFS	52	27	39
FCM	59	31	46
FMP	77	53	71
HMA	80	**56**	71
LNS	**87**	30	**75**
PFL	65	22	58
SFC	69	46	52
Average	**69,85**	**37,86**	**58,86**
Standard Deviation	**11,39**	**12,55**	**12,89**

Table 10 presents the results obtained from the cross-validation processes for each one of the speakers. While the average value from each speaker is not far from the ones obtained from the SVMs based on all the speakers (see Table 8), ViKi accuracy can vary significantly depending on the speaker (more than 30% of recognition rates differences are found between them). These differences prove the variability between speakers, and the hard task of creating a recognition system based exclusively on visual features. In fact, the lower results obtained by some of the speakers may not only be related with the reduced articulation of some of them, but also to the variations between the recordings of the same word.

Figure 34 presents two frames from two different recordings of the same word. Both show the minimum width of lips that the speaker achieved during these recordings. Although the word is the same, the range of movement was clearly different, presenting a rounder appearance on the right frame, with what seems to be a higher protrusion too. Despite all the normalization processes, these variations between recordings of the same speaker, associated with a different speed of speech, can reduce the recognition accuracy.

Figure 34 - Intra-speaker variability between two recordings of the word "Fotografias". The two frames present the differences between the minimum width achieved during the recordings.

Regarding the results obtained by the different types of features used for the classification, the SVM based on geometric features still achieved the best results on every speaker, while the SVM that used the AF obtained the lower accuracies. Observing the confusion matrix of the speaker with the best results (LNS with the SVM based on geometric features) on Appendix E, we can see that the results obtained from the digits set was still below the results of the set of words of the AAL scenario, matching what happened with the results obtained using all the speakers.

Regarding related work, Hilder et al. (Hilder, et al., 2009) also tested a shape-based approach on a speaker-dependent VSR. They obtained an average accuracy of 74.29%, however their dataset consisted only of one repetition of the English letters A to F and 5 speakers.

4.3. Analysis of the Individual Sets

Since the results obtained using the complete vocabulary showed some differences between the sets of words (the set of digits always included the worst two accuracies results with every

SVM), ViKi was trained with each of these sets, separately and for each approach, to assess this phenomena. The obtained results are presented on Table 11.

Despite the higher number of words in the AAL set, the results show that it obtained higher accuracies with the SVM using geometric features and the SVM using both geometric and AF. Its accuracy was lower than the one obtained by the digits set with the articulatory approach, but the difference was only 2% lower. These results confirm that the accuracy of a VSR system is influenced either by the number and the type of words included in its vocabulary.

Table 11 - Accuracies average of each word set with each SVM

	Accuracy Average (%)		
Word Set	**Geometric Features**	**Articulatory Features**	**Geometric + Articulatory Features**
Digits	69	**43**	55
AAL	**76**	41	**58**

Some of the EP words for the digits from 0 to 9 have the same or similar viseme, being frequently confused. If we take a look at the Figure 32 and Appendix D, we can observe that the word "Oito" is often confused with the word "Quatro", as well as the word "Seis" with the word "Três", contributing for the lower results of this set when compared with the AAL set.

Using an appearance-based approach, Potamianos at al. (Potamianos, et al., 2003) also recognized a 10-digit word dataset with 10 speakers and a total of 200 recordings, obtaining a 65% accuracy. Lee & Park (Lee & Park, 2008) obtained approximated results (63.9%) with a Korean isolated digits dataset (from 0 to 9), 56 speakers and three repetitions of each word. Comparing the results from the digits set with the previous ones, we verify that ViKi compares well with the literature and even achieved a higher word accuracy.

4.4. SUMMARY

In this chapter, the accuracy results obtained from the various classification approaches, in a word recognition task (using our visual speech recognition techniques), pronounced by a set of 8 speakers, were analyzed and discussed. Several aspects were evaluated, namely the global word accuracy of ViKi for each one of the SVMs, the influence of the depth data on the accuracy, the accuracy on speaker-independent and speaker-dependent scenarios and the accuracy for each set of words from the vocabulary.

Regarding the global word recognition accuracy, the best average result achieved by ViKi was 68%, based on geometric features. The approaches based on AF and both features performed worse, reaching 34% and 49% of word recognition accuracy, respectively. These results proved that an articulatory approach doesn't outperform a shape-based approach, contrary to what happens with appearance approaches. Despite the differences between the accuracies achieved by each approach, these are encouraging results since the classification by guessing, considering the selected vocabulary, is approximately 4%. The used of the combined RGB and depth data also proved to benefit the classifiers accuracy, for the shape-based (geometric) and articulatory approaches, compared with the use of unimodal appearance approach based only on RGB.

When ViKi was tested individually for each speaker, a wide range of results were obtained, reaching more than 30% of accuracy differences between speakers, on the SVM based on geometric features and the one based on AF features. These results are important to show the intra- and inter-speaker variability. In fact, it is not uncommon for speakers to pronounce the same word differently on consecutive recordings, leading to a larger dispersion of the extracted features values and, consequently, more confusion on the word classification process.

Considering that the words with higher classification errors belong to the digits set, an independent evaluation of both sets was performed, using the same methodology. Despite the smaller number of words on the digits set, its accuracy was lower than the obtained from the AAL set. However, the achieved accuracy of the digits set is still higher that some of the obtained on similar works reported in the literature.

5. Conclusions and Future Work

In this thesis, VSR has been explored to create acoustic-independent systems, or simply lip-reading systems, since the beginning of the 90's. For more than 15 years, studies on VSR were based on traditional video cameras and RGB frames processing. However, the release of the Microsoft Kinect for Windows which included, in addition of a standard RGB camera, a depth sensor, brought new possibilities to this field. Furthermore, its SDK allowed to track dozens of facial points, helping to locate the lips. Recently, Microsoft released the second version of the Kinect (Kinect One for Windows), which include, among others, a Full HD camera and a depth sensor with a more accurate technology (TOF). To explore the contribution of the depth information we set out to develop a VSR system using exclusively the new Kinect.

5.1. Conclusions

Using the collected RGB and depth information from the Kinect One, we have created our VSR system – ViKi – that achieve a 68% and 34% recognition accuracy with our shape-based approach and articulatory approach, respectively. Our first hypothesis (**H1**) suggested that the skeleton tracking associated with the RGB and depth information, could improve the word recognition accuracy, when compared to unimodal approaches based on RGB data only. As it was shown on the section 3.4, the skeleton tracking made possible to quickly locate the speaker head and the region of the lips. Furthermore, this Kinect SDK ability, ensured the correct tracking of the head, even on challenging scenarios, helping in reducing segmentation errors. To assess the depth data influence in the word recognition, we tested our approaches with and without depth data. On both

the shape-based approach and the articulatory approach, better word recognition results were obtained when the depth information was included (+2% and +8%, respectively). These results showed its relevancy for the field of VSR and proved our second hypothesis (**H2**), which affirms that articulatory VSR approaches based on RGB and depth data outperform articulatory approaches based solely only on RGB imaging.

Our last hypothesis (**H3**) stated that articulatory approaches outperform shape-based approaches on VSR systems. While it has been demonstrated in the literature that an articulatory approach is superior to a viseme VSR technique (Saenko, et al., 2004), for word and phrase recognition (Saenko, et al., 2009), this was not verified in our VSR system. Saenko et al. compared the word and phrase recognition results, using AF and visemes for the classification and an appearance based approach to identify them previously. They concluded that AF are more discriminative than the visemes, since one frame can only have one classification in terms of visemes but it can have several ones in terms of AF (as much as the extracted features). However, this was not verified in this thesis, proving that it may not be true for shape-based approaches. While geometric features, like width and height, can assume continuous values, articulatory approaches values are discrete to identify specific features (e.g. opening or rounding). If we consider the opening feature, for example, the articulatory approach is not capable to distinguish all the opening variations of the lips, even if several levels are considered. For example, the opening articulatory feature doesn't distinguish between a 0.5 cm opening and 1.5 cm opening, only classifying the lips as "opened". However, the shape-based approach uses the exact values extracted from the frames, considering all the variations for the classification.

Regarding the obtained results, ViKi outperformed some of the results of previous studies. Using our entire dataset, we were able to surpass the recognition accuracy (68% vs. 44.39%) of the appearance approach of Galatas et al. (Galatas, et al., 2012). The results achieved only with our digits set (69%) also outperformed the 65% recognition accuracy achieved by Potamianos at al. (Potamianos, et al., 2003) and the 63.9% achieved by Lee & Park (Lee & Park, 2008). However, differences in both the number of speakers and the size of the dataset makes difficult a direct comparison. On a speaker-dependent context, ViKi achieved encouraging results, reaching almost 70% of average word recognition accuracy. The wide range of results is explained by the variability between speakers. While the sound of each word must be very similar for every speaker, the range

of movement of the lips, tongue and jaw varies a lot. Considering that the lips movements are small, these variations explain the differences between speaker-dependent accuracies.

5.2. Future Work

The Microsoft Kinect has brought new possibilities for several speech recognition fields (e.g. ASR, VSR or SSI), showing promising results. The new Kinect One presents major improvements in its sensors compared with the previous version and, although it was only released in beta version at the beginning of the production of this thesis, and only a few weeks ago, in RTM (Released to Manufacturing) version, at the time of writing of this dissertation, it will certainly give rise to a new series of research and applications in these fields.

Considering that we had the opportunity to explore the new Kinect One at first-hand, some future work could be suggested. It would be interesting to test an appearance-based or hybrid-based approaches using the new Kinect One and its SDK capabilities to assess how they perform in a VSR system, by comparing them with our articulatory shape-based approach. The IR sensor and the TOF technology of the depth sensor could also be explored to create a light-independent VSR system. In fact, one of the main issues faced by VSR systems, is the variation of light. By using only the IR and depth sensors, it should be possible to create a VSR system that would work even in dark scenarios. Since the IR data was collected for our dataset, it could be used for the development, training and testing of a visual speech recognition system of this kind.

The versatility to be used on a wide range of applications makes it easier to merge different Kinect based applications or even to integrate these applications on other systems. As a concrete example, ViKi could easily be integrated on an ASR system (developed with the Kinect One SDK), to create an ASVR (Audio-Visual Speech Recognition) system. It could also be used with other applications, even not related directly with speech recognition. In fact, the first steps to merge ViKi with a Portuguese Sign Language (PSL) Recognition system have been made. The PSLR system was developed to identify signs (gestures and postures) from the PSL language (Sotelo, 2014). Merging it with ViKi would allow to add a lip expression recognition to the system, since it is common for sign language speakers to associate facial expressions with hand gestures in complex

signs. Our developed data collection subsystem, is actually able to collect the necessary data from the hands and the lip region, to be then used for training and classification.

BIBLIOGRAPHY

1st Provider's Choice, 2014. *Medical Voice Recognition & EMR Software.* [Online] Available at: http://www.1stproviderschoice.com/medical-voice-recognition-software.php [Accessed 14 08 2014].

Anusuya, M. A. & Katti, S. K., 2009. Speech Recognition by Machine: A Review. *International Journal of Computer Science and Information Security,* 6(3), pp. 181-205.

Baalke, R. & Goodall, K., 1997. *Mars Polar Lander Official Website.* [Online] Available at: http://mars.jpl.nasa.gov/msp98/lidar/microphone/index.html [Acedido em 4 janeiro 2013].

Baldwin, J. F., Martin, T. P. & Saeed, M., 1999. *Automatic Computer Lip-Reading Using Fuzzy Set Theory.* Santa Cruz, Auditory-Visual Speech Processing.

Berg, M. d., Kreveld, M. v., Overmars, M. & Schwarzkopf, O., 2000. *Computational Geometry Algorithms and Applications.* 2nd ed. Eindhoven: Springer.

Bernstein, L. E. & Benoît, C., 1996. *For Speech Perception by Humans or Machines, Three Senses are Better Than One.* Philadelphia, 4th International Conference on Spoken Language.

Bogert, B. P., Healy, M. J. R. & Tukey, J. W., 1963. *The Quefrency Analysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Safe Cracking.* New York, Proceedings of the Symposium on Time Series Analysis.

Chitu, A. & Rothkrantz, L. J. M., 2012. Automatic Visual Speech Recognition. In: S. Ramakrishnan, ed. *Speech Enhancement, Modeling And Recognition.* Rijeka: InTech, pp. 95-120.

Davis, K. H., Biddulph, R. & Balashek, S., 1952. Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America,* 24(6), pp. 637-642.

Estellers, V. & Thiran, J.-P., 2012. Multi-pose Lipreading And Audio-Visual Speech Recognition. *EURASIP Journal on Advances in Signal Processing,* 2012(51).

Freitas, J., Teixeira, A. & Dias, M. S., 2014. *A Multimodal Corpora for Silent Speech Interaction.* Reykjavik, Language Resources and Evaluation Conference.

Galatas, G. et al., 2011. *Bilingual Corpus for AVASR using Multiple Sensors and Depth Information.* Volterra, 11th International Conference on Auditory-Visual Speech Processing.

Galatas, G., Potamianos, G. & Makedon, F., 2012. *Audio-Visual Speech Recognition Incorporating Facial Depth Information Captured by The Kinect.* Bucharest, Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European.

Galatas, G., Potamianos, G. & Makedon, F., 2012. *Audio-Visual Speech Recognition Using Depth Information From The Kinect in Noisy Video Condition.* New York, Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environment.

Galatas, G., Potamianos, G. & Makedon, F., 2013. *Robust Multi-Modal Speech Recognition in Two Languages Utilizing Video and Distance Information from the Kinect.* Las Vegas, 15th international Conference on Human-Computer Interaction: Interaction Modalities and Techniques.

Gick, B., Wilson, I. & Derrick, D., 2013. *Articulatory Phonetics.* 1st ed. Oxford: Blackwell Publishing.

Gold, B., Morgan, N. & Ellis, D., 2011. *Speech and Audio Signal Processing.* 2nd ed. New Jersey: John Wiley & Sons, Inc..

Gonzalez-Mora, J. et al., 2007. *Bilinear Active Appearance Models Jose.* Malaga: s.n.

Granström, B. & House, D., 2004. *Audiovisual Representation of Prosody in Expressive Speech Communication.* Nara, Speech Prosody.

Gurban, M., 2009. *Multimodal Feature Extraction and Fusion for Audio-Visual Speech Recognition,* Lausanne: École Politéchnique Fédérale de Lausanne.

Hasegawa-Johnson, M., Livescu, K., Lal, P. & Saenko, K., 2007. *Audiovisual Speech Recognition with Articulator Positions as Hidden Variables.* Saarbrücken, International Congress of Phonetic Science.

Hassanat, A. B. A., 2011. Visual Speech Recognition. In: I. Ipsic, ed. *Speech Language Technologies.* Rijeka: InTech, pp. 279-302.

Hassanieh, H., Indyk, P., Katabi, D. & Price, E., 2012. *Simple and Practical Algorithm for Sparse Fourier Transform.* Kyoto, ACM-SIAM Symposium On Discrete Algorithms.

Hickok, G., 2013. The Cortical Organization of Speech Processing: Feedback Control and Predictive Coding the Context of a Dual-Stream Model. *Journal of Communication Disorders,* 45(6), pp. 393-402.

Hilder, S., Harvey, R. & Theobald, B.-J., 2009. *Comparison of Human and Machine-Based Lip-Reading.* Norwich, International Conference on Auditory-Visual Speech Processing.

Holt, L. L. & Lotto, A. J., 2010. Speech Perception as Categorization. *National Institutes of Health,* 72(5), pp. 1218-1227.

Honda Motor Corporation, 2014. *History of ASIMO Robotics.* [Online] Available at: http://asimo.honda.com/asimo-history/ [Accessed 26 August 2014].

IBM, 2003. *IBM Archives: IBM Shoebox.* [Online] Available at: http://www-03.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html [Accessed 22 August 2014].

Jody Kreiman, D., Vanlancker-Sidtis & Gerratt, B. R., 2005. Perception of Voice Quality. In: D. B. Pisoni & R. E. Remez, eds. *The Handbook of Speech Perception.* Oxford: Blackwell Publishing Ltd, pp. 338-362.

Juneja, A., 2012. A Comparison of Automatic and Human Speech Recognition in Null Grammar. *The Journal of the Acoustical Society of America,* 131(3), pp. 256-261.

Lan, Y., Harvey, R. & Theobald, B.-J., 2012. *Insights Into Machine Lip Reading.* Kyoto, International Conference on Acoustics, Speech, and Signal Processing.

Lee, B. et al., 2004. *AVICAR: Audio-Visual Speech Corpus in a Car Environment.* Jeju Island, s.n.

Lee, J.-s. & Park, C. H., 2008. Robust Audio-Visual Speech Recognition Based on Late Integration. *IEEE Transactions on Multimedia,* 10(5), pp. 767-779.

Levelt, W. J. M., 1999. Models of Word Production. *Trends in Cognitive Sciences,* 3(6), pp. 223-232.

Lippmann, R. P., 1997. Speech Recognition by Machines and Humans. *Speech Comunication,* Volume 22, pp. 1-15.

McGurk, H. & MacDonald, J., 1976. Hearing Lips and Seeing Voices. *Nature,* 264(5588), pp. 746-748.

McKay, P. et al., 2013. *Read My Lips: Towards Use of the Microsoft Kinect as a Visual-Only Automatic Speech Recognizer.* Newcastle, Ninth Symposium On Usable Privacy and Security.

Meyer, B. et al., 2006. A Human-Machine Comparison in Speech Recognition Based on a Logatome Corpus. *Speech Recognition and Intrinsic Variation*, 5 May, pp. 1-6.

Moore, R. K. & Cutler, A., 2001. *Constraints on Theories of Human vs. Machine Recognition of Speech.* Nijmegen, SPRAAC Workshop on Speech Recognition as Pattern Classification.

Moura, A., 2005. *Reconhecimento Automático da Fala com Processamento Simultâneo de Características Acústicas e Visuais,* Porto: Dissertação de Mestrado, Faculdade de Engenharia da Universidade do Porto.

Moura, A., Pêra, V. & Freitas, D., 2005. *Leitura Labial Automática em Sistemas de Reconhecimento Automático da Fala: Desenvolvimento de um Sistema para o Português Europeu.* Lisbon, Jornadas de Engenharia de Electrónica e Telecomunicações e de Computadores.

Newman, J. L., Theobald, B.-J. & Cox, S. J., 2010. *Limitations of Visual Speech Recognition.* Hakone, International Conference on Auditory-Visual Speech Processing.

Nuance Communications, Inc., 2008. *Speech Recognition: Accelerating the Adoption of Electronic Medical Records.* [Online] Available at: http://www.nuance.com/ucmprod/groups/dragon/documents/webasset/nd_004916.pdf [Accessed 14 08 2014].

Nuance Communications, Inc., 2014. *Dragon - Dragon NaturallySpeaking - Nuance.* [Online] Available at: http://www.nuance.com/dragon/index.htm [Accessed 14 08 2014].

Olson, H. & Belar, H., 1957. Phonetic Typewriter. *IRE Transactions on Audio,* AU-5(4), pp. 90-95.

Patterson, E. K., Gurbuz, S., Tufekci, Z. & Gowdy, J. N., 2002. *CUAVE: A New Audio-Visual Database for Multimodal Human-Computer Interface Research.* Orlando, s.n.

Pera, V., Moura, A. & Freitas, D., 2004. *LPFAV2: a New Multi-Modal Database for Developing Speech Recognition Systems for an Assistive Technology Application.* St. Petersburg, s.n.

Petajan, E., Bischoff, B. & Bodoff, D., 1988. *An Improved Automatic Lipreading System To Enhance Speech Recognition.* Washington, CHI - Computer-Human Interaction Conference.

Petajan, E. D., 1984. *Automatic Lipreading to Enhance Speech Recognition.* Urbana: University of Illinois.

Potamianos, G. & Neti, C., 2001. *Improved ROI and Within Frame Discriminant Features for Lipreading.* Thessaloniki, International Conference on Image Processing.

Potamianos, G. et al., 2003. Recent Advances in the Automatic Recognition of Audio-Visual Speech. *Procedding of the IEEE,* 91(9), pp. 1306 - 1326.

Rabiner, L. & Biing-Hwang, J., 1993. *Fundamentals of Speech Recognition.* New Jersey: Prentice-Hall.

Rabiner, L. R. & Juang, B.-H., 2000. *The Digital Signal Processing handbook - Video, Speech, and Audio Signal Promcessing and Associated Standards.* 1st ed. Boca Raton: CRC Press.

Rao, R. R. & Mersereau, R. M., 1994. *Lip Modeling for Visual Speech Recognition.* Pacific Grove, Conference Record of the Twenty-Eighth Asilomar Conference on Signals, Systems and Computers.

Robjohns, H., 2000. A Brief History of Microphones. In: C. Woolf, ed. *Microphone Data.* s.l.:Human-Computer Interface Ltd.

Rosenblum, L. D., 2005. Primacy of Multimodal Speech Perception. In: D. B. Pisoni & R. E. Remez, eds. *The Handbook of Speech Perception.* Oxford: Blackwell Publishing Ltd, pp. 51-78.

Rosenblum, L. D., 2013. Speech Perception as a Multimodal Phenomenon. *National Institutes of Health,* 17(6), pp. 405-409.

Rosen, C. A. & Simpson, B., 2008. Anatomy and Physiology of the Larynx. In: *Operative Techniques in Laryngology.* Berlin: Springer Berlin Heidelberg, pp. 1-8.

Saenko, K., Darrell, T. & Glass, J., 2004. *Articulatory Features for Robust Visual Speech Recognition.* Pennsylvania, ICMI - International Conference on Multimodal Interfaces.

Saenko, K., Livescu, K., Glass, J. & Darrell, T., 2009. Multistream Articulatory Feature-Based Models for Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Learning,* 31(9), pp. 1700-1707.

Saenko, K. et al., 2005. *Visual Speech Recognition with Loosely Synchronized Feature Streams.* Beijing, Tenth IEEE International Conference on Computer Vision.

Schaller, C., 2011. *Time-of-Flight - A New Modality for Radiotherapy.* Erlangen: s.n.

Seeley, R. R., Stephens, T. D. & Tate, P., 2008. *Anatomy & Physiology.* 8th ed. New York: McGraw-Hill.

Seikel, J. A., King, D. W. & Drumright, D. G., 2010. *Anatomy & Physiology for Speech, Language, and Hearing.* 4th ed. New York: Delmar.

Shaikh, A. A., Kumar, D. K. & Yau, W. C., 2010. *Lip Reading Using Optical Flow and Support Vector Machines.* Yantai, 3rd International Congress on Image and Signal Processing.

Shin, J., Lee, J. & Kim, D., 2011. Real-time Lip Reading System for Isolated Korean Word Recognition. *Pattern Recognition,* 44(3), pp. 559-571.

Smedt, K. d., 1996. Computational Models of Incremental Grammatical Encoding. In: *Computational Psycholinguistics: AI and Connectionist Models of Human Language Processing.* London: Taylor & Francis, pp. 279-307.

Song, M. G. et al., 2012. A Robust and Real-Time Visual Speech Recognition for Smartphone Application. *International Journal of Innovative Computing, Information and Control,* 8(4), pp. 2837-2853.

Sony Corporation, 2000. *Sony Announces Sale of 2nd Generation Autonomous Entertainment Robot "AIBO" [ERS-210].* [Online] Available at: http://www.sony.net/SonyInfo/News/Press_Archive/200010/00-050A/ [Accessed 26 janeiro 2014].

Sotelo, C., 2014. *Portuguese Sign Language Recognition from Depth Sensing Human Gesture and Motion Capture.* Braga: Universidade do Minho.

Souza, C., 2013. *Accord.NET Framework.* [Online] Available at: http://accord-framework.net/ [Accessed 07 October 2014].

Sun, C., Xu, C., Bao, B.-K. & Mei, T., 2012. *Kinect-based Visual Communication System.* New York, ICIMCS - Proceedings of the 4th International Conference on Internet Multimedia Computing and Service.

Teixeira, V. D. et al., 2012. *Towards Elderly Social Integration Using a Multimodal Human-Computer Interface.* Vilamoura, s.n.

Vocapia Research SAS, 2014. *Speech-to-text VoxSigma Software Suite.* [Online] Available at: http://www.vocapia.com/voxsigma-speech-to-text.html [Accessed 14 08 2014].

VoltDelta, 2014. *Telecomm Customer Service Call Center.* [Online] Available at: http://www.voltdelta.com/how-we-help/by-industry/telecommunications [Accessed 14 08 2014].

Werda, S., Mahdi, W. & Hamadou, A. B., 2007. Lip Localization and Viseme Classification for Visual Speech Recognition. *International Journal of Computing & Information Sciences,* 5(1), pp. 62-75.

Yargıç, A. & Doğan, M., 2013. *A Lip Reading Application on MS Kinect Camera.* Albena, INISTA - IEEE International Symposium on Innovations in Intelligent Systems and Applications.

Yau, W. C., Weghorn, H. & Kumar, D. K., 2007. *Visual Speech Recognition and Utterance Segmentation Based on Mouth Movement.* Glenelg, 9th Biennial Conference of the Australian Pattern Recognition Society on Digital Image Computing Techniques and Applications.

Zhang, L., Curless, B. & Seitz, S. M., 2002. *Rapid Shape Acquisition Using Color Structured Light and Multi-pass Dynamic Programming.* Padova, s.n.

APPENDIX

A. CONSENT FORM (IN EP)

FOLHA DE INFORMAÇÃO AOS PARTICIPANTES

Eu, Hélder Abreu, investigador do Centro de Desenvolvimento da Linguagem e estudante de mestrado da Universidade do Minho, venho solicitar a sua colaboração para uma recolha de dados no âmbito da tese com o tema *Visual Speech Recognition for European Portuguese*. A sua participação representa um importante contributo, não só para o estudo em curso, mas também para o desenvolvimento do conhecimento na área da Interação Homem-Máquina e Ambientes Assistidos.

O principal objetivo desta recolha de dados é o desenvolvimento de um sistema de reconhecimento visual da fala, não invasivo, com base em múltiplas modalidades, para o Português Europeu. Através da gravação de 9 repetições de 24 palavras distintas, previamente definidas, serão recolhidas imagens a cores, imagens infravermelhas e a respetiva informação de profundidade. O tempo médio, previsto, para a sua participação será de 1h30 a 2h. Ao longo da gravação estará sempre presente o investigador que o ajudará em toda a experiência. Para recolher os dados ser-lhe-á pedido que se sente em frente a uma camara (Microsoft Kinect). De seguida, o investigador responsável irá ajustar e posicionar a mesma. O Microsoft Kinect contém:

- 1 Camara RGB (cores) para a captura de vídeo
- 1 Camara de infravermelhos para capturar informação de profundidade através da tecnologia "time-of-flight"

A sua participação é voluntária e, assim sendo, poderá a qualquer momento optar por desistir de participar no estudo, sem que daí advenham quaisquer prejuízos ou consequências. Será garantida a confidencialidade e o anonimato de todas as informações recolhidas, sendo apenas utilizados os dados referentes à idade e género para efeitos de análise estatística. Os sinais recolhidos não serão utilizados para outros fins que não os definidos no âmbito da investigação. Deste estudo não advirá qualquer risco físico, emocional e/ou financeiro para si. O investigador está ao seu dispor para responder a qualquer dúvida ou questão que considere que não foi esclarecida com este documento informativo.

Está interessado em participar neste estudo (marque com um X a sua resposta)?

Sim ☐ **Não** ☐

Género:

Masculino ☐ **Feminino** ☐

Idade: _____

Nº de participante _____

TERMO DE CONSENTIMENTO LIVRE E INFORMADO

Por favor, responda às questões que se seguem desenhando uma cruz na coluna apropriada.

	Sim	Não
Eu li a Folha de Informação aos Participantes.		
Eu recebi toda a informação adequada a este estudo.		
Foi-me permitido colocar questões sobre o estudo.		
Todas as minhas dúvidas sobre o estudo foram esclarecidas.		
Eu compreendi que posso desistir do estudo em qualquer altura, sem que qualquer consequência advenha dessa desistência.		
Eu compreendi que todos os dados serão tratados de forma confidencial garantindo o anonimato.		

Esta experiência tem um carácter voluntário. O participante tem a possibilidade, por motivos éticos, de negar a participação ou de se retirar do estudo, a qualquer momento, sempre que assim o entender.

De acordo com as normas da Comissão de Proteção de Dados, os dados recolhidos são totalmente anónimos.

Tendo tomado conhecimento sobre a informação disponível do estudo, declaro que aceito participar neste estudo.

Nome do participante:
_____

Assinatura do participante:
_____

Nome do investigador:
_____

Assinatura do investigador:
_____

Data: _____ - _____ - _____

B. Cross-Validation Results

- **Geometric Features**

Fold	Training (%)	Validation (%)
1	97	68
2	97	63
3	97	70
4	97	70
5	97	64
6	97	67
7	97	68
8	97	**74**
9	97	64
Average (Standard Deviation)	**97.0 (0.0)**	**67.56 (3.37)**

- **Geometric Features (without depth information)**

Fold	Training (%)	Validation (%)
1	91	67
2	91	62
3	90	66
4	90	68
5	91	60
6	90	67
7	89	**70**
8	90	70
9	91	65
Average (Standard Deviation)	**90.33 (0.67)**	**66,11 (3.18)**

- ARTICULATORY FEATURES

Fold	Training (%)	Validation (%)
1	47	40
2	47	31
3	47	35
4	47	33
5	47	28
6	47	36
7	46	32
8	46	37
9	46	34
Average (Standard Deviation)	46.67 (0.5)	34.0 (3.54)

C. Confusion Matrix - Articulatory Features

	Vídeos	Ligar	Contactos	Mensagens	Voltar	Pesquisar	Anterior	Fotografias	Família	Ajuda	Seguinte	Lembretes	Calendário	Email	Zero	Um	Dois	Três	Quatro	Cinco	Seis	Sete	Oito	Nove	Word Accuracy
Vídeos	39	6	5	0	1	0	0	0	1	2	2	2	2	0	1	3	2	1	0	4	0	1	0	0	**0,54**
Ligar	1	45	0	0	1	2	2	0	1	1	6	1	5	0	1	0	0	1	0	1	0	0	2	2	**0,63**
Contactos	8	4	35	0	1	1	1	4	2	0	0	1	2	1	0	1	0	0	6	0	0	1	2	2	**0,49**
Mensagens	3	31	0	1	1	11	1	0	0	0	15	0	3	1	1	0	0	0	0	0	0	3	0	1	**0,01**
Voltar	1	18	1	0	14	1	0	1	9	4	2	0	1	0	1	2	9	0	5	0	0	0	3	0	**0,19**
Pesquisar	1	17	2	1	3	18	1	3	6	0	9	0	1	0	0	0	2	0	0	1	0	5	0	2	**0,25**
Anterior	5	21	4	2	0	2	15	2	0	7	8	0	3	0	2	0	0	0	0	0	0	0	0	1	**0,21**
Fotografias	1	0	6	0	4	2	2	41	4	0	2	0	2	0	0	0	6	0	1	0	0	0	1	0	**0,57**
Família	0	6	3	1	2	4	0	1	23	0	1	7	0	14	0	0	1	1	1	0	0	1	1	5	**0,32**
Ajuda	3	9	0	0	2	1	5	0	0	40	1	0	1	1	1	2	0	0	2	0	0	0	3	1	**0,56**
Seguinte	3	23	1	1	0	2	1	4	0	0	24	0	4	0	0	0	0	0	0	1	0	2	4	2	**0,33**
Lembretes	9	4	2	1	0	4	2	0	6	0	5	25	3	6	1	0	0	0	1	0	0	1	2	0	**0,35**
Calendário	5	17	3	3	0	0	1	0	0	2	7	0	25	0	2	0	0	1	1	2	0	1	1	1	**0,35**
Email	2	4	6	0	1	2	0	2	4	0	1	9	2	22	0	2	1	1	1	0	0	7	1	4	**0,31**
Zero	1	16	0	0	0	0	0	0	0	0	1	0	9	0	15	0	1	1	0	15	0	13	0	0	**0,21**
Um	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	63	1	0	2	1	0	1	2	0	**0,89**
Dois	0	4	4	1	4	0	0	5	1	1	0	0	2	0	0	0	29	0	6	1	0	6	5	3	**0,4**
Três	6	41	1	1	0	1	0	0	3	0	3	0	4	0	1	0	0	0	0	0	0	8	1	2	**0**
Quatro	2	0	2	0	2	0	0	1	1	2	0	0	0	0	0	3	4	0	45	0	0	0	6	4	**0,63**
Cinco	6	2	3	0	0	1	0	1	1	2	0	0	2	1	11	1	1	1	0	25	0	9	3	2	**0,35**
Seis	8	30	3	2	1	1	1	1	0	0	9	0	5	0	2	0	0	1	0	1	0	6	0	1	**0**
Sete	2	33	0	0	0	2	0	0	2	0	9	0	3	0	3	0	1	1	0	0	1	9	3	3	**0,13**
Oito	0	0	2	1	3	2	0	3	0	4	0	0	0	1	0	6	5	1	14	0	0	5	21	4	**0,29**
Nove	3	6	1	0	2	2	3	2	4	3	2	0	1	3	1	1	6	0	4	0	0	5	11	12	**0,17**

D. Confusion Matrix - Geometric and Articulatory Features

	Vídeos	Ligar	Contactos	Mensagens	Voltar	Pesquisar	Anterior	Fotografias	Família	Ajuda	Seguinte	Lembretes	Calendário	Email	Zero	Um	Dois	Três	Quatro	Cinco	Seis	Sete	Oito	Nove	Word Accuracy
Vídeos	35	0	4	1	0	0	2	0	0	1	2	3	0	0	1	8	0	1	1	10	0	1	1	1	**0,49**
Ligar	0	44	0	1	3	3	0	0	1	1	0	1	1	0	3	0	0	4	2	2	1	0	2	3	**0,61**
Contactos	9	0	38	0	0	0	0	1	1	0	0	2	0	1	2	5	3	0	6	0	1	0	2	1	**0,53**
Mensagens	2	3	2	29	5	10	1	0	1	0	6	1	1	0	0	0	0	2	1	0	5	1	0	2	**0,4**
Voltar	1	0	1	0	26	2	0	3	9	3	2	0	1	2	1	0	9	1	9	0	2	0	0	0	**0,36**
Pesquisar	0	1	2	3	2	37	0	4	5	0	5	1	0	0	0	0	0	0	2	3	1	1	2	3	**0,51**
Anterior	8	0	3	2	1	1	39	1	0	8	0	1	1	0	0	0	0	2	0	1	1	1	1	1	**0,54**
Fotografias	0	0	3	0	3	1	0	51	4	0	0	0	0	0	0	0	3	0	2	0	0	0	4	1	**0,71**
Família	1	0	2	0	2	1	1	2	46	0	0	4	0	4	0	0	0	0	2	1	0	0	1	5	**0,64**
Ajuda	2	0	0	0	0	0	11	0	0	45	0	0	1	1	0	3	0	0	1	0	0	0	7	1	**0,63**
Seguinte	4	1	1	2	1	4	1	2	0	0	25	0	2	0	1	0	0	6	4	1	4	6	4	3	**0,35**
Lembretes	7	0	6	0	1	0	3	1	7	0	0	30	1	6	0	0	0	0	2	1	0	3	3	1	**0,42**
Calendário	1	3	3	0	0	1	1	0	0	3	3	0	35	1	3	0	0	2	3	6	0	0	5	2	**0,49**
Email	1	0	8	0	0	1	1	2	7	1	0	10	0	26	1	0	0	4	2	0	0	1	4	3	**0,36**
Zero	2	0	1	2	0	1	0	0	0	0	0	0	2	1	32	1	0	2	2	16	4	2	1	3	**0,44**
Um	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	64	1	0	3	0	0	0	1	1	**0,9**
Dois	0	0	1	0	4	0	0	7	0	1	0	0	1	1	0	0	36	0	8	1	1	2	5	4	**0,5**
Três	7	6	0	0	0	0	1	0	0	0	3	0	1	0	1	0	2	32	2	2	6	3	3	3	**0,44**
Quatro	2	0	2	0	0	0	0	0	1	1	0	0	0	0	0	3	2	0	54	0	0	0	6	1	**0,75**
Cinco	7	0	3	1	0	0	0	0	0	0	1	0	1	1	7	4	0	2	1	31	4	3	5	1	**0,43**
Seis	6	2	5	2	1	2	1	0	2	0	3	1	3	0	2	0	0	11	0	2	22	5	1	1	**0,31**
Sete	2	0	1	3	0	3	1	0	1	0	6	0	0	0	4	0	0	4	2	1	9	29	1	5	**0,4**
Oito	0	0	3	0	0	2	0	3	0	4	0	0	0	1	0	13	2	0	19	0	0	1	18	6	**0,25**
Nove	4	0	2	0	0	0	1	1	1	3	0	0	0	2	0	0	2	1	5	2	0	4	16	28	**0,39**

E. Confusion Matrix – Geometric Features – Speaker LNS

	Vídeos	Ligar	Contactos	Mensagens	Voltar	Pesquisar	Anterior	Fotografias	Família	Ajuda	Seguinte	Lembretes	Calendário	Email	Zero	Um	Dois	Três	Quatro	Cinco	Seis	Sete	Oito	Nove	Word Accuracy
Vídeos	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**0,89**
Ligar	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Contactos	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Mensagens	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0,89**
Voltar	0	0	0	0	8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0,89**
Pesquisar	0	0	0	0	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0,89**
Anterior	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Fotografias	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Família	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Ajuda	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**1**
Seguinte	0	0	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	2	0	0	0	1	0	0	**0,67**
Lembretes	0	0	0	0	0	0	0	0	0	0	0	8	0	1	0	0	0	0	0	0	0	0	0	0	**0,89**
Calendário	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	0	0	**1**
Email	0	0	1	0	0	0	0	0	0	0	0	0	0	7	0	0	0	1	0	0	0	0	0	0	**0,78**
Zero	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	0	0	0	1	1	0	0	0	**0,78**
Um	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	0	0	0	0	1	**0,89**
Dois	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	1	0	0	0	0	1	**0,78**
Três	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	0	0	5	0	0	0	1	0	0	**0,56**
Quatro	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	0	0	1	0	**0,89**
Cinco	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	1	0	0	0	**0,78**
Seis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	0	0	0	**0,89**
Sete	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	7	0	0	**0,78**
Oito	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	1	**0,89**
Nove	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	6	**0,67**