Universidade de Aveiro Departamento de Eletrónica, Telecomunicações e Informática
Ano 2015

João Dinis Colaço de Freitas

Interfaces de Fala Silenciosa Multimodais para Português Europeu com base na Articulação

Articulation in Multimodal Silent Speech Interface for European Portuguese

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Informática, realizada sob a orientação científica do Professor Doutor António Joaquim da Silva Teixeira, Professor Associado do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e do Professor Doutor José Miguel de Oliveira Monteiro Sales Dias, Diretor do Centro de Desenvolvimento da Linguagem da Microsoft Portugal e Professor Associado Convidado do Instituto Universitário de Lisboa (ISCTE-IUL).

To my Mother and to Diana for their immeasurable patience, support and unconditional love.

To the loving memory of Carlos, I am forever grateful.

o júri

presidente

Doutora Anabela Botelho Veloso
Professora Catedrática da Universidade de Aveiro

Doutora Isabel Maria Martins Trancoso
Professora Catedrática do Instituto Superior Técnico da Universidade de Lisboa

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

Doutor José Miguel de Oliveira Monteiro Sales Dias
Professor Associado Convidado do ISCTE-IUL – Instituto Universitário de Lisboa (Coorientador)

Doutor Luís Manuel Dias Coelho Soares Barbosa
Professor Associado da Escola de Engenharia da Universidade do Minho

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

Doutor Carlos Jorge da Conceição Teixeira
Professor Auxiliar da Faculdade de Ciências da Universidade de Lisboa

Doutor Cengiz Acartürk
Assistant Professor, Informatics Institute, Middle East Technical University (METU), Ankara - Turkey

Agradecimentos

I would like to thank to all the people whose contributions, assistance and support made the end result possible.

In my case, both my supervisors, Prof. António Teixeira and Prof. Miguel Sales Dias, head the list. I want to thank them for their patience, continuous encouragement, support and understanding. Additionally, I also want to thank them for always finding the time to discuss new ideas and for helping me in the search for the right path. Their contributions for this work are invaluable and it has been an immense privilege to be their student. I am forever in their debt for all the knowledge they have passed me.

I also want to thank the Microsoft Language Development Center for supporting this research and to all my colleagues that work or worked there for their encouragement, understanding and support. In particular, I want to thank António Calado, Daan Baldewijns, Pedro Silva, Manuel Ribeiro and Filipe Gaspar, for their reviews and friendship.

To the remarkable researchers, Samuel Silva and Catarina Oliveira, from the University of Aveiro, and Artur Ferreira, from the Lisbon Engineering Institute, thank you all your help and advices. Your collaboration in several studies here described and your contributions add an enormous value to this thesis. I also want to thank Prof. Dr. Francisco Vaz for his contribution in designing the Ultrasonic Doppler device and in the Ultrasonic Doppler studies.

To my colleagues from MAP-i, Carlos Pereira, Nuno Oliveira, Nuno Luz and Ricardo Anacleto, thank you for the great times in Braga during the first year of the PhD programme.

I would also like to deeply thank all the people that have taken their time to participate in the data acquisition sessions. To all, thank you for donating your (silent) speech and for the valuable feedback.

To my friends and family, thank you for all the get-away moments. It was after those moments that the best solutions and ideas were created in my mind. In particular, I want to thank my godfather, Carlos Faria, and his family, for being such wonderful hosts, for their continuous encouragement, belief, and for making my weekly trips to Braga worthy. Carlos was like an older brother to me, with great influence in my decision to pursue higher education. I wish he could be here to share this moment.

To my mother, thank you for all the sacrifices, for always being present and for supporting me in every way.

To Diana, my deepest thanks, I surely could not have done this without her love and understanding. In that sense, this thesis is also hers. Her friendship, support and patience are infinite.

Last but not least, a word to my son, Pedro. Your smile, your ability for making me laugh even in the worst moments, are all the motivation I could ever ask. I hope that one day, the hours of absence, can be compensated by passing what I have learnt to you.

Palavras-chave

Fala Silenciosa, Multimodal, Português Europeu, Articulação, Nasalidade, Interação Humano-Computador, Electromiografia de superfície, Video, Informação de profundidade, Medição de Doppler em ultrassons

Resumo

O conceito de fala silenciosa, quando aplicado a interação humano-computador, permite a comunicação na ausência de um sinal acústico. Através da análise de dados, recolhidos no processo de produção de fala humana, uma interface de fala silenciosa (referida como SSI, do inglês *Silent Speech Interface*) permite a utilizadores com deficiências ao nível da fala comunicar com um sistema. As SSI podem também ser usadas na presença de ruído ambiente, e em situações em que privacidade, confidencialidade, ou não perturbar, é importante.

Contudo, apesar da evolução verificada recentemente, o desempenho e usabilidade de sistemas de fala silenciosa tem ainda uma grande margem de progressão. O aumento de desempenho destes sistemas possibilitaria assim a sua aplicação a áreas como Ambientes Assistidos. É desta forma fundamental alargar o nosso conhecimento sobre as capacidades e limitações das modalidades utilizadas para fala silenciosa e fomentar a sua exploração conjunta.

Assim, foram estabelecidos vários objetivos para esta tese: (1) Expansão das linguagens suportadas por SSI com o Português Europeu; (2) Superar as limitações de técnicas de SSI atuais na deteção de nasalidade; (3) Desenvolver uma abordagem SSI multimodal para interação humano-computador, com base em modalidades não invasivas; (4) Explorar o uso de medidas diretas e complementares, adquiridas através de modalidades mais invasivas/intrusivas em configurações multimodais, que fornecem informação exata da articulação e permitem aumentar a nosso entendimento de outras modalidades.

Para atingir os objetivos supramencionados e suportar a investigação nesta área procedeu-se à criação de uma plataforma SSI multimodal que potencia os meios para a exploração conjunta de modalidades. A plataforma proposta vai muito para além da simples aquisição de dados, incluindo também métodos para sincronização de modalidades, processamento de dados multimodais, extração e seleção de características, análise, classificação e prototipagem. Exemplos de aplicação para cada fase da plataforma incluem: estudos articulatórios para interação humano-computador, desenvolvimento de uma SSI multimodal com base em modalidades não invasivas, e o uso de informação exata com origem em modalidades invasivas/intrusivas para superar limitações de outras modalidades.

No trabalho apresentado aplica-se ainda, pela primeira vez, métodos retirados do estado da arte ao Português Europeu, verificando-se que sons nasais podem causar um desempenho inferior de um sistema de fala silenciosa. Neste contexto, é proposta uma solução para a deteção de vogais nasais baseada num único sensor de eletromiografia, passível de ser integrada numa interface de fala silenciosa multimodal.

Keywords

Silent Speech, Multimodal, European Portuguese, Articulation, Nasality, Human-Computer Interaction, Surface Electromyography, Video, Depth Information, Ultrasonic Doppler Sensing.

Abstract

The concept of silent speech, when applied to Human-Computer Interaction (HCI), describes a system which allows for speech communication in the absence of an acoustic signal. By analyzing data gathered during different parts of the human speech production process, Silent Speech Interfaces (SSI) allow users with speech impairments to communicate with a system. SSI can also be used in the presence of environmental noise, and in situations in which privacy, confidentiality, or non-disturbance are important.

Nonetheless, despite recent advances, performance and usability of Silent Speech systems still have much room for improvement. A better performance of such systems would enable their application in relevant areas, such as Ambient Assisted Living. Therefore, it is necessary to extend our understanding of the capabilities and limitations of silent speech modalities and to enhance their joint exploration.

Thus, in this thesis, we have established several goals: (1) SSI language expansion to support European Portuguese; (2) overcome identified limitations of current SSI techniques to detect EP nasality (3) develop a Multimodal HCI approach for SSI based on non-invasive modalities; and (4) explore more direct measures in the Multimodal SSI for EP acquired from more invasive/obtrusive modalities, to be used as ground truth in articulation processes, enhancing our comprehension of other modalities.

In order to achieve these goals and to support our research in this area, we have created a multimodal SSI framework that fosters leveraging modalities and combining information, supporting research in multimodal SSI. The proposed framework goes beyond the data acquisition process itself, including methods for online and offline synchronization, multimodal data processing, feature extraction, feature selection, analysis, classification and prototyping. Examples of applicability are provided for each stage of the framework. These include articulatory studies for HCI, the development of a multimodal SSI based on less invasive modalities and the use of ground truth information coming from more invasive/obtrusive modalities to overcome the limitations of other modalities.

In the work here presented, we also apply existing methods in the area of SSI to EP for the first time, noting that nasal sounds may cause an inferior performance in some modalities. In this context, we propose a non-invasive solution for the detection of nasality based on a single Surface Electromyography sensor, conceivable of being included in a multimodal SSI.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABREVIATIONS AND ACRONYMS

AF	*Articulatory Features*	48
API	*Application Programming Interface*	125
ASIO	*Audio Stream Input/Output*	126
ASR	*Automatic Speech Recognition*	1
AV-ASR	*Audio-Visual Automatic Speech Recognition*	36
BCI	*Brain Computer Interface*	32
DARPA	*Defense Advanced Research Projects Agency*	28
DCT	*Discrete Cosine Transform*	37
DFT	*Discrete Fourier Transform*	78
DTW	*Dynamic Time Warping*	41
DWT	*Discrete Wavelet Transformation*	37
EEG	*Electroencephalography*	18
EGG	*Electroglottograph*	25
EMA	*Electromagnetic Articulography*	4
EMG	*Electromyography*	26
EP	*European Portuguese*	3
EPG	*Electropalatography*	25
EU	*European Union*	5
FACS	*Facial Action Coding System*	84
FD	*Feature Discretization*	49
FFT	*Fast Fourier Transform*	42
FIRST	*Fast Invariant to Rotation and Scale Transform*	38
fMRI	*Functional magnetic resonance imaging*	25
FR	*Feature Reduction*	49
FS	*Feature Selection*	49
GEMS	*General electromagnetic motion sensor*	25
HCI	*Human-Computer Interaction*	2
HMM	*Hidden Markov Model*	31
ISCTE-IUL	*University Institute of Lisbon*	56
kNN	*k-Nearest Neighbor*	117
LDA	*Linear Discriminant Analysis*	37
LPC	*Linear Predictive Coding*	41
LSDA	*Locality Sensitive Discriminant Analysis*	37

MEG	*Magnetoencephalography*	25
MEV	*Minimum Eigenvalues*	66
MRI	*Magnetic Resonance Imaging*	4
NAM	*Non-Audible Murmur*	19
OPG	*Optopalatography*	25
PCA	*Principal Component Analysis*	37
PMA	*Permanent-Magnetic Articulography*	18
P-MIC	*Physiological microphone*	25
RGB	*Additive color model based on red, green and blue*	9
RGB-D	*RGB plus depth information*	48
ROI	*Region of Interest*	37
RT-MRI	*Real-Time Magnetic Resonance Imaging*	3
SAMPA	*Speech Assessment Methods Phonetic Alphabet*	4
SDK	*Software Development Kit*	92
SEMG	*Surface Electromyography*	9
SHOT	*Signature of Histograms of OrienTations*	84
SIFT	*Scale Invariant Feature Transform*	38
SPIN	*Spin Image local surface descriptor*	84
SSI	*Silent Speech Interface*	2
STD	*Standard Deviation*	63
SURF	*Speeded Up Robust Features*	38
SVM	*Support Vector Machine*	117
UDS	*Ultrasonic Doppler Sensing*	9
UML	*Unified Modeling Language*	124
US	*Ultrasound Imaging*	19
VSR	*Visual Speech Recognition*	36
WER	*Word-Error Rate*	42
YIQ	*Luminance, In-phase, Quadrature (Color space used by NTSC TV system)*	105

CHAPTER I

Introduction

"Silence is one of the great arts of conversation."

Marcus Tullius Cicero

Contents

Since the dawn of mankind, speech has been and still is the dominant mode of human communication and information exchange. For this reason, spoken language technology has suffered a significant evolution in the last years. Nonetheless, nowadays conventional Automatic Speech Recognition (ASR) systems still rely on a single source of information – the audio signal. When this audio signal becomes corrupted in the presence of environmental noise or assumes variations, like the ones verified in elderly and children speech, speech recognition performance degrades, leading users to opt by a different communication modality, or to not use the system at all.

ASR systems have also revealed to be inadequate for users without the ability to produce an audible acoustic signal or in situations where non-disturbance or privacy is required. Using audible speech in public environments where silence is required such as in meetings, cinema or seminars, is usually considered annoying. Thus, providing the ability to perform an urgent call or to issue commands in these situations has become a point of common interest. Additionally, the need for privacy often occurs when interacting with mobile devices using speech. An example is the disclosure of private conversations by performing a phone call in a public place, which may lead to embarrassing situations to the caller or even to information leaks.

With the evolution and globalization of the telecommunication industry and associated technologies in the last decades, a quest for a system able to handle these issues and capable of adapting to speech-impaired users without the capability of creating an audible speech signal, started to gain momentum. This necessity gave origin to the Silent Speech Interface (SSI) concept, in parallel with many other research initiatives targeting better acoustic models and speech recognition enhancement. The term silent speech, when applied to Human-Computer Interaction (HCI), describes a system which allows for speech communication in the absence of an acoustic signal. It can be used as an HCI modality in high-background-noise environments such as in living rooms, or in aiding speech-impaired individuals (Denby et al., 2010). Commonly, this type of systems, acquire sensor data from elements of the human speech production process – from glottal and articulators' activity, their neural pathways or the brain itself – and creates an alternative digital representation of speech, which can be recognized and interpreted, synthesized directly, or routed into a communications network. Informally, one can say that an SSI extends the human speech production process (Levelt, 1995) using the signals measured by ultrasonic sensing devices, vision technologies and other sources of information. This provides a more natural approach than currently available speech pathology solutions like, electrolarynx (Weiss et al., 1979), tracheo-oesophageal speech (Debruyne et al., 1994), or a cursor-based text-to-speech solution (Denby et al., 2010).

1.1. Motivation

An increasing number of commercial ASR solutions become available in the last years. Speech recognition solutions can nowadays be found in a wide range of scenarios that include personal assistants in mobility, desktop, automotive and living-room entertainment. Nevertheless, being speech the dominant mode of social communication, the adoption of speech interfaces for our daily tasks is still scarce, mostly because of situations that cause an inferior performance of the system, such as environmental noise or physical disabilities concerning the vocal tract. Examples of this fact can also be found in the literature (Wilpon and Jacobsen, 1996) and in several studies where the author was also involved (Júdice et al., 2010; Oliveira et al., 2013), which corroborate that the available ASR solutions are not adapted to elderly speech.

The main motivation for this work is to contribute for an alternative universal communication system based on the concept of SSI, which holds a potential solution for a more natural interface and is theoretically able to address the issues inherent to ASR technology based on the acoustic signal.

For innumerous reasons that go from performance to available technology, current SSI systems are not yet seen as a viable technological solution. To take these systems to the next level, a comprehensive understanding of the underlying technologies, their capabilities and limitations and how to increase their performance is required. Addressing a multimodal[1] human-computer interaction approach to SSI, allows us to tackle the limitations and advantages of each HCI modality and establish the grounds to achieve higher goals and improve the results of research in this field. This knowledge may provide the basis for a combined exploitation of modalities, where eventually weaknesses of one modality can be minored by the strengths of other(s), and for exploring less invasive solutions, acceptable for all users.

In addition, revealing and informative technologies such as Real-Time Magnetic Resonance Imaging (RT-MRI), although not adequate and usable for an SSI in an HCI context, hold the potential to provide more direct measures of the articulators and the human speech production process in general. This information could eventually be used for evolving other modalities, consequently increasing their performance. Also, by having reliable ground truth data about the articulators' movements, we can probably increase our understanding of the capabilities and limitations of each modality.

Beyond finding an alternative universal communication system, this will only be relevant if the interface is adapted to the user and the user enjoys and takes benefit in using it. Thus, it is important to provide a speech interface in the users' native language, consequently leading to a more comfortable and natural interaction. We believe that this aspect applies to SSI, in particular to elderly users more reluctant to use technologies. However, little is known in terms of language adoption and its influence in the performance of an SSI system. Given that European Portuguese (EP) is the native language of the author and has been the focus of previous studies and collaborations, we decided to study the adoption of EP for this type of interfaces. No SSI existed for EP before this thesis and, as a consequence, we are unaware if the particular characteristics of this language affect existing state-of-the-art techniques. Amongst the distinctive characteristics of EP, the one we are most interested for this work is nasality, a known challenge in SSI research (Denby et al., 2010). In the following

[1] Multimodal can have different meanings depending on the research discipline. In this thesis we refer to multimodal as the use of multiple input modalities for SSI.

subsection we have included a more detailed insight of why language adoption constitutes an important motivational aspect for the work here presented.

Additionally, the elderly being one of the target groups for using an SSI, they are also part of the motivation for this work. Thus, we have also included a more detailed perspective (in section 1.1.2) of elder people limitations when using other interfaces, particularly those based on conventional ASR.

1.1.1. Adoption of European Portuguese

The first publications in the area of SSI focusing on EP were made by the author in the last years (Freitas et al., 2012b, 2011). Nonetheless, we can find in the literature previous research on related areas, such as the use of Electromagnetic Articulography (EMA) (Rossato et al., 2006) and Magnetic Resonance Imaging (MRI) (Martins, 2014; Martins et al., 2008) for speech production studies, articulatory synthesis (Teixeira and Vaz, 2000) and multimodal interfaces involving speech (Dias et al., 2009; Teixeira et al., 2005).

There are also some studies on lip reading systems for EP that aim at robust speech recognition based on audio and visual streams (Pera et al., 2004; Sá et al., 2003). However, these do not address EP distinctive characteristics, such as nasality.

1.1.1.1. European Portuguese characteristics

According to Strevens (1954), when one first hears EP, the characteristics that distinguish it from other Western Romance languages are: the large amount of diphthongs, nasal vowels and nasal diphthongs, frequent alveolar and palatal fricatives and the dark diversity of the l-sound. Additionally, although EP presents similarities in vocabulary and grammatical structure to Spanish, the pronunciation significantly differs (Martins et al., 2008). In terms of co-articulation, i.e. phenomena that describes the articulatory or acoustic influence of a speech segment on another, results show that EP stops are less resistant to co-articulatory effects than fricatives (Magen, 1997).

In the work here presented we focus on the nasal characteristics of the language, more particularly on the EP nasal vowels. Nasality is present in a vast number of languages around the world, however, only 20% have nasal vowels (Rossato et al., 2006). In EP there are five nasal vowels[2]: [i~] (e.g. *tinto*), [e~] (e.g. *penta*), [6~] (e.g. *manto*), [o~] (e.g. *bom*), and [u~] (e.g. *umbigo*); three nasal consonants ([m] (e.g. *mar*), [n] (e.g. *nada*), and [J] (e.g. *vinho*)); and several nasal diphthongs [w6~] (e.g. *quando*), [we~] (e.g. *aguentar*), [ja~] (e.g. *fiando*), [wi~] (e.g. *ruim*) and

[2] Phonetic transcriptions use the Speech Assessment Methods Phonetic Alphabet (SAMPA - http://www.phon.ucl.ac.uk/home/sampa (Wells et al., 1992)).

triphthongs [j6~w~] (e.g. *peão*) (Teixeira, 2000). Nasal vowels in EP diverge from other languages with nasal vowels, such as French, in its wider variation in the initial segment and stronger nasality at the end (Lacerda and Head, 1966; Trigo, 1993). Additionally, differences were also detected at the pharyngeal cavity level and velum port opening quotient when comparing EP and French nasal vowels articulation (Martins et al., 2008).

For a more detailed description of the European Portuguese characteristics we forward the reader to Martins (2014) and Strevens (1954).

1.1.2. Limitations and requirements imposed by the Elderly

Elderly population individuals have developed resistance to conventional forms of HCI, like the keyboard and mouse, therefore making it necessary to test new natural forms of interaction such as speech, silent speech, touch and gestures (Dias et al., 2012; Phang et al., 2006). In addition, elderly people often have difficulties with motor skills due to health problems such as arthritis. Therefore, small and difficult to handle equipment such as smartphones, are seen as disadvantageous. It is also known that due to ageing, senses like vision become less accurate, hence difficulties in the perception of details or important information in conventional graphical interfaces may arise since current mainstream interfaces, most notably in the mobility area, are not designed with these difficulties in mind. There is also evidence that the European Union (EU) population is ageing rapidly. The European Commission estimates that by 2050 the elderly population in the EU will be around 29% of the total population. This means that it is hastily becoming necessary to create solutions that allow overcoming the difficulties age brings to people who want to use new technologies in order to remain socially active. Elderly people who are connected to the world through the internet are less likely to become depressed and have greater probability of becoming socially integrated (Cisek and Triche, 2005). However, despite being the population group that is more rapidly going online (Fox, 2006), technological and interaction barriers still do not allow seniors to take full advantage of the available services and content (Dias et al., 2012; Oliveira et al., 2013; Stephanidis et al., 1998).

Collaborative research initiatives in the Portuguese R&D support frameworks, such as QREN 7900 LUL - Living Usability Lab ("QREN 7900 LUL - Living Usability Lab," n.d.) and QREN 13852 AAL4ALL ("QREN 13852 AAL4ALL," n.d.), where the author has contributed, have been paving the way to close such gap, with Ambient Assisted Living (AAL) solutions for home and mobility scenarios that have been positively evaluated with elderly populations. One of the motivations for investigating SSI is precisely related with such AAL initiatives. We see SSI systems as a potential alternative solution for HCI usable by elderly speakers, a group of users which has been found to prefer speech interfaces in the mentioned scenarios (Freitas et al., 2009; V. Teixeira et al., 2012), but also facing limitations in its use due to the inability of these systems to accurately model this population group, as detailed in the following section.

1.1.2.1. Elderly speech characteristics

Literature draws a divergent picture about how to characterize elderly speech, however, observations considering the voice of elderly people have proved that it is possible to state differences between elderly speech and teenagers or adults speech, on an acoustic phonetic level (Helfrich, 1979). The absence of a single deterministic phonetic cue, existent, for example, in gender determination, makes elderly speech classification inexact. Since aging increases the difference between biological age and chronological age and considering that biological aging can be influenced by factors such as abuse or overuse of the vocal folds, smoking, alcohol consumption, psychological stress/tension, or frequent loud/shouted speech production without vocal training (Jessen, 2007; Linville, 2001), it is not possible to determine an exact age limit for speech to be considered as elderly. Nonetheless, most of the studies consider ages between 60 and 70 as the minimum age for the elderly age group (Wilpon and Jacobsen, 1996).

With increasing age there is a deprivation of chest voice, general changes in frequencies, in the voice quality and the timbres. Changes in the heights of vowel formant frequencies particularly appear in older men, not only for biological reasons, but also because of social changes. According with (Helfrich, 1979; Pellegrini et al., 2013; Stover and Haynes, 1989), a slower speech-rate, greater use of pauses, elimination of articles and possessive pronouns, and lower volume of speech were detectable. Many studies also agree that utterances get overall shorter with increased age, that seniors produce less correct verb tenses and also other correct morphological forms.

These differences influence the performance of human-computer interfaces based on speech (Pellegrini et al., 2012; Schultz, 2007). Although being a stable characteristic when compared with the awareness and emotional state of a speaker, age influences the acoustic signal and the performance of a SR engine, as several parameters of the speech wave form are modified, such as fundamental frequency, first and second formants (Albuquerque et al., 2014), jitter, shimmer and harmonic noise ratio (Xue and Hao, 2003). In brief, ASR systems trained with young adults speech, perform significantly worse when used by the elderly population, due to the various mentioned factors (Wilpon and Jacobsen, 1996).

The typical strategy to improve ASR performance under these cases is to collect speech data from elderly speakers in the specific domain of the target application and train elderly-only or adapted acoustic models (Anderson et al., 1999; Baba et al., 2002; Vipperla et al., 2009). However, there is a considerable cost and effort associated with these collections (Hämäläinen et al., 2012).

Specifically for EP, recent initiatives from the research community to improve speech technologies can be found in the literature (Hämäläinen et al., 2012; Oliveira et al., 2013).

1.2. Problem Statement

To the best of our knowledge, before this thesis, no SSI system existed for EP, leaving native speakers with speech impairments unable to interact with HCI systems based on speech. Furthermore, no study or analysis has been made regarding the adoption of a new language with distinctive characteristics, to this kind of systems, and the problems that may arise from applying existent work to EP remain unknown. A particularly relevant characteristic of EP are the nasal sounds, which may pose problems to several SSI modalities.

Current research on SSIs has shown that it is possible to achieve (some) communication in the absence of an acoustic signal by extracting information from other sources of the human speech production process. The notion of an SSI system entails that no audible acoustic signal is available, requiring speech information to be extracted from articulators, facial muscle movement or brain activity. Today´s researchers of SSI struggle to design modalities that allow them to achieve satisfactory accuracy rates without a high degree of invasiveness. Considering a real world scenario, this often leads to unpractical and invasive solutions due to the difficulty in extracting silent speech information using current technologies.

Nonetheless, invasive solutions such as RT-MRI, allows us to extract more direct measures of the human speech production process. As such, this group of modalities should not be discarded, because they can provide valuable ground truth information for better understanding less invasive modalities. However, combining distinct modalities poses additional problems, such as synchronous acquisition of data. Furthermore, after acquiring the data, it is necessary to extract relevant information from high-dimensional datasets and find the means to take advantage of complementary strengths to solve the weaknesses of other modalities.

In short, the problem addressed by this thesis can be defined as follows: SSI systems usually rely on a single modality incapable of handling all types of users. Thus, how can we leverage not only existing SSI approaches, but other technologies (which allow measuring speech production) as well, to reduce existing limitations and to create a multimodal non-invasive interface adapted for EP, an unsolved challenge in SSI.

1.3. Thesis Hypotheses

For this thesis the following hypotheses were considered:

1) Our first and most obvious hypothesis is that it is possible to extend/adapt the work on SSI for languages, such as English, to European Portuguese.
2) Our second hypothesis is that, although nasal sounds, particularly relevant in EP, pose problems to some of the considered SSI modalities, we believe that their detection with less invasive input HCI modalities is possible.
3) The third hypothesis is that a multimodal HCI approach, based on less invasive modalities, has the potential to improve recognition results when compared with results from a single HCI modality.
4) Our fourth hypothesis is that supplementary direct measures acquired from more invasive HCI modalities in multimodal setups, can be used as ground truth to enhance our comprehension and explore information provided by less invasive modalities.

1.4. Objectives

In order to satisfy and demonstrate the 4 hypotheses stated above, the following objectives were defined:

- **For Hypothesis 1: SSI language expansion to support European Portuguese (Objective 1 or O1)** – The adaptation of SSI to a new language and the procedures involved constitutes by itself an advance to the current scientific knowledge in this area. To satisfy the first hypothesis, we have addressed the challenges of developing an SSI for EP, the first approach for this language in the Portuguese and international research community. Since no SSI existed for European Portuguese, it was not possible to perform a comparison evaluation based on linguistic terms. Using the techniques described in literature and adapting them to a new language, provides novel information useful for language independence and language adoption techniques.
- **For Hypothesis 2: Detect EP nasality with the multimodal SSI, by overcoming identified limitations of current techniques (Objective 2 or O2)** – One of the areas of research to address is the problem of recognizing nasal sounds of EP, as pointed out by (Denby et al., 2010). Considering the particular nasal characteristics associated with EP, we were expecting to see performance deterioration in terms of recognition rates and accuracy using existent approaches. Upon occurrence of such phenomenon, the root of the system performance deterioration have been identified and new nasality detection techniques based on that information have been thought. For example, in a given HCI modality, by adding a sensor that is able to capture the missing information required for nasality detection. This

allowed us to conclude on the particular aspects that influence language expansion, language independency, limitations of SSIs for the EP case and to address the second hypothesis.

- **For Hypothesis 3: Develop a Multimodal HCI approach for SSI (Objective 3 or O3)** – From the recognition perspective, an SSI can be implemented using several types of sensors working separately to feed independent and concurrent input HCI modalities, or adopting a multimodal combination of them, in order to achieve better recognition results. For this work we have preferably adopted the less invasive approaches and sensors that are able to work both in silent and noisy environments and work for elderly, post-laryngectomy patients or similar. The input HCI modalities considered in this thesis were, Surface Electromyography (SEMG), Ultrasonic Doppler Sensing (UDS), (RGB) Video and Depth. The corresponding input data streams were captured by appropriate sensory devices, being the Ultrasonic Doppler sensory equipment (unavailable in the market), developed in-house in the context of this thesis. Further investigation was also conducted on silent speech processing, respectively on data acquisition of the mentioned input data streams collected by appropriate sensing devices, followed by feature extraction, and classification, as well as on combining techniques through data collection with multiple sensory devices, data fusion and solving asynchrony issues verified in different signals (Srinivasan et al., 2010) in order to complement and overcome the inherent shortcomings of some approaches without compromising the usability of the system. To measure the performance and accuracy of the developed system, typical measures found in the literature of this area and obtained during the experiments where used, such as recognition accuracy, word error rate and mutual information or correlation measures. The comparison with existent systems depended on the used sensors and techniques. In the comparison of accuracy results several factors have been taken into account. Thus, our comparison analysis, considered aspects such as the type of sensors, the type of corpora used to train the classification model, the classification models being used, if we were considering murmurs, regular, silent or unspoken speech, the characteristics of the test set, etc.
- **For Hypothesis 4: Explore more direct measures in the Multimodal SSI for EP (Objective 4 or O4) –** The human speech production process is composed by several stages that go from intention to articulation effects (Levelt, 1995). A complete representation of these stages allows for a comprehensive assessment of the speech process and consequently, for an improvement of the performance of the envisaged SSI systems. Nonetheless, to obtain more direct measures of the structures involved in speech production (e.g. articulators), more invasive modalities are often required. Our aim has been to use some of these invasive modalities, such as RT-MRI or Ultrasonic Imaging, which are not appropriate for being

deployed in mainstream universal HCI, but do help us increase the knowledge about less invasive modalities, provide ground-truth information and enhance our comprehension of the limitations and potential of such sources of information.

1.5. Publications

The outcomes in the context of, or related with this thesis, can be grouped into three main areas: 1) state-of-the-art update; 2) individual SSI modalities applied to EP; 3) multimodal SSI for EP. The dissemination of these achievements was made through the publication and presentation of scientific papers in peer-reviewed conferences and journals, as described below.

1. **State-of-the-art update**
   - A state-of-the-art overview of SSI, including the latest related research and a new taxonomy that associates each type of SSI to a stage of the human speech production process was published in the book chapter below and also presented as a key note at a symposium in Brazil. The remaining publications also include a continuous update of the state-of-the-art analysis.
     - **Freitas, J.**, Teixeira, A., Dias, M.S., Bastos, C., 2011. "Towards a Multimodal Silent Speech Interface for European Portuguese", in: Speech Technologies. Ivo Ipsic (Ed.), InTech. doi:10.5772/16935 (Book Chapter) http://www.intechopen.com/articles/show/title/towards-a-multimodal-silent-speech-interface-for-european-portuguese
     - Dias, M.S., **Freitas, J.**, 2011. "Tendências da pesquisa em realidade virtual e aumentada e interação", SVR 2011, XIII Symposium on Virtual and Augmented Reality, Uberlândia, Brasil, Maio 23-26 2011.

2. **Single SSI modalities applied to EP**
   - In the context of this thesis, several SSI modalities such as SEMG and UDS were applied for the first time to EP. These preliminary studies allowed to detect accuracy performance degradation between minimal pairs of words that only differ on nasality of one of the phones, using state-of-the-art techniques.
     - **Freitas, J.**, Teixeira, A., Dias, M.S., 2012. "Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge", in: Int. Conf. on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2012). pp. 91–100. http://www.scitepress.org/DigitalLibrary/Index/DOI/10.5220/0003786100910100

- ○ **Freitas, J**., Teixeira, A., Vaz, F., Dias, M.S., 2012. "Automatic Speech Recognition Based on Ultrasonic Doppler Sensing for European Portuguese", in: Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 227–236. doi:10.1007/978-3-642-35292-8_24 (Book Chapter)
  http://link.springer.com/chapter/10.1007/978-3-642-35292-8_24

3. Multimodal SSI for EP

- In our research for the best combination of SSI modalities for EP, we have developed a multimodal framework that enables obtaining synchronous data from several modalities, simultaneously promoting their joint usage. With this framework we have collected several datasets, some of them containing up to six distinct sources of information synchronously acquired.
  - ○ **Freitas, J.**, Teixeira, A., Dias, M.S., 2013. "Multimodal Silent Speech Interface based on Video, Depth, Surface Electromyography and Ultrasonic Doppler: Data Collection and First Recognition Results", in: Int. Workshop on Speech Production in Automatic Speech Recognition. Lyon.
    http://ttic.uchicago.edu/~klivescu/SPASR2013/final_submissions/freitas_SPASR2013.pdf
  - ○ **Freitas, J.**, Teixeira, A., Dias, M.S., 2014. "Multimodal Corpora for Silent Speech Interaction", in: 9th Language Resources and Evaluation Conference. pp. 1–5.
    http://www.lrec-conf.org/proceedings/lrec2014/pdf/264_Paper.pdf

- Published applications of the developed framework include: (1) study the use of feature selection techniques for handling high-dimensionality scenarios such as the one found when dealing with multiple data streams; (2) analysis of tongue movement detection with surface EMG using ultrasound imaging as ground information; (3) nasality detection based on less invasive modalities such as SEMG and UDS using velum information extracted from RT-MRI to interpret other modalities data.
  - ○ **Freitas, J.**, Ferreira, A., Figueiredo, M., Teixeira, A., Dias, M.S., 2014. "Enhancing Multimodal Silent Speech Interfaces with Feature Selection", in: 15th Annual Conf. of the Int. Speech Communication Association (Interspeech 2014). pp. 1169–1173 Singapore.
  - ○ **Freitas, J.**, Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014. "Assessing the Applicability of Surface EMG to Tongue Gesture Detection", to appear in: Proceedings of IberSPEECH 2014 and Lecture Notes in Artificial Intelligence (LNAI) 8854, pp. 189-198, published by Springer. (Book Chapter)
  - ○ **Freitas, J.**, Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014. "Velum Movement Detection based on Surface Electromyography for Speech Interface", in: Int. Conf. on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2014). pp. 13–20.

http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0004741100130020

- **Freitas, J.**, Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014. "Velar Movement Assessment for Speech Interfaces: An exploratory study using Surface Electromyography", to appear in: Communications in Computer and Information Science (CCIS) published by Springer-Verlag. (Book Chapter)
- **Freitas, J.**, Teixeira, A., Dias, M.S., 2014. "Silent Speech for Human-Computer Interaction", in: Doctoral Consortium at Int. Conf. on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2014).
- **Freitas, J.**, Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014. "Detecting Nasal Vowels in Speech Interfaces based on Surface Electromyography", revised version under review in: Plos ONE (impact factor 2013 - 3.534).
- **Freitas, J.**, Teixeira, A., Dias, M.S., 2014. "Can Ultrasonic Doppler Help Detecting Nasality for Silent Speech Interfaces? - An Exploratory Analysis based on Alignement of the Doppler Signal with Velum Aperture Information from Real-Time MRI", in: Int. Conf. on Physiological Computing Systems (PhyCS 2014). pp. 232 – 239. http://www.scitepress.org/DigitalLibrary/Index/DOI/10.5220/0004725902320239

Other outcomes related with this thesis include, establishing elderly users' requirements, design and development of a multimodal SSI prototype and a collaborative project between academia and industry, as follows:

- Elderly are amongst the probable users of an SSI due to the inherent problems related with aging and the characteristics of elderly speech. To cope with this aspect, we have studied the limitations and requirements imposed by elderly when using a multimodal HCI, where speech is one of the analyzed and most important interaction modalities.
    - Júdice, A., **Freitas, J.**, Braga, D., Calado, A., Dias, M.S., Teixeira, A., Oliveira, C., 2010. "Elderly Speech Collection for Speech Recognition Based on Crowd Sourcing", in: Proceedings of DSAI Software Development for Enhancing Accessibility and Fighting Info-exclusion (DSAI 2010), Oxford, UK. http://www.academia.edu/2665364/Elderly_Speech_Collection_for_Speech_Recognition_Based_on_Crowd_Sourcing
    - Hamalainen, A., Pinto, F., Dias, M.S., Júdice, A., **Freitas, J.**, Pires, C., Teixeira, V., Calado A., Braga, D., 2012. "The First European Portuguese Elderly Speech Corpus", IberSPEECH 2012 Conference, Madrid, Spain. http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/The_First_European_Elderly_Speech_Corpus_v1.pdf

  - Teixeira, V., Pires, C., Pinto, F., **Freitas, J.**, Dias, M.S., Rodrigues, E.M., 2012. "Towards elderly social integration using a multimodal human-computer interface", in: Workshop on AAL Latest Solutions, Trends and Applications (AAL 2012). http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/Paper_Towards_elderly_social_integration_using_MMUI.PDF
  - Dias, M.S., Pires, C., Pinto, F., Teixeira, V., **Freitas, J.**, 2012. "Multimodal user interfaces to improve social integration of elderly and mobility impaired", in: Studies in Health Technology and Informatics, IOS Press, vol. 177, pHealth 2012, pp. 14-25.
    http://ebooks.iospress.nl/publication/21439
  - Oliveira, C., Albuquerque, L., Hämäläinen, A., Pinto, F.M., Dias, M.S., Júdice, A., **Freitas, J.**, Pires, C., Teixeira, V., Calado, A., Braga, D., Teixeira, A., 2013. "Tecnologias de Fala para Pessoas Idosas", in: Laboratório Vivo de Usabilidade (Living Usability Lab). ARC Publishing, pp. 167–181.
  - Albuquerque, L., Oliveira, O., Teixeira, T., Sá-Couto, P., **Freitas, J.**, Dias, M.S., 2014. "Impact of age in the production of European Portuguese vowels", in: 15th Annual Conf. of the Int. Speech Communication Association (Interspeech 2014). Singapore.

- A more concrete outcome of this thesis consists in a multimodal prototype designed for SSI research, including for EP. This multi-platform solution incorporates some of the published results but also allows for a rapid testing of new algorithms.

- The idea behind this thesis also gave origin to a research project involving industry and academic partners from 3 different countries – Portugal (Microsoft, University of Aveiro, FaceInMotion), Spain (University of Zaragoza) and Turkey (Middle East Technical University) - in the area of natural communication using SSI has also been proposed, accepted for funding and it is now in progress. The lifespan of the project goes beyond this thesis and will provide the means for the continuity of this line of research.
  - Marie Curie actions, Industry-Academia Partnerships and Pathways, IRIS: Towards Natural Interaction and Communication (FP7-PEOPLE-2013-IAPP, ref. 610986).

1.6. Thesis Outline

The remainder of this document is organized into six chapters. The contents of each chapter are the following:

After describing our motivation, problem, hypotheses, objectives and related publications in chapter I, in **Chapter II – Background and Related Work –** we start with fundamental knowledge about the human speech production process, the technologies used to measure it, followed by a summary of SSI history and related work. This chapter includes an updated version of the information included in (Denby et al., 2010; Freitas et al., 2011) and an analysis of existing multimodal approaches for this area. Additionally, we also provide some background on two different approaches for extracting and modelling the most relevant information from SSI modalities – articulatory features and feature selection. We finalize the chapter with a critical assessment of the state-of-the-art techniques used in SSI.

The background knowledge and state-of-the-art analysis lead to the definition of a set of preliminary experiments with distinct modalities, chosen particularly, for their non-invasive characteristics. As described in **Chapter III - Single Modality Experiments for Portuguese -** these investigations constituted the first approach to (RGB) Video, UDS and SEMG, allowed for an initial assessment of these modalities with the previously defined objectives in mind. The experiments share similar methods and the same task of recognizing isolated words in EP. The considered word set can be divided into minimal pairs differing only by the presence or absence of nasality in one of its phones, aligned with our goal of exploring the adoption of EP and the detection of the nasality phenomena.

The results of the experiments using a single modality allowed not only to acquire hands-on experience, but also to better understand the capabilities and limitations of each modality and how they could eventually be combined. Thus, to support our objective of a multimodal HCI approach for SSI, **Chapter IV – A Multimodal Framework for Silent Speech Research –** presents a multimodal framework that enables collecting synchronous data from several input modalities, simultaneously promoting their joint usage and exploration. Along with the data acquisition processes, the framework also provides a set of methods for processing modalities, including how to extract and select the best features, with the aim of leveraging the use of the data for SSI experiments.

In **Chapter V – Multimodal SSI - Analysis, Classification and Prototyping –** we present several examples of experimental application of the framework, in line with our goal of combining and exploring multiple modalities. Thus, we include several experiments, namely, a study on tongue gestures assessment using SEMG; an example of combining modalities for movement detection and characterization of external articulators; and an analysis about the application of multiple feature selection techniques to SSI. Also in this chapter, to demonstrate the versatility of the framework, a multimodal SSI prototype is included.

The nasality detection challenge and also one of the goals of this thesis is addressed in **Chapter VI - The Challenge of Nasality Detection in Silent Speech**. This chapter can be seen as a case study which details two experiments using the framework described in Chapter 4 for different sensors – SEMG and UDS. In these experiments we explore the possibility of detecting nasality with less invasive modalities, relying on offline information extracted from RT-MRI to know essential ground truth information about the velum movement.

We conclude this thesis with **Chapter VII – Conclusions and Future Work** – where we present a brief overview of the thesis work, a discussion on how we have satisfied our 4 hypotheses and what were the key contributions of this thesis. Finally, we end this work with some directions for future work and final remarks.

CHAPTER II

Background and Related Work

"Silence surpasses speech."

Japanese proverb

Contents

In the last decades, propelled by the evolution and globalization of the telecommunication industry and related technologies, a quest for a system able to recognize speech in noisy environments and capable of dealing with speech-impaired users started to receive more attention. Among other research initiatives, such as the proposal of better acoustic models and speech enhancement, these needs originated the silent speech concept.

This chapter introduces this concept by providing some essential background on speech production, like the physiology involved in speech production, facial muscles and articulation. Each of these topics could be explored, by itself, up to great lengths, thus, we will focus on the necessary knowledge to understand the following chapters of this thesis. Afterwards, we will introduce some of the technologies that can be used for measuring and sensing speech production beyond the acoustic signal. Additionally, we will present a summary of the history of SSI and an overview of several modalities and their latest results. We will also describe some of the multimodal initiatives existing in the literature, as well as some methods to extract information from these modalities, used in posterior chapters. Finally, we end this chapter with a critical assessment of the existing SSI modalities, discussing their limitations and advantages for a multimodal approach.

2.1. The Speech Production Process

Speech production is a complex mechanism and its understanding requires knowledge from diverse fields of research such as anatomy, physiology, phonetics and linguistics. Providentially, essential and exhaustive descriptions can be found in the literature (Hardcastle, 1976; Seikel et al., 2009; The UCLA Phonetics Laboratory, 2002).

The speech production process can be divided into several stages. According to Levelt (1995), the first stage of the human speech production process starts in the brain, converting patterns of goals into messages, followed by the grammatical encoding of the preverbal message to surface structure. The next phase of the speech production is the passage from the surface structure to the phonetic plan which, stated simply, is the sequence of phones that are fed to the articulators. This phonetic plan can be divided into the electrical impulse fed into the articulators and the actual process of articulation. The final phase consists on the consequent effects of the previous phases, which results in the acoustic speech signal.

The existent experimental SSI systems described in the literature cover all the stages of speech production, from intention, to articulation, to effects, as depicted in Figure 1.

. The modalities currently in use cover all stages of human speech production, as follows:

- **Conceptualization and Formulation** (brain / Central Nerve System): Interpretation of signals from implants in the speech-motor cortex (Brumberg et al., 2010), Interpretation of signals from Electroencephalography (EEG) sensors (Porbadnigk et al., 2009);
- **Articulation control** (muscles): Surface Electromyography of the articulator muscles (Heistermann et al., 2014; Jorgensen and Dusan, 2010; Wand et al., 2013a);
- **Articulation** (movement): Capture of the movement of fixed points on the articulators using Permanent-Magnetic Articulography (PMA) sensors (Fagan et al., 2008; Hofe et

al., 2013a); Real-time characterization of the vocal tract using ultrasound (US) and optical imaging of the tongue and lips; Capturing the movements of a speaker's face through Ultrasonic Doppler sensing (Freitas et al., 2012b; Srinivasan et al., 2010);

- **Articulation effects** (vocal tract): Digital transformation of signals from a Non-Audible Murmur (NAM) microphone (Nakajima et al., 2003a; Toda, 2010); Analysis of glottal activity using electromagnetic (Holzrichter et al., 2009; Quatieri et al., 2006) or vibration (Patil and Hansen, 2010) sensors.

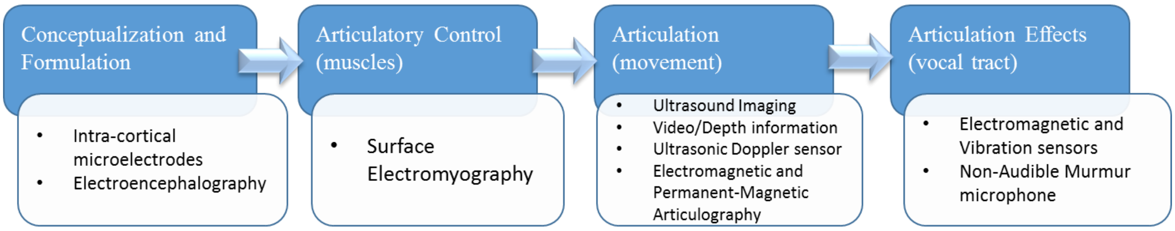

Conceptualization and Formulation
Intra-cortical microelectrodes
Electroencephalography
Articulatory Control (muscles)
Surface Electromyography
Articulation (movement)
Ultrasound Imaging
Video/Depth information
Ultrasonic Doppler sensor
Electromagnetic and Permanent-Magnetic Articulography
Articulation Effects (vocal tract)
Electromagnetic and Vibration sensors
Non-Audible Murmur microphone

Figure 1. Overview of the existing Silent Speech Interface technologies mapped into several stages of the human speech production process.

The taxonomy presented above and illustrated in Figure 1 allows us to associate each type of SSI to a stage of the human speech production process, providing a better understanding from where the speech information is extracted.

Each of these stages entails extensive background knowledge (Hardcastle, 1976; Seikel et al., 2009). Thus, in the following subsections, following the order of the various stages, we present a selective description of the speech process, focusing on the relevant and necessary topics to understand the work presented in the following chapters. Hence, we start by briefly explaining how motor control occurs for speech production. Then, we describe some of the muscles for articulation control, with an emphasis on relevant and useful muscles for SSI. Afterwards, we introduce the articulators, their position and their function.

2.1.1. Speech Motor Control

Speech production requires a particularly coordinated sequence of events to take place, being considered the most complex sequential motor task performed by humans (Seikel et al., 2009). After an intent or idea that we wish to express has been developed and coded into a language, we map it into muscle movements. This means that the motor impulse received by the primary motor cortex is the result of several steps of planning and programming that already occurred in other parts of the brain, such as Broca's area, the supplementary motor area and the pre-motor area.

The nervous system controls the activation of a motor unit and the associated activation rate. Nerve impulses are carried from anterior horn cells of the spinal column to the end of the nerve via motor neurons. The motor neurons send the signals from the brain to the exterior body parts through the axon (see Figure 2a). The motor axons are then divided into several branches that end with a neuro-muscular junction known as the motor endplate, meaning that a single motor neuron innervates several muscle fibers, as depicted in Figure 2b. The muscle fibers coalesce among several motor units.

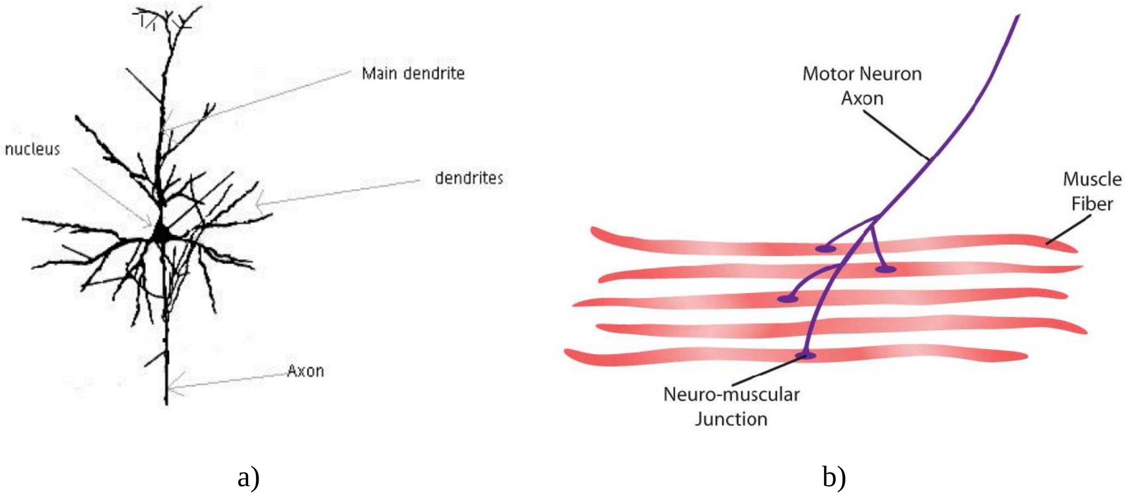


Figure 2. a) Typical neuron structure (Calliess and Schultz, 2006). b) Motor neuron innervation of muscle fibers (Betts et al., 2006).

When the nerve impulse reaches the neuromuscular junction, the neurotransmitter acetylcholine is released. This causes sodium and potassium *cation* (i.e. an ion with a positive charge) channels in the muscle fiber to activate, subsequently causing an action potential propagation (i.e. a short-lasting event in which the electrical membrane potential of a cell rapidly rises and falls) from the endplate to the muscle-tendon junction. The depolarization process and ion movement generate an electromagnetic field in the area surrounding the muscle fibers. This time-varying potential is referred to in literature as the myoelectric signal (De Luca, 1979). These electrical potential differences generated by the resistance of muscle fibers lead to voltage patterns that, when speaking, occur in the region of the face and neck and when measured at the articulatory muscles, provide means to collect information about the resultant speech. This myoelectric activity occurs independently of the acoustic signal, i.e. it occurs whether the subject produces normal, murmured or silent speech.

2.1.2. Articulatory Muscles in Speech Production

In the speech production process the facial muscles represent a vital role since they help to shape the air stream into recognizable speech. Thus, muscles related with lip movement, tongue and mandibular movement will have a strong influence on speech production. Below, some of the main muscles of the face and neck used in speech production are described (Hardcastle, 1976):

- *Orbicularis oris*: This muscle can be used for rounding and closing the lips, pulling the lips against the teeth or adducting the lips. It is considered the sphincter muscle of the face. Since its fibers run in several directions, many other muscles blend in with it;
- *Levator anguli oris*: This muscle is responsible for raising the upper corner of the mouth and may assist in closing the mouth by raising the lower lip for the closure phase in bilabial consonants;
- *Zygomaticus major:* this muscle is used to retract the angles of the mouth. It has influence on the production of labiodental fricatives and the on the production of the [s] sound;
- *Platysma:* The *platysma* is responsible for aiding the *depressor anguli oris* muscle, lowering the bottom corners of the lips. The platysma is the closest muscle to the surface in the neck area as depicted in Figure 3.
- *Tongue*: The tongue plays a fundamental role in speech articulation and is divided into intrinsic and extrinsic muscles. The intrinsic muscles (*Superior and Inferior Longitudinal; Transverse*) mostly influence the shape of the tongue, aiding in palatal and alveolar stops, in the production of the [s] sound by making the seal between the upper and lower teeth, and in the articulation of back vowels and velar consonants. The extrinsic muscles (*Genioglossus; Hyoglossus; Styloglossus; and Palatoglossus*) are responsible for changing the position of the tongue in the mouth as well as its shape, and are important in the production of most of the sounds articulated in the front of the mouth, in the production of the vowels and velars, and in the release of alveolar stop consonants. They also contribute to the subtle adjustment of grooved fricatives.
- *Anterior Belly of the Digastric*: this is one of the muscles used to lower the mandible. Its function is to pull the hyoid bone and the tongue up and forward for alveolar and high frontal vowel articulations and raising pitch.

Figure 3 illustrates these and other muscles found in the areas of the face and neck, providing a glimpse of the complexity in terms of muscle physiology in this area.

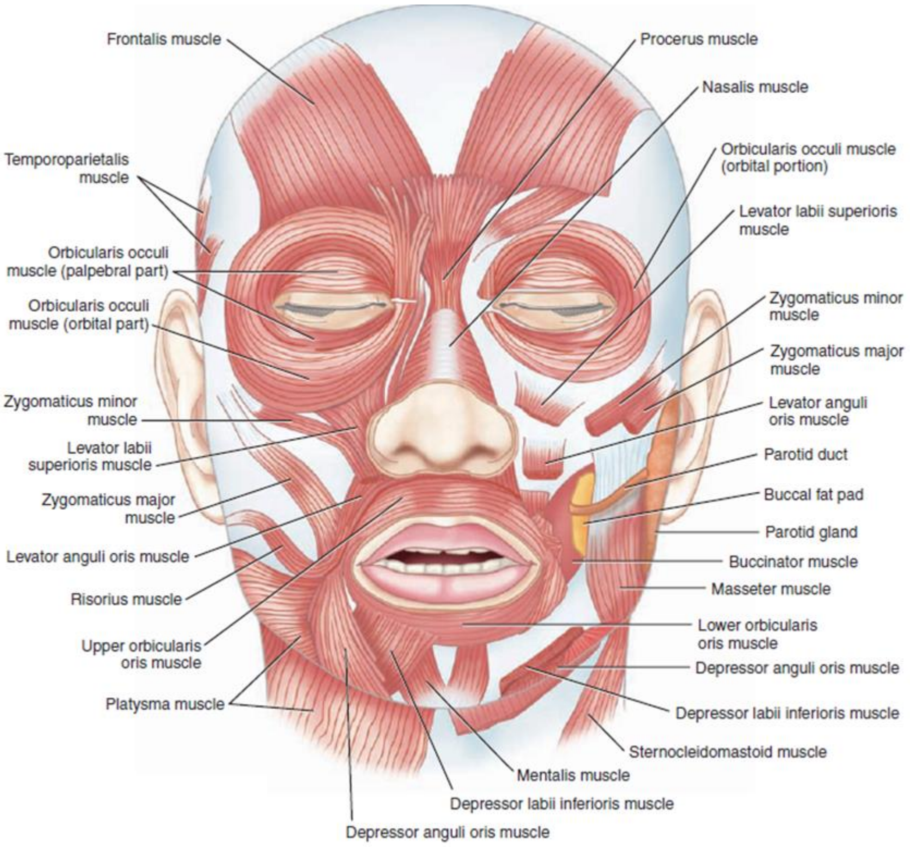

Frontalis muscle
Procerus muscle
Nasalis muscle
Temporoparietalis muscle
Orbicularis occuli muscle (orbital portion)
Levator labii superioris muscle
Orbicularis occuli muscle (palpebral part)
Orbicularis occuli muscle (orbital part)
Zygomaticus minor muscle
Zygomaticus major muscle
Zygomaticus minor muscle
Levator anguli oris muscle
Levator labii superioris muscle
Parotid duct
Zygomaticus major muscle
Buccal fat pad
Parotid gland
Levator anguli oris muscle
Buccinator muscle
Risorius muscle
Masseter muscle
Lower orbicularis oris muscle
Upper orbicularis oris muscle
Depressor anguli oris muscle
Platysma muscle
Depressor labii inferioris muscle
Sternocleidomastoid muscle
Mentalis muscle
Depressor labii inferioris muscle
Depressor anguli oris muscle

Figure 3. Human facial muscles (Seikel et al., 2009).

2.1.2.1. Velum Related Muscles

As some of the objectives of this thesis are concerned with nasality, this section provides a notion of how this process occurs and what muscles are involved.

In general, the production of a nasal sound involves air flow through the oral and nasal cavities. This air passage through the nasal cavity is essentially controlled by the velum which, when lowered, allows for the velopharyngeal port to be open, enabling resonance in the nasal cavity, which causes the sound to be perceived as nasal. The production of oral sounds occurs when the velum is raised and the access to the nasal cavity is closed (Beddor, 1993).

The process of moving the soft palate involves the following muscles (Fritzell, 1969; Hardcastle, 1976; Seikel et al., 2009), also depicted in Figure 4:

- *Levator veli palatini*: This muscle has its origin in the inferior surface of the apex of the petrous part of the temporal bone and its insertion in the superior surface of the palatine aponeurosis. Its main function is to elevate and retract the soft palate achieving velopharyngeal closure;

- *Musculus uvulae:* This muscle is integrated in the structure of the soft palate. In speech it helps velopharyngeal closure by filling the space between the elevated velum and the posterior pharyngeal wall (Kuehn et al., 1988);
- *Superior pharyngeal constrictor*: Although this is a pharyngeal muscle, when it contracts it narrows the pharynx upper wall, which elevates the soft palate;
- *Tensor veli palatini*: This muscle tenses and spreads the soft palate and assists the *levator veli palatine* in elevating it. It also dilates the Eustachian tube. This muscle is innervated by means of the mandibular nerve of the V trigeminal, and not by the XI accessory nerve, as the remaining muscles of the soft palate. It is the only muscle of the soft palate that is innervated by a different nerve;
- *Palatoglossus*: Along with gravity, relaxation of the above-mentioned muscles and the *Palatopharyngeous*, this muscle is responsible for the lowering of the soft palate.

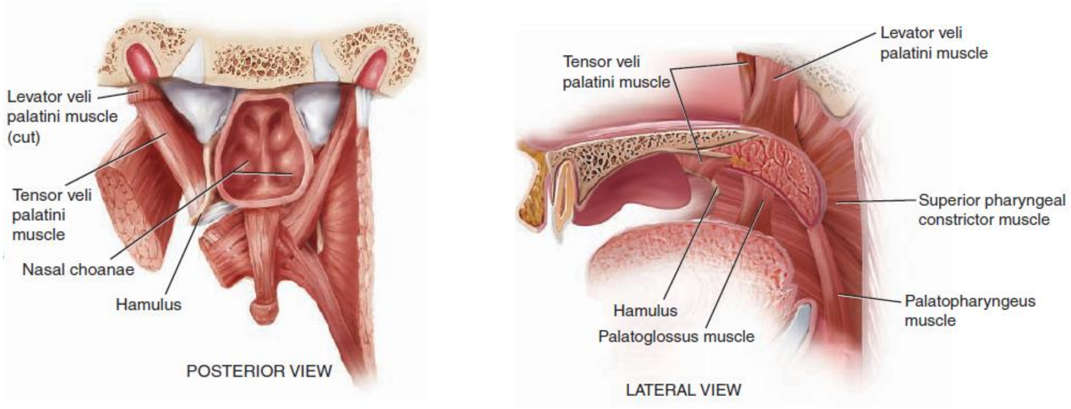

Levator veli palatini muscle (cut)
Tensor veli palatini muscle
Nasal choanae
Hamulus
POSTERIOR VIEW
Tensor veli palatini muscle
Levator veli palatini muscle
Superior pharyngeal constrictor muscle
Palatopharyngeus muscle
Hamulus
Palatoglossus muscle
LATERAL VIEW

Figure 4. Muscles of the soft palate from posterior (left), and the side (right) view (Seikel et al., 2009).

2.1.3. Articulators

Articulation describes how humans produce speech sounds and which speech organs are involved in this process. The positioning of the articulators defines the articulatory and resonant characteristics of the vocal tract. Although all surfaces and cavities of the vocal tract are contributors to the production of speech, some parts of the articulatory system are more influential than others.

In the literature we find articulators split into two groups: mobile and passive. The mobile articulators are usually positioned in relation to a passive articulator, through muscular action, to achieve different sounds. Mobile articulators are the tongue, lower jaw, velum, lips, cheeks, oral cavity (fauces and pharynx), larynx and the hyoid bone. The passive articulators are the alveolar ridge of the upper jaw, the hard palate and the teeth (Seikel et al., 2009). Both groups of articulators are depicted in Figure 5.

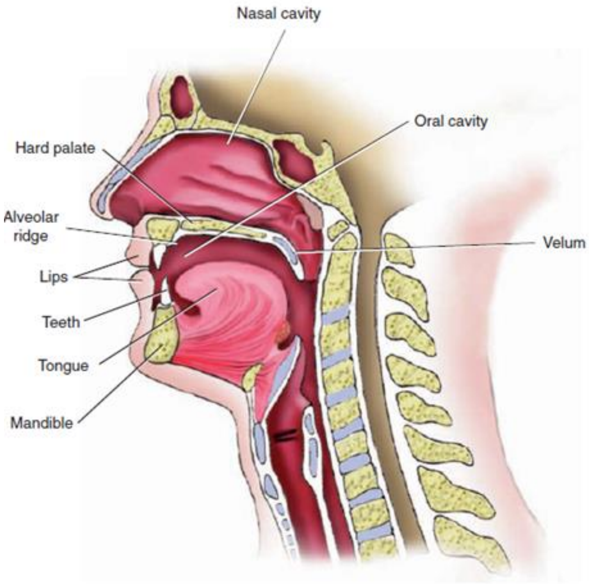


Figure 5. Sagittal view of the articulators (Seikel et al., 2009).

2.2. Measuring Speech Production

An SSI obtains information from one or more parts of the human speech production process and to capture this information, one needs adequate technology. Considering research in SSI and related topics from the last two decades, we see in the literature an increasing number of technologies, which act as non-acoustic sensors, used to implement an SSI system. Also, in part motivated by speech production research, considerable amount of groundwork can be found for more invasive technologies with the ability to extract detailed real-time information about the human speech production process such as RT-MRI. These technologies, although not appropriate for the HCI context, provide us the possibility of acquiring direct measures of the articulators, holding an enormous potential for SSI research in the sense that a more comprehensive understanding of other modalities can be achieved.

Each of these technologies captures speech information at different levels, beyond the conventional acoustic manifestations, as summarized in Table 1, showing that a complete representation of the speech production process is in our reach, as suggested by (Jou, 2008).

Table 1. Non-exhaustive list of technologies able to capture speech related information.

Technology	Information Captured
Intra-cortical microelectrodes	Electrical brain signals from the speech-motor cortex
Electroencephalography (EEG)	Brain activity usually acquired from Broca's and Wernicke's areas by recording of electrical activity
Magnetoencephalography (MEG)	Brain activity by recording magnetic fields produced by electrical currents occurring in the brain
Functional magnetic resonance imaging (fMRI)	Brain activity detected via associated changes in blood flow
Surface Electromyography (SEMG)	Electrical muscle activity from facial and neck muscles
Ultrasound Imaging (US)	Mainly tongue position and other structures such as the hyoid bone, the position of the short tendon, visible muscles such as the *genioglossus* and also fat below the tongue
RGB Video cameras	Visual information of the lips, chin, cheeks and mandible
Depth cameras	Depth information of the face
Ultrasonic Doppler sensing (UDS)	Doppler shifts caused by articulators movement
Electromagnetic Articulography (EMA) or Permanent-Magnetic Articulography (PMA)	3D position information of fixed points on the articulators during speech
Optopalatography (OPG)/Electropalatography (EPG)	Tongue activity, by measuring the contact between the tongue and an artificial hard palate
Real-time Magnetic Resonance (RT-MRI)	Real-time imaging information of the anatomy and function of the articulators
Non-Audible Murmur (NAM) microphone	Low amplitude sounds generated by laryngeal airflow noise and its resonance in the vocal tract
Electromagnetic sensors (e.g. General electromagnetic motion sensor (GEMS), Electroglottograph (EGG))	Glottal/Tissue movement during vocal cords vibration
Vibration sensors (e.g. Throat microphones, Physiological microphone (P-MIC), Bone, In-Ear, Vibrocervigraphic microphones)	Glottal vibrations transmitted through tissue and bone
Respiration sensor	Information on abdominal or thoracic respiration, or nose/mouth air flow

More technologies can be used for measuring speech production in general, such as video-based fluoroscopy systems (i.e. an imaging technique that uses X-rays to obtain real-time moving images) or similar. For an exhaustive review of speech technologies the reader is forwarded to (Hardcastle et al., 2012).

2.2.1. Direct measures of hidden articulators

Some of the technologies listed in Table 1 allow direct measurements of the articulators. This is true for most image-based technologies such as RT-MRI, Ultrasound Imaging (US), and Video, and also for EMA or similar tracking technologies usually employed in speech production studies (Kroos, 2012). Nevertheless, it is hard to envisage a usable and natural interface with the current RT-MRI technology.

One of the notions presented in this study is to combine modalities that provide direct measurements of the articulators with less-invasive and more natural modalities. By leveraging the information captured by such modalities, we place ourselves in a privileged position to analyze signals such as EMG and UDS, which blend information (and noise) from multiple muscles or articulators, respectively. The latest work related with SSI, where more invasive modalities are used to study less-invasive modalities, include using RT-MRI to analyze velum movement with SEMG (Freitas et al., 2014c) and UDS (Freitas et al., 2014b).

In the concrete case of SEMG, one of the most important positions for placing the EMG electrodes is in the areas of neck and beneath the chin (Jorgensen and Dusan, 2010; Wand and Schultz, 2011a) capturing data related with tongue movement. However, it is not clear which movements are actually being captured. As such, a possible solution would be to understand if the information extracted from modalities such as US or EMA, which are capable of obtaining more direct measurements of the tongue, provide the necessary ground knowledge to better understand EMG signals.

In the past, in the field of phonetics, other studies using intra-oral EMG electrodes attached to the tongue have provided valuable information about the temporal and spatial organization of speech gestures. These studies have analyzed different cases such as vowel articulation (Alfonso and Baer, 1982) or defective speech gestures that happen in pathologies such as aphasia (Shankweiler et al., 1968). These studies also benefit from the fact that they are able to acquire more accurate measures, the sensors being directly placed in the articulator using intra-oral or intra-muscular EMG electrodes. The position of these sensors avoids some of the muscle cross-talk and superposition.

There are also studies of the tongue that use other technologies such as RT-MRI (Bresch et al., 2008; Narayanan et al., 2011), Cinefluorography (Kent, 1972), Ultrasound, using a headset to permit natural head movement (Scobbie et al., 2008) and ElectroPalatoGraphy (EPG) (Stone and Lundberg, 1996), which allow us to get a very good understanding of the tongue shapes and movements during speech.

2.2.2. Measuring Velum Activity

Measuring the velum state or its movement is of great interest to phonetics and speech production research. Some of the techniques mentioned in Table 1, such as RT-MRI, or radiography-based techniques provide accurate information about the velum state, however, they are, in general, complex and costly to use. However, more specific techniques for clinical environments that provide information about the velum can be found in the literature.

Existing techniques can be grouped into three types: optical, mechanical and acoustic (Birkholz et al., 2014). Optical solutions are based on the transmission of light via the velopharyngeal port, using a fiber optic tube placed through the nasal cavity into the pharynx (Dalston, 1982, 1989). Mechanical techniques include techniques that allow the measurement of the pressure made by velar movement. This can be captured by an internal lever resting on the upper side of the velum (Bell-Berti et al., 1993; Horiguchi and Bell-Berti, 1987); a thin spring wire and a resistance strain gauge that is fixed to the molars and touches the velum at the bottom side (Moller et al., 1971); or using an older hydrokinetic technique that measures the nasal and oropharyngeal pressure difference (Warren and DuBois, 1964). Acoustic methods have the advantage of being less invasive when compared with the previous types, and as such several implementations can be found. Part of these systems (Bressmann, 2005; Watterson et al., 2005) work by measuring the acoustic output from the oral and nasal cavities' openings, and calculating the ratio between them. Another technique, called "acoustic rhinometry", which is also included in this category, consists of measuring the acoustic reflection to detect changes in velar positioning (Hilberg et al., 1989; Seaver et al., 1995). A more recent study presented a frustum-shaped capsule containing a miniature speaker that emits low-frequency ultrasound bursts and a microphone that captures the echo modelled by the nasal cavity (Birkholz et al., 2014). Other similar approaches can be found in the literature, such as measuring the nasal vibration with an accelerometer microphone attached to the outer part of the nose (Tronnier, 1998).

In previous studies, EMG has also been applied to measure the level of activity of the velum muscles by means of intramuscular electrodes (Bell-Berti, 1976; Fritzell, 1969) and surface electrodes positioned directly on the oral surface of the soft palate (Kuehn et al., 1982; Lubker, 1968). No literature exists in terms of detecting the muscles involved in the velopharyngeal function with surface EMG electrodes placed on the face and neck. Previous studies in the lumbar spine region have shown that if proper electrode positioning is considered, a representation of deeper muscles can be acquired (McGill et al., 1996). This raises a question that remains as of yet unanswered: is SEMG positioned in the face and neck regions able to detect activity of the muscles related to nasal port opening/closing and consequently detect the nasality phenomenon? Another related question that can be raised is how we can show, with some confidence, that the signal we are seeing is in fact the myoelectric signal generated by the velum movement and not by spurious movements caused by neighboring muscles unrelated to the velopharyngeal function.

2.3. A Brief History of Silent Speech

In this section a brief summary of SSI history will be presented. For a more detailed history framework the reader is pointed to Denby et al. (2010). Silent speech interpretation through an electronic system or computer came to the attention of the community as early as 1968. The idea of visual speech recognition was spread by Stanley Kubrick's 1968 science-fiction film "2001 – A Space Odyssey", where a spaceship computer – "HAL 9000" – discovers a plot by analyzing a conversation between two astronauts with a video camera. However, it took more than a decade for real solutions to appear. An example of this is the automatic visual lip-reading by Petajan (1984) and the patents registered for lip-reading equipment by Nakamura (1988), which were proposed as an enhancement to ASR in noisy environments. In 1985 an SEMG-based speech prosthesis was developed by Sugie and Tsunoda (1985) and achieved an average correction rate of 64% when recognizing 5 Japanese vowels. Almost simultaneously, Morse and O'Brien (1986) applied four EMG steel surface electrodes for recognizing two words at first, and a few years later applied the same technique on a ten word vocabulary problem, with accuracy rates around 60% using a neural networks based classifier (Morse et al., 1991). Working on a similar problem, Hasegawa and Ohtani (1992) achieved a 91% recognition rate using a video of the speaker's face, from which lip and tongue features were extracted.

In the 90's with the massive adoption of cellular telephones, SSIs started to appear as a possible solution for problems such as privacy in personal communications, and for users who had lost their capacity to produce voiced speech. In the early 2000's DARPA (Defense Advanced Research Projects Agency) focused on recovering glottal excitation cues from voiced speech in noisy environments, with the Advanced Speech Encoding Program and in 2002, in Japan, an NTT DoCoMo (Japan mobile service provider) press release announced a silent cellphone prototype using EMG and optical capture of lip movement (Fitzpatrick, 2002), specially targeting environment noise and speech-impaired users.

In recent years, the SSI concept became more prominent in the speech research community and diverse modalities were used to drive SSI research. Among the chosen modalities we can find more invasive modalities such as intra-cortical electrodes and non-obtrusive modalities such as Video or UDS. In the section 2.4, a more detailed overview of the state-of-the-art of each modality is presented.

2.4. An Overview of Existing Modalities for SSI

Considering our objective of finding the best modalities to implement a multimodal solution, we first searched in the literature for the most suitable ones. Table 2 summarizes the approaches found in the literature chosen for designing and implementing an SSI. This table enables a comparison between the modalities based on their main advantages, limitations, if they are invasive or not and if they are obtrusive or not. What is an invasive or not invasive modality strongly depends on the perspective of the reader. For this thesis, we define as invasive modalities those that require medical expertise or permanent attachment of sensors, and as obtrusive those that require "wearing" or equipping the sensor in a non-ubiquitous way. For example, EMA/PMA may be considered a borderline case, since it requires the permanent attachment of sensors (typically using dental or chirurgical glue), but depending on the location (e.g. velum vs. lips) it may (not) require a high degree of medical expertise. As such, in this thesis, although we refer to EMA/PMA as an invasive modality, when compared with others that require a chirurgical intervention it is acceptable to classify it as non-invasive.

Table 2. Single modalities overview with the respective advantages, limitations and whether or not they are considered to be invasive and obtrusive.

Modality	Main Advantages	Limitations	Invasive/ Obtrusive
Interpretation of signals from implants in the speech-motor cortex (Brumberg et al., 2010)	Better signal-noise ratio; More accurate and durable positioning;	Highly invasive; Requires medical expertise;	Yes / No
Interpretation of signals from electro encephalographic (EEG) sensors (Porbadnigk et al., 2009)	Recognizes unspoken speech; Setup is far less complex when compared with other BCI's;	Low recognition rates; Reduced vocabularies; Requires a learning process;	No / Yes
Surface Electromyography of the articulator muscles (Heistermann et al., 2014; Jorgensen and Dusan, 2010; Wand et al., 2013b)	Achieved promising results in the literature; Captures information related with the control of visible and hidden articulators, such as the tongue or even the velum;	Sensitive to positioning and user physiology; Facial electrodes connected through wires (may be mitigated through the use of a facemask); Noise caused by the superposition of facial muscles;	No / Yes
Video (optical imaging) and Depth Information from 3D cameras	Widely available; Does not require glottal activity;	Only captures visible articulators (e.g. lips);	No / No

Real- time characterization of the tongue using ultra-sound (Hueber et al., 2008)	Achieved good results when combined with video;	Probe stabilization; Only allows for tongue (and surrounding structures) visualization;	No / Yes
Capture movements of a talker's face through ultrasonic sensing devices (Freitas et al., 2012b; Srinivasan et al., 2010);	Low cost; Accessible equipment; Can be easily incorporated in other devices such as smartphones;	Sensitive to movement; Speech generates short frequency variations; Sensitive to distance and speaker variations;	No / No
Capture of the movement of fixed points on the articulators using Electromagnetic Articulography (EMA) sensors or permanent magnets detected by magnetic sensors positioned around the user's head, referred as Permanent Magnetic Articulography (PMA) (Fagan et al., 2008; Hofe et al., 2013b)	Accurate tracking of the articulators; Provides direct measures of articulators' movement;	Requires permanent fixing of the magnetic beads; Some users may experience uncomforted with magnetic beads in more hidden articulators such as the velum; Complex setup;	Yes / Yes
Digital transformation of signals from a Non-Audible Murmur (NAM) microphone (a type of stethoscopic microphone) (Nakajima et al., 2003b; Toda, 2010),	Low cost; Small and discrete device;	Requires an external vibrator to work with users that have had their larynx surgically removed by undergoing a laryngectomy. Susceptible to eaves-dropping; Sensitive to noise caused by clothing, hair, respiration, etc.; Glottal activity required;	No / Yes
Analysis of glottal activity using electromagnetic (Holzrichter et al., 2009; Quatieri et al., 2006), or vibration sensors (Patil and Hansen, 2010)	Works well in noisy environments; Captures information of the vocal cords' movement and vibration;	Radar waves may raise ethical issues; Does not work well with speech-impaired users that have suffered a laryngectomy; Glottal activity required;	No/ Depends on the type of sensor

Several approaches have been employed in the development of experimental SSI systems, nonetheless it is difficult to make a fair comparison between them because their results are dependent on the chosen modality, the extracted features and the selected classification models and characteristics that strongly influence the performance of the SSI. Furthermore, other factors need to be considered when comparing SSI results. In terms of accuracy, rates present a great variation depending on the following additional factors:

- **Language**: To the best of our knowledge, Silent Speech prototypes have been mainly developed and designed for English, with some exceptions for Portuguese (Freitas et al., 2012b), French (Tran et al., 2009), Japanese (Toda et al., 2009) and Arabic (Fraiwan et al., 2011). However, language characteristics such as nasality have been proven to influence performance (Freitas et al., 2012a).
- **Vocabulary and speech units' size**: The size of the vocabulary and the speech units also influence the performance of an SSI. In an initial stage, SSI approaches tend to start by recognizing isolated words (usually digits) (Betts et al., 2006; Florescu et al., 2010; Zhu et al., 2007), later evolving for continuous speech recognition scenarios using larger vocabularies and phoneme-based acoustic models (Hueber et al., 2012; Wand and Schultz, 2011b).
- **Corpus size and number of repetitions**: Another important aspect to consider is the number of repetitions of each speech unit. For global-data models, such as Hidden Markov Models (HMM), in order to obtain a more complete representation of each unit, many representations are required. However, when example-based approaches (De Wachter et al., 2007) are considered, one of the advantages is the small size of the required dataset, which consequently reduces the cost associated with the respective data collections. This becomes of extreme importance with novel approaches using early prototypes, where many variables that need to be defined in data acquisition sessions are yet unclear.
- **Speaker independence**: Many of the SSI techniques depend on the physiology and anatomy of the speaker. The acquired signals vary strongly between speakers, speaker independence being one of the challenges targeted by the research community in the last years (Denby et al., 2010; Wand and Schultz, 2011a).
- **Acoustic feedback and user experience with SSI**: Error rates tend to improve substantially when considering audible speech articulation with modalities like SEMG (Wand et al., 2011) or UDS (Freitas et al., 2012b), as opposed to silent speech articulation. It is also relevant for performance whether or not a user has experience with the SSI, i.e. knows how the modality works and it is accustomed to silent articulation of speech (Wand et al., 2011).
- **Acquisition setup**: Within the same modality several types of hardware devices can be used. For example, in SEMG, there are techniques that use ring-shaped electrodes wrapped around the thumb and two fingers (Manabe, 2003), array-based (Wand et al., 2013b) and more

classic techniques using bipolar and monopolar configurations of electrodes pairs that range from 7 (Maier-Hein et al., 2005) to a single pairs of electrodes (Betts et al., 2006).

In the following subsections, a brief description and current status is provided for each modality, with a stronger emphasis on the ones selected for the individual experiments here presented (i.e. SEMG, Video and UDS). The presented technologies for silent speech recognition are ordered according to the several stages of the speech production process and it is explained in what way information can be collected at all stages.

2.4.1. Brain Computer Interfaces

The goal of a Brain Computer Interface (BCI) is to interpret thoughts or intentions and convert them into a control signal. With the evolution of cognitive neuroscience, brain-imaging and sensing technologies, means have been provided to understand and interpret the physical processes in the brain. BCI’s have a wide scope of application and can be applied to several problems, like assistance to subjects with physical disabilities (e.g. mobility impairments), detection of epileptic attacks, strokes, or to control computer games (Nijholt and Tan, 2008). A BCI can be based on several types of changes that occur during mental activity, such as electrical potentials, magnetic fields, or metabolic/hemodynamic recordings.

An SSI based on unspoken speech is particularly suited for subjects with physical disabilities such as the locked-in syndrome. The term unspoken speech refers to the process where the subject imagines speaking a given word without moving any articulatory muscle or producing any sound.

Current SSI approaches have been based on electrical potentials, more exactly on the sum of the postsynaptic potentials in the cortex. Two types of BCI’s have been used for unspoken speech recognition, one invasive approach based on the interpretation of signals from intra-cortical microelectrodes in the speech-motor cortex and a non-invasive approach based on the interpretation of signals from electroencephalographic sensors. Unspoken speech recognition tasks have also been tried based on Magneto-Encephalograms (MEG), which measures the magnetic fields caused by current flows in the cortex. However, this approach requires a shielded room, no metal on the patient is allowed, and is extremely expensive considering that the results have shown no significant advantages over EEG-based systems (Suppes et al., 1997). An overview of BCI can be found in (Brumberg et al., 2010; Nijholt and Tan, 2008).

2.4.1.1. Intra-cortical microelectrodes

This technique consists of the implantation of an extracellular recording electrode and electrical hardware for amplification and transmission of brain activity. Relevant aspects of this procedure include: the location for implanting the electrodes; the type of electrodes; and the decoding modality. Due to the invasive nature, increased risk and medical expertise required, this approach is only applied as a solution to restore speech communication in extreme cases such as subjects with the locked-in syndrome which are medically stable and present normal cognition. When compared with EEG sensors this type of systems presents a better performance enabling real-time fluent speech communication and the recordings are not affected by motor potentials caused by inadvertent movements (Brumberg et al., 2010). Results for this approach in the context of a neural speech prosthesis show that a subject is able to correctly perform a vowel production task with an accuracy rate up to 89% after a training period of several months (Brumberg et al., 2009) or that classification rates of 21% (above chance) in a 38 phonemes classes problem can be achieved (Brumberg et al., 2011).

The latest achievements with this modality, in a subject with lock-in syndrome, include the control of a speech synthesizer in real-time, which eliminates the need of a typing process (Brumberg et al., 2013).

2.4.1.2. Electroencephalographic sensors

In this approach, EEG sensors are externally attached to the scalp, as depicted in Figure 6. These sensors capture the potential in the respective area, which during brain activity can go up to 75µV and during epileptic seizure can reach 1*m*V (Calliess and Schultz, 2006). Results from this approach have achieved accuracies significantly above chance (Lotte et al., 2007) and indicate that the Broca's and Wernicke's areas as the most relevant in terms of sensed information (Kober et al., 2001).

Other studies (DaSalla et al., 2009) using vowel speech imagery were performed regarding the classification of the vowels /a/ and /u/ and achieved overall classification accuracies ranging from 68 to 78%, indicating the use of vowel speech as a potential speech prosthesis controller.

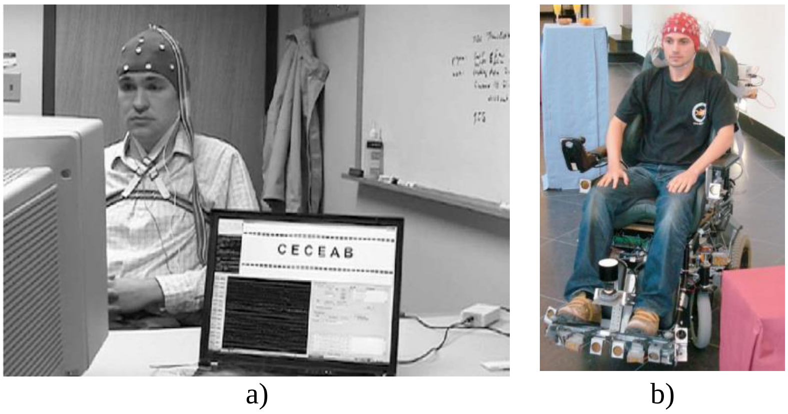


Figure 6. a) EEG-based recognition system for unspoken speech (Wester and Schultz, 2006). b) A brain-actuated wheelchair (Nijholt and Tan, 2008).

2.4.2. Surface Electromyography

During the human speech production process, in the phase before the actual articulation, myoelectric signals are sent and can be measured at the correspondent muscles, providing means to collect information about the resultant speech. The articulator's muscles are activated through small electrical currents in the form of ion flows, originated in the central and peripheral nervous systems. The electrical potential differences generated by the resistance of muscle fibers leads to patterns that occur in the region of the face and neck and which can be measured by this bioelectric technique. The process of recording and evaluating this electrical muscle activity is then called Electromyography.

Currently there are two sensing techniques to measure electromyography signals: invasive indwelling sensing and non-invasive sensing. The work here presented will focus on the second technique, which is when the myoelectric activity is measured by non-implanted electrodes and which is referred to in the literature as Surface Electromyography. Surface electrodes are non-invasive; however their attachment to the subject is usually done on some adhesive basis, which can obstruct movement, especially in facial muscles. By measuring facial muscles, the SEMG electrodes will measure the superposition of multiple fields (Gerdle et al., 1999) and for this reason the resulting EMG signal should not be attributed to a single muscle and should consider the muscle entanglement verified in this part of the human body. The sensor presence can also cause the subject to alter his/her behavior, be distracted or restrained, subsequently altering the experiment result. The EMG signal is also not affected by noisy environments, however differences may be found in the speech production process in the presence of noise (Junqua et al., 1999). Muscle activity may also change in the presence of physical apparatus, such as mouthpieces used by divers, medical conditions such as laryngectomies, and local body potentials or strong magnetic field interference (Jorgensen and Dusan, 2010). Surface EMG-based speech recognition overcomes some of the major limitations found on automatic speech recognition based on the acoustic signal such as: non-disturbance of bystanders, robustness in acoustically degraded environments, privacy during spoken conversations and as such constitutes an alternative for speech-handicapped subjects (Denby et al., 2010). This technology has also been used for solving communication in acoustically harsh environments, such as the cockpit of an aircraft (Chan et al., 2002) or when wearing a self-contained breathing apparatus or a hazmat suit (Betts et al., 2006).

Relevant results in this area were first reported in 2001 by Chan et al. (2001) where five channels of surface Ag-AgCl sensors were used to recognize ten English digits. In this study accuracy rates as high as 93% were achieved. The same author (Chan, 2003) was the first to combine conventional ASR with SEMG with the goal of robust speech recognition in the presence of environment noise. In 2003, Jorgensen et al. (2003) achieved an average accuracy rate of 92% for a vocabulary with six distinct English words, using a single pair of electrodes for non-audible speech. However, when increasing the vocabulary to eighteen vowel and twenty-three consonant phonemes in later studies (Jorgensen and Binsted, 2005) using the same technique the accuracy rate decreased to 33%. In this study problems in the alveolar pronunciation and subsequently recognition using non-audible speech were reported and several challenges identified such as sensitivity to signal noise, electrode positioning, and physiological changes across speakers. In 2007, Jou et al. (2007a) reported an average accuracy of 70.1% for a 101-word vocabulary in a speaker dependent scenario. In 2010, Schultz and Wand (2010) reported similar average accuracies using phonetic feature bundling for modelling coarticulation on the same vocabulary and an accuracy of 90% for the best-recognized speaker. In 2011, the same authors achieved an average of 21.9% on a 108-word vocabulary task (Wand and Schultz, 2011a).

In the last year several issues of EMG-based recognition have been addressed, such as investigating new modeling schemes towards continuous speech (Jou et al., 2007; Schultz and Wand, 2010), speaker adaptation (Maier-Hein et al., 2005; Wand and Schultz, 2011a); the usability of the capturing devices (Manabe and Zhang, 2004; Manabe, 2003) and even trying to recognize mentally rehearsed speech (Meltzner et al., 2008). The latest research in this area has been focused on the differences between audible and silent speech and how to decrease the impact of different speaking modes (Wand et al., 2012, 2011); the importance of acoustic feedback (Herff et al., 2011); analysis of signal processing techniques for SEMG-based SSI (Meltzner et al., 2010); EMG-based phone classification (Wand and Schultz, 2011b); session-independent training methods (Wand and Schultz, 2011a); removing the need for initial training before actual use (Wand and Schultz, 2014); EMG recording systems based on multi-channel electrode arrays (Heistermann et al., 2014; Wand et al., 2013a); reducing the number of sensors and enhancing SEMG continuous speech modelling (Deng et al., 2014); and EMG-based synthesis (Zahner et al., 2014).

In parallel with the topics described above, we are also seeing an increasing number of EMG resources available for the scientific community (Wand et al., 2014).

Other related work includes using EMG to control an electrolarynx (Heaton et al., 2011) and using SEMG to recognize disordered speech (Deng et al., 2009), which has the potential of being applied in speech therapy.

2.4.3. Visual Speech Recognition using RGB Information

The human speech perception is bimodal in nature, and the influence of the visual modality over speech intelligibility has been demonstrated by the McGurk effect (McGurk and MacDonald, 1976; Stork and Hennecke, 1996), which states the following: Vision affects the performance of the human speech perception because it permits to identify the source location; it allows a better segmentation of the audio signal; and it provides information about the place of articulation, facial muscle and jaw movement (Potamianos et al., 2003). This fact has motivated the development of Audio-Visual ASR (AV-ASR) systems and Visual Speech Recognition (VSR) systems, composed by the stages depicted on Figure 7.

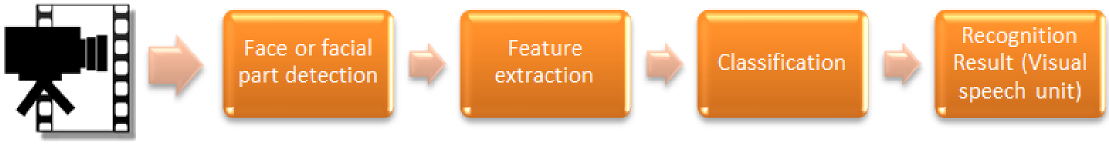

Face or facial part detection
Feature extraction
Classification
Recognition Result (Visual speech unit)

Figure 7. Typical Visual Speech Recognition system pipeline.

In visual-only ASR systems, a video composed of successive RGB frames is used as an input for the system. Relatively to the commonly used audio front-end, the VSR system adds a new step before the feature extraction, which consists of segmenting the video and detecting the location of the speaker's face, including the lips. After this estimation, suitable features can be extracted. The majority of the systems that use multiple simultaneous input channels such as audio plus video have a better performance when compared to systems that depend on a single visual or audio only channel (Yaling et al., 2010). This has revealed to be true for several languages such as English, French, German, Japanese and Portuguese; and for various cases such as nonsense words, isolated words, connected digits, letters, continuous speech, degradation due to speech impairment, etc. (Potamianos et al., 2003). In the last years we have watched VSR research being applied to several contexts (e.g. isolated digits recognition under whispered and neutral speech (Tao and Busso, 2014)); to different problems (e.g. analysis of dyslexic readers (Francisco et al., 2014)); to different techniques and classifiers (Noda et al., 2014; Shaikh et al., 2010) and to other languages besides English (Shin et al., 2011). There is also a noticeable trend towards sharing resources, with more and larger databases being published (Alghowinem et al., 2013; Burnham et al., 2011; Tran et al., 2013).

In existent SSI approaches, VSR is mostly used as a complement to other approaches, such as ultrasound imaging. Furthermore, many times only lips are considered as the Region of Interest (ROI), not taking into account information which could be extracted from jaws and cheeks.

2.4.3.1. RGB-based Features

According to the literature (Potamianos et al., 2003; Yaling et al., 2010) there are three basic methodologies to extract features in a VSR system: appearance-based; shape-based; or a fusion of both.

The first method is based on the information extracted from the pixels in the whole image or from some regions of interest. This method assumes that all pixels contain information about the spoken utterance, leading to high dimensionality issues.

Shape-based approaches base the extraction of features in the lip’s contours and also parts of the face such as the cheeks and jaw. This method uses geometrical and topological aspects of the face in order to extract features, like the height, width and area of the mouth image moment descriptors of the lip contours, active shape models or lip-tracking models. Shape-based methods require accurate and reliable facial and lip feature detection and tracking, which prove to be complex in practice and hard at low image resolution (Zhao et al., 2009).

The third method is a hybrid version of the first and second method and combines features from both methodologies either as a joint shape appearance vector or as a cooperative statistical model learned from both sets of features. Appearance-based methods, due to their simplicity and efficiency, are the most popular (Yaling et al., 2010).

The challenge in extracting features from video resides in collecting required information from the vast amounts of data present in image sequences. Each RGB frame contains a large number amount of pixels and is obviously too large to model as a feature vector. In order to reduce dimensionality and to allow better feature classification, techniques based on linear transformations are commonly used. Examples are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Discrete Cosine Transform (DCT) and Discrete Wavelet Transformation (DWT), Haar transforms, Locality Sensitive Discriminant Analysis (LSDA), or a combination of these methods (Potamianos et al., 2003; Yaling et al., 2010). A comprehensive overview of these methodologies can be found in (Potamianos et al., 2003).

2.4.3.2. Local Feature Descriptors

An alternative approach to the methods already mentioned, and adopted for our preliminary experiments was to explore the use of local feature descriptors (Carvalho et al., 2013). We have followed an appearance-based approach using feature extraction and tracking. This type of approach has proliferated in the area of computer vision due to its intrinsic low computational cost, allowing real time solutions.

To fully address our problem, we need a robust feature extraction and tracking mechanism and the Computer Vision community provides us various alternatives, such as Harris & Stephens (Harris and Stephens, 1988), SIFT – Scale Invariant Feature Transform (Lowe, 2004), PCA-SIFT (Ke and Sukthankar, 2004), SURF – Speeded Up Robust Features (Bay et al., 2006) or FIRST – Fast Invariant to Rotation and Scale Transform (Bastos and Dias., 2009). In terms of VSR, the concepts behind these techniques have been used for the elimination of dependencies on affine transformations by Gurbuz et al. (2001) and promising results, in terms of robustness, have been achieved. These methods have shown high matching accuracy on the presence of affine transformations. However, some limitations in real-time applications were found. For example, an objective analysis in Bastos and Dias (2009) showed that SURF took 500 milliseconds to compute and extract 1500 image features on images with resolution of 640x480 pixels, while PCA-SIFT took 1300ms, SIFT took 2400ms and FIRST only 250ms (half of the second most efficient, SURF). As for matching the same 1500 features against themselves, the figures for SURF, PCA-SIFT, SIFT and FIRST were, respectively, 250ms, 200ms, 800ms and 110ms, as observed by Bastos and Dias (2009).

Given these results, we have selected FIRST as the technique to be used in our preliminary experiments to extract and match features, since the real-time requirement is essential in practical VSR systems.

2.4.4. Ultrasound imaging

One of the limitations found in the VSR process described earlier is the visualization of the tongue, an essential articulator for speech production. In this approach, this limitation is overcome by placing an ultrasound transducer beneath the chin, thus providing a partial view of the tongue surface in the mid-sagittal plane (Hueber et al., 2010). This type of approach is commonly combined with frontal and side optical imaging of the user’s lips. For this type of system the ultrasound probe and the video camera are usually fixed to a table or to a helmet to ensure that no head movement is performed or that the ultrasound probe is correctly oriented with regard to the palate and that the camera is kept at a fixed distance (Florescu et al., 2010), as depicted on Figure 8a.

The latest work using US/Video relies on a global coding approach in which images are projected onto a more fit space regarding the vocal tract configuration – the EigenTongues. This technique encodes not only tongue information but also information about other structures that appear in the image such as the hyoid bone and the surrounding muscles (Hueber et al., 2010) (see Figure 8b). Results for this technique show that for an hour of continuous speech, 60% of the phones are correctly identified in a sequence of tongue and lip images, showing that better performance can be obtained using more limited vocabularies or using isolated word in silent speech recognition tasks, still considering realistic situations (Hueber et al., 2009). Other approaches include the use of ultrasound for articulatory-to-acoustic (Hueber et al., 2012) and animation of articulatory models (Fabre et al., 2014).

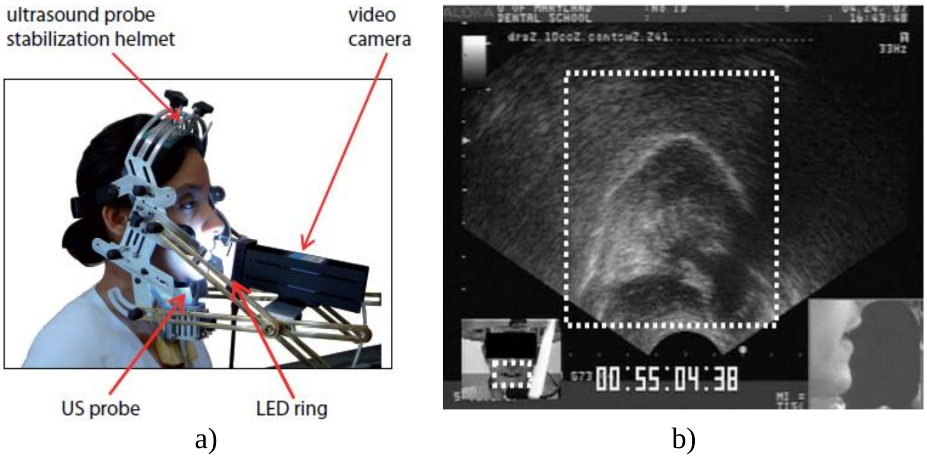


Figure 8. a) Portable acquisition US plus Video system (Florescu et al., 2010). b) Example of an US vocal tract image with embedded lateral and frontal lip view (Hueber et al., 2010).

2.4.5. Ultrasonic Doppler Sensing

Ultrasonic Doppler Sensing (UDS) of speech is one of the approaches reported in the literature that is suitable for implementing as an SSI (Srinivasan et al., 2010; Zhu, 2008). An SSI performs ASR in the absence of an intelligible acoustic signal and can be used to tackle problems such as environmental noise, privacy, information disclosure and aiding users with speech impairments. This technique is based on the emission of a pure tone in the ultrasound range towards the speaker's face that is received by an ultrasound sensor tuned to the transmitted frequency. The reflected signal will contain Doppler frequency shifts proportional to the movements of the speaker's face. Based on the analysis of the Doppler signal, patterns of movements of the facial muscles, lips, tongue, jaw, etc., can be extracted (Toth et al., 2010).

2.4.5.1. The Doppler Effect

The doppler Effect is the modification of the frequency of a wave when the observer and the wave source are in relative motion. If v_s and v_o are the speed of the source and the observer measured on the direction and sense observer-source, c is the propagation velocity of the wave on the medium and f_0 the source frequency, the observed frequency will be:

$$f = \frac{c + v_o}{c + v_s} f_0 \tag{1}$$

Considering a standstill observer $v_o = 0$ and $v_s \ll c$ the following approximation is valid:

$$f = \left(1 - \frac{v_s}{c}\right) f_0 \text{ or } \Delta f = -\frac{v_s}{c} f_0 \tag{2}$$

We are interested in echo ultrasound to characterize the moving articulators of a Human speaker. In this case a moving body with a speed v (positive when the object is moving towards the emitter/receiver) reflects an ultrasound wave which frequency is measured by a receiver placed closely to the emitter. The observed Doppler shift will then be the double:

$$\Delta f = \frac{2v}{c} f_0 \tag{3}$$

Considering $c = 340\text{m/s}$ as the sound air speed, a maximum articulator speed of 1m/s and a 40kHz ultrasound primary wave, the maximum frequency shift will be 235Hz.

To put it simply, we could consider a scenario (as the one depicted in Figure 9) where a source *T* emits a wave with frequency *f₀* that is reflected by the moving object, in this case the speaker's face. Since articulators move at different velocities when a person speaks, the reflected signal will have multiple frequencies each one associated with the moving component (Toth et al., 2010).

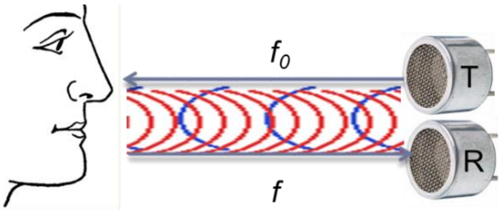


Figure 9. Doppler Effect representation (T – Transmitter, R- Receptor).

Below, a first example of the spectrogram of the Doppler signal and the correspondent audio signal applied to an EP word canto [k6~tu] (corner) is depicted on Figure 10.

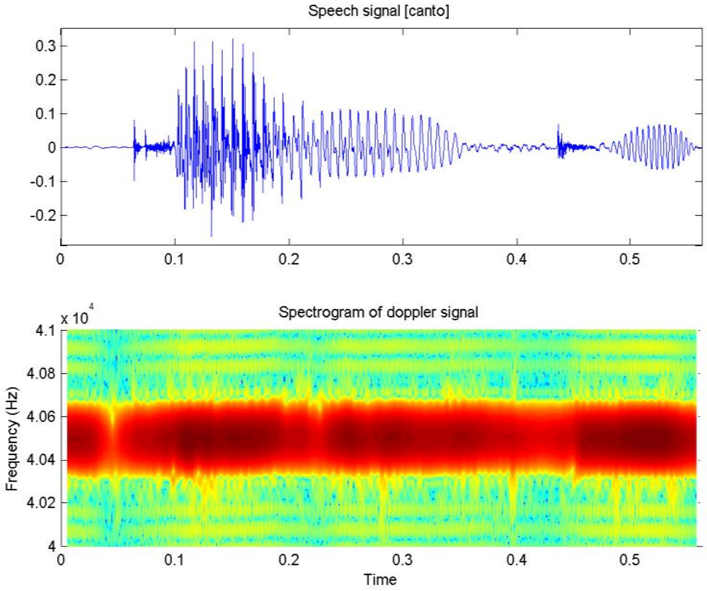


Figure 10. Audio signal (above) and spectrogram of the Doppler signal (below) for the word canto.

2.4.5.2. Related work

Ultrasonic sensors are used in a variety of applications that range from industrial automation to medical ultrasonography, with new developments also being applied to distinct areas of HCI (Raj et al., 2012). This modality has been applied to the characterization and analysis of human gait (Kalgaonkar and Raj, 2007), voice activity detection (Kalgaonkar et al., 2007; McLoughlin, 2014), gesture recognition (Kalgaonkar and Raj, 2009), speaker recognition (Kalgaonkar and Raj, 2008), speech synthesis (Toth et al., 2010), and speech recognition (Freitas et al., 2012b; Srinivasan et al., 2010).

Regarding speech recognition, ultrasonic devices were first applied to ASR in 1995 using an ultrasonic lip motion detector by Jennings and Ruck (1995). In this work, an experiment where the “Ultrasonic Mike”, as the authors call it, is used as an input to an automatic lip reader with the aim of improving ASR in noisy environments by combining it with a conventional ASR system. The used hardware is consists of an emitter, a receiver based on piezoelectric material and a 40 kHz oscillator to create a continuous wave ultrasonic signal. In the feature extraction phase, 10 Linear Predictive Coding (LPC) cepstral coefficients are extracted from the acoustic signal. The classification is based on Dynamic Time Warping (DTW) distances between the test utterances and the ones selected as ground truth. The best results for this work include an accuracy of 89% for the ultrasonic input alone using 4 template utterances, in a speaker dependent isolated digit recognition task, considering 5 test sessions and each session containing 100 utterances. For the cross-session scenario no higher than a 12.6% accuracy was achieved.

It was only a few years later, in 2007, that UDS was again applied to speech recognition by Zhu et al. (2007). In their work an ASR experiment was conducted based on a statistical approach and a continuous speech recognition task was considered. In terms of hardware, Zhu used an ultrasonic transmitter and a receiver tuned to a resonant frequency of 40 kHz. The received signal was then multiplied by a 35.6 kHz sinusoid causing it to be centered at 4.4 kHz. This study collected 50 sequences of ten random digits of twenty speakers at a 15.2 cm distance relative to the sensors. As far as feature extraction was concerned, the authors split the signal in frequency and magnitude sub bands and then features based on energy-band frequency centroids and frequency sub-band energy averages were extracted for each frame. The features were later projected to a lower dimensional space using PCA. The experiments were conducted using a landmark-based speech recognizer. The accuracy results for the ultrasonic approach were very similar across multiple noise levels, with a best result of 70.5% Word-Error Rate (WER).

In terms of UDS signal analysis Livescu et al. (2009) studied the phonetic discrimination in the UDS signal. In this study the authors tried to determine a set of natural sub-word units, concluding that the most prominent groupings of consonants include both place and manner of articulation classes and that, for vowels, the most salient groups included close, open and round vowels.

In 2010, Srinivasan et al. (2010), were able to improve upon previous results and achieved an overall accuracy of 33% also on a continuous digit recognition task. In this work Srinivasan et al. used hardware similar to the setup previously described, however, the synchronization of the two-channel (audio and ultrasound) output was added, the carrier was located at 8 kHz and the sensor was positioned at 40.5cm from the speaker. In terms of features, the authors applied a Fast Fourier Transform (FFT) over the pre-processed signal and applied a Discrete Cosine Transform to the bins corresponding to the frequencies between 7 kHz and 9.5 kHz. For classification, HMM models with 16 states and one Gaussian per state were used. The best results for fast speech presented an accuracy of 37.75% and 18.17% for slow speech.

When compared with other secondary sensors the Ultrasonic Doppler sensors have the advantage of not needing to be mounted on the speaker, and although their measurements were not as detailed as in P-mics or GEMS (Hu and Raj, 2005), the results for mutual information between UDS and acoustic speech signals were very similar to the ones reported for the other secondary devices (Hu and Raj, 2005). When compared with vision devices such as cameras, these sensors presented a much lower cost, since an ultrasonic sensing setup can be had for less than $10.

The results for ultrasound-only approaches are still far from audio-only performance. Nonetheless, the latest studies reveal viability and a margin for improvement for this approach. If an analysis using the same criteria as used in (Denby et al., 2010) is performed, the following can be concluded:

- **Works in noise**: the ultrasound signal is not affected by environment noise in the audible frequency range;
- **Works in silence**: since this technique is based on the signal that contains Doppler frequency shifts caused by facial movements no acoustic audio signal is required.
- **Works for laryngectomy**: Based on what was stated before, no glottal activity is required;
- **Non-invasive**: The device is completely non-obtrusive and it has been proven to work at a distance of 40.0cm without requiring any attachments;
- **Ready for market**: Results for this approach are still preliminary;
- **Low cost**: The hardware used on this approach is commercially available and is very inexpensive.

This approach, as was stressed before, still has a margin for improvement, especially when applied to Silent Speech recognition. The potential for detecting characteristics such as nasality is still unknown. Future research in this area will need to address issues such as changes in pose and distance of the speaker variation, both of which affect the ultrasound performance.

2.4.6. Electromagnetic and Permanent-Magnetic Articulography

Using the principle of magnetic sensing, this technique monitors the movement of fixed points in the articulators, collecting information from the articulation stage (referred on Figure 1) of the speech production process. A variant of this approach are the standard Electromagnetic Articulography (EMA) systems (Perkell et al., 1992) (e.g. Carstens AG500 (“Carstens,” n.d.)) that use glued coils, which can be electrically connected to external equipment (Kroos, 2012). However, although this approach enables accurate access Cartesian positions of the articulators, the necessary electrical connections for this approach make it hard to use in an SSI context. Nevertheless, a pilot study using EMA for capturing movements of the lips, jaw and tongue during speech, shows interesting results of 93.1%, 75.2% and 78.7% for vowel, consonant and phoneme classification experiments, respectively (Heracleous et al., 2011). Other studies using EMA systems include sentence (Wang et al., 2012a) and whole-word recognition (Wang et al., 2012b). The first achieved error rates of 94.9% across ten subjects and the latter 60.0% in a speaker-dependent scenario. Later the same author improved the results to an average speaker-dependent recognition accuracy of 80.0% (Wang et al., 2013). The most recent results include normalization methods for speaker-independent silent speech recognition (Wang et al., 2014).

A recent variant of this concept consists of using magnets attached to the vocal apparatus (see Figure 11), coupled with magnetic sensors positioned around the user’s head (Fagan et al., 2008;

Gilbert et al., 2010; Hofe et al., 2013b), referred to as Permanent-Magnetic Articulography (PMA) (Hofe et al., 2013a). In contrast to EMA, PMA does not provide exact locations of the markers, as it is not yet possible to separate the signals of individual magnets from the overall magnetic field. Instead, the authors (Hofe et al., 2013a) look at the detected patterns of magnetic field fluctuations, being associated with specific articulatory gestures through statistical modelling.

For example, in Fagan (Fagan et al., 2008), magnets were placed on the lips, teeth and tongue of a subject and were tracked by 6 dual axis magnetic sensors incorporated into a pair of glasses. Results from this laboratory experiment show an accuracy of 94% for phonemes and 97% accuracy for words, considering very limited vocabularies (9 words and 13 phonemes). More recently, studies that consider a larger vocabulary of 57 words, maintain accuracy rates above 90% (Gilbert et al., 2010), achieving in some cases a 98% accuracy rate (Hofe et al., 2010). Latest results, considering an additional axis, five magnetic sensors, and 71 words vocabulary, show the best result, for an experienced speaker, of 91.0% word accuracy rate (Hofe et al., 2013a). On a phonetic level, these studies also show the capability of this approach to detect voiced and unvoiced consonants (Hofe et al., 2013a, 2013b). Latest results investigated other aspects of speech production, such as voicing, place of articulation and manner of articulation (Gonzalez et al., 2014). Results found in Gonzalez et al. (2014) show that PMA is capable of discriminating the place of articulation of consonants, however it does not provide much information regarding the voicing and manner of articulation.

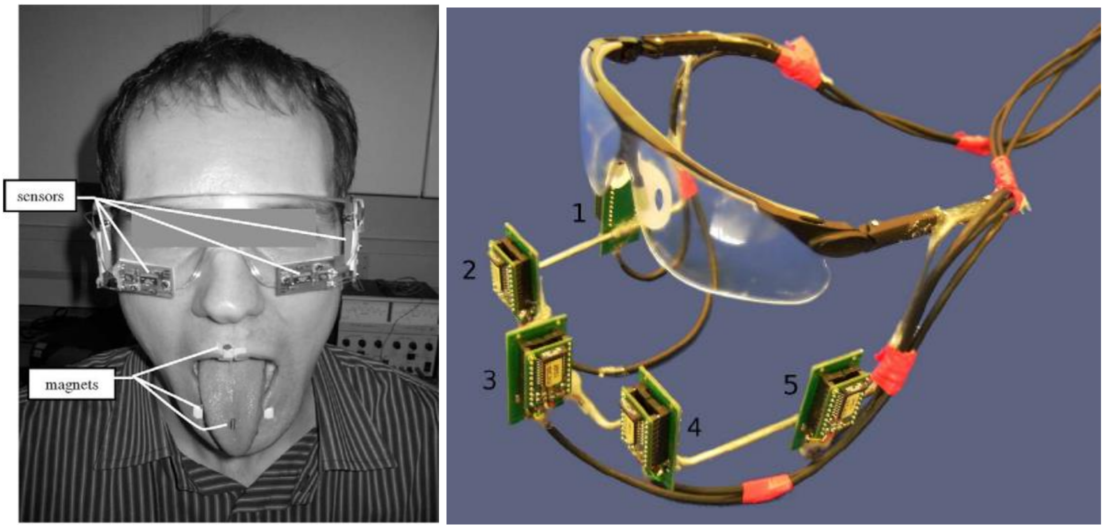


Figure 11. On the left, placement of magnets and magnetic sensors for a PMA-based SSI from Fagan et al. (2008). On the right, the MVOCA prototype from Hofe et al. (2013a), with five magnetic field sensors attached to a pair of glasses.

2.4.7. Electromagnetic and vibration sensors

The development of this type of sensors was motivated by several military programs in Canada, EUA and European Union to evaluate non-acoustic sensors in acoustically harsh environments such as interiors of military vehicles and aircrafts. In this case, by non-acoustic we mean that the sound is propagated through tissue or bone, rather than air (Denby et al., 2010). The aim of these sensors is then to remove noise by correlating the acquired signal with the one obtained from a standard close-talk microphone.

These types of sensors can be divided into two categories, electromagnetic and vibration (Denby et al., 2010). Regarding electromagnetic sensors the following types can be found: Electroglottograph (EGG); General Electromagnetic Motion System (GEMS) and Tuned Electromagnetic Resonating Collar (TERC). In terms of vibration microphones the following types can be found: Throat Microphone, Bone microphone, Physiological microphone (PMIC) and In-ear microphone.

These sensors have presented good results in terms of noise attenuation with gains up to 20db (Quatieri et al., 2006) and significant improvements in word error rate (WER) (Jou et al., 2004). It has also been presented by (Quatieri et al., 2006) that these sensors can be used to measure several aspects of the vocal tract activity such as low-energy, low-frequency and events such as nasality. Based on these facts, the use of these technologies is being considered by Advanced Speech Encoding program of DARPA for non-acoustic communication (Denby et al., 2010). When comparing the PMIC with Close-Talk Microphones (CTM) for stress detection and speaker recognition, the PMIC perform better or similar than the CTM (Patil and Hansen, 2010).

The concept behind these sensors also gave origin to successful commercial products, such as the Jawbone (Jawbone, n.d.), which uses a sensor to detect glottal vibrations from the jaw, cheek bones, and skin to conduct noise-cancellation tasks.

The disadvantage found in these sensors is that they usually require minimum glottal activity (also referred by Holzrichter et al. (2009) as Pseudo-Silent Speech). Eventually, as suggested by Holzrichter et al. (2009), under appropriate conditions and with enough electromagnetic sensors, one could apply this technique to silent speech by measuring the shapes and shape changes in the vocal tract.

2.4.8. Non-audible Murmur microphones

Non-audible murmur is the term given by research community to low amplitude speech sounds produced by laryngeal airflow noise resonating in the vocal tract (Nakajima et al., 2003b). This type of speech is not perceptible to nearby listeners, but can be detected using the NAM microphone (see

Figure 12), introduced by Nakajima et al. (Nakajima et al., 2003a). This microphone can be used in the presence of environmental noise, enables some degree of privacy and can be a solution for subjects with speaking difficulties or laryngeal disorders, such as the elderly. The device consists of a condenser microphone covered with soft silicone or urethane elastomer, which helps to reduce the noise caused by friction to skin tissue or clothing (Otani et al., 2008). The microphone diaphragm is exposed and the skin is in direct contact with the soft silicone. This device has a frequency response bandwidth of about 3 kHz with peaks at 500-800 Hz. Some problems concerning small spectral distortions and tissue vibration have been detected. However, the device remains an acceptable solution for robust speech recognition (Denby et al., 2010).

The best location for this microphone was determined by Nakajima (Nakajima, 2005) to be on the neck surface, more precisely below the mastoid process on the large neck muscle. In 2003, Heracleous et al. (2003) reported values of around 88% using an iterative adaptation of normal-speech to train a Hidden-Markov Model (HMM), requiring only a small amount of NAM data. This technology has also been tried in a multimodal approach in (Tran et al., 2009), where this approach is combined with a visual input and achieves recognition rates of 71%.

More recent work using NAM includes fusing NAM data with audio and visual information (Heracleous and Hagita, 2010; Tran et al., 2009); improving training methods transforming normal speech data into NAM data (Babani et al., 2011); the use of a stereo signal from two NAM microphones to reduce noise through blind source separation (Ishii et al., 2011; Itoi et al., 2012); and voice conversion methods from NAM to normal speech (Kalaiselvi and Vishnupriya, 2014; Toda, 2012; Tran et al., 2010).

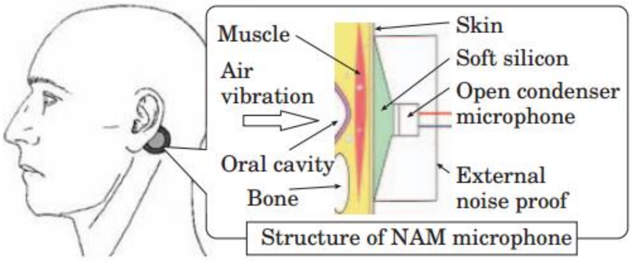


Figure 12. NAM microphone positioning and structure from (Ishii et al., 2011)

2.4.9. Challenges to Develop an SSI

In the work presented by Denby and coworkers (Denby et al., 2010) several challenges to the development of SSI can be found, such as:

- **Intra- and inter speaker adaptation**: The physiological and anatomic characteristics of the speaker are of major importance for most SSI. These differences found between speakers require robust modelling of the acquired signals and may also require large datasets in a generic statistical approach. In order to achieve speaker independence and higher usability rates an accurate method for positioning the sensors must also be found. Currently, the results of several approaches such as EMA, EMG and EEG show high sensitivity in sensor positioning, requiring previous training/adaptation or very accurate deployment.

- **Lombard and silent speech effects**: The effects adverting from silent speech articulation and the Lombard effect (different articulation when no auditory feedback is provided) are not yet clear and require further investigation. This effect can be minimized depending on the user experience and proficiency with SSI, however, relying on this would introduce a highly subjective requirement.
- **Prosody and nasality**: The extraction of speech cues for prosody in systems where output synthesis is envisaged and systems that want to detect emotions in speech is a major challenge in SSI and as yet an unexplored one. As for nasality, due to the modified or absent speech signal, the information for these parameters must be obtained by other means, as the ones described in this thesis.

These categories represent areas in this field where further research is required in order to reach an optimal SSI solution. In the work here presented we focus on a different challenge of adapting an SSI to a new language – European Portuguese – which involves addressing some of the issues referred to above, such as nasality.

2.5. Multimodal SSI

Not many studies can be found using more than one modality (excluding audio) in SSI research. However, analyzing the trend of recent years, there is an increasing amount of multimodal initiatives. For example, in 2006, a patent for a helmet with RGB, audio and ultrasonic data input was designed to increase transcription accuracy and/or to process silent speech (Lahr, 2006).

Results in general show that the use of more than one stream of information improves the results of individual modalities. Specifically in the SSI field, a first experiment was reported in 2004 by Denby and Stone (2004), where 2 input modalities, in addition to speech audio, were used to develop an SSI. The authors employed US of the tongue, lip profile video and acoustic speech data with the goal of developing an SSI. These two approaches (Video and US) are highly complementary since each modality captures articulatory information that the other one lacks. However, US still requires a complex and uncomfortable setup. Considering future advances in the technology, a possibility would be to have a cell phone with an incorporated US probe that the user can press against his or her chin, as envisaged by Denby (2013). More recently, Florescu et al. (2010), using these same modalities achieved a 65.3% recognition rate only considering silent word articulation in an isolated word recognition scenario with a 50-word vocabulary using a DTW-based classifier. The reported approach also attributes substantially more importance to the tongue information, only considering a 30% weight during classification for the lip information.

For other modalities, Tran et al. (2009) reported a preliminary approach using information from two streams of information: whispered speech acquired using a NAM and visual information of the face using the 3D position of 142 colored beads glued to the speaker's face. Later, using the same modalities, the same author (Tran et al., 2010) achieved an absolute improvement of 13.2% when adding the visual information to the NAM data stream. The use of visual facial information combined with SEMG signals has also been proposed by Yau et al. (2008). In this study, Yau et al. presented an SSI that analyses the possibility of using SEMG for unvoiced vowels recognition and a vision-based technique for consonant recognition.

There is also recent work using RGB-D (i.e., RGB plus depth information) information (Galatas et al., 2012a), showing that the depth facial information can improve the system performance over audio-only and traditional audio-visual systems.

On a related topic, but with a different aim and using totally different modalities, Dubois et al. (2012) compared audio-visual speech discrimination using EEG and fMRI in different phonological contrasts (e.g. labialization of the vowels, place of articulation and voicing of the consonants) to investigate how "visemes" are processed by the brain. Some conclusions from the authors are that visual perception of speech articulation helps discriminating phonetic features, and that visual dynamic cues contribute to an anticipated speech discrimination.

2.6. Extracting Information from SSI modalities

All these modalities can be seen as powerful instruments for extracting information about the human speech production process. However, such a complex sequence of events is not easily modelled and reaching a recognition result becomes a challenging task in the absence of an acoustic signal.

A rather promising and attractive approach, for the speech community in general, is to extract the underlying articulatory process by exploring the concept of Articulatory Features (AF) (Kirchhoff et al., 2002; Livescu et al., 2007; Papcun et al., 1992; Saenko et al., 2004) and as mentioned in section 2.2, the available technology provides the capabilities for capturing a detailed view of the articulatory process. According to Kirchhoff et al. (2002), the articulatory information can be modeled in different ways: articulatory parameters recovered from the acoustic signal by inverse filtering; use of articulatory classes' probabilities (usually according to the categories of the sounds or their articulation), which represent properties of a given phoneme, such as the place or the manner of articulation (also referred as phonetic features by Schultz and Wand (2010)); and direct measures of the articulators obtained with imaging technologies such as cineradiography (Papcun et al., 1992). In this thesis, we will focus on the last one, since that by obtaining a comprehensive set of measures in a multimodal approach, a more complete representation of the articulation process can be achieved.

However, when combining multiple modalities, due to the fact there being several information streams, the dimensionality of the feature space rapidly increases, yielding the well-known “curse of dimensionality”. As a consequence, in order to extract useful information from this data, one has to resort to techniques that provide better data representations such as Feature Selection (FS), Feature Reduction (FR) and Feature Discretization (FD). These techniques allow us to find adequate subsets (in the case of FS and FR) or to reach discretized versions of the data (in the case of FD) based on supervised and unsupervised information theory and statistics concepts (Ferreira, 2014). In this thesis we will use FS and FR methods, analyzing the pros and cons of each technique as we go along.

The following subsections provide a brief overview of the state-of-the-art in articulatory features in SSI and an overview of the machine learning techniques explored in this thesis for selecting the best features of each modality.

2.6.1. Articulatory Features

Articulatory features can be extracted using different modalities. Examples include video (Gan et al., 2007; Livescu et al., 2007; Saenko et al., 2009, 2004) and SEMG (Jou et al., 2006a; Schultz and Wand, 2010).

Visual approaches, such as video, allow us to extract fairly accurate information of visible articulators, obtaining measures regarding lip opening, closing, and rounding. However, if only a vision-based approach is used, we are limited by the inherent capabilities of the modality and can only capture some tongue tip visualizations.

Articulatory features have also been used in audiovisual recognition systems, providing a rather detailed description of co-articulation phenomena, since they are related to both the acoustic signal and the higher level of linguistic information (Gan et al., 2007). The asynchrony between articulatory features reflects the inherent pronunciation mechanism of human speech, thus, by modelling the pronunciation process, we can better predict and represent the co-articulation phenomenon.

Another approach found in literature is using SEMG sensors to monitor the articulatory muscles and to model their phonetic properties (Jou et al., 2006a; Schultz and Wand, 2010). Jou et al. successfully built a phoneme-based system and analyzed it by studying the relationship between SEMG and articulatory features on audible speech. Articulatory features are usually described as abstract classes, which capture relevant characteristics of the speech signal in terms of articulatory information. They are expected to be robust because they represent articulatory movements, which are less affected by speech signal differences or noise.

Experiments in acoustic and visual speech recognition have shown that articulatory-feature systems can achieve superior performance under clean and adverse environments (Ghosh and Narayanan, 2011; Saenko et al., 2009). However, the context in which articulatory features are used must be taken into account, since it has been previously shown that significant differences exist between articulation in normal vocalized speech and in silent speech (Wand et al., 2011).

2.6.2. Selecting the best features

Nowadays, there are a few datasets available for SSI research, including multimodal scenarios (Hueber et al., 2007). When multiple input modalities are considered, the large dimensionality of the feature space augments the complexity of the recognition task. To address these problems, the SSI literature presents many approaches that rely on FR techniques (Lee and Verleysen, 2007), such as LDA (Galatas et al., 2012b; Wand and Schultz, 2011a). However, the use of FR techniques such as LDA does not allow us to unveil and directly interpret which modalities and/or features contribute more significantly to better recognition performance. FR techniques often generate a new set of features, which are functions of the original features that correspond to the physical process. Thus, for SSI data it may be preferable to apply a FS method in order to filter and keep a subset of the original features. Moreover, it has been found that many FR techniques such as LDA may not perform well with low amounts of data for high-dimensional spaces (Qiao et al., 2009).

Feature Selection (FS) techniques aim at finding adequate subsets of features for a given learning task (Guyon and Elisseeff, 2003). The use of FS techniques may improve the accuracy of a classifier learnt from data by helping to avoid the so-called “curse of dimensionality” and may speed up the training time while improving the test (generalization) processes. In a broad sense, FS techniques are classically grouped into four main types of approach: wrapper; embedded; filter; and hybrid methods (Das, 2001; Guyon and Elisseeff, 2003; Guyon et al., 2006). Among these four types, filter approaches are characterized by assessing the adequacy of a given subset of features solely using characteristics of the data, without resorting to any learning algorithm or optimization procedure. It is often the case that for high-dimensional data, such as SSI data, the filter approach is the only one that produces acceptable results in terms of their running-time (Yu and Liu, 2003). For this reason, despite the different types of approaches, in this thesis we consider solely filter FS methods. There are decades of research on FS, for different problems and domains (Das, 2001; Guyon and Elisseeff, 2003; Guyon et al., 2006). However, in the context of multimodal signal processing, the research for FS methods has received little attention. Recently, an approach based on information theory, with a focus on audio–visual speech recognition has been proposed (Gurban and

Thiran, 2009); the proposed methods check for redundancy among features, yielding better performance than LDA.

2.7. Discussion

Each modality presented in this chapter has its advantages and limitations. Thus, following the same order, we will first discuss each modality individually and then analyze its joint use, starting by the modalities that look at the brain signals to recognize intent.

Understanding how the brain works has received a strong focus and great investment of the research community in the last years (e.g. Human Brain Project ("Human Brain Project," n.d.)), thus, outcomes of these initiatives will influence the state of SSI based on brain signals and allow for enhanced technologies and better modelling of the signals. Currently we have two main techniques in this area used for recognizing unspoken speech: intra-cortical electrodes and EEG.

The intra-cortical electrodes are currently used in extreme cases with some success, however, they still require a high degree of medical expertise. Nevertheless, from a theoretical point of view, it remains one of the most interesting concepts for SSI, which eventually could be explored in a long-term future, assuming parallel advances of medical procedures, particularly in the implantation procedure. Electroencephalographic sensors also allow for the analysis of brain signals, but without the medical implications of the previous approach. However, the amount of noise in the signal is also higher. Both approaches entail an adaptation process where the users learn to control their brain activity (e.g. users' imagine speaking a word to create known patterns).

One of the most promising technologies used for implementing an SSI is SEMG, which, according to previous studies, has achieved interesting performance rates (i.e. Word Error Rate (WER) of 21.9% on a 108-word vocabulary (Heistermann et al., 2014; Wand and Schultz, 2011a)) in a continuous speech recognition scenario, and it requires no glottal activity, working in both in silent and noisy environments. Additionally, although this technique is capable of capturing information about articulation control, which precedes the movement itself, we are still unaware of its full potential for measuring deeper muscles, such as the muscles responsible for the velar movement.

Measuring the pros and cons of each modality, RGB-D data extracted from 3D cameras, such as the Microsoft Kinect, emerge as important modalities to achieve the proposed objectives since they gather interesting characteristics in terms of silent speech interaction (i.e. low cost, not invasive, works in noisy environments, etc.). Video is also one of the most explored modalities found in the literature, with a vast amount of studies not only in VSR but also in other areas of computer vision. However, it is limited to visible articulators such as the lips and complementary information about other articulators such as the tongue and the velum would be of extreme value for improving recognition performance.

We consider Ultrasound Imaging to be a powerful approach in the sense that it provides information of hidden articulatory structures. However, the current technology, besides being considerably expensive, still requires a cumbrous setup that, in order to obtain quality data, entails the use of either a complex helmet or even more complex solutions to fix the associated probe. For that reason, and also from personal experience, we do not consider it to be a modality usable in daily tasks by, for example, an elderly subject. However, new developments considering miniature probes (Denby, 2013) and dry-electrodes may allow this modality to become more user-friendly.

Ultrasonic Doppler Sensing is an interesting approach and an attractive research topic in the area of HCI (Raj et al., 2012), mainly due to advantages such as its non-obtrusive nature and its low cost. Still, several issues which can be found in the state-of-the-art remain unsolved: speaker dependence, sensor distance sensitivity, spurious movements made by the speaker, silent articulation, amongst others. Since Doppler shifts capture the articulators' movement, we believe that some of these problems can be attenuated or even solved if information about each articulator can be extracted. However, it is still unclear if articulators such as the tongue or the velum are actually being captured by UDS.

Permanent-Magnetic Articulography has also been one the approaches strongly focused upon in the SSI research community, but since it needs permanent attachment of the magnetic beads it becomes more appropriate for speech production studies than for natural daily HCI.

Electromagnetic and vibration sensors have the drawback of needing to be mounted on the jaw bone or the speaker's face or throat, which may restrain their applicability or leave the user uncomfortable. However, they have been used with success as commercial applications, particularly in noise-cancelation scenarios. When compared with other SSI technologies the biggest disadvantage is that these sensors require glottal activity.

Overall, for the envisaged scenario of natural HCI, modalities such as SEMG, Video, Depth and UDS, although distinct, hold important and combinable characteristics, like being non-invasive and capturing different stages of speech production, which are important characteristics to achieve the established objectives. It can nevertheless be argued whether PMA is actually invasive, for example, when compared with SEMG. Also, an interesting question is where to draw the line between what is invasive and what is not. Our point of view is that PMA, although interesting, in order to obtain its full potential requires permanent attachment of sensors with chirurgical (or similar) glue, and if an articulator like the velum is considered, then the placement of such sensors already entails some medical concerns and expertise.

Additionally, despite the efforts and the available technology, the performance attained by SSI is still low when compared with ASR based on the acoustic signal with some of the technologies presented achieving promising and encouraging results (Denby et al., 2010). Nonetheless, for better performance we need to increase our understanding of the capabilities and limitations of each modality. It is important to note that in many usage contexts we need to consider only non-invasive modalities, in order to further motivate user-acceptance and to improve the usability of the interface. To assess these modalities it becomes essential to have complementary information and more direct measures regarding the phenomena we wish to capture. For example, for SEMG it is yet unclear which tongue movements are actually being detected and there is not enough information to allow accurately inferring, from prompts, which tongue movements are occurring during speech.

Therefore, to better understand the capabilities of each modality, we need reliable production of data. Taking the tongue as an example, one of the main articulators in the human speech production process (Seikel et al., 2009), several alternatives capable of gathering said data can be found: RT-MRI (Narayanan et al., 2011), US (Scobbie et al., 2008), among others (Stone and Lundberg, 1996) and (Hofe et al., 2013b). The resulting information could potentially be used to provide the grounds for further exploring the capabilities of EMG and other modalities.

Hence, the joint exploration of multiple non-invasive modalities must be addressed, with each modality potentially benefiting from complementary data obtained from other modalities and from the acquisition of "ground truth" data, commonly available by more invasive, obtrusive or less user-friendly modalities such as RT-MRI or US. Still, the joint exploration of modalities raises several challenges and requirements as follows:

- Reach a complementary and richer set of modalities, exploring as much as possible the complementary strengths and solving the weaknesses of individual modalities;
- Synchronize acquisition and ensure the proper conditions for conducting correlational studies;
- Include modalities that provide direct measures during different stages of speech production and the movement of articulators';
- Extract the relevant data from the different modalities considering the large amounts of data involved;
- Find the best way to fuse or use the information;
- Classify high-dimensional cross-modality data.

2.8. Summary

In this chapter we have provided a brief introduction to the human speech production process and its relation with the SSI approaches found in the literature. We have also shown that the available technologies allow to obtain direct and accurate measures of a great portion of this process.

In terms of related work, we presented a brief description about the origin and history of SSI, an overview of the existing modalities and some of the current challenges for research in this topic. Due to the numerous existent approaches, an emphasis was placed on the approaches selected for the studies described in this thesis. This chapter also complements and updates the overview provided by Denby et al. (2010) on SSI with the latest results, including an approach based on ultrasonic waves, here referred to as UDS.

Complementing this overview, we presented some of the existing studies that explore the combined use of modalities and potential techniques to model and select the best features coming from these data sources.

Finally, we discussed and compared the advantages and limitation of each modality and some ideas for their combined use, anticipating some requirements. From this analysis, three modalities emerged as the ones best fitted to fulfill our objectives - SEMG, Video and UDS – motivating their use for further experiments, presented in the next chapter.

CHAPTER III

Single Modality Experiments for Portuguese

> *"All life is an experiment. The more experiments you make the better."*
>
> Ralph Waldo Emerson
> Journals of Ralph Waldo Emerson, with Annotations - 1841-1844

Contents

The work presented in this chapter aims at creating the conditions to explore more complex combinations of HCI input modalities for SSI, starting by exploring non-invasive and recent modalities, such as UDS, or minimally intrusive modalities, such as SEMG, and to make an analysis of the first experimental results. Thus, we have selected a few input HCI technologies based on: the possibility of being used in a natural manner without complex medical procedures, low cost, tolerance to noisy environments, ability to work with speech-handicapped users and cost. Given these requirements, the following specifications were defined for initial experiments:

- Explore the myoelectric information of the articulator muscles using EMG;

- Acquire facial information from visual sensors;
- Capture of facial movements during speech using UDS.

Besides studying the selected modalities applied to EP, another important research question is addressed in this chapter: the capability of these different approaches to distinguish nasal sounds from oral ones. With this objective in mind, we have designed a scenario where we want to recognize/distinguish words differing only by the presence or absence of nasality in one of its phones. In EP, nasality can distinguish consonants (e.g. the bilabial stop consonant [p] becomes [m]) with nasality creating minimal pairs such as [katu]/[matu] and vowels, in minimal pairs such as [titu]/[ti~tu].

This chapter presents three experiments, one for each chosen modality. It starts by reporting an experiment in the area of SEMG, applied for the first time to EP and used in an isolated word recognition task. Next, we explore a visual approach based on RBG Video, where FIRST (Bastos and Dias, 2008) features are extracted for a similar recognition task. Afterwards, we report our first steps with UDS in similar scenarios, also for EP, to assess the potential of this modality. The chapter ends with a discussion of the overall results and a summary of what was reported. As in Chapter 2, the experiments are ordered according to their stage of the human speech production process.

An important aspect for any study involving human participants is to obtain the approval of a regulated ethics committee. As such, a detailed description of the data collections and associated studies/experiments were submitted for ethics approval. In the case of the experiments described in this thesis, all data collections participants gave informed written consent and were properly informed of the purpose of the experiment, its main features and that they could quit at any time. The experiments here reported have been evaluated and approved by the ethics committee of the University Institute of Lisbon (ISCTE-IUL), regulated by the dispatch nº7095/2011.

3.1. SSI based on Surface Electromyography for European Portuguese

In our research we designed an experiment to analyze and explore the SEMG silent speech recognition applied to EP. In this experiment, the following important research questions were addressed: (1) Is the SEMG approach capable of distinguishing EP nasal sounds from oral ones? (2) Can we achieve similar performance to what was reported for other languages?

3.1.1. Method

To address this research problem, we considered two scenarios. In the first scenario, we recognized arbitrary Portuguese words in order to validate our system and to assess the potential of SEMG for Portuguese. In the second scenario, we recognized/distinguished words differing only by the presence or absence of nasality in the phonetic domain.

3.1.1.1. Acquisition setup

The acquisition system hardware that was used (from ("Plux Wireless Biosignals," n.d.)) consisted of 4 pairs of SEMG of Ag/Ag-Cl surface electrodes connected to a device that communicates with a computer via Bluetooth. These electrodes measured the myoelectric activity using bipolar surface electrode configuration, thus the result was the amplified difference between the pair of electrodes. An additional reference electrode was placed in a location with low or none muscle activity.

As depicted on Figure 13, the sensors were attached to the skin using 2.5cm diameter clear plastic self-adhesive surface. Their position followed the recommendations from previous studies found in the literature (Jorgensen and Dusan, 2010; Maier-Hein et al., 2005; Schultz and Wand, 2010) with a minimum of 2.0cm spacing between the electrodes center, as recommended by Hermens et al. (2000). The 4 electrodes pairs and their corresponding muscles are presented on Table 3. A reference electrode was placed on the mastoid portion of the temporal bone.

Table 3. Approximate electrode pair/muscle correspondence.

Electrode pair	Muscle
1	*Tongue and Anterior belly of the digastric*
2	*Zygomaticus major*
3	*Lower orbicularis oris*
4	*Levator angulis oris*

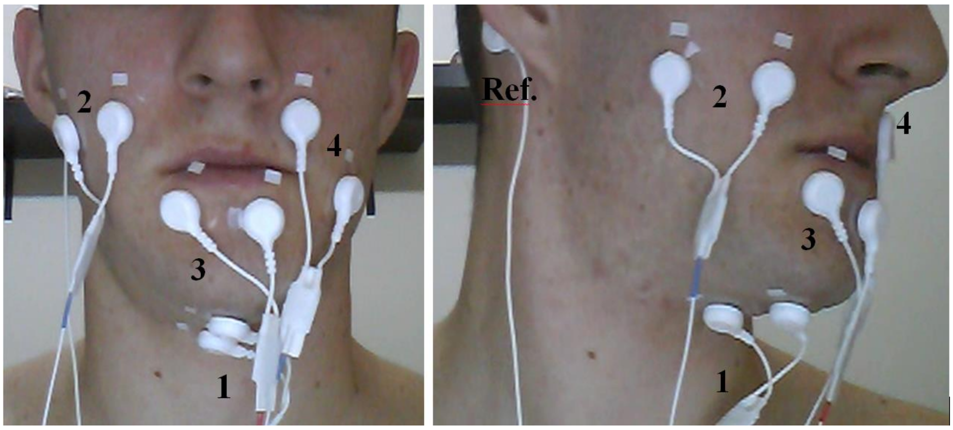


Figure 13. Surface EMG electrode positioning and the respective channels (1 to 4) plus the reference electrode (Ref.)

The technical specifications of the acquisition system included snaps with a diameter of 14.6mm and 6.2mm of height, a voltage range that went from 0V to 5V and a voltage gain of each EMG sensor (i.e. ratio between the output signal and input EMG signal) of 1000. The recording signal was sampled at 600Hz and 12 bit samples were used.

3.1.2. Corpora

For this experiment we created two corpora - PT-EMG-A and PT-EMG-B – containing respectively 96 and 120 observation sequences of the EMG signals. All observations were recorded by a single speaker (the author) on a single recording session (no electrodes repositioning was considered). The PT-EMG-A consisted of 8 different European Portuguese words with 12 different observations of each word, 4 of the words were part of minimal pairs where the presence or absence of nasality in one of its phones was the only difference, and 4 were digits, as described in Table 4. The PT-EMG-B corpus consisted also of 8 different words in EP, but with 15 different observations of each word. For this corpus, the words represented four minimal pairs of words containing oral and nasal vowels (e.g. cato/canto) and sequences of nasal consonant followed by nasal or oral vowel (e.g. *mato/manto*). Table 5 lists the pairs of words used in the PT-EMG-B corpus and their respective SAMPA phonetic transcription.

Table 4. Words that compose the PT-EMG-A corpus and their respective SAMPA phonetic transcription.

Word List	Phonetic Transcription
Cato	[katu]
Peta	[pet6]
Mato	[matu]
Tito	[titu]
Um	[u~]
Dois	[dojS]
Três	[treS]
Quatro	[kuatru]

Table 5. Minimal pairs of words used in the PT-EMG-B corpus and their respective SAMPA phonetic transcription. These pairs differ only by the presence or absence of nasality in one of its phones.

Word Pair	Phonetic Transcription
Cato/ Canto	[katu] / [k6~tu]
Peta / Penta	[petɐ] / [pe~t6]
Mato / Manto	[matu] / [m6~tu]
Tito / Tinto	[titu] / [ti~tu]

3.1.3. Processing

For processing the raw data we followed a typical recognition pipeline, divided into two stages: (1) feature extraction and (2) classification.

3.1.3.1. Feature Extraction

For feature extraction we used a similar approach to the one described by Jou et al. (2006b) based on temporal features, instead of spectral ones or even a combination of spectral with temporal features, since it has been shown that time-domain features present better results (Jou et al., 2006b). The extracted features were frame-based. Thus, for any given SEMG signal, $s[n]$ (where n is the sample index), frames of 30.0ms and a frame shift of 10.0ms were considered (i.e. signal capturing window of 30.0ms, in 10.0ms intervals, with 20ms overlap across consecutive frames). Denoting $x[n]$ as the normalized mean of $s[n]$ and $w[n]$, as the nine-point double-averaged signal, high-frequency signals $p[n]$ and $r[n]$ can be defined as:

$$x[n] = \frac{(s[n] - \bar{s}[n])}{\sigma_{s[n]}} \tag{4}$$

$$v[n] = \frac{1}{9}\sum_{n=-4}^{4} x[n] \tag{5}$$

$$w[n] = \frac{1}{9}\sum_{n=-4}^{4} v[n] \tag{6}$$

$$p[n] = x[n] - w[n] \tag{7}$$

$$r[n] = |p[n]| \tag{8}$$

A feature *f* will then be defined as:

$$f = [\bar{w}, P_w, P_r, \mathrm{z_p}, \bar{r}\ ] \tag{9}$$

where $\bar{w}$, and $\bar{r}$ represent the frame-based time-domain means, P_w and P_r the frame-based powers, and $\mathrm{z_p}$ the frame-based zero-crossing rate as described below.

$$\bar{w} = \frac{1}{N}\sum_{n=0}^{N-1} w[n] \tag{10}$$

$$P_w = \frac{1}{N}\sum_{n=0}^{N-1} |w[n]|^2 \tag{11}$$

$$P_r = \frac{1}{N}\sum_{n=0}^{N-1} |r[n]|^2 \tag{12}$$

$$z_p = \text{zero-crossing of } p[n] \tag{13}$$

$$\bar{r} = \frac{1}{N}\sum_{n=0}^{N-1} r[n] \tag{14}$$

The feature vector also considers the concatenation of *j* adjacent frames as formulated below

$$FV(f,j) = \left[f_{i-j}, f_{i-j+1}, \dots, f_{i+j-1}, f_{i+j}\right] \tag{15}$$

where *i* is the current frame index. Recent studies (Schultz and Wand, 2010; Wand and Schultz, 2011b) shown that *j* =15 yields the best results.

In the end, a feature vector of 620 dimensions was built by stacking the frame-based features of the four channels. In order to address the dimensionality of the resultant feature vector, PCA was applied reducing it to a vector of 32 coefficients per frame.

3.1.3.2. Classification

For classification we used an algorithm that measured the similarity between two temporal sequences (i.e. time series) independent of their non-linear variations in time. Dynamic Time Warping (DTW) was used in the early stages of ASR research (Rabiner et al., 1978) and for example-based approaches (De Wachter et al., 2007), with promising results. Thus, considering our limited dataset, we selected the DTW to explore this modality.

To reach the final classification results and to split the data into a train and a test dataset, the following algorithm was applied:

1. Randomly select *K* observations from each word in the corpus (see Table 5) that will be used as the reference (training) pattern, while the remaining ones will be used for testing.
2. For each observation from the test group:
    a. Compare each observation with the representative (training) examples.
    b. Select the word that provides the minimum distance.
3. Compute WER, which is given by the number of incorrect classifications over the total number of observations considered for testing.
4. Repeat the procedure *N* times.

3.1.4. Experimental Results

In terms of results, we start by presenting for the first time EMG speech patterns in EP. In Figure 14 and Figure 15, observations of the EMG signal in the four channels for the minimal pair *cato/canto*, can be depicted. Based on a subjective analysis we can see that the signals in both figures present similar patterns, where contraction of the muscles related with tongue movement is first evidenced on channel 1, followed by the remaining muscles movement on the other channels.

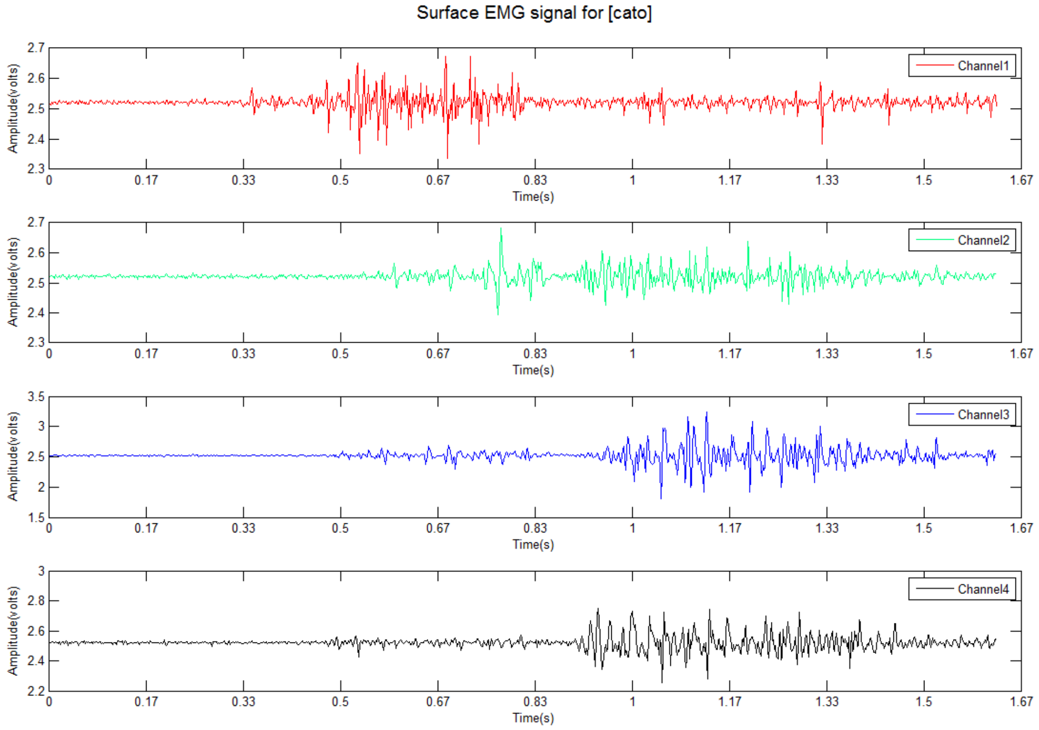


Figure 14. SEMG signal for the word *cato*. The four rows represent signals from the four EMG channels ordered from top to bottom in ascending order.

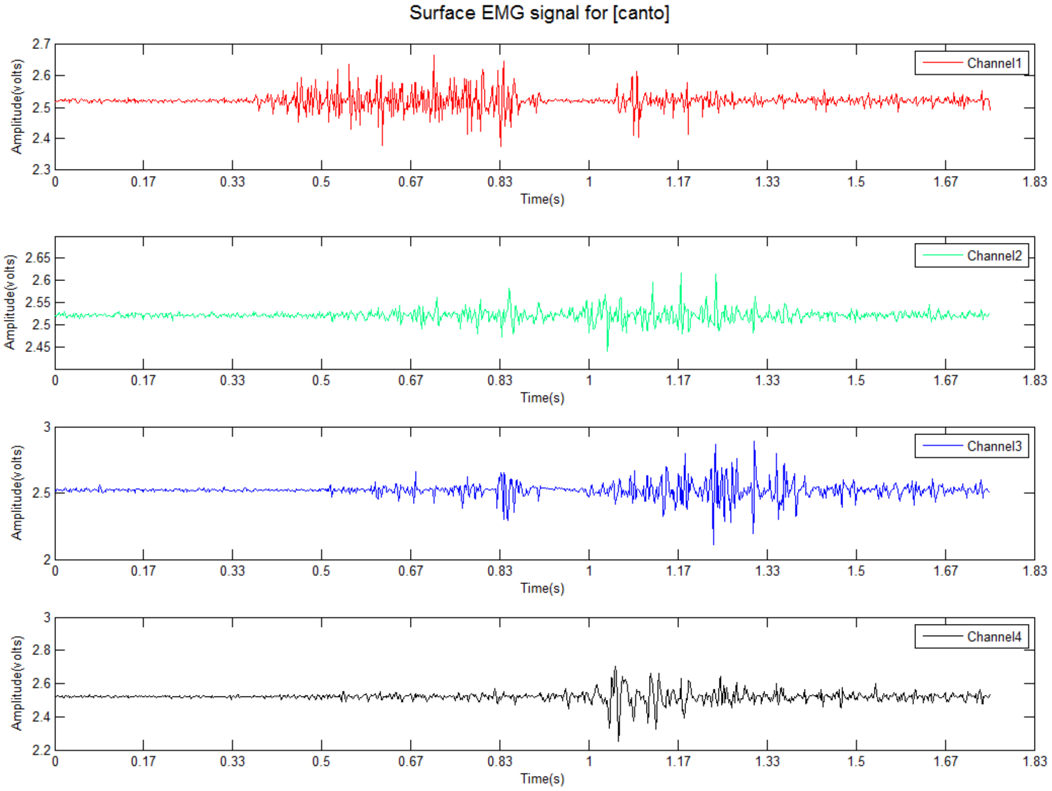


Figure 15. SEMG signal for the word *canto*. The four rows represent signals from the four EMG channels ordered from top to bottom in ascending order

Regarding the classification results for the corpus PT-EMG-A, the achieved values, using 20 iterations and *K* varying from 1 to 9, are listed on Table 6.

Table 6. Surface EMG WER classification results for the PT-EMG-A corpus considering 20 trials (N = 20).

K	***Mean***	***σ***	***Min***	***Max***
1	47.73	6.34	32.95	59.09
2	40.50	6.90	27.50	51.25
3	34.24	5.56	27.78	48.61
4	30.70	6.21	20.31	40.63
5	26.61	4.33	16.07	35.71
6	26.67	6.33	18.75	39.58
7	25.25	6.43	15.00	35.00
8	**22.50**	7.42	**9.38**	37.50
9	25.62	6.38	12.50	37.50

Based on Table 6, we found the best result for *K* = 8 having an average WER of 22.50%. The best run was achieved for *K* = 8 with a 9.38% WER.

The classification results for the PT-EMG-B corpus are described in Table 7. These values were achieved using 20 iterations (N = 20) and K varying from 1 to 11.

Table 7. Surface EMG word error rate classification results for 20 trials (N = 20).

K	***Mean***	***σ***	***Best***	***Worst***
1	64.10	5.55	50.89	75.00
2	56.87	5.97	44.23	65.38
3	53.38	6.09	43.75	63.54
4	51.19	5.10	43.18	61.36
5	50.50	6.89	36.25	66.25
6	50.06	6.50	40.27	61.11
7	47.89	4.33	39.06	53.12
8	47.14	6.61	35.71	57.14
9	45.72	5.42	33.33	54.16
10	**42.87**	4.53	35.00	52.50
11	43.13	6.92	**31.25**	56.25

Based on Table 7, we found the best result for *K* = 10 having an average WER of 42.87% and a Standard Deviation (STD) of 4.53%. The best run was achieved for *K* = 11 with a 31.25%. However, the STD result of almost 7% indicates a high discrepancy between iterations, probably caused by the low number of test observations. By analyzing the WER values across *K*, the linear trend depicted in Figure 16 indicates that increasing the amount of observations in the training set might be beneficial to the applied technique.

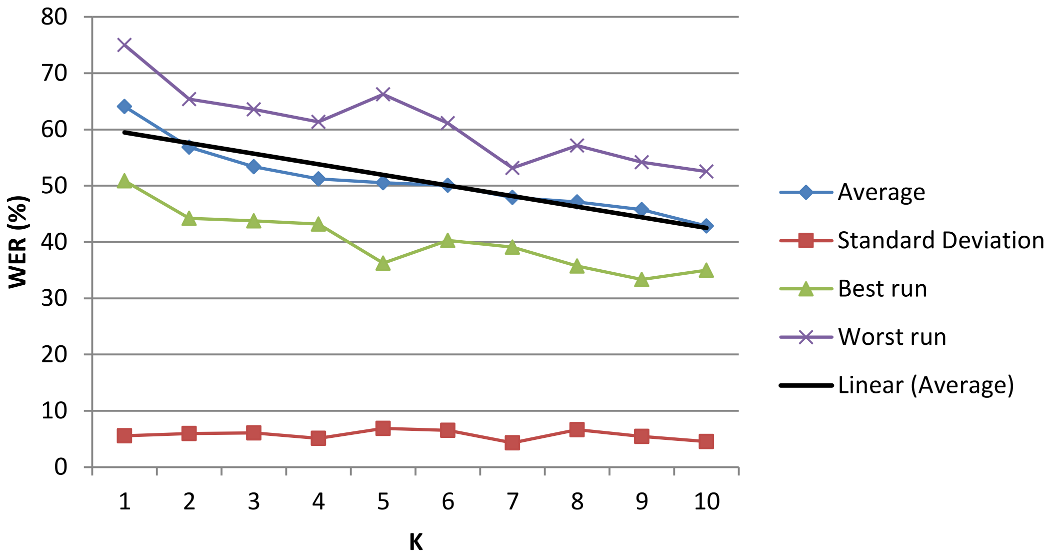


Figure 16. Average, standard deviation, best and worst run WER for K training sets, and the average results trend for the SEMG experiment.

Regarding the difference in the results with the two corpora, an absolute difference of almost 20.0% and a relative difference of 46.80% were verified between the best mean results. If we compare the best run results, then an absolute difference of 21.87% and a relative difference of 70.0% were verified.

The major discrepancy between the results from both corpora motivated an error analysis. In Table 8 we present for each value of *K* the average error percentage for all minimal pairs listed in Table 5. The confusion matrix for all considered runs is depicted in Figure 17. These results showed that 25.4% of the results, more than 50% of the total error, occurred between the analyzed minimal pairs. This can be seen as an indicator that the current techniques for silent speech recognition based on SEMG, such as the ones used in this experiment, may present a degraded performance when dealing with languages with nasal characteristics like EP. For example, if we examine the error represented by the confusion matrix of the best run for the PT-EMG-B corpus (depicted in Figure 18), we verify that most of the misclassifications occurred in the nasal pairs. In this case errors can be found between the following pairs: [k6~tu] as [katu] and vice-versa; [pe~t6] as [pet6]; [m6~tu] as [matu]; and [titu] as [ti~tu].

Table 8. Error analysis for the several values of *K*. The “Correct” column contains the percentage of correct classifications, e. g. observation *cato* was classified as *cato*. The “Minimal Pair Error” column represents the percent of observations classified as its pair, e. g. observation *cato* was classified as *canto*. The “Remaining Error” column presents the remaining classification errors.

K	Correct (%)	Minimal Pair Error (%)	Remaining Error (%)
1	35.89	21.42	42.67
2	43.12	24.18	32.69
3	46.61	24.58	28.80
4	48.80	26.07	25.11
5	49.50	27.31	23.18
6	49.93	27.29	22.77
7	52.10	26.01	21.87
8	52.85	26.69	20.44
9	54.27	26.25	19.47
10	57.12	25.00	17.87
11	56.87	24.53	18.59
Total Mean	**49.74**	**25.40**	**24.87**

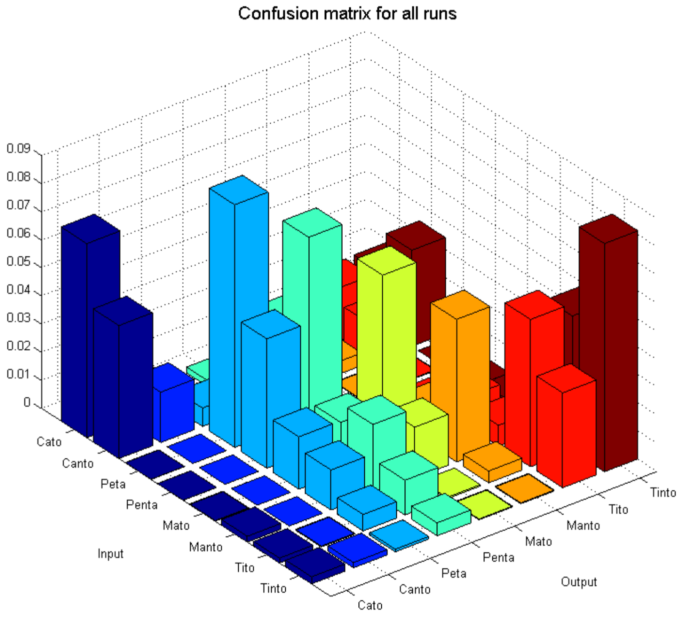


Figure 17. Confusion matrix for all considered runs

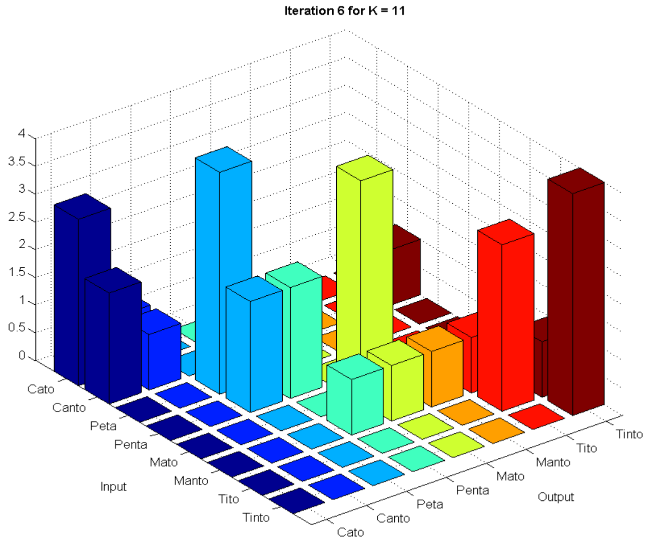


Figure 18: Confusion matrix for the best run (K=11 and N=6).

3.2. Visual Speech Recognition based on Local Feature Descriptor

The visual speech experiment here presented aims at demonstrating the hypothesis that the skin deformation of the human face caused by the pronunciation of words, can be captured by a video camera, processed in the feature domain and can drive a word classifier. For this purpose we studied the local time-varying displacement of skin surface features distributed across the different areas of the face, where the deformation occurs. We can abstract these image features as particles and here we were interested in studying the kinematic motion of such particles and specially, its displacements in time, in relation to a given reference frame. Differently from authors that focused their research solely in the analysis of lip deformation (Zhao et al., 2009), in our approach we were interested in other areas of the face, in addition to the lips area. It is worth recalling that this experiment was designed having in mind its further applicability in real-time speech recognition.

Due to previous evidence of FIRST advantage, when compared with SIFT, PCA-SIFT and SURF in applications like augmented reality, that require real-time behavior while keeping sufficient precision and robust scale, rotation and luminance invariant behavior, we have decided to use the FIRST features in our experiment. The selected approach, FIRST, can be classified as a corner-based feature detector transform for real-time applications. Corner features are based on points and can be derived by finding rapid changes in edge’s direction and analyzing high levels of curvature in the image gradient. FIRST features were extracted using Minimum Eigenvalues (MEV) and were made scale, rotation and luminance invariant (up to some extent of course, since with no luminance, no vision in the visible domain is possible), using real-time computer vision techniques. The FIRST feature transform was mainly developed to be used in application areas of Augmented Reality, Gesture Recognition and Image Stitching. However, to the best of our knowledge this approach had not yet been explored in VSR.

3.2.1. Method

To put our experiment into practice, we have specified, developed and tested a prototype VSR system. The VSR system receives an input video containing a spoken utterance. We first asked the speaker to be quiet for a few moments, so that we were able to extract FIRST features from the video sequence. After just some frames, the number of detected features stabilized. We refer to those as the calibration features, storing their position. The number of calibration features remained constant for the full set of the pronounced words, which, in our case was 40. After calibration, we assumed that the speaker was pronouncing a word. Therefore, in each frame, we needed to track the new position of each feature in the image plane.

In each frame, FIRST feature extraction and subsequently template matching with the calibration features was performed. Further optimizations towards real-time behavior are possible, by using the tracking approach of (Bastos and Dias., 2009), which uses optical flow and feature matching in smaller image regions. If the template matching normalized cross correlation is higher than a predefined threshold, then we assumed that the feature was matched and its new *u*, *v* image position was updated. Afterwards, the Euclidian distance between the updated feature position in the current frame and its position in the previous frame, was computed. For each feature, the resulting output was a law in time of the displacement (distance), relatively to the calibration position.

During the feature matching process several outliers may occur. These were later removed in a post-processing phase. In each frame, we were able to compute the displacement of each of the human face surface features that we were tracking. These feature displacements were then used as input feature vectors for a following machine classification stage. By analyzing these feature vectors during the full story of the observed word pronunciation and comparing this analysis with the remaining examples, we could choose the one with the closest distance, consequently being able to classify that observation as a recognized word. The distance was obtained by applying DTW (Rabiner and Juang, 1993).

In the following subsections, we provide a detailed description of the process.

3.2.1.1. Acquisition setup

The input videos were recorded by the author using a webcam video of 2 megapixels during daytime with some exposure to daylight. In these videos the user exhibits some spaced facial markings in areas surrounding the lips, chin and cheeks. These fiduciary markers were made with a kohl pencil.

3.2.2. Corpora

For this experiment a database containing 112 video sequences was built from scratch. The videos contained a single speaker uttering 8 different words in European Portuguese, with 14 different observations of each word. For this experiment the same minimal pairs described in Table 5 were used.

3.2.3. Processing

The data processing of this experiment was divided into the phases depicted in Figure 19 and described in detail in the following subsections.

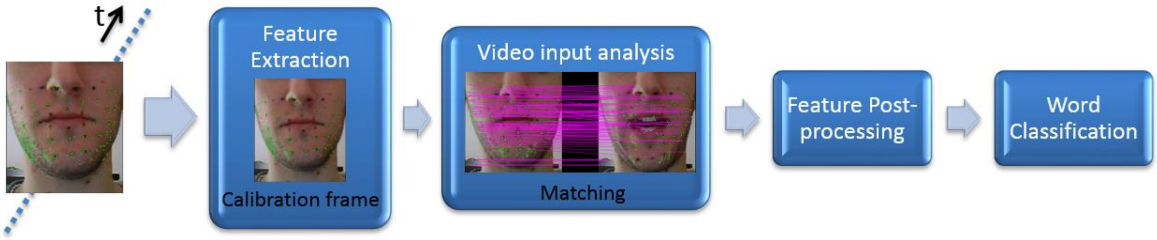


Figure 19. The phases of our Visual Speech Recognition experiment.

3.2.3.1. Feature extraction

The feature extraction process followed the work of Bastos and Dias (2008) and was performed using Minimum EigenValues (MEV), as described in the Shi and Tomasi (1994) detector. The reason for choosing this detector was related with its robustness in the presence of affine transformations. For feature extraction, the RGB image was first converted to gray scale. Then, a block of 3x3 pixels was taken at every image position and the first derivatives in the direction of x (D*x*) and y (D*y*) were computed using the Sobel operators O*x* and O*y* (Eq.16), for convolution with the 3x3 pixels block. The convolution resulted in evaluation of the mentioned first derivatives in direction of *x* and *y*. With the computed derivatives, we could construct matrix C, where the sum was evaluated in all elements of the 3x3 block. The Eigen Values were found by computing Eq. 18 where *I* is the identity matrix and λ the column vector of Eigen Values.

$$O_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} O_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (16)$$

$$C = \begin{bmatrix} \sum D_x^2 & \sum D_x D_y \\ \sum D_x D_y & \sum D_y^2 \end{bmatrix} \quad (17)$$

$$det(\,C - \lambda I) = 0 \quad (18)$$

Two solutions (λ_1, λ_2) resulted from the equation and the minimum Eigen Value ($min(\lambda_1, \lambda_2)$) was retained. In order to perform feature identification and determine strong corners, a threshold was applied to the resulting MEVs. Only features that satisfied the threshold value of 1% of the global maximum in the current MEV spectrum were selected. Non-maximum suppression was also performed by evaluating if the candidate corner's MEV was the maximum in a neighborhood of 3x3 pixels. After the features position in the image plane were found, several computer vision techniques were applied in order to make such features scale, rotation and luminance invariant, while at the same time maintaining the efficiency requirements. The algorithm for this procedure was described in (Bastos and Dias, 2008), and for this reason we will only refer which techniques were used. To make the FIRST features scale invariant it was assumed that every feature had its own intrinsic scale factor. Based on the results of the Sobel filters, the edges length could be directly correlated with the zooming distance. By finding the main edge length of the derivatives, an intrinsic scale factor could be computed. The scale factor was then enabled for the intrinsic feature patch to be normalized and consequently make it scale invariant. Only a surrounding area of 7x7 pixels relatively to the feature center was considered in order to deal with other derivatives that may appear resultant from zooming in/out. In order to achieve rotation invariance the highest value of the feature's data orientation was determined. Assuming that the feature's data was an *n* x *n* gray scale image patch (g_i) centered at (c_x, c_y), already scale invariant, the function that finds the main orientation angle of the feature g_i is given by:

$$\theta\,(g_i) = b \max(H(g_i)) \quad (19)$$

where $H(g_i)$ gives the highest value of orientation of g_i based on an orientation histogram composed by *b* elements (each element corresponds to 360°/*b* degrees interval). The *max* function returns the $H(g_i)$ histogram vector index. After obtaining the result of Eq. 19, a rotation of $\theta(g_i)$ degrees was performed to the g_i gray scale patch. Luminance Invariance was accomplished by using a template matching technique that used invariant image gray scale templates (Bastos and Dias., 2009). This technique was based on the image average and standard deviation to obtain a normalized cross correlation value between features. A value above 0.7 (70%) was used as correlation factor.

3.2.3.2. Feature matching

The FIRST feature transformation here presented is not as distinctive as SIFT or PCA-SIFT. For that reason, we used a method based on feature clustering. The method groups features into clusters through a binary identification value with low computation cost. The identification signature was obtained by evaluating three distinct horizontal and vertical regions of Difference of Gaussians patches (Davidson and Abramowitz, 2006). Difference of Gaussians is a gray scale image enhancement algorithm, which involves the subtraction of one blurred version of an original gray scale image from another, which is a less blurred version of the original. The blurred gray scale images were obtained by convolving the original image with Gaussian kernels with different standard deviations.

The signature was composed by 8 digits and only features that correspond to a specified feature's binary signature were matched, thus reducing the overall matching time. For positive or null regions a value of 1 was assigned. Negative regions were assigned with 0. When a FIRST feature patch was processed and created, this evaluation was performed and this feature was inserted in the corresponding cluster using the obtained binary identification. When matching a feature, we also computed the binary identification of the candidate feature, which allowed us to only match with potential candidates instead of matching with all the calibration features collected in a previous phase.

For each feature, when a matching was found, the displacement (standard Euclidian distance (Eq. 20) was computed between the updated feature position and the initial (calibrated) position, given by:

$$\mathrm{d}(\mathrm{p},\mathrm{q}) = \sqrt{(\mathrm{p}_\mathrm{x} - \mathrm{q}_\mathrm{x})^2 + (\mathrm{p}_\mathrm{y} - \mathrm{q}_\mathrm{y})^2} \qquad (20)$$

The result from this process was a bi-dimensional pattern, such as the one depicted on Figure 20 for the words "cato" [katu] (in English: cact) and *canto*. The horizontal axis represents the video frame (i.e. time) and the vertical axis represents the 40 features displacements, observed in the respective examples of word pronunciation. These images provide us with a view of how features vary in time for different words.

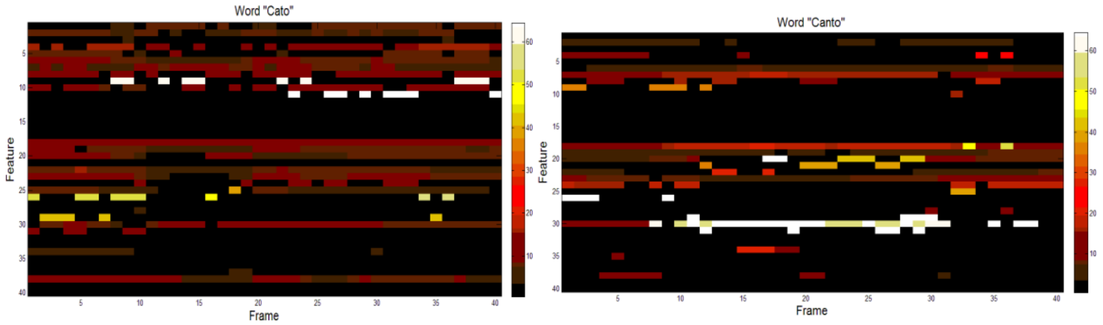


Figure 20. Image graph for the words *cato* (left) and *canto* (right).

The input videos for these results were recorded under approximately the same conditions and a clear difference between these two words could be noticed in the patterns. As depicted in Figure 21, each feature has a different behavior in its displacement across time, which shows two features from different human face regions for the same word.

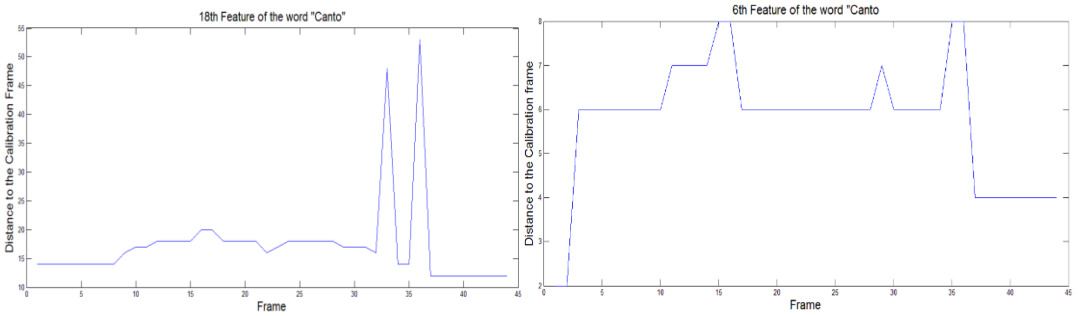


Figure 21. Temporal behavior of the feature number 18 (left) and number 6 (right) for the word *canto*.

3.2.3.3. Post-processing

In the post-processing phase outliers were removed using Chauvenet's criterion (Chavuenet, 1871) with a 0.5 threshold. This approach, although being simple, has shown good results demonstrated by empirical measures. A value was considered to be an outlier when the matching between two features was incorrect, i.e. the matching feature is not the original one. This situation is highlighted on Figure 22 were the outlier is marked with a red circle on both frames.

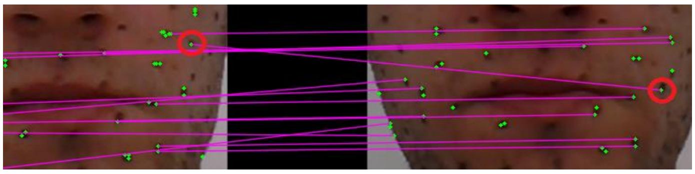


Figure 22. Matching between two frames with an outlier highlighted by a red circle.

3.2.3.4. Classification

For this initial stage of research the DTW technique was used to find an optimal match between the observations. DTW was chosen considering the number of observations and also because it addressed very well one of the characteristics of our problem: it provides temporal alignment to time varying signals that have different durations. This was precisely our case, since even observations of the pronunciation of the same word had different elapsed times.

In Figure 23 the DTW was applied to several pairs of words observations and we depict the DTW distance results by means of gray scale coding of such results. For the minimal pair *cato/canto* DTW computation we have, in the horizontal axis, the number of frames of canto production, whereas in the vertical axis, we have the number of frames for *cato* pronunciation. These diagrams can be simply interpreted as follows: The similarity between two words is given by the smallest DTW distance between them across time, thus when two words are the same, as shown in the comparison between canto and canto (upper-left graph), the lowest distance will lay in the image's diagonal. The red line represents the lower DTW distances found across time. In the upper-left panel the control case is represented by comparing a word observation with itself originating a straight diagonal line (i.e. all DTW distances in the diagonal are zero). Furthermore, as expected, a certain similarity between Cato and Canto (upper-right panel) can be noticed, since the only difference relies on a nasal phoneme instead of an oral one. It is also visible that the word *tinto* [ti~tu] (red) (bottom-left) presents the highest discrepancy regarding canto.

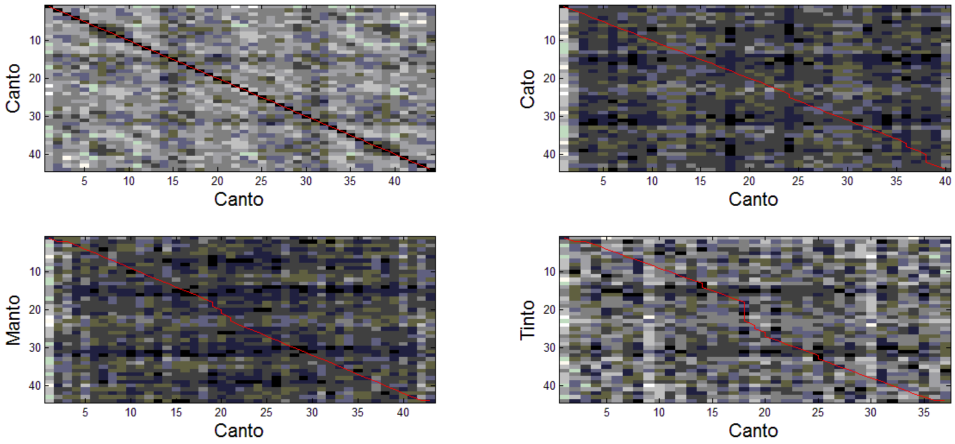


Figure 23. Distance comparison between the words canto [k6~tu], cato [katu], manto [m6~tu] and tinto [ti~tu].

In order to classify the results and to split the data into train and test, we used a similar procedure to the one described in the previous experiment (see section 3.1.3.2).

3.2.4. Experimental Results

Considering the classification algorithm described in section 3.1.3.2 with N = 20 and K varying from 1 to 10, the following results (listed in Table 9) in terms of WER were achieved:

Table 9. WER classification results for 20 trials (N = 20).

K	Mean	σ	Best	Worst
1	32.98	5.43	25.00	45.19
2	26.93	4.84	19.79	41.67
3	22.95	5.20	15.91	36.36
4	17.94	4.73	11.25	26.25
5	16.04	3.99	9.72	22.22
6	12.81	3.53	7.81	20.31
7	13.04	4.26	3.57	19.64
8	12.08	3.68	6.25	22.92
9	**8.63**	4.01	**2.50**	17.50
10	9.22	3.28	3.13	15.63

For this experiment, based on the results from Table 9, the best result was achieved when K=9, having an average WER of 8.63% and 2.5% WER for the best run. When analyzing the mean WER values across the K values, a clear improvement can be noticed, when the amount of representative examples of each word increased, suggesting that increasing the training set can be beneficial for our technique. Additionally, the discrepancy found between the best and worst values suggests that further research is required on how to select the best representation for a certain word. Analyzing the results from a different perspective, a value stabilization of WER when K=6 can be observed in the boxplot from Figure 24. However, considering the available corpora it is important to highlight that when K was higher the amount of test data was also reduced. For example, when K=10 only 4 observations from each word were considered. In this graph outliers can also be observed for K=2, K=7 and K=8.

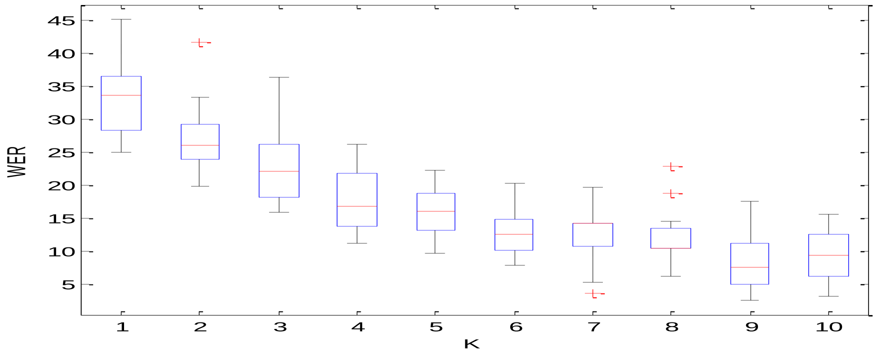
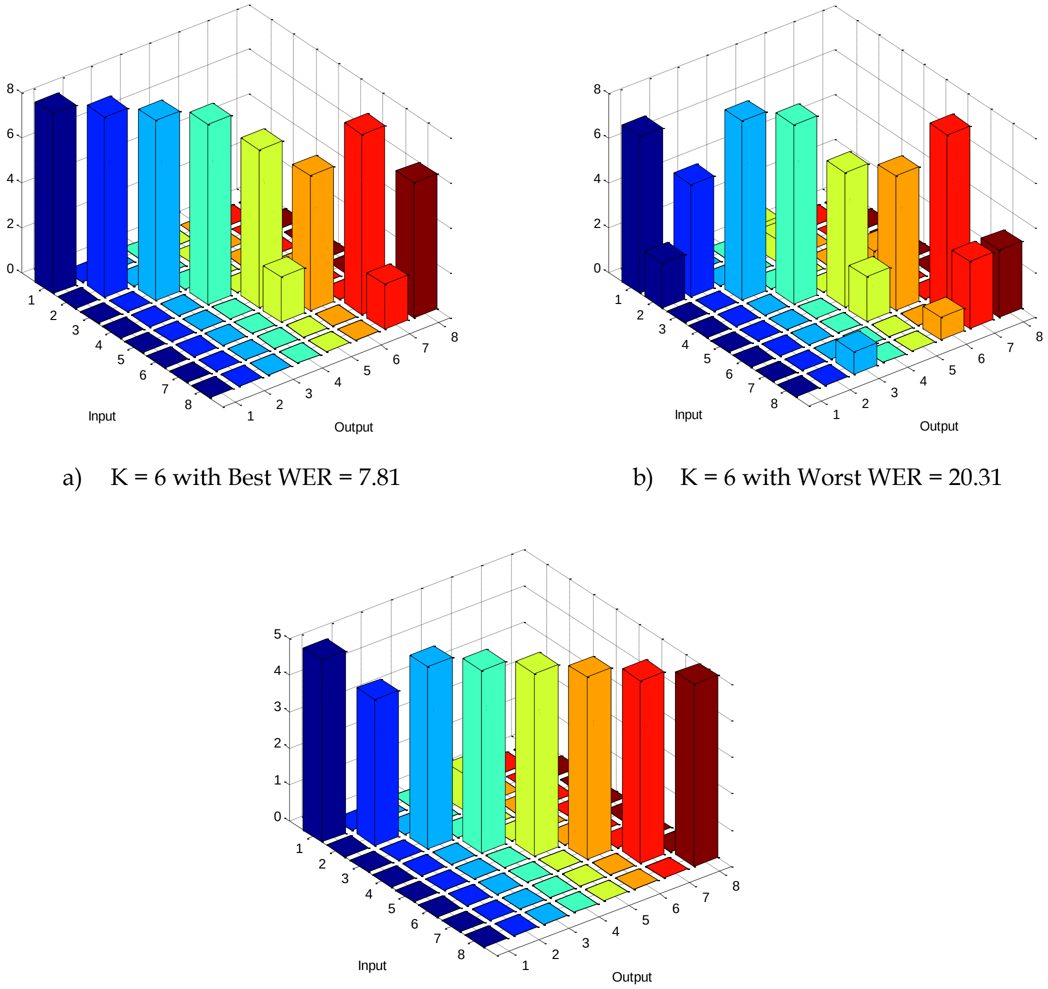


Figure 24. Boxplot of the WER results for the several K values.

In order to further analyze the quality of the experiment several confusion matrixes are depicted in Figure 25 for the most relevant runs. Each input represents the actual word and each output represents the classified word. The order presented in Table 5 was applied for each word. When analyzing the confusion matrixes for the several trials, errors can be found between the following pairs: [k6~tu] as [katu] (Figure 25b); [matu] as [m6~tu] and vice-versa (Figure 25a and b); [ti~tu] as [titu] (Figure 25a and b); and [k6~tu] as [matu] (Figure 25c). As expected, confusion is more often between words where the only difference relies in the nasal sounds (consonants and vowels).



a) K = 6 with Best WER = 7.81

b) K = 6 with Worst WER = 20.31

c) K=9 with Best WER = 2.5%

Figure 25. Confusion matrix for best and worst run of K = 6 and best run of K = 9. Input and output axis values have the following correspondence with the words from the Corpus: cato = 1, canto =2, peta = 3, penta = 4, mato = 5, manto = 6, tito = 7 and tinto = 8. The vertical axis corresponds to the number of classifications.

3.3. SSI based on Ultrasonic Doppler Sensing for European Portuguese

The characteristics of UDS (e.g. non-invasive, non-obtrusive, low cost, etc) make it very appealing for research in SSI. Therefore, Ultrasonic Doppler Sensing was one of the modalities chosen for the first set of experiments reported in this chapter. Since the necessary device is not commercially available, in order to study the Doppler effect of a speaker a dedicated circuit board was developed, specifically for the purposes of this thesis research.

In terms of experiment definition, we decided to study the following:

1. Assess if the custom built device prototype was working as expected and reported in the literature;
2. Analyze the performance of this modality in an isolated word recognition task (much like in the two previously reported experiments - VSR and SEMG);
3. Investigate the capability of distinguishing minimal pairs where the only difference relies in having nasal or oral sounds;
4. Understand some of the limitations reported in the literature for this modality such as the distance between the speaker and the device.

In the following subsections we start by describing the hardware and how the simultaneous acquisition of speech and the Doppler signal is achieved. Afterwards, we describe a recognition experiment where we analyze the use of the sensor in two distinct positions and in nasality detection.

3.3.1. Method

In terms of methodology, we considered again for this experiment an isolated word recognition task. However, when compared with the previous two, the main difference was that multiple speakers and difference distances between the speaker and the device were considered. Since this was our first experiment using this modality, we collected the minimal pairs described earlier in Table 5 and also 10 digits in EP for pipeline validation. In terms of experimental results, we started with an exploratory analysis of the signals and then proceeded with a classification experiment.

3.3.1.1. Acquisition setup

The custom built device, depicted in Figure 26 and based on the work of Zhu et al. (2007), includes a dedicated circuit board with: 1) the ultrasound transducers (400ST and 400SR working at 40 kHz) and a microphone to receive the speech signal; 2) a crystal oscillator at 7.2MHz and frequency dividers to obtain 40 kHz and 36 kHz; 3) all amplifiers and linear filters needed to process the echo signal and the speech. To capture the signal, the board was placed in front of the speaker.

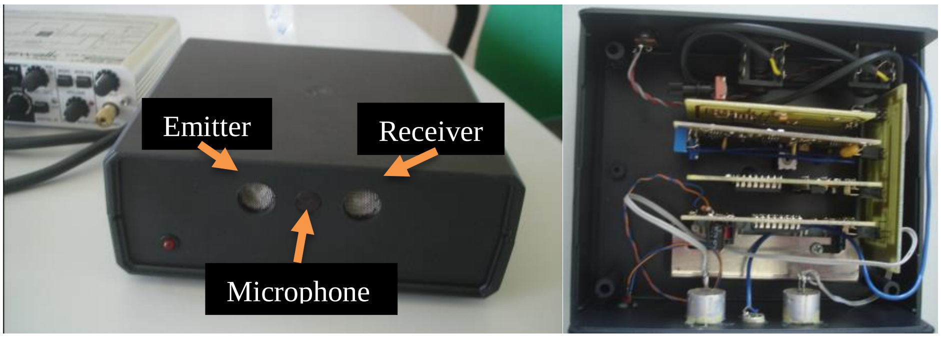


Figure 26. Ultrasonic Doppler Sensing device. On the left the outer aspect of the device is displayed and on the right the electronics behind it.

Looking at the resulting echo signal as the sum of the contributions of all the articulators, if the ultrasound generated is a sine wave $sin2\pi f_0\, t$, an articulator with a velocity v_i will produce an echo wave characterized by:

$$x_i = a_i \sin 2\pi f_0 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{21}$$

a_i, φ_i are parameters defining the reflection and are function of the distance. Although they are also function of time they are slow varying and are going to be considered constants. The total signals will be the sum for all articulators and the moving parts of the face of the speaker, as follows:

$$x = \sum_i a_i \sin 2\pi f_0 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{22}$$

The signal is a sum of frequency modulated signals. We have decided to make a frequency translation: multiplying the echo signal by a sine wave of a frequency $f_a = 36kHz$ and low passing the result to obtain a similar frequency modulated signal, centered at $f_1 = f_0 - f_a$,i.e., $f_1 = 4kHz$

$$d = \sum_i a_i \sin 2\pi f_1 \left(t + \frac{2}{c} \int_0^t v_i \, d\tau + \varphi_i \right) \tag{23}$$

This operation was made on the board and we have used an analog multiplier AD633. The Doppler echo signal and speech were digitized at 44.1 kHz. The following process was implemented in Matlab.

3.3.2. Corpora

The European Portuguese UDS data collected in this study was split into 2 corpora hereon referred as: (1) PT-DIGIT-UDS and (2) PT-NW-UDS. The first corpus is similar to what was used in previous studies (Srinivasan et al., 2010; Zhu, 2008) that addressed ASR based on UDS. It consists of ten digits (listing the digits and their phonetic transcription) - um [u~], dois [dojS], três [treS], quatro [kwatru], cinco [si~ku], seis [s6jS], sete [sEt@], oito [ojtu], nove [nOv@], dez [dES] - with the difference that we are using EP digits instead of English ones and that only isolated digits are considered. The second corpus is similar to what was used in the previous experiments of this chapter (section 3.1 and 3.2). It consists in 4 pairs of EP common words that only differ on nasality of one of the phones (minimal pairs, e.g. Cato/Canto [katu]/[k6~tu] or Peta/Penta [pet6]/[p6~t6] – see Table 5 for more details).

For the first corpus we recorded 6 speakers – 4 male and 2 female - and each speaker recorded an average of 6 utterances for each prompt. For the second corpus we recorded the same 6 speakers and each speaker recorded 4 observations for each prompt, giving a total of 552 utterances recorded at 40cm, for both datasets, as summarized in Table 10.

Most of the recordings occurred at a distance of approximately 40.0cm from the speaker to the sensor with exception for an extra session of a single female speaker that recorded 40 utterances using the PT-DIGIT-UDS prompts at a distance of 12.0cm for comparison and analysis.

Table 10. Summary of the UDS corpora collected for preliminary experiments.

Corpus	*Word set*	*Speakers*	*Utterance count*	*Distance speaker-device*
PT-DIGIT-UDS	10 EP digits	4 male and 2 female	360	40.0 cm
Extra session (PT-DIGIT-UDS)	10 EP digits	1 female	40	12.0 cm
PT-NW-UDS	Minimal Nasal Pairs (8 words)	4 male and 2 female	192	40.0 cm
Total	*18 words*	*4 male and 2 female*	*592*	*N/A*

3.3.3. Processing

The data processing was divided into the following phases (each corresponding to a subsection below): pre-processing, feature extraction and classification. The signal pre-processing and feature extraction procedures followed a similar approach to the one described by Srinivasan et al. (2010).

3.3.3.1. Signal pre-processing

After some exploratory analysis, we started by subtracting the signal mean. Then, the signal was passed through a third order sample moving average filter to suppress the 4 kHz carrier. Afterwards, a difference operator was applied. Figure 27 shows the resulting spectrogram after this pre-processing.

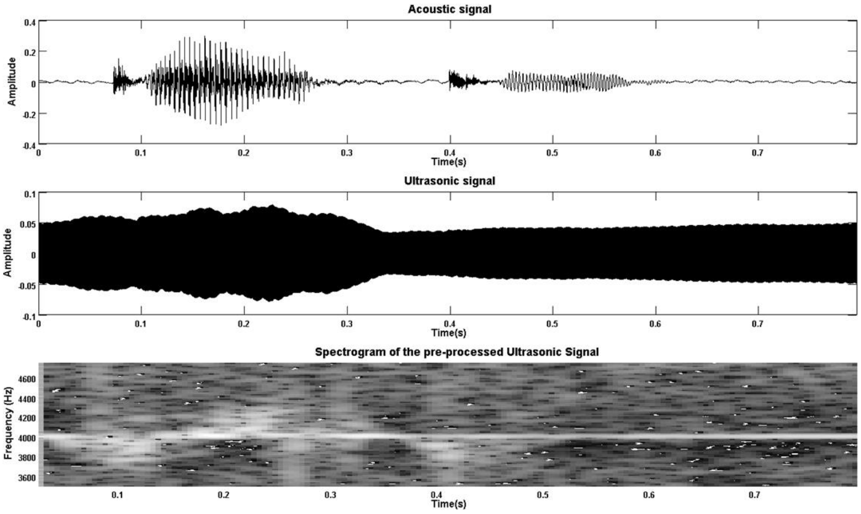


Figure 27. Signals for the EP word *Cato*, from top to bottom: Acoustic signal, raw ultrasonic signal and spectrogram of the pre-processed ultrasonic signal.

3.3.3.2. Feature extraction

Before conducting feature extraction, we pre-processed the signal as described earlier. After the pre-processing stage, we sampled the signal into 50ms frames. Then, we calculated a Discrete Fourier Transform (DFT) using a second-order Goertzel algorithm over the pre-processed signal for the interval of 3500 Hz to 4750 Hz. Then, a DCT was applied to the DFT results to de-correlate the signal. Finally, we extracted the first 38 DCT coefficients, which contain most of the signal energy, reducing the dimensionality of the final feature vector.

3.3.3.3. Classification

After the pre-processing and the feature extraction phase we needed to classify to which class the feature vector belonged. Much like in the previous experiments, based on the number of available observations and considering the limited vocabulary, we have chosen DTW as the classification technique, which was also employed by Jennings and Ruck (1995) to classify this type of signals. The classification algorithm used a 10-fold cross-validation for partitioning the data into train and test sets. Then, each observation from the test set was compared with the observations from the train set. The selected class was the one that provided the minimum distance in the feature domain.

3.3.4. Experimental Results

The following section describes the first recognition experiments based on UDS which are not applied to English. These experiments analyze the recognition of EP digits, the minimal pairs described in section 3.3.2, and a combination of both. Results regarding the effect of the sensor distance to the speaker are also reported.

3.3.4.1. Exploratory analysis

After the pre-processing stage, a first exploratory analysis of the signal showed a clear difference between EP digits. After performing a discriminative analysis of the signal (depicted in Figure 28) we noticed that the digits that require more movement from visible articulators presented distinguishable differences towards digits that require less facial movements. For example, if we compare an observation of "*um*" (one) with an observation of "*quarto*" (four) a clear magnitude difference is visible across time in Figure 28.

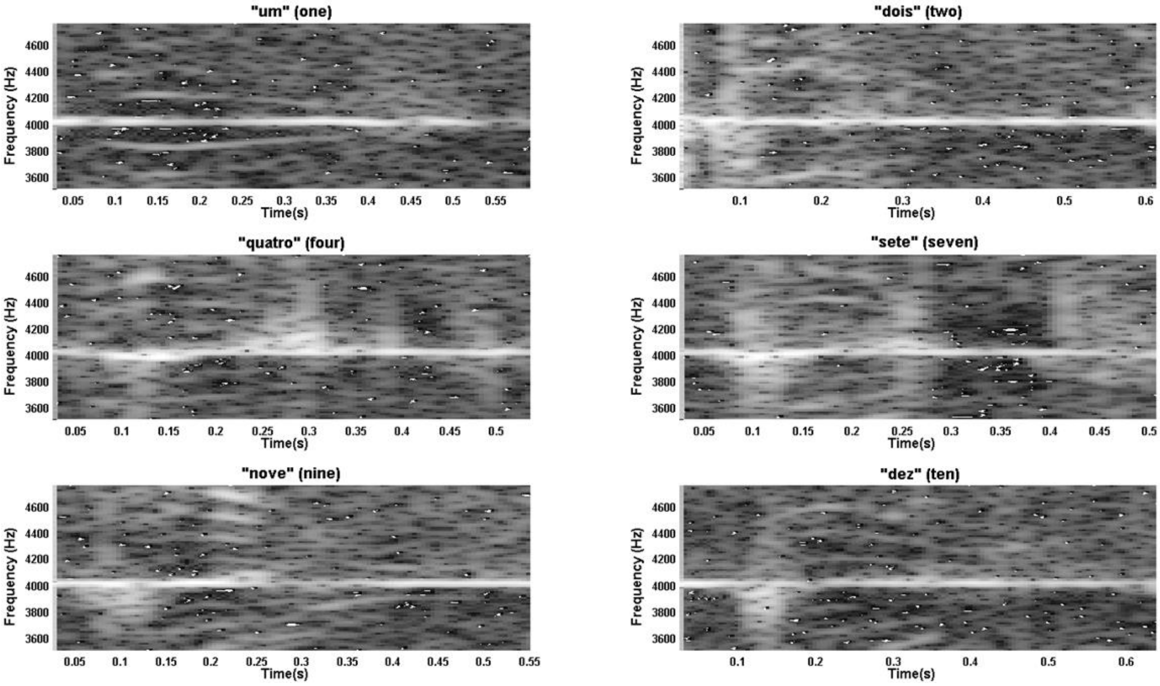


Figure 28. Spectrogram of the pre-processed signal for 6 EP digits and frequencies between 3500 and 4750 for a single speaker.

Figure 29 shows a similar analysis performed for the words "*cato*" and "*canto*", a minimal pair where the only difference is the presence of nasality. In this case, dissimilarities are more subtle, nonetheless, they seem to be present between the two words. Regarding the cross speaker signals, we found relevant differences between them (as depicted in Figure 29).

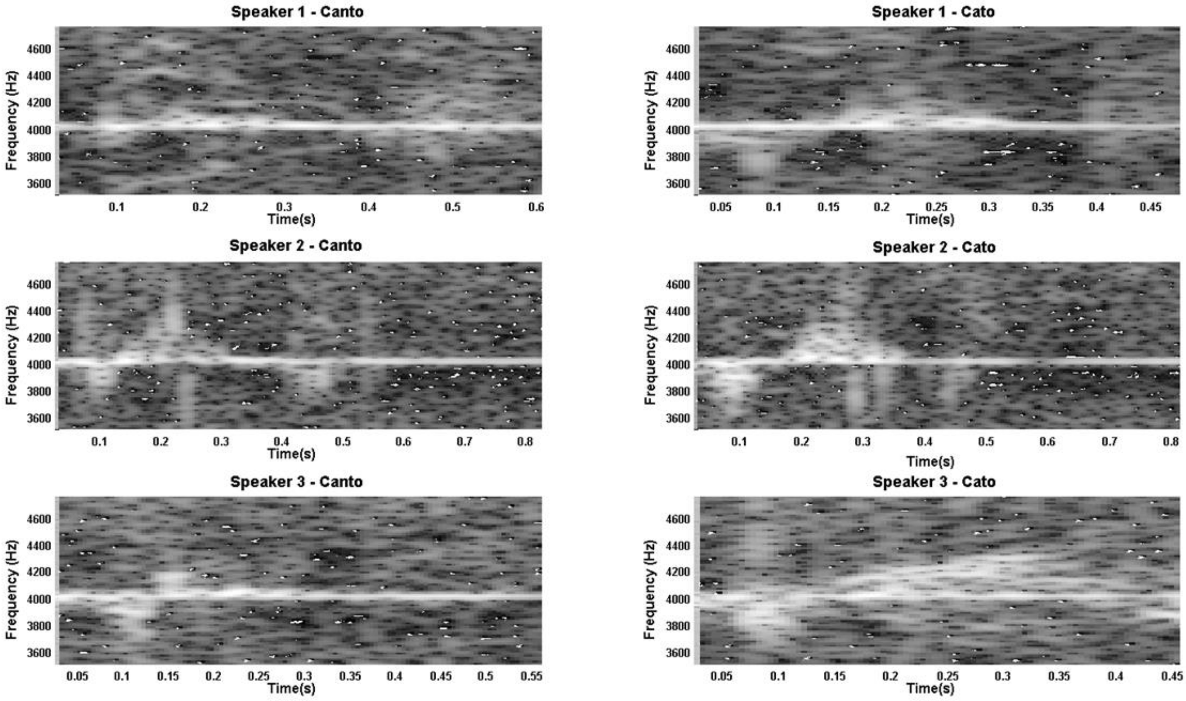


Figure 29. Spectrogram of the words "Cato" and "Canto" for 3 speakers.

When analyzing the signal using Dynamic Time Warping (DTW), which provides temporal alignment to time varying signals that have different durations, differences between minimal pairs could be noticed. In Figure 30 we depict the DTW applied to several pairs of words observations and the respective distance results by means of gray scale coding. The similarity between two words is given by the smallest DTW distance between them, across time. Thus, when two words are the same, the lowest distance will lay in the images' diagonal.

The highest discrepancy in the results was found when we compared different speakers, again showing that the signal is highly dependent of the speaker. We also noticed that these differences were more accentuated when we compared different genders.

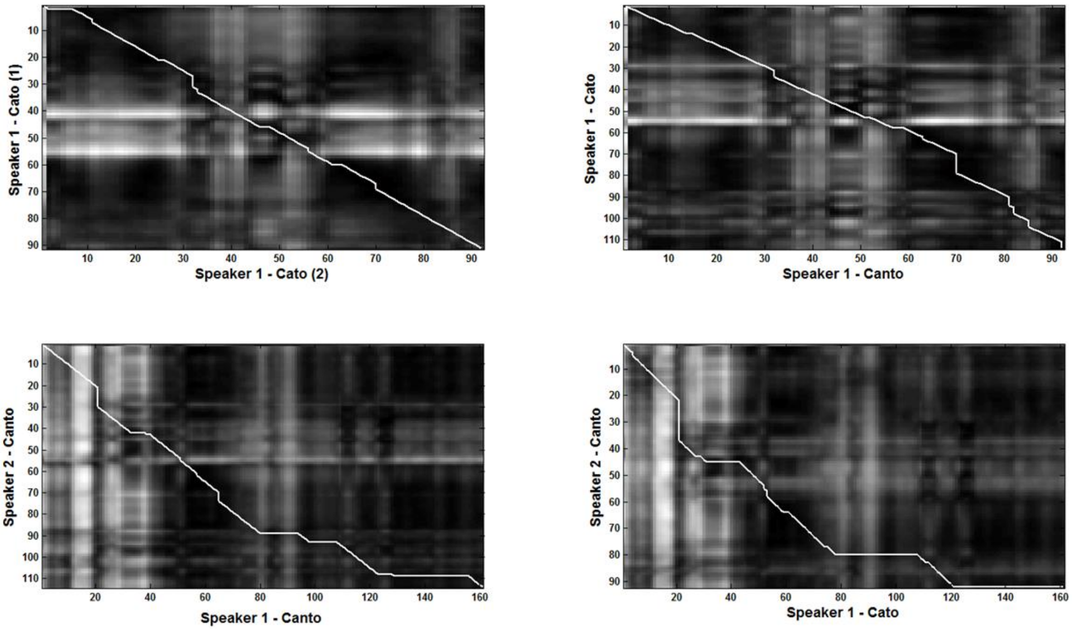


Figure 30. Distance comparison between the word *canto* and *cato* for different speakers. The white line represents the lower DTW distances found across time.

3.3.4.2. Recognition results

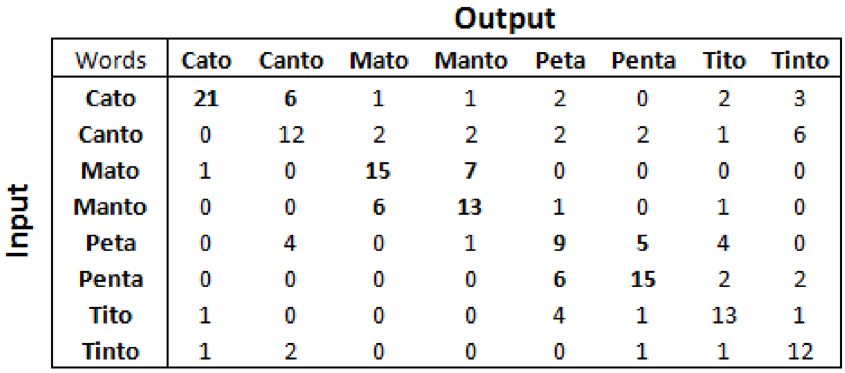
Table 11 presents the results of three test conditions: using the PT-DIGIT-UDS, the PT-NW-UDS and using all the available data at the distance of 40.0cm.

Table 11. Classification results for the following sets of data: PT-DIGITS-UDS, PT-NW-UDS and a combination of both.

	PT-DIGITS-UDS	PT-NW-UDS	Both
Word Error Rate (WER)	**36.1%**	42.7%	45.3%

Considering an isolated word recognition problem we were able to achieve the best WER of 36.1% for a vocabulary of 10 digits (PT-DIGITS-UDS). It was also noticeable that when we considered a smaller vocabulary with only 8 words based on minimal pairs (PT-NW-UDS) the error rate increased to 42.7%. When analyzing the confusion matrix for this run, depicted in Table 12, a large error incidence could be found in the minimal pairs. For instance, in the case of the word *mato*, 87.5% of the incorrect observations were classified as *manto*. For the reverse case (*manto* being classified as *mato*) 75% of the incorrect observations were classified as *mato*. This problem was also evident for the case of *cato* being confused with *canto* and for the pair *peta/penta*. Nonetheless, for the minimal pair *tito*/*tinto* this was not verified.

Table 12. Confusion matrix for the recognition experiment with the PT-NW-UDS corpus.



		Output							
	Words	Cato	Canto	Mato	Manto	Peta	Penta	Tito	Tinto
Input	Cato	**21**	**6**	1	1	2	0	2	3
	Canto	0	12	2	2	2	2	1	6
	Mato	1	0	**15**	**7**	0	0	0	0
	Manto	0	0	**6**	**13**	1	0	1	0
	Peta	0	4	0	1	**9**	**5**	4	0
	Penta	0	0	0	0	**6**	**15**	2	2
	Tito	1	0	0	0	4	1	13	1
	Tinto	1	2	0	0	0	1	1	12

Additionally, we also ran an experiment using a joint vocabulary of 18 words, merging the previous two vocabularies, and obtained a slight worse error rate of 45.3%. A considerable part of the error still occurred in the *mato/manto* and *peta/penta* minimal pairs. For the case of *mato*, 42.9% of the incorrect observations were classified as *manto*. As for the case of *penta*, 55.6% of the incorrect observations were classified as *peta*.

3.3.4.3. Distance effect

As mentioned before, we also recorded an extra session considering a closer distance with the device positioned 12cm from the speaker. The goal was to investigate the mismatch effect caused by the distance change in the traininig patterns and the test conditions, further analysing distance limitations. Thus, we have ran the previous experiment using the following data distributions: 1) Use only the PT-DIGIT-UDS data recorded at 12cm; 2) Use the PT-DIGIT-UDS data recorded at 12cm as a test group, creating a mismatch between train (recorded at 40cm) and test; 3) Use the PT-DIGIT-UDS data from the previous experiment (recorded at 40cm) plus the PT-DIGIT-UDS data recorded at 12cm for train and test. The obtained results are presented in Table 13.

Table 13. Classification results for three data sets: 1) Only PT-DIGIT-UDS data recorded at 12cm. 2) Use only PT-DIGIT-UDS data recorded at 12m in the test group and data recorded at 40cm in the train group. 3) Use all PT-DIGIT-UDS data for classification.

	1) 12cm data only	2) 12cm data as test	3) All data
Word Error Rate	35.0%	35.0%	**27.8%**

For the first two data distributions, no differences were found. For results considering all PT-DIGIT-UDS data (i.e. 40cm + 12cm data), a relative improvement of 23.0% (absolute of 8.3%) was found when compared with the results achieved considering only PT-DIGIT-UDS data recorded at 40cm.

3.4. Discussion

Overall, the achieved results give support to the demonstration of the first hypothesis (it is possible to extend/adapt the work on SSI to EP) and highlight the problem behind the statement of the second hypothesis (nasal sounds pose problems to some of the considered SSI modalities and their detection with less invasive modalities is possible). This chapter has shown that it is possible to adapt state-of-the art SSI approaches to a new language, EP. However, this language expansion comes with a performance cost in situations where detection of nasality is necessary. Hence, as implied in the second hypothesis, we have noticed performance issues related with the detection of nasal sounds, which still need to be tackled.

When comparing the three analyzed techniques, using as a metric the Word Error Rate, (RGB) Video achieved the best performance. However, as mentioned in section 2.4, several factors influence the performance of an SSI and the use of small datasets in our data driven classification approach, along with differences found between experiments can influence the final results. For example, in the UDS experiment, the used corpora contains data from several speakers, while, in the Video experiment, we considered a single speaker. Therefore, these results should be prudently analyzed before drawing final conclusions.

It should be noted that the required resources, to accomplish most SSI studies, have a higher cost than speech data collections, thus, a large data collection is strongly unadvisable without an adequate set of exploratory studies.

These initial studies have also provided a better understanding of the analyzed input HCI modalities, particularly SEMG, which, in our opinion, is a modality where previous hands-on experience is relevant. Moreover, this initial approach has also contributed for a more comprehensive understanding of the modalities as a whole, and to create a vision of how our third objective – multimodal HCI framework development and analysis – could eventually be accomplished. The section below discusses in more detail the results of each modality.

3.4.1.1. Individual HCI Modalities

Surface Electromyography

Analyzing single modality results, the SEMG experiment showed similar accuracy when compared with the best state-of-the-art results for EMG-based recognition (Wand and Schultz, 2011a). However, in the latter case, a much larger dataset and vocabulary was used. This difference may be explained by the low number of observations, as demonstrated by the improvement verified when *K* increases, by the differences in the number of sensors used and its positioning configuration (monopolar). In our study, we have used 4 pairs of EMG sensors; other studies found in the literature used 5 (Wand and Schultz, 2011a) and 7 (Maier-Hein et al., 2005) pairs of EMG sensors, or multi-channel electrode arrays (Wand et al., 2013b). In posterior studies reported in the following chapters, these equipment limitations have been tackled by working together with the manufacturer and changing the hardware accordingly.

Also for SEMG, differences between the results of the two considered datasets (absolute difference of 20.4% between the best average WER of PT-EMG-A and PT-EMG-B) indicate that the presence of nasality may be a potential error source, as corroborated by the error analysis. These results suggest that tackling the nasality phenomena verified in EP, may be beneficial for the performance of an SSI based on SEMG, for EP. Hence, a possibility could be to analyze if the muscles involved in the nasal process can eventually be detected by the SEMG sensors.

(RGB) Video

The VSR experiment achieved quite interesting WER results, however if, for example, we compare it with the results achieved by Galatas et al. (2012b), it should be remarked that only a single speaker was considered. Also, an evident limitation are the visual marks that were placed in the facial area, prior to the data collection, to achieve a more accurate feature matching. A possibility to minimize this issue could be to capture only half of the face to remove issues caused by facial symmetry. Another possibility could be to perform facial segmentation into several regions of interest. Additionally, one could consider the use of depth information registered in the 2D RGB image to obtain a 3D point cloud defined in the sensor reference frame, from where relevant geometrical and topological information can be obtained, such as facial features (using an established standard such as FACS – Facial Action Coding System (Ekman and Rosenberg, 1997)), lip regions and lip protrusion, along with a parallel research of using the new generation of 3D image feature descriptors, such as SPIN (Spin Image local surface descriptor (Johnson and Hebert, 1999)) or SHOT (Signature of Histograms of OrienTations (Tombari et al., 2010)). In terms of nasality detection, errors were also verified among the minimal pairs when using the RBG information.

Ultrasonic Doppler Sensing

For UDS, the exploratory analysis of the signal has shown differences between the selected words, especially in those where the articulators' movement is more intense. It is also visible a difference across speaker, which corroborates the results achieved by Jennings and Ruck (1995), where the WER performance of the system has a drastic reduction when cross-speaker recognition is considered.

Previous recognition experiments have achieved a WER of 67% in a continuous speech recognition task of 10 English digits (Srinivasan et al., 2010). Although in our case we are considering an isolated digit recognition task on the same vocabulary size and the tests conditions are not the same, if a direct comparison was made with the best result of 27.8% WER, we find a relative improvement of 58.6% (absolute of 39.2%). Additionally, the error analysis seems to indicate that minimal pairs such as *mato/manto* and *peta*/*penta* may cause recognition problems, for an interface based on this approach, which again is related with the nasality phenomena.

Based on Table 13 (that presents the WER results of combining datasets recorded at 12cm and 40cm), the close WER results for train sets recorded at different distances (40cm and 12 cm) indicate similar characteristics of both signals. This line of though is reinforced by the best result (WER 27.8%) being achieved when both datasets are used to train the classification model.

3.5. Summary

This chapter described three experiments based on the same number of input HCI modalities for unimodal SSI for EP: SEMG, Video, and UDS. The overall aim of these experiments was to identify possible technological solutions and methodologies capable to drive the next step of this thesis research, aiming at achieving with success objective 3 (that envisages the demonstration of hypothesis 3): to explore the development of a multimodal SSI, specially targeted for EP.

In the first experiment, we applied SEMG for the first time to EP. We collected two corpora, the first composed by 8 EP words and the second composed by 4 minimal pairs of EP words, with 12 and 15 observations of each word, respectively. For feature extraction, we used an existing technique based on time-domain features. Afterwards, we were able to classify the silent pronunciation of the selected prompts, using a classification scheme based on the Dynamic Time Warping (DTW) technique. The overall results showed a performance discrepancy between the two datasets, with the first showing better accuracy results with a mean WER of 22.5% and the second 42.9% WER. The error analysis of these results allowed us to determine that for the second corpus, more than 50% of the error was introduced by the minimal pairs of words that only differ on nasal sounds, confirming the existence of problems at the level of the nasality phenomenon, when applying state-of-the-art techniques to EP. In chapter VI of this thesis, we address this problem in detail and propose a solution, capable of satisfying our hypothesis number 2 (the possibility of nasality detection with a suitable SSI).

The second experiment introduced a novel technique for feature extraction and classification in the VSR domain. This technique was based on an existing robust scale, rotation and luminance invariant feature detection transform in computer vision (FIRST), which has been applied in the past to real-time applications, such as Augmented Reality. The novelty of our approach was the application of FIRST to extract skin features spread across different regions of a speaker’s face and to track their displacement, during the time that elapses while the speaker utters a word. For this experiment, we collected a corpus, similar to the second corpus of the first experiment (i.e. 4 minimal pairs of EP words), but with 14 different observations of each word. For classification, the same process based on DTW was used. DTW was used to find an optimal match between a sufficient number of observations and, in our experiments, we were able to calculate a mean WER of 8.63% (STD 4.01%) with the best figure of 2.5% WER for the best run. As expected, error analysis detected recognition problems between similar words, where the only difference is a nasal phoneme instead of an oral one (again, the nasality phenomenon popping up). This demonstrates the difficulty in distinguishing pairs of words that only differ on nasal sounds. However, in this experiment many word classes were successfully discriminated, supporting our hypothesis that the skin deformation of the human face, caused by the pronunciation of words, can be captured by studying the local time-varying displacement of skin surface features using FIRST, and that this framework can be applied to a successful vision-based SSI system for EP.

The last reported experiment of UDS-based speech recognition for EP, describes the device used in the acquisition of this type of data and an analysis to the signal that shows viability for using this type of approach in speech recognition. The conclusions and results are in line with the state-of-the-art, achieving a best WER of 27.8%. However, the experiment also shows that much is yet to be explored for this modality, particularly, understanding what articulators are exactly being captured. For example, does the Doppler signal contains velum or tongue movement information (question addressed in chapter VI)?

Overall, the results of the experiments are promising and motivate the development of a multimodal SSI based on these technologies. However, to achieve our goals we need to create the means to promote the joint exploitation of modalities. That is the major topic to be addressed in the next chapter.

CHAPTER IV

A Multimodal Framework for Silent Speech Research

““Correm rios, rios eternos por baixo da janela do meu silêncio.”

Fernando Pessoa, Livro do Desassossego

Contents

In the previous chapter we have explored single modalities, selected based on their non-invasive characteristics and promising results. In this chapter, we describe the grounds for exploring other hypotheses, namely the third (a multimodal HCI approach, based on less invasive modalities, has the potential to improve recognition results) and fourth (supplementary measures acquired from more invasive modalities can be use as ground truth) hypotheses, related with the combined use of modalities.

Based on the known limitations and potential of each modality, it is possible that (some) single modalities do not acquire sufficient information to develop successful interfaces, particularly those that aim at natural HCI. Thus, by acquiring knowledge from multiple modalities, we gain access to a more comprehensive view of the articulation process. Furthermore, by knowing the movements of many different articulators or having different perspectives of the same articulator might allow the development of interfaces that better model the articulation.

The main objective of the work presented in this chapter is the creation of a framework that enables acquiring the synchronized data from distinct modalities, promoting their joint usage and exploration and also providing the grounds for achieving our goals. The proposed framework goes beyond the data acquisition process itself, providing an adequate set of methods that facilitate further SSI research. Examples of application of the framework include articulatory and speech production studies, and HCI systems.

We present different instantiations of the framework that illustrate multiple stages of research and demonstrate the potential, flexibility and extensibility of the proposed solution. The stages go from defining the requirements and the corresponding setups and methods to the analysis and classification of the processed data. Our aim is to demonstrate how the use of multiple modalities can be maximized, in a way such that the weakest points of one modality can be mitigated by other(s).

For that purpose, we focus on the following premises aligned with the third and fourth hypotheses: (1) the use of non-invasive, low cost and promising input modalities such as (RGB) Video, Depth, SEMG and UDS to understand which information stream achieves the best results in a speech recognition task and how the existing redundancy, among these modalities, influence the overall results; (2) the use of more invasive and obtrusive modalities, like RT-MRI and US, to obtain direct measures of articulators, such as the tongue and the velum, that can be used as ground truth data to better understand other modalities and take further advantage of their capabilities.

For better readability, the entire framework is divided between chapters IV, V and VI. In this chapter we present a general and conceptual overview of the framework. Then, for each part of the framework, concrete instantiations are presented, not including the analysis and classification part.

The last stage (analysis and classification) and a more complete case study are presented, respectively, in chapters V and VI.

4.1. A Multimodal SSI Framework Overview

A framework supporting research in multimodal SSI can be defined by a set of 5 stages, as depicted in Figure 31: (1) Data Collection method and General Setup for Acquisition; (2) Online and Offline Synchronization; (3) Collected Corpora; (4) Data Processing and Feature Extraction, Selection and

Fusion; (5) Analysis, Classification and Prototyping of single and multiple modalities. The final result can be applied to HCI and to address existing problems, such as nasality detection.

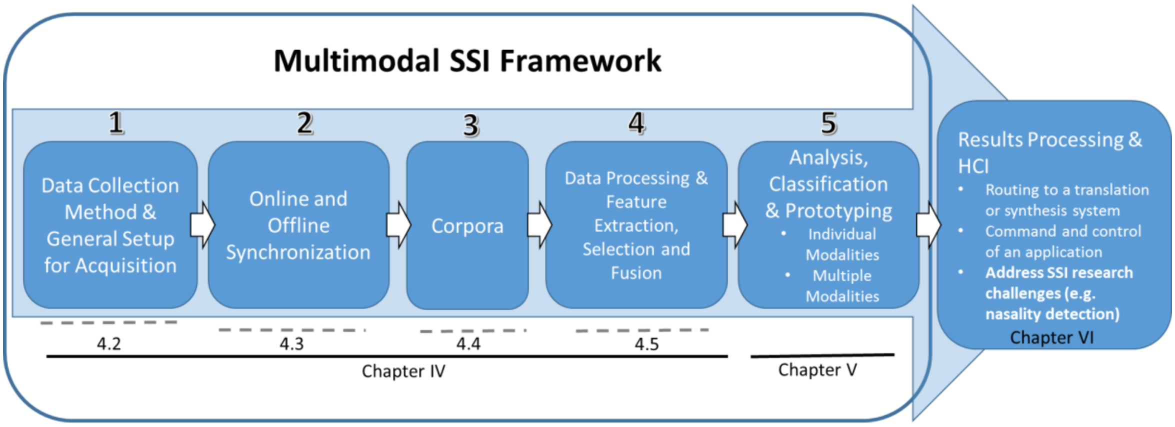

Multimodal SSI Framework
1
Data Collection Method & General Setup for Acquisition
2
Online and Offline Synchronization
3
Corpora
4
Data Processing & Feature Extraction, Selection and Fusion
5
Analysis, Classification & Prototyping
• Individual Modalities
• Multiple Modalities
Results Processing & HCI
• Routing to a translation or synthesis system
• Command and control of an application
• Address SSI research challenges (e.g. nasality detection)
Chapter VI
4.2
4.3
4.4
4.5
Chapter IV
Chapter V

Figure 31. Overview the proposed multimodal SSI framework with multiple stages and possible outcomes.

The stages that constitute the framework are structured in a similar manner to conventional machine learning or ASR systems pipelines. The difference towards these systems relies on its multimodal nature, described as follows.

When first approaching the silent speech recognition problem (represented as stage 1 in Figure 31), we start by defining the requirements and the methods to solve the problem, including which technologies can provide an adequate response. In many situations, we need to deal with the limitations of each technology and to select a single modality is not sufficient to tackle the interaction barriers.

After defining our method and the different components of our setup, we need to acquire data samples. This is usually a cumbersome and time costly procedure, particularly for global-data techniques that try to generalize the data observations. For avoiding future acquisitions, it is important to ensure a synchronous acquisition of modalities (represented as stage 2 in Figure 31), but also that the maximum of data is extracted. Thus, an extendable approach that allows for additional modalities to be effortlessly included is desirable.

After acquiring the corpora, some operations related with the storage and processing the respective metadata (e.g. annotation) may be necessary. This is depicted as intermediate stage 3 in Figure 31.

At a posterior stage, after storing/processing the corpora, we need to transform the data into usable information from a research perspective. This is depicted in Figure 31 as stage 4 and includes for each modality: processing the raw data; aligning in time, enabling a synchronous relation between modalities; extracting relevant information and discarding the redundant one; selecting the best features, consequently reducing the dimensionality for posterior classification stages; and, when applicable, fuse cross-modalities information.

In the last stage, we can analyze and classify the resulting information from individual modalities using other modalities as ground truth, or consider fusing multiple streams of information in a multimodal scenario.

In the following sections we describe in more detail the stages 1, 2, 3 and 4 of the framework with concrete examples for each stage.

4.2. Data Collection Method and General Setup for Acquisition

After determining which modalities are going to be collected, the main goal of this stage is to define the requirements for each modality and to define the protocol for data acquisition. This includes analyzing the restrictions imposed by the simultaneous use of different devices. For example, during a RT-MRI acquisition it is not possible, with the current technology, to acquire SEMG; or when collecting SEMG and US, if the ultrasound probe is placed beneath the chin, the space for placing SEMG sensors in the neck region will be limited.

When defining the data collection method, it is also crucial, for future use, that the data streams are correctly synchronized. For the data acquisition setup we propose two paths for synchronizing the data streams, here referred as online and offline synchronization (described in detail in section 4.3). In the first case and the most common, we explore the possibility of using hardware or software markers to synchronize the data across streams. However, in some cases, such as those including data from RT-MRI, it becomes impossible to conduct a simultaneous multimodal acquisition due to technological restrictions. Therefore, the proposed framework takes advantage of an offline synchronization method, as depicted in Figure 32.

Then, after acquiring, processing, and aligning all the streams we are finally able to join them for a posterior analysis and classification.

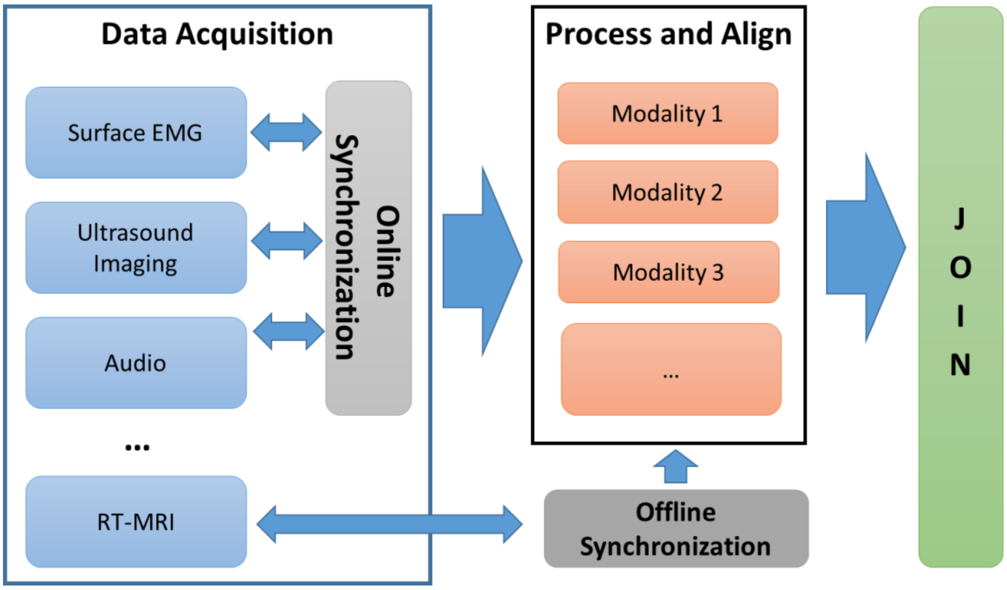


Figure 32. Conceptual model of a multimodal data collection setup scenario and joint exploration of the selected modalities.

The following subsections present two concrete examples of acquisition setups. These examples illustrate the instantiation of tangible multimodal solutions using the proposed framework.

4.2.1. Data Collection Setup

Given our aim to develop a multimodal SSI, a concrete data collection setup (in line with Figure 32) for a novel SSI based on the following modalities was defined as our target: (1) Facial information acquired from Visual and Depth sensors; (2) surface EMG of muscles related with speech articulation; (3) capture of facial movements during speech using UDS.

After assembling all the necessary data collection equipment which, in the case of Ultrasonic Doppler, led us to the in-house development of custom built equipment (Freitas et al., 2012b) based on the work of Zhu et al. (2007), we have created a first version of the multimodal data collection setup depicted in Figure 33, hereon referred as Setup A.

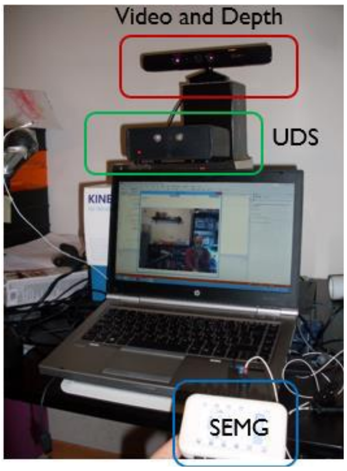


Figure 33. Setup A: Acquisition devices and laptop with the data collection application running for setup A based on (RGB) Video, Depth information (provided by infrared reflection), Ultrasonic Doppler sensing and Surface Electromyography.

The devices employed in Setup A were the following:

- Microsoft Kinect for Windows that acquires visual and depth information;
- Surface EMG sensor acquisition system from Plux (“Plux Wireless Biosignals,” n.d.), that captures the myoelectric signal from the facial muscles;
- Custom built dedicated circuit board (referred to as UDS device), that includes: 2 ultrasound transducers (400ST and 400SR working at 40 kHz), a crystal oscillator at 7.2 MHz and frequency dividers to obtain 40 kHz and 36 kHz, and all amplifiers and linear filters needed to process the echo signal (Freitas et al., 2012b);

For the multiple modalities included in the setup instantiation described above, some additional information is due regarding particular aspects of its configuration.

The Kinect sensor was placed at approximately 70.0cm from the speaker. It was configured, using Kinect Software Development Kit (SDK) 1.5, to capture a 24-bit RGB color video stream at 30 frames per second with a resolution of 640x480 pixel, and an 11-bit depth stream to code the Z dimension, with a resolution of 640x480 pixel, also at 30 frames per second. Kinect was configured to use the Near Depth range (i.e. range between 40.0cm and 300.0cm) and to track a seated skeleton.

The utilized SEMG acquisition system was the one previously described in section 3.1 and used five pairs of EMG surface electrodes connected to a device that communicates with a computer via Bluetooth. As depicted in Figure 34, each sensor was attached to the skin using a single use 2.5cm diameter clear plastic self-adhesive surface and also considering approximately 2.0cm spacing between the electrodes center for bipolar configurations. As performed earlier, before placing the surface EMG sensors, the sensor location was previously cleaned with alcohol. While uttering the prompts any movements besides the ones associated with speech production were kept to a minimum.

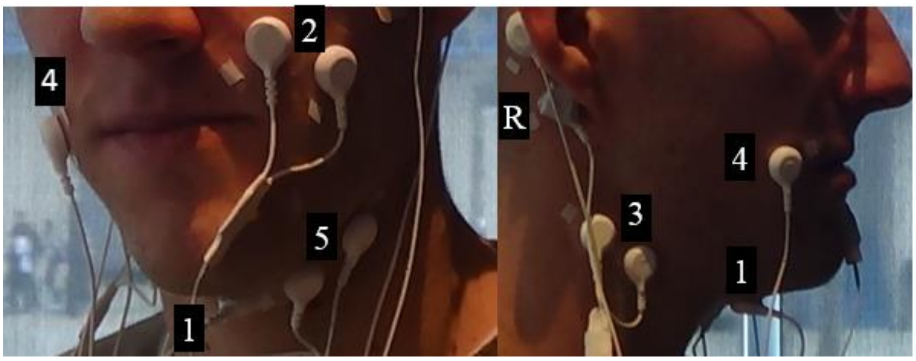


Figure 34. Surface electromyography electrodes positioning and the respective channels (1 to 5) plus the reference electrode (R) in setup A.

For this setup (setup A), the five electrode pairs were placed in order to capture the myoelectric signal from the following muscles: the *zygomaticus major* (channel 2); the tongue (channel 1 and 5), the *anterior belly of the digastric* (channel 1); the *platysma* (channel 4) and the last electrode pair was placed below the ear between the mastoid process and the mandible. The surface EMG channels 1 and 4 used a monopolar configuration (i.e. placed one of the electrodes from the respective pair in a location with low or negligible muscle activity), with the reference electrodes placed on the mastoid portion of the temporal bone. The positioning of the EMG electrodes 1, 2, 4, and 5 was based on previous work (e.g. (Wand and Schultz, 2011a)) and EMG electrode from channel 3 was placed according to findings by the author about the detection of nasality in SSI (Freitas et al., 2014c), described in detail in Chapter VI.

4.2.2. Augmenting the data collection setup with “ground truth” modalities

Motivated by the need to further analyze the selected modalities and to obtain direct measures of articulators such as the velum and the tongue, to serve as ground truth data, we have decided to propose an extended setup (setup B) depicted in Figure 35, which features two new modalities added to the framework: audio and ultrasound imaging. The available technology allowed for these two modalities to be acquired simultaneously with the modalities from Setup A, giving origin to Setup B. Ultrasound Imaging was added to the setup to acquire further information about the tongue movement.

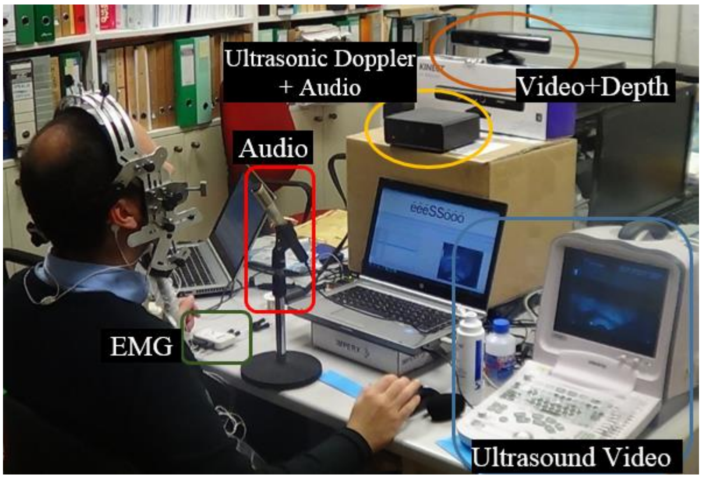


Figure 35. Setup B: Augmented acquisition setup based on Video (RGB), Depth information (provided by infrared reflection), Ultrasonic Doppler sensing, Surface Electromyography, Ultrasound Imaging and Audio.

For Setup B we added the following devices to Setup A:

- Mindray DP6900 ultrasound system with a 65EC10EA transducer, an Expresscard|54 Video capture card, to capture the ultrasound video;
- SyncBrightUp Audio-Video synchronization unit
- Directional microphone

To support all the devices listed in Setup A and B we used two sound cards, a TASCAM US-1641 (main sound board), a Roland UA-25 EX (secondary sound board) and 2 laptops. The hardware connection of all devices is mapped into the scheme depicted in Figure 36.

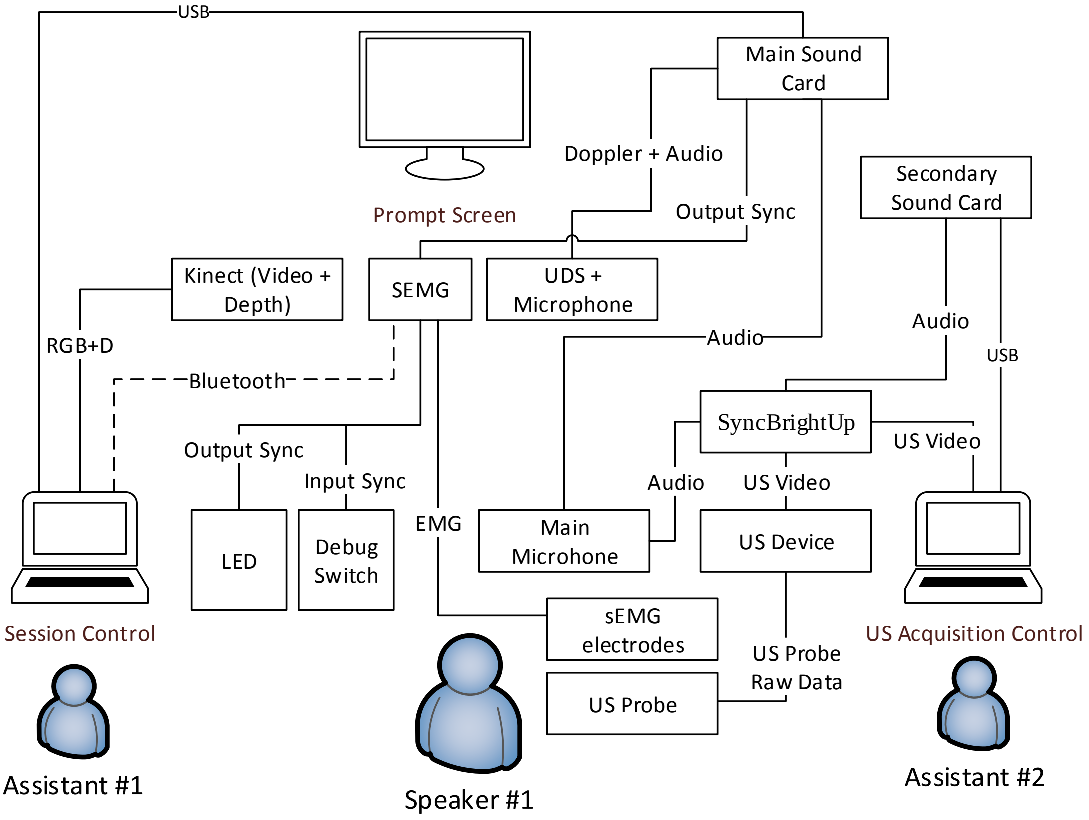


Figure 36. Device scheme for multimodal data acquisition of (RGB) Video, Depth information (provided by infrared reflection), Ultrasonic Doppler sensing, surface Electromyography, Ultrasound Imaging and audio (2 microphones).

For setup B, in order to capture tongue information, the same 5 pairs of surface EMG electrodes were placed in the areas of the neck beneath the chin, somewhat limited by the Ultrasound probe placed beneath the chin, as depicted in Figure 37.

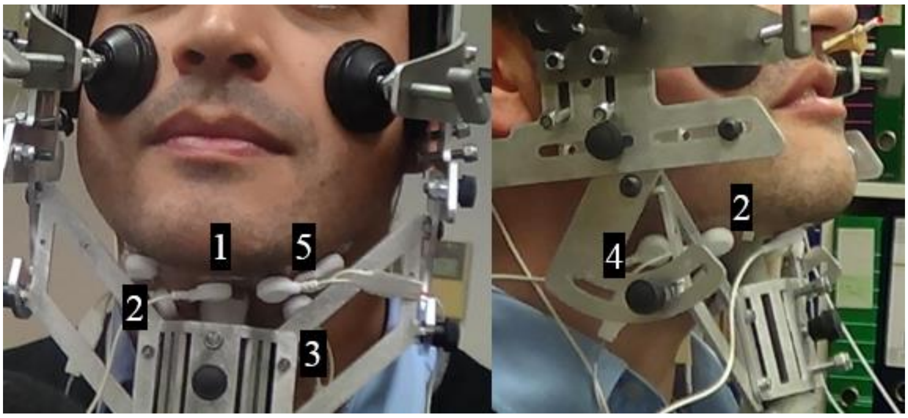


Figure 37. Frontal (left image) and lateral (right image) view of the surface electromyography sensors (channels 1 to 5) in setup B.

The UDS device was placed at approximately 40.0cm from the speaker and was connected to the main sound board, which in turn was connected to the laptop through a USB connection. The

Doppler echo and the synchronization signals were sampled at 44.1 kHz and to facilitate signal processing, a frequency translation was applied to the carrier by modulating the echo signal by a sine wave and low passing the result, obtaining a similar frequency modulated signal centered at 4.0 kHz.

The ultrasound setup comprises the Mindray DP6900 ultrasound system to capture the ultrasound video, a microphone, connected to a Roland UA-25 external soundcard and a SyncBrightUp unit, which allows synchronization between the audio and ultrasound video, recorded at 30 frames per second. A stabilization headset was used (Scobbie et al., 2008) to ensure that the relative position of the ultrasound probe towards the head was kept during the acquisition session, also securing the ultrasound probe below the participant's chin.

In terms of acquisition protocol, we started by placing the stabilization headset in the participant’s head and the ultrasound probe in place, bellow the chin, but left at some distance. This was to serve as a reference for placing the SEMG sensors outside the probe region. Since the US probe, when tightly secured in place against the chin, causes some discomfort and the SEMG sensor placing takes some time, we opted for not doing so beforehand. After SEMG sensor placement the headset was properly secured. Finally, the participant was asked to push the ultrasound probe against the chin as much as possible, keeping it within comfort levels, while the ultrasound was monitored to check for proper tongue imaging.

The prompts were presented in a computer display, and the participant instructed to read them when signaled (prompt background turned green). For each recorded sequence, EMG recording was started before US recording and stopped after the US was acquired.

4.2.3. Scaling the setup to additional modalities

With setup B, we already support capturing seven streams of information. However, if we would like to extend our setup with additional modalities, the cost would be minimal. For example, if one would like to capture information from the vibration of the vocal cords, it would simply require connecting the additional acquisition device to the main sound board, as depicted in Figure 38.

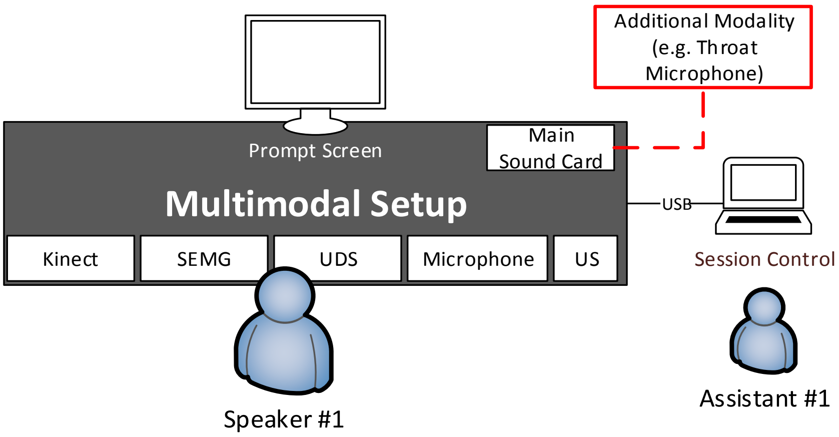


Figure 38. Device scheme with an additional modality (dashed red line)

More restrictive solutions such as RT-MRI require using a separate setup and need to recur to an offline synchronization method, described in section 4.3.1.

4.3. Synchronization

For a successful joint use of modalities we need to ensure the necessary conditions to record all signals with adequate synchronization. The challenge of synchronizing all signals resides in the fact that an effective synchronization event needs to be captured simultaneously by all input modalities.

Setup A

In order to synchronize all input modalities from setup A (Video, Depth, SEMG and UDS) via time alignment between all corresponding input streams, we used an I/O bit flag in the SEMG recording device, which has one input switch for debugging purposes and two output connections, as depicted in Figure 36. Synchronization occurred when the output of a synch signal, programmed to be automatically emitted by the surface EMG device at the beginning of each prompt, was used to drive a Light-Emitting Diode (LED) and to provide an additional channel in an external sound card. The alignment between the video and depth streams was ensured by the Kinect SDK.

Using the information from the LED and the synchronization signal, the remaining signals were time aligned after data acquisition. To align the RGB video and the depth streams with the remaining modalities, we used an image-based standard template matching technique that automatically detected the LED position on each color frame.

With the external sound card channel configured to maximum sensitivity, the activation of the output bit flag of the SEMG recording device generated a small voltage peak on the recorded synchronization signal. To enhance and detect that peak, the second order derivative was computed on the signal, followed by an amplitude threshold. Then, after automatically detecting the peak, we removed the extra samples (before the peak) in all channels. The time-alignment of the EMG signals was ensured by the SEMG recording device, since the I/O flag was recorded in a synchronous way with the samples of each channel.

Setup B

In setup B, the acquisition of ultrasound related data was managed by Articulate Assistant Advanced (AAA) ("Articulate Assistant Advanced Ultrasound Module User Manual, Revision 212," n.d.), which was also responsible for recording audio and ultrasound video and triggering the SynchBrightUp unit. The SyncBrightUp unit, when triggered, introduced synchronization pulses in the audio signal and, for each of those, a white square on the corresponding ultrasound video frames of the sequence. The synchronization between the audio and the US video was tuned after the recording session, in AAA, by checking the pulses in the audio signal and aligning them with the frames containing bright squares.

For synchronizing the EMG signals (and remaining modalities recorded synchronously with the EMG using the main soundcard) with the US video, we used the audio signal provided by the SyncBrightUp unit, which contained the synchronization pulses, and also recorded it using the main sound card along with the pulse emitted by the EMG device. The result was that we got the audio signal with the synchronization pulses recorded synchronously with the remaining data in both settings (SEMG and US). The audio signal was then used for synchronization between them.

To measure the delay between the two settings (US and remaining modalities), we performed a cross-correlation between the audio tracks. After resampling the signal with the lower sample rate, we used the maximum value of the cross-correlations between them, yielding the time lag between the two, and then removed the necessary samples from the EMG signals (for which the recording was always started first).

4.3.1. Offline synchronization

Offline synchronization comes into play when it is not possible, with current technology, to have a simultaneous acquisition of modalities that could enable a more straightforward synchronization. This more exclusive case happens for example with RT-MRI that entails the use of a machine on which no metal is allowed.

Hence, motivated by the existing RT-MRI information, we have designed a method that allows aligning data from asynchronous collections. This method relies on the use of a common data source between the target setups: speech data. In the case of RT-MRI, it is possible to collect sound by using a fiber optic microphone. Having the two distinct audio recordings, we need to establish the bridge between the two audio recordings. For that purpose, we use the DTW, this time not for classification (as in chapter III), but to find the optimal match between the two audio sequences. Thus, based on the DTW result, we are able to extract a warping function applicable in both directions to the respective time axis.

This method, although not applicable in all situations, allows to take advantage of more reliable data such as the images from RT-MRI. A case study and a more concrete example of this process is described in Chapter VI.

4.4. Collected Corpora

This section describes two corpora, collected at different times, using the framework. The first corpus vocabulary consists of common EP words for different command and control scenarios. The second was designed to include different tongue movements that occur during EP speech.

4.4.1.1. A command and control SSI corpus

The corpus was collected using setup A with the aim of building an SSI for EP. It includes 32 Portuguese words, which can be divided into 3 distinct sets.

The first set, adopted from previous work found in the literature for other languages (e.g. (Srinivasan et al., 2010)) and for EP in prior work of the author (Freitas et al., 2012b), consists of 10 digits from zero to nine.

The second set contains the 4 minimal pairs of common words in EP that only differ on nasality of one of the phones (minimal pairs regarding this characteristic, e.g. (using SAMPA) Cato/Canto [katu]/[k6~tu] or Peta/Penta [pet6]/[pe~t6], used in the experiments described in the third chapter.

Finally, Table 14 shows the third set, with 14 common words in EP, taken from context free grammars of an Ambient Assisted Living (AAL) application that supports speech input. The set was chosen based on past experiences of the author (V. Teixeira et al., 2012).

A total of 99 prompts per session were randomly presented to the speaker with each prompt being pronounced individually, in order to allow isolated word recognition. All prompts were repeated 3 times per recording session. Three additional silence prompts were also included at the beginning, middle and at the end of the session, to capture a non-speaking state of each user.

Table 14. Set of words of the EP vocabulary, extracted from Ambient Assisted Living contexts.

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

The data collection was split in two rounds. For the first round, the participants silently articulated the words and for the second round an audible acoustic signal was also acquired. The first round of recordings comprised 9 sessions of 8 native EP speakers (one speaker recorded two sessions) – 2 female and 6 male – with no history of hearing or speech disorders, with an age range from 25 to 35 years old and an average age of 30 years old. In this first round, considering the interest in studying differences found between silently articulated speech and audible uttered speech, related with the lack of acoustic feedback (Herff et al., 2011), we have only recorded silent speech. Thus, no audible acoustic signal was produced by the speakers during the recordings and only one speaker had past experience with silent articulation.

In the second round, we collected data from three EP native speakers, one male that also participated in the previous data collection, aged 31, and two female elderly speakers, aged 65 and 71, without any history of speech disorders known so far. In this second round of data collection, each speaker recorded two sessions without removing the SEMG electrodes or changing the recording position.

4.4.1.2. An SSI Corpus for Tongue Movement Studies

The second example is a dataset designed to explore the use of ground truth modalities. Using setup B, we have synchronously acquired Video, Depth, UDS, and EMG data along with Ultrasound (US) imaging, a modality that is able to provide essential ground truth information about tongue movement.

For that purpose, we have created a corpus where we cover several tongue transitions in the anterior-posterior axis (e.g. front-back and vice-versa) and also elevation and depression of several tongue parts. After acquiring the data and synchronizing all data streams, we determined and characterized the segments that contain tongue movement, based on the US data.

To define the corpus for our experiment we considered the following goals: (1) record tongue position transitions; (2) record sequences where the movement of articulators other than the tongue is minimized (lips, mandible, velum). Considering these goals, we selected several /vCv/ contexts for the consonants: [k, l, L, t, s] (using SAMPA). Context-wise, we varied the backness and the closeness of the vowels (e.g. [aka, iki, uku, EsO, itu, eLe]). In order to explore the tongue transitions in vowels, we considered the combination of several vowel sequences (/V1V2V1/). These combinations include the tongue transitions in terms of vowel closeness and backness as well. The selected sequences are composed by the transition /V1V2/ and its opposite movement /V2V1/. For instance, the prompt "iiiuuuiii" is composed by both the tongue transition from [i] to [u] and from [u] to [i]. In order to minimize movement of other articulators than the tongue, we have not included bilabial and labio-dental consonants.

Based on pilot recordings, we noticed that the speaker would get uncomfortable after a long time using the US headset (see Figure 37). As such, we focused our aim on the most relevant transitions in order to minimize the length of the recordings.

For each prompt in the corpus, three repetitions were recorded and, for each recording, the speaker was asked to say the prompt twice, e.g. "iiitiii…iiitiii", with around one second of interval, yielding a total of 6 repetitions per prompt. The prompts were recorded in a random order. To facilitate movement annotation we asked the speakers to sustain each syllable for at least one second. The prompts were presented on a computer display and the participant was instructed to read them when signaled (prompt background turned green). For each recorded sequence, the US recordings started before the remaining modalities where acquired.

The three speakers participating in this study were all male native speakers of EP, aged 28, 31, and 33 years old. No history of hearing or speech disorders were known for any of them at the time of the data collection. Each speaker recorded a total of 81 utterances containing 2 repetitions each, giving a total of 486 observations (3 speakers x 81 utterances x 2 repetitions of each utterance).

4.5. Data Processing, Feature Extraction and Feature Selection

After the multimodal data acquisition stage, we needed to process our data streams, extract their characteristics and prepare them for analysis and classification, addressing issues such as high-dimensionality in the feature space.

In the following subsections, we describe the processing methods made available for the several modalities supported in our approach.

4.5.1. Processing US data for ground truth definition

The ultrasound video data was processed to identify the segments where tongue movement was present. The goal was to examine the video sequence and annotate (tag) those segments where the tongue is in motion. Given the large amount of US data, it was important to ensure that the same criteria were used for movement annotation in all sequences, which is hard to accomplish when considering manual annotation. This led us to consider an automatic annotation[3] approach as follows.

The inter-frame difference is obtained, between each pair of video frames in the sequence, by computing the difference between corresponding pixels. The pixel-wise differences are added and the result used as an indication of the amount of movement happening between the two frames. Computing these differences along the whole US sequence, when the tongue moves, the inter-frame difference rises resulting in local maxima (refer to Figure 39 for an illustrative example). Starting from these maxima, the second derivative is considered, left and right, to expand the annotation. In order to identify the repetitions, in each recorded prompt, we used the envelope of the speech signal.

Considering, for example, one repetition of the sequence "iiiuuuiii", at least four tongue movements are expected (Figure 39): one movement at the start, one backward movement for the transition from [i] to [u], one forward movement for the transition from [u] to [i] and one final at the end when the tongue goes into the resting position. Therefore, the two annotated regions on the middle of each repetition correspond to the transitions between sounds (Figure 39). Since we know which sounds are involved in each transition, besides annotating tongue movement we can also add further information. For instance, from [i] to [u] the tongue moves backwards, and from [u] to [i] the tongue moves forward. This allows filtering the data by the type of tongue movement and therefore, explore the SEMG signals assessing both their applicability to detect tongue movement in general, as well as, specific movements.

For more details about the proposed method we forward the reader to Silva and Teixeira (2014).

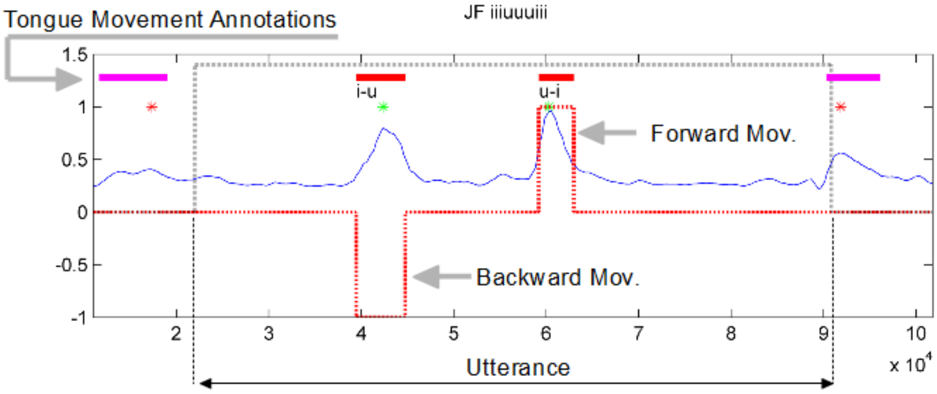


Figure 39. Inter-frame difference curve (in blue) for a speaker uttering "iiiuuuiii". The automatic annotation identifies all segments with tongue movement and movement direction, for those inside the utterance.

[3] Work made in collaboration with Dr. Samuel Silva from the University of Aveiro, who has implemented the described approach, in the context of the project IRIS (ref. 610986, FP7-PEOPLE-2013-IAPP).

4.5.2. Feature Extraction

For preparing the data for posterior analysis and classification, we selected recent Feature Extraction (FE) techniques for each modality, especially those that reported good results in the existing literature. For Video, we focused on two feature extraction techniques: the first based on appearance methods and a second one dedicated to the extraction of articulatory information.

4.5.2.1. Surface Electromyography

For feature extraction from SEMG, we used an approach which is based on temporal features, similar to the one described in section 3.1, without applying the feature reduction (FR) technique LDA. For the particular case of the tongue gesture data (from Setup B), the EMG signals were first normalized, then a 50Hz notch filter was applied to the signals and they were also filtered using Single Spectrum Analysis (SSA). The features were extracted for each given EMG signal frame of 30ms and a frame shift of 10ms was considered. A context width of 15 frames was also used (as described in Eq. 15, Chapter III), generating a final feature vector of 155 dimensions per signal channel. Finally, we stacked all the channels in a single feature vector of 775 dimensions (5 channels x 155).

4.5.2.2. Ultrasonic Doppler Sensing

For UDS, we followed the same approach to that of section 3.3. Additionally, since in some cases no acoustic signal existed (e.g. first acquisition round for setup A, described in section 4.4.1.1), we use the UDS signal to understand if facial movement occurred. For this purpose, we used a movement detection algorithm, inspired in the work of Kalgaonkar et al. (2007), that uses the energy of the UDS pre-processed spectrum information around the carrier (Srinivasan et al., 2010). After obtaining the spectrum information, we applied a third order moving average filter and obtain the energy contour, as depicted on the top plot of Figure 40. Then, we applied a threshold to the resulting signal (depicted in the center plot of Figure 40). The threshold value was calculated using the mean of the energy contour of the signal and the silence prompts of each speaker. The variations associated with the facial movement in the resulting bit signal were then grouped under the assumption that only one word was uttered. This method allowed us to segment the part of the utterance where silent speech actually occurred and to remove artifacts in the beginning and end of the utterance (bottom plot of Figure 40).

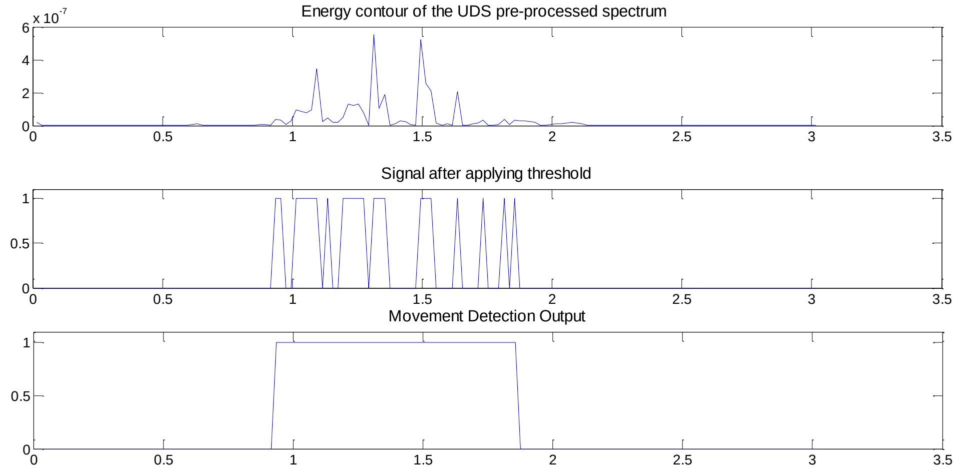


Figure 40. Movement detection processing phases (from top to bottom) for the prompt "Voltar" [voltar]: energy contour of the Ultrasonic Doppler sensing pre-processed signal; signal after applying a threshold value and the signal output.

4.5.2.3. Video and Depth

For (RGB) Video, we extracted two types of features; the first used appearance-based methods applied to the RBG image, while in the second we extracted articulatory information from the lips (i.e. lip movement). For Depth, we have only used the extraction method based on appearance, applied to the "grayscale" depth image.

Appearance-based features

For the first type of features, we started by establishing a Region of Interest (ROI) containing the lips and surrounding areas. Using real-time Active Appearance Models (AAM) (Cootes et al., 2001), we were able to obtain a 64x64 pixel ROI centered at the speaker's mouth. Then, we applied an appearance based method, which, due to variations in illumination, skin color, facial hair and other factors, is usually preferred to shape based methods. In this context, one of the most classical approaches is to use a DCT transform (Oppenheim et al., 1999) in the ROI. Following previous studies (Gurban and Thiran, 2009), we compressed the pixel information by computing the DCT and keeping the low spatial frequencies by selecting the first 64 coefficients contained in the upper left corner of the 64x64 coefficients matrix. We only considered the odd columns of the DCT, in order to take advantage of the facial symmetry and imposing horizontal symmetry to the image. After applying the 2D DCT, the first and second temporal derivatives were appended to the feature vector to capture visual speech dynamics (in line to what is used in the literature (Galatas et al., 2012b; Potamianos et al., 2003)), generating a final feature vector of 192 dimensions per frame. The existing variations between speakers and recording conditions were attenuated by using Feature Mean Normalization (FMN) (Potamianos et al., 2003).

Articulatory features

Additionally, we also processed Video to extract information related with articulation. The RGB image stream of the video modality, provides a stream of data from which it is possible to track the movement and shape variation of the external articulators such as the lips. Using this stream of information it is possible to follow the movement and shape variation of the external articulators such as the lips.

We developed a simple application that performs lip tracking[4], using 4 points of the external contour - top, bottom and corners. These points allowed us to extract the measures of rounding and spreading of the lips on any frame and were obtained after several image processing stages, as follows.

The first step to obtain the four external points is to crop the original image to the lips ROI, using the tracked points of the lips as references. Since the tracked points are estimated based on AAM (Cootes et al., 2001), it takes some time (i.e. frames) to converge correctly, not being able to keep up with some lip movements during speech. Therefore, the tracked points although useful to derive a ROI, are not appropriate to determine their coordinates.

The next step consists in the lips segmentation process (depicted in Figure 41), which starts by converting the ROI to the YIQ (Luminance, In-phase, Quadrature) color space.

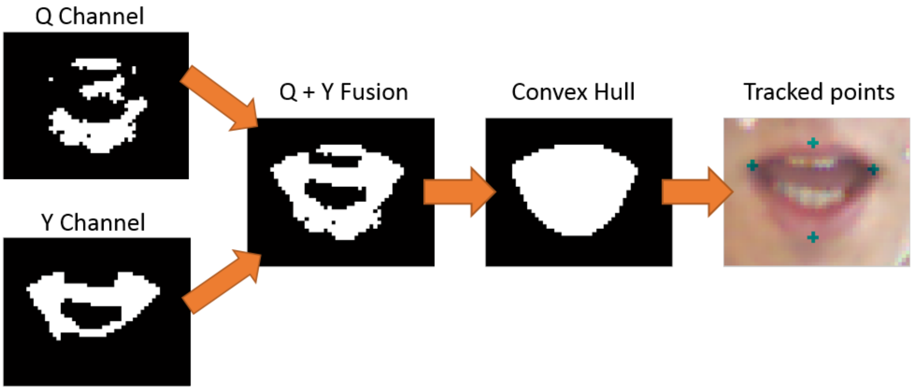


Figure 41. Lips segmentation process based on the YIQ (Luminance, In-phase, Quadrature) color space of the region-of-interest. The two initial images on the left are a binary representation of the Q and Y channel after noise removal. The third image is the fusion of the previous. The fourth image is the result after computing the convex hull of the third image. The fifth and last image (on the right) is the ROI in the RGB color space with 4 2D points extracted from the edges of image 4.

[4] Work done in collaboration with Hélder Abreu, MSc student and intern at the Microsoft Language Development Center.

The Y channel, that conveys the achromatic energy of the image, allows a consistent extraction of the horizontal middle region of the lips (good for corner points), while the Q, one of the chrominance channels, provides a cloud of points with higher density in the lips (good for top and bottom points). After some noise removal, using techniques such as opening or erosion, with subsequent removal of small groups of pixels, the image is stabilized using previous frames. Then both binary images are joined into a single one. The resulting image is used to obtain the coordinates of the 4 points and to extract the characteristic width and height of the lips, computing the Euclidean distances between them.

As an example, Figure 42 shows three plots of the lips' width and height variations for the words "Voltar" [voltar], "Cato" [katu] and "Ligar" [ligar]. Each one of these plots has several pictures corresponding to the indicated frame number (gray line). Observing the variations of the plots, we may predict the appearance of the lips in a given frame. For instance, considering the word "Voltar" and knowing that each utterance starts in a non-speaking state, we can assume that in the 19th frame the lips are closed, while in the 33rd the corners of the lips are closer that in the 19th (since the width value is smaller). We can then expect lips with a rounder appearance in the 33rd frame of Figure 42.

A lower width between the corners of the lips indicates a rounding appearance (e.g. frame number 43 of the word "Cato"), while a higher width associated to a lower height indicates that the lips are closed (e.g. frame 85 of the word "Ligar"). Therefore, using the width and height values of the lips in each frame, we were able to extract articulatory parameters, such as lip opening and lip rounding.

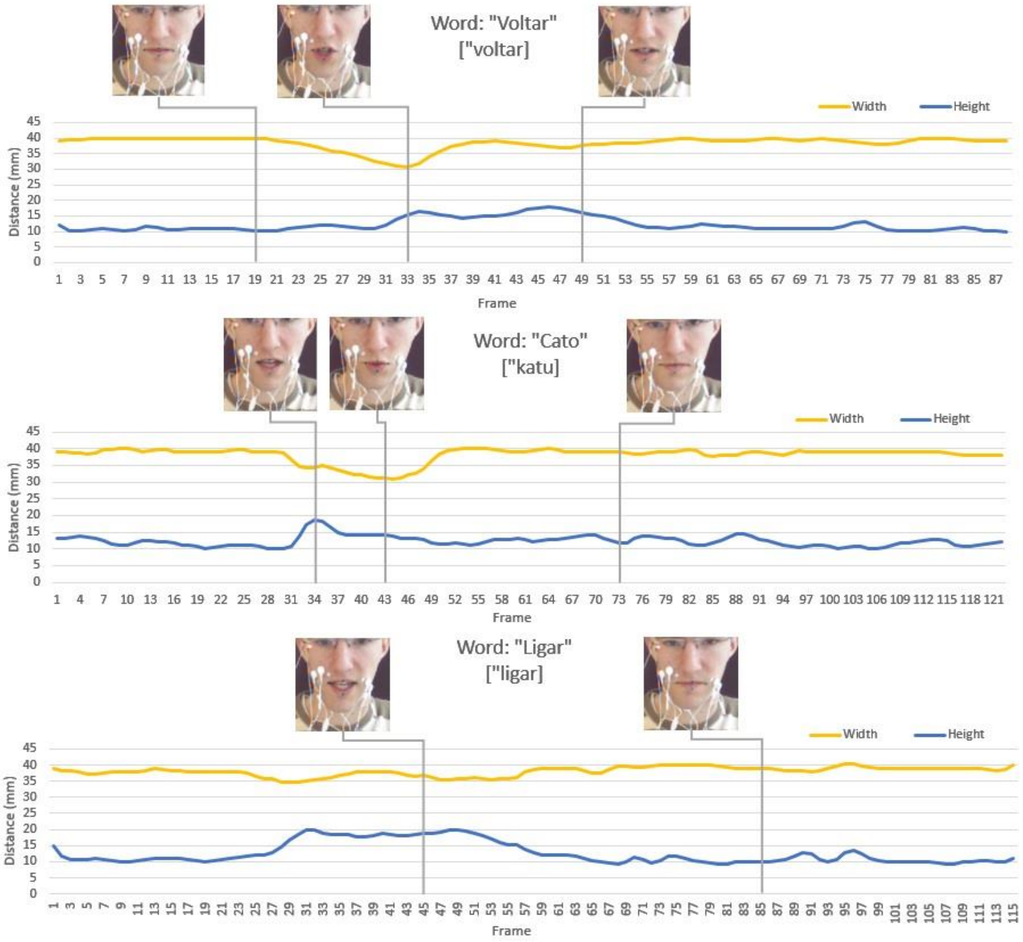


Figure 42. Width and height characteristic values for the lips computed along the production of the words "Voltar", "Cato" and "Ligar". The presented image frames depict frames showing notable lip configurations.

4.5.3. Feature Selection

Given the large amount of data extracted to build the corpora, after the feature extraction phase, there was the need to apply a Features Selection (FS) method in order to reduce the dimensionality of the input space (for the subsequent classification phase of our method). This allowed us to achieve better learning performance with this data.

Regarding FS, several techniques have been made available in the literature, from which we selected two unsupervised and two supervised relevance measures (Freitas et al., 2014a) based on their running-time and previous results with high-dimensional data (Ferreira and Figueiredo, 2012). For the unsupervised case, we considered: i) the Mean-Median (MM), that is, the absolute value of the difference between the mean and the median of a feature (an asymmetry measure) (Ferreira and Figueiredo, 2012); and ii) the quotient between the Arithmetic Mean and the Geometric Mean (AMGM) of each feature, after exponentiation (a dispersion measure) (Ferreira and Figueiredo, 2012). For the supervised case, we considered two well-known measures: i) the (Shannon's) Mutual Information (MI) (Cover and Thomas, 2005), which measures the dependency between two random variables; and ii) the Fisher's Ratio (Fisher, 1936), which measures the dispersion among classes.

For finding the most relevant features, we considered the Relevance-Redundancy FS (RRFS) filter method proposed in (Ferreira and Figueiredo, 2012). In a nutshell, RRFS uses a relevance measure (one of the four mentioned above) to sort the features in decreasing order, and then performs a redundancy elimination procedure on the most relevant ones. At the end, it keeps the most relevant features exhibiting up to some Maximum Similarity (MS) between themselves. The similarity between features is assessed with the absolute cosine of the angle between feature vectors, say Xi and Xj, given by:

$$AC_{ij} = \left|\cos\theta_{ij}\right| = \left|< X_i\,,\ X_j\ >/\|X_i\|\ \left\|X_j\right\|\right|, \tag{24}$$

where <.,.> denotes inner product between vectors and ||.|| is the L2 norm of a vector. AC_{ij} yields 0 for orthogonal feature vectors and 1 for collinear ones. RRFS has been applied successfully to other types of high-dimensional data (Ferreira and Figueiredo, 2012). For additional details the reader is pointed to Freitas et al. (2014a).

4.6. Discussion

With the proposed framework we believe to have created the conditions to address some of the stated hypotheses (3 - a multimodal HCI approach, based on less invasive modalities, has the potential to improve recognition results; and 4 - supplementary measures acquired from more invasive modalities can be use as ground truth) and respective objectives that motivated its design and development. The effort of designing a framework to support the research in multimodal SSI is important to address the challenges concerning a lack of understanding on how the different modalities can be used jointly. Furthermore, it is also an essential step in the research agenda, as it allows a faster deployment of new experiments and data collection setups, synchronized acquisition and processing of several modalities, and reusing of previously acquired data. Moreover, this flexible and extensible contribution to the state-of-the-art, lays the bricks for other studies beyond SSI. An example would be to apply this framework to speech production studies or, at an application level, to develop enhanced speech therapy solutions.

The proposed multimodal framework was instantiated using different setups that share the important requirement of synchronous acquisition of all data streams. This is an evidence of the versatility of the framework to tackle the challenges identified as the motivation for our work. These setups were designed to serve our current research goals and should not be understood as completely defining the scope of possible modalities. In fact, other modalities such as EEG and Vibration/Electromagnetic sensors (Holzrichter et al., 1998), could be also easily included, as long as synchronization is ensured.

The methods herein presented to process the acquired data, considering the different particularities of each modality, are an important component of the framework. These methods allow focusing on key aspects (e.g. a specific articulator), reducing the data dimensionality and adding “meaning” to the data with annotations of their relevant segments. For instance, extract velum movement curves from RT-MRI image sequences or characterizing the movement of lips. These methods are not meant to be interpreted as the optimal and/or most versatile and efficient approaches to deal with the acquired data. Instead, they are a baseline approach for extracting notable information that is useful to address concrete research questions. For example, regarding the ultrasound data if, instead of computing inter-frame differences, the tongue was segmented along the sequences, data concerning tongue height and backness could be easily extracted and used to further explore the SEMG signals. The rationale is to keep evolving the processing stage, as needed, to increasingly provide additional features.

The proposed framework allows for the joint exploration of more direct measures of the movements of the articulators. After processing the data, we gain further insight over the applicability of the considered modalities to capture different aspects of speech. This is performed by using the reference data gathered from modalities, such as US, that help to model the target phenomenon, using data from one or more of the non-invasive modalities (e.g. SEMG) and then perform classification.

4.7. Summary

This chapter described a framework that fosters leveraging modalities and combining information, supporting research in multimodal SSI. It includes the first four stages (of five) of the framework: data collection method and general setup for acquisition; online and offline synchronization; collected corpora; multimodal data processing, feature extraction and feature selection. Examples of application, provided for each stage, include the development of a multimodal SSI based on less invasive modalities; and the use of ground truth information coming from more invasive/obtrusive modalities, to overcome the limitations of other modalities.

In this chapter we demonstrated the adaptive characteristics of the framework by presenting examples where, combined devices, allowed capturing up to 7 streams of data. Using the defined data collection setups we acquired two distinct datasets, designed according to the requirements of previously established scenarios. Our framework supports acquiring simultaneous modalities in a synchronous way, but it also provisions a method for aligning data collected off-line at different moments in time, tackling latent obstacles of some modalities such as RT-MRI. We also provided examples of data processing, using the US imaging data, feature extraction methods for different modalities and techniques based on FS to deal with high-dimensionality cases, like the ones verified in multimodal approaches.

The chapter concludes with remarks about the present framework and a discussion of its characteristics.

CHAPTER V

Multimodal SSI - Analysis, Classification and Prototyping

"Night, when words fade and things come alive. When the destructive analysis of day is done, and all that is truly important becomes whole and sound again."

Antoine de Saint-Exupéry

Contents

In this chapter, we present four applications of the multimodal SSI framework described in Chapter IV, completing the description started in the previous chapter. Namely, three experiments concerning analysis and classification and one example of design and implementation of an SSI application prototype.

It is clear from the literature and the results in Chapter III that SEMG is an interesting SSI modality. However, there is not much information about which tongue gestures are actually being detected by SEMG sensors in the neck region. Using the multimodal corpora made possible by our multimodal SSI framework, we created the conditions to explore the use of ground truth information provided by other modalities. Hence, the first experiment of this chapter is related to the problem of acquiring deeper knowledge on the existing individual modalities, such as SEMG, and on the challenge of characterizing tongue gesture, an important articulator in many speech sounds, particularly vowels. In more concrete terms, our SSI framework, by providing synchronized information on SEMG and other modalities such as Ultrasound (US) (which can be used as ground truth), allows for research on the detection of tongue movements with the non-invasive SEMG modality.

The second experiment explored the need and importance of also extracting articulatory information from visible articulators, such as the lips (to which we have direct access). This experiment used the combination of two modalities (Video and UDS) to characterize the visible articulators' movement. Based on the information extracted from Video and UDS, we were able to estimate when movement occurred and to automatically detect the type of movement made by the lips.

With the feature extraction methods presented in the previous chapter and the high-level articulatory features extracted from the analysis in the first two experiments, high-dimensional information of several modalities becomes available. As such, it becomes essential to understand its potential when used in an isolated and jointly manner. To pursue this goal, in the third experiment we assessed the use of different feature selection techniques, in a classification experiment using single and multiple modalities.

There is also interest in integrating the acquired knowledge into an SSI. Thus, the last part of the chapter describes the design and implementation of an SSI application prototype, which allows to test different feature extraction and classification methods with live data input.

In terms of chapter organization, the experiments descriptions, follow a similar structure as presented in Chapter III, with the difference that the corpus of each experiment was already described in Chapter IV.

5.1. Tongue Gesture Detection: Assessing EMG potential using other input modalities

The most promising approaches for SEMG based speech interfaces commonly target tongue muscles (Schultz and Wand, 2010). These studies used SEMG electrodes positioned in the facial muscles responsible for moving the articulators during speech, including electrodes in the upper neck area to capture possible tongue movements. Using these approaches, features extracted from the EMG signals were directly applied to the classification problem, in order to distinguish between different speech units (i.e. words or phonemes).

Despite the interesting results in small vocabularies tasks, it is yet unclear which tongue movements are actually being detected, and there is a lack of information about different tongue movements during speech. Finally, we don’t know whether these movements can be correctly identified using SEMG.

To address these complex aspects, in this particular study we explored a novel method, based on synchronous acquisition of SEMG and US of the tongue, to assess the applicability of EMG to tongue gesture detection. In this context, the US image sequences allowed us to gather data concerning tongue movement over time, providing the grounds for the EMG analysis and, in time, develop more informed EMG classifiers that could be easily integrated in a multimodal SSI with an articulatory basis. This way, information about an articulator which is normally hidden by the lips in the case of a visual approach, or is very difficult to extract, could be provided. Ideally, one could consider a neckband that detects tongue gestures integrated with other modalities such as video, all of which contribute to a less invasive approach to the characterization of speech production.

5.1.1. Method

Using the synchronized data, made available from the combined usage of the proposed general acquisition setup and processing methods integrating the SSI framework, we investigated the detectability of tongue gestures based solely on SEMG sensors. For this experiment, we explored the corpora described in section 4.4.1.2 and the use of the “ground truth” modality Ultrasound Imaging. Regarding EMG, the result of the processing described in section 4.5.2.1 was used. Ultrasound video was processed as described in section 4.5.1.

In order to investigate the SEMG potential to tackle this problem, we conducted a detection experiment based on the probability of movement. The adopted method consisted in modeling the classes’ distribution. Then, based on these models, we derived the probability of movement given the measured EMG signal. The considered classes were: “Movement 1”, “Movement 2”, and “Non movement”. The first two represented the tongue movements (front to back and vice-versa) found in each utterance and the third class denoted no tongue movement. The processing of information for analysis is detailed in the following section.

5.1.2. Processing

Using random utterances of each speaker and the automatic annotations from US, the probability mass functions for three classes were computed and compared. Based on preliminary experiments, we noticed that the statistics for each speaker stabilize after a few utterances. A Gamma distribution was adopted based on the shape of the histograms. The two parameters of this distribution were estimated using Matlab ("MATLAB, Statistics Toolbox," n.d.). As depicted in Figure 43, differences in distributions between forward movement and non-movement were found for all speakers with some variations within EMG channels.

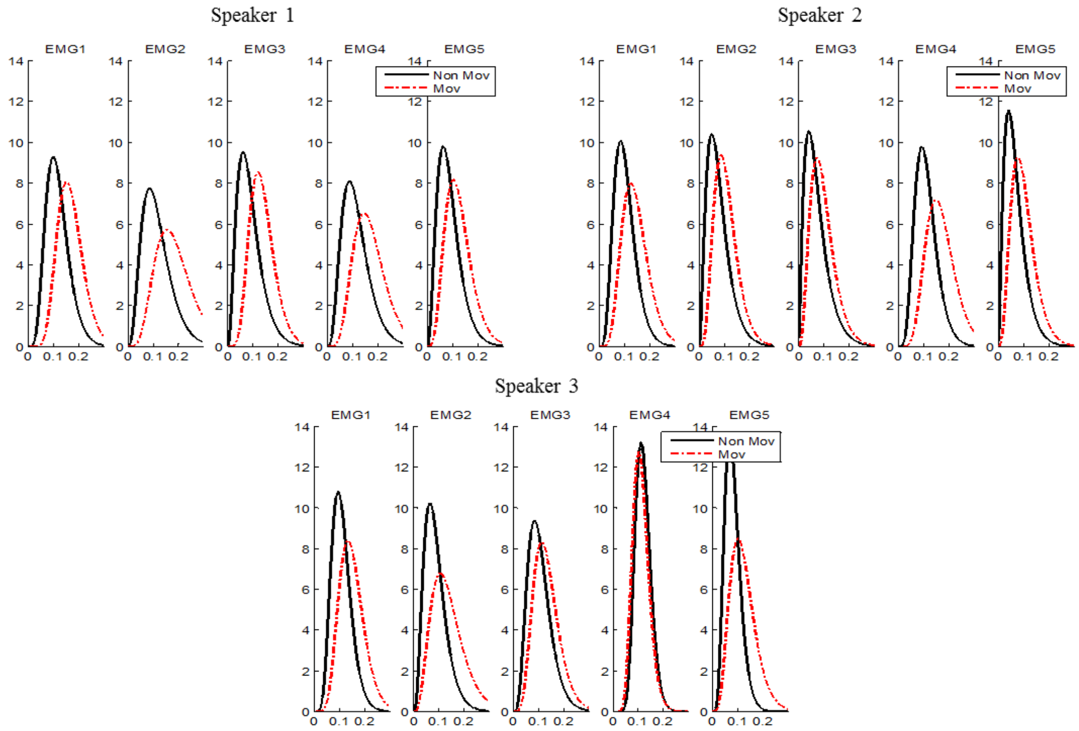


Figure 43. Density functions for the five Electromyography channels of the three speakers, including curves for one of the movements (forward) and non-movement.

Based on the probability distribution functions described in the previous section, we estimated the probability of each movement. Hence, considering the probability of the measured EMG *meas* given a movement *mov*, $p(meas|mov)$, we can apply Bayes rules as follows:

$$p(mov|meas) = \frac{p(meas|mov)\,p(mov)}{p(meas)} \qquad (25)$$

Using the law of total probability to expand $p(meas)$ we obtain:

$$p(mov|meas) = \frac{p(meas|mov)\ p(mov)}{(1-p(mov))\ p(meas|nonmov)+p(mov)\ p(meas|mov)} \tag{26}$$

The detection threshold was set to 0.5 and $p(mov)$ to 0.3, which, based on an empirical analysis, presented a good balance between detections and false positives. Figure 44 presents an example of the results obtained considering SEMG channel 1 for one of the utterances. As can be observed, the movement detected based on the probability distribution provides promising results regarding forward movements.

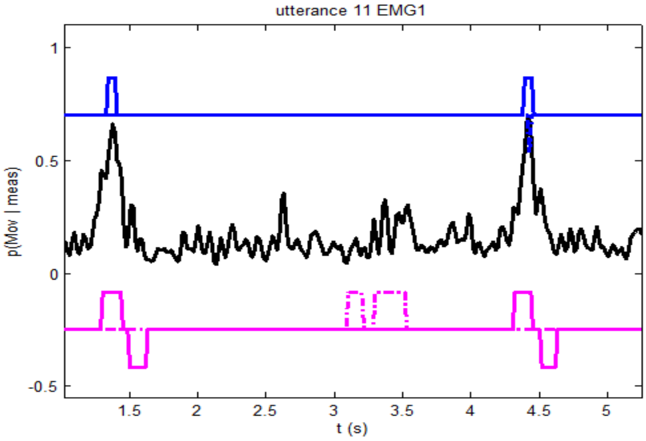


Figure 44. Example of movement detection: top, the detected movements; middle, the probability of movement; and bottom, the US based movement annotation, where forward movement is depicted by the segments represented above the middle line.

To assess the applied technique, we compared the detection results with the US annotation in order to obtain information on correct detections, failures in detection, and false detections. As this processing was done for each sample, the outputs were manually analyzed to determine the number of correct detections and number of failures. For False Positive (FP) detections, only a qualitative assessment was done, quantifying it into three classes (1 = a few or none; 2 = some; 3 = many).

5.1.3. Experimental Results

The results as function of speaker and sensor are summarized in Table 15. The best results were attained for Speaker 1 in channel 3 with a detection of 80.0% and an average of 67.1% for the 5 channels. The other 2 speakers obtained the best detection result of 68.6% and 66.8% in SEMG channels 4 and 5, respectively.

Table 15. Percentage of correct movement detections, by speaker, by SEMG sensor and the corresponding averages.

	SEMG Channel					*Average*
	1	*2*	*3*	*4*	*5*	
Speaker 1	60.3	77.4	**80.0**	51.5	66.2	*67.1*
Speaker 2	56.9	39.2	47.5	68.6	26.4	*47.7*
Speaker 3	46.7	51.7	35.4	1.5	66.8	*40.5*
Average	*54.7*	*56.1*	*54.3*	*40.5*	*53.1*	

Looking into the results, in more detail, there is also a strong variation of results across prompts, as depicted in Figure 45. The best results, in terms of prompts, were achieved for [L, s and t] in a /vCv/ context.

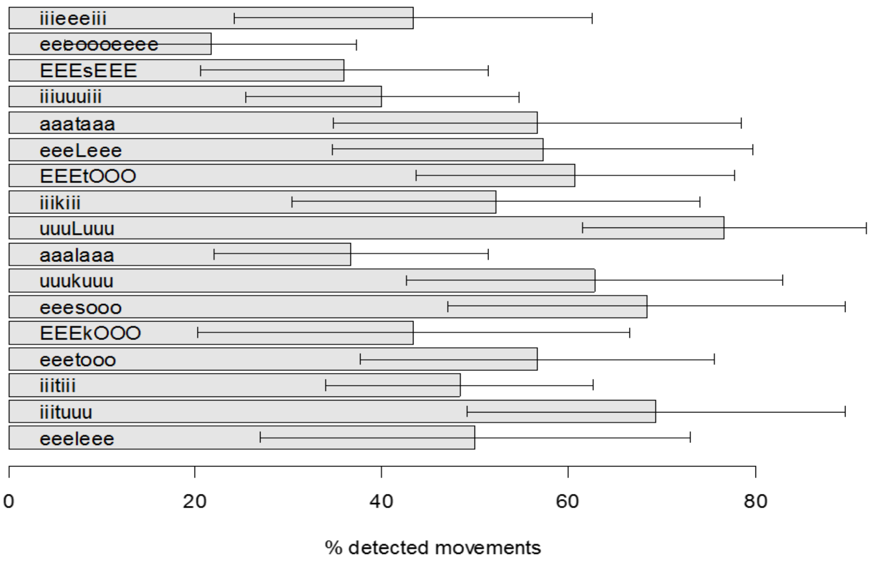


Figure 45. Detection accuracy results with 95% confidence interval for some of the prompts.

In terms of FP, we noticed that, although speaker 1 presented the best results, it also had a high rate of FP with 38.8% of the utterances having many FP. In that sense, speaker 2 presented the best relation between correct detections and FP with 47.1% of the utterances presenting none or few FP. In terms of sensors the best relation between correct detections and FP was found for channels 2 and 5. Table 16 summarizes the attained results.

Table 16. False positive results (% of possible cases) by speaker and by SEMG sensor for 3 classes (1 = a few or none; 2 = some; 3 = many).

	Speaker			**SEMG Channel**				
	1	*2*	*3*	*1*	*2*	*3*	*4*	*5*
Class 1: few or none	32.9	47.1	37.6	29.4	39.2	31.4	54.9	41.2
Class 2: some	28.2	32.9	45.9	33.3	41.2	45.1	29.4	29.4
Class 3: many	38.8	20.0	16.5	37.3	19.6	23.5	15.7	29.4

5.2. Multimodal analysis and classification of visible articulators

In this section, we present two types of results based on the extracted lip measures: (1) a time sequence with relevant articulatory events such as lip opening, lip rounding, and facial movement extracted from UDS; and (2) a classification experiment based on the RGB information.

5.2.1. Method

As presented in previous chapter (section 4.5.2.3, page 104), we were able to automatically obtain the width and height of the lips, by tracking four of its points (top, bottom and corners) and measuring the Euclidean distances between them. These distances allowed us to extract articulatory features such as lip opening and lip rounding.

Another source of information about the visible articulators was the Doppler signal collected using UDS and, as shown in section 4.5.2.2. Based on the energy of the signal, we could establish if facial movement was detected.

Since our SSI framework ensured synchronization between Video and UDS, we consequently gained the possibility of crossing the information of these modalities. In this case, since we were capturing the same stage of speech production in both data sources, most of the practical benefits resided in obtaining additional robustness for processing and further validation of facial events.

For classification and in order to explore different techniques, we considered the k-Nearest Neighbor (kNN) (Aha et al., 1991), Support Vector Machine (SVM) (Burges, 1998), and DTW classifiers (Müller, 2007). For training and testing the classifiers we used the first round of recordings of the corpus for command control, containing recordings of 8 speakers and described in section 4.4.1.1. The dataset was split into train and test using a stratified (i.e. using the same class proportions) 10-fold strategy, splitting it into 216 train utterances and 24 test utterances for 10-word datasets and 173 train utterances and 19 test utterances for the 8-word datasets, in each run. The error rate herein reported was estimated from the average error rate of the 10 folds.

5.2.2. Processing

For the extraction of articulatory features, no additional image processing was needed, since we used the width and height values previously measured. Knowing that an utterance always started in a non-speaking state, we used the average width and height values of the first ten frames as a reference to define each feature's threshold. A lip opening state was assumed when the lips height value of a frame was above the defined threshold, while the lip rounding state required the width value to be under the threshold. Each feature was then tagged with the value 0 or 1 (not present or present, respectively) for each of the articulatory features. Finally, we applied a third order moving average filter to remove isolated artifacts in the signal.

5.2.2.1. Classification

For assessing the value of the obtained information, we configured the classifiers as follows. The kNN classifier used the Euclidean distance for prediction, an inverse distance weighting function (i.e. each point of the model had a weight equal to the inverse of its distance) and the number of neighbors, *k*, was dynamically determined as the square root of *n*, the number of training instances (De Wachter et al., 2007). The SVM classifier from LIBSVM (Chang and Lin, 2011) was configured with a linear kernel. The DTW classifier used the distance between the test word samples and each sample of the training set choosing the class with the minimum distance. The SVM and kNN classifier were fed with the lip rounding and lip opening bit signal, while the DTW received the width and height measures after being processed with a third order moving average filter.

5.2.3. Experimental Results

Using the extracted values of each frame it is possible to observe the evolution of the lips appearance, regarding the lip opening and lip rounding features. Figure 46 depicts the results obtained from the words "Voltar" [voltar], "Cato" [katu] and "Ligar" [ligar], and several frames of these words corresponding to the respective frame number (gray lines). While the lip opening state is always expected in a word, the presence of a lip rounding state is normally associated in EP to the backness and the closeness of its vowels. The lip rounding state can be found in both "Voltar" and "Cato" words, due to the [o] vowel. In the word "Cato", the lip opening state is only flagged in five frames (33rd to 37th) due to the opening of the [a] vowel. For the word "Ligar" no lip rounding is required to pronounce it.

The presented tool for articulatory features extraction showed consistent results in words with different kinds of pronunciation and different speakers.

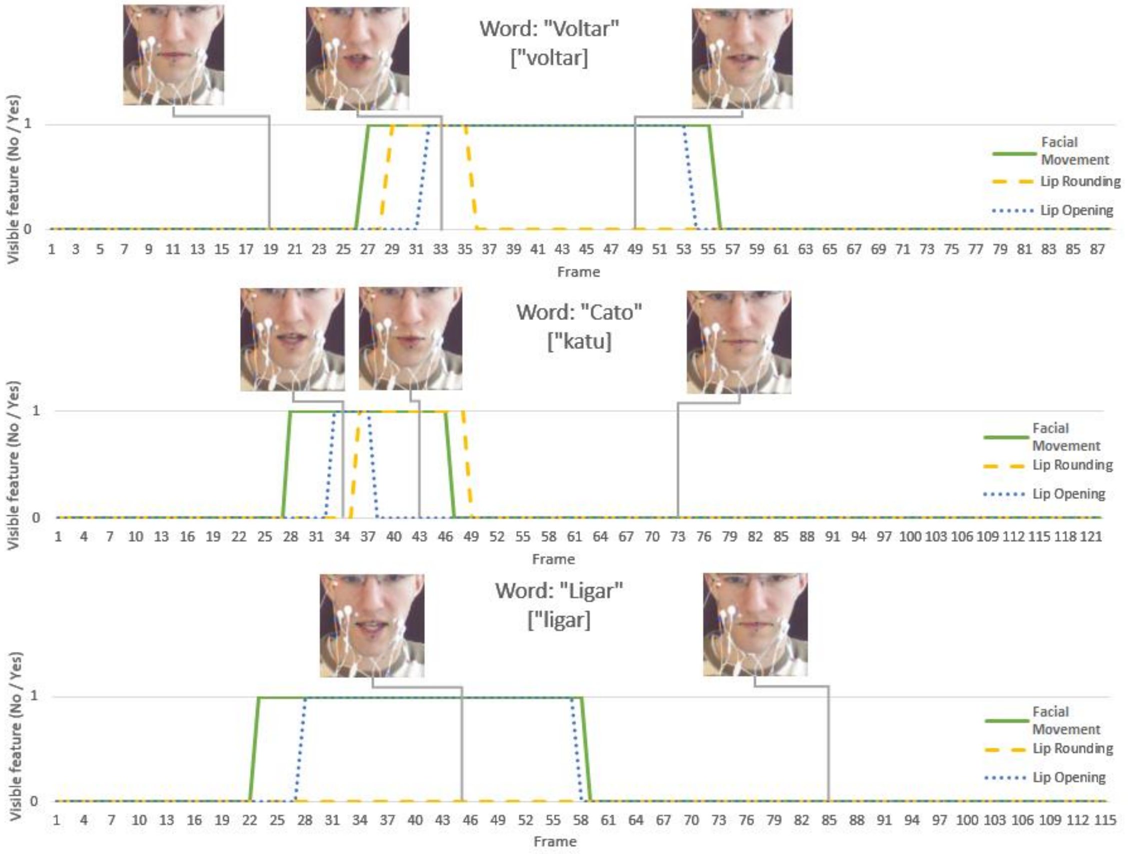


Figure 46. Lip opening, lip rounding (extracted from Video) and movement detection events (extracted from Ultrasonic Doppler sensing) along the frames of the words "Voltar", "Cato" and "Ligar". For each word, several frames are presented, corresponding to the pointed frame number (gray lines).

Besides articulatory events detection results, we have also assessed these measures by conducting a classification experiment. Results from this experiment are shown in Table 17. When comparing classifiers, the best result was achieved by the kNN classifier with an average recognition error rate of 62.5%.

Table 17. Average isolated word recognition error rate (in percentage) with 95% confidence interval of the 10-fold using 4 different vocabularies (including a random mixed selection of 8 words based on the other vocabularies) for the kNN, SVM and DTW classifiers.

Classifier	Vocabulary Mix	Nasal Pairs	Digits	AAL words	Average Error Rate
kNN (lips events)	**56.9 ± 4.5**	60.8 ± 7.0	66.1 ± 5.0	66.3 ± 5.2	**62.5**
SVM (lips events)	64.7 ± 7.7	67.1 ± 8.9	70.2 ± 6.6	75.0 ± 6.2	69.3
DTW (height + width)	65.8 ± 5.3	65.8 ± 4.4	67.2 ± 5.2	68.7 ± 5.5	66.9

5.3. Classification based on single and multiple modalities

After synchronously acquiring data from several input HCI modalities, our aim for this experiment and one of our main goals for our research in this area, has been to understand which modality or combination of modalities achieves the best results in a recognition task and how redundancies, among the considered modalities, affect the overall results.

5.3.1. Method

In order to assess the benefits of applying Feature Selection (FS) techniques, we started by estimating the recognition error of each modality without applying FS methods, using the most common features of each modality and an SVM classifier, to establish a baseline error rate. Afterwards, we applied FS techniques to each individual modality and also to multiple modalities, considering a fusion scenario.

From our command and control oriented dataset (described in section 4.4.1.1), we selected only the 10 digit subset from the second round of recordings, with an audible acoustic signal. Then, we divided the data into train and test using a stratified 9-fold strategy, splitting it into 160 training utterances and 20 test utterances (2 test utterances per class), in each run. Since we had 10 classes and 180 total observations, splitting the data into 9 folds allowed us to have an equal number of utterances per class in each fold. The error rate herein reported was estimated from the average error rate of the 9 folds.

The following sections report the results for single modalities and their joint use. For FS purposes, we considered the RRFS method with the following four relevance measures (two unsupervised and two supervised), mentioned in section 4.5.3: Mutual Information; Fisher’s Ratio; Mean-Median (MM); and Arithmetic Mean with the Geometric Mean (AMGM).

5.3.2. Processing

The applied processing was similar for all modalities and selected FS measures. In a first stage we extracted the features of each modality according to the techniques described in section 4.5.2. In a second stage, we have applied the RRFS filter method to the feature set of each modality, varying the Maximum Similarity (MS) values of RRFS between (not including) 0 and 1 and the FS technique to find the best combination.

5.3.3. Experimental Results

In this section we first present classification results for single modalities, where baseline results (i.e. before FS or Feature reduction - FR) and results considering the four mentioned FS techniques are presented. Afterwards, we present the results for multiple modalities in a feature fusion scenario with and without using FS techniques.

5.3.3.1. Single modality

The results obtained regarding WER are presented in Table 18. The second column of the table shows the error rate for the baseline using all the features without applying any FS or FR technique (representing a baseline for comparison). Among the four modalities, the best result was obtained by SEMG, with 46.7%, followed by UDS with 50.6%. Depth-only information presented the worst result, with 70.6%.

Regarding the considered FS techniques, on average, supervised techniques performed better than unsupervised ones; the best results were achieved by Fisher's ratio on three modalities; for the UDS modality, MM achieved the best result.

Table 18. Average word recognition error rate (%) with 95% confidence interval (9-fold), for each modality and different FS measures.

		Supervised		*Unsupervised*	
Modality	**Baseline (No FS/FR)**	**Mutual Info**	**Fisher's Ratio**	**MM**	**AMGM**
Video	53.9±6.3	41.1±4.8	**34.4±5.3**	42.8±7.1	40.6±4.5
Depth	70.6±5.8	70.6±4.1	64.4±6.6	68.3±3.7	67.2±7.5
SEMG	46.7±4.9	46.7±6.5	45.0±3.7	46.1±5.1	46.1±3.6
UDS	50.6±5.3	50.6±5.3	50.0±6.1	47.8±5.2	50.6±5.3

In terms of compression ratio, the best result was achieved for RGB Video with a compression rate of 95.1%, followed by Depth with 59.8%. Surface EMG achieved a compression rate of 26.7% and UDS only 5%.

The Maximum Similarity (MS) parameter values of RRFS, for the best results, varied according to the modality, the FS technique and the classifier. For example, as depicted in Figure 47 for Video using the SVM classifier, we have the largest improvement when using Fisher's ratio technique for selecting the best features, with MS set to 0.3. In the second best case achieved using SEMG and Fisher's ratio, MS was 0.8.

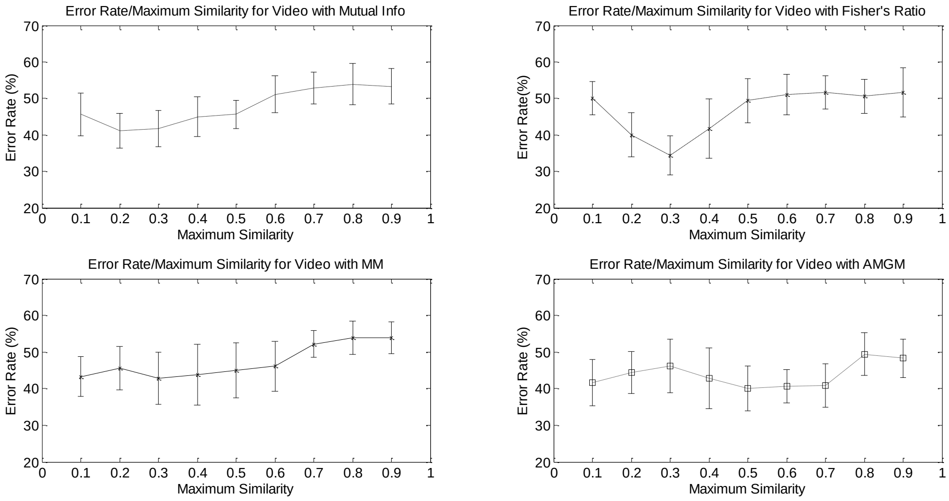


Figure 47. Average word recognition error rate (%) and the respective 95% confidence interval using different Maximum Similarity values between 0.1 and 0.9, for Video and multiple feature selection techniques (Mutual Information – upper left; Fisher's ratio – upper right; Mean-Median – bottom left; Arithmetic Mean and the Geometric Mean – bottom right).

For comparison purposes, we have also applied Feature Reduction - FR using Linear Discriminant Analysis LDA to this dataset. In the achieved results only Depth with the SVM classifier, improved accuracy with an absolute performance gain of 16.7% (relative 23.7%). The results for the remaining modalities were either similar or worse.

5.3.3.2. Multiple modalities

A very important question was to understand if combining multiple modalities could improve the achieved results. When fusing the feature vectors of the studied modalities in different combinations, improvements were noticed for some modalities groupings, when compared with the baseline results in the second column of Table 18. For example, Video combined with SEMG improved the results of baseline Video by an absolute value of 7.8% and, the results of baseline SEMG, by an absolute value of 0.6%. Similar improvements were noticed for UDS when combined with the SEMG.

When applying RRFS prior the combinations of two modalities, improvements could be noticed for almost all combinations, using the SVM classifier. The best results were achieved for the cases of Video combined with UDS using Fisher's ratio and, of Video combined with Depth using AMGM, with absolute performance gains in the error rate gain, of 19.4% and 13.3%, respectively. For the combination of Video+SEMG and SEMG+UDS, we observed absolute performance improvements in the same metric, of 1.1% and 1.7%, respectively. Regarding the combination of three and four modalities, we noticed an interesting pattern where the previous results are somewhat replicated. For example, adding UDS to the combination of Video+Depth showed the same improvement, as compared to Video+Depth alone. However, adding SEMG (and consequently including all streams), improved the baseline results of the individual modalities found in Table 18, but the FS results presented no improvements. Table 19 reports the most relevant results for the SVM classifier (the one that reported the best results in this analysis).

The overall results showed an average absolute improvement of 7.1% achieved across all possible combinations of two modalities and 3.9% for combinations of three and four modalities.

Table 19. Average word recognition error rate with 95% confidence interval (9-fold), before and after FS (only the best result is presented), using SVM. The rightmost column is the absolute improvement between the fusion results before and after FS. The best combination are *Video+UDS* with FS with the SVM classifier.

Modalities	Before FS	After FS (best technique)	Improvement
Video+Depth	69.4±6.8	56.1±5.1 (AMGM)	13.3%
Video+UDS	53.7±6.1	**34.3±5.2** (Fisher)	**19.4%**
Video+SEMG	46.1±4.6	45.0±3.7 (Fisher)	1.1%
SEMG+UDS	46.7±4.9	45.0± 3.7 (Fisher)	1.7%
Video+Depth+UDS	69.4±6.8	56.1±5.1 (AMGM)	13.3%
All Modalities	46.1±4.6	46.1±4.6 (AMGM)	0%

5.4. Multimodal SSI Prototyping

In the previous three experiments we used the multimodal framework for analysis and classification. This last part of this chapter follows an application oriented approach, with a proposal of a modular solution for developing and testing a multimodal SSI system with live data. We first present the solution design and then provide a concrete implementation example that uses different development environments.

5.4.1. System Design

The envisaged system (depicted in Figure 48) is divided in a front-end, which handles the synchronized input SSI modalities, as well as manages any relevant output modalities and a back-end, where the actual data processing occurs.

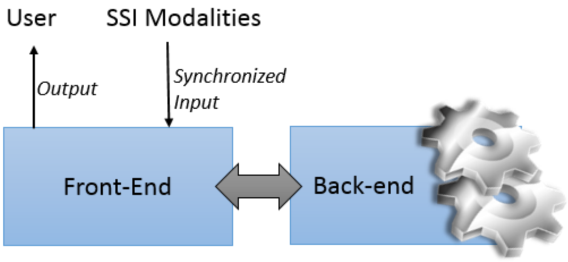


Figure 48. High-level system overview.

The front-end is responsible for collecting and sending data from/to the user and also from the input devices, in a synchronized way. The back-end of our multimodal SSI framework, includes data processing, extracting and selecting the best features or reducing the feature vector dimensionality and finally, the stages of analyzing and classifying the data. The communication between both parts of the system is made, in a loosely coupled manner, through an asynchronous service, agnostic of the underlying technology.

The Unified Modeling Language (UML) sequence diagram in Figure 49, provides a high-level example of an HCI scenario, where a prototype is used for word recognition, capturing its dynamics.

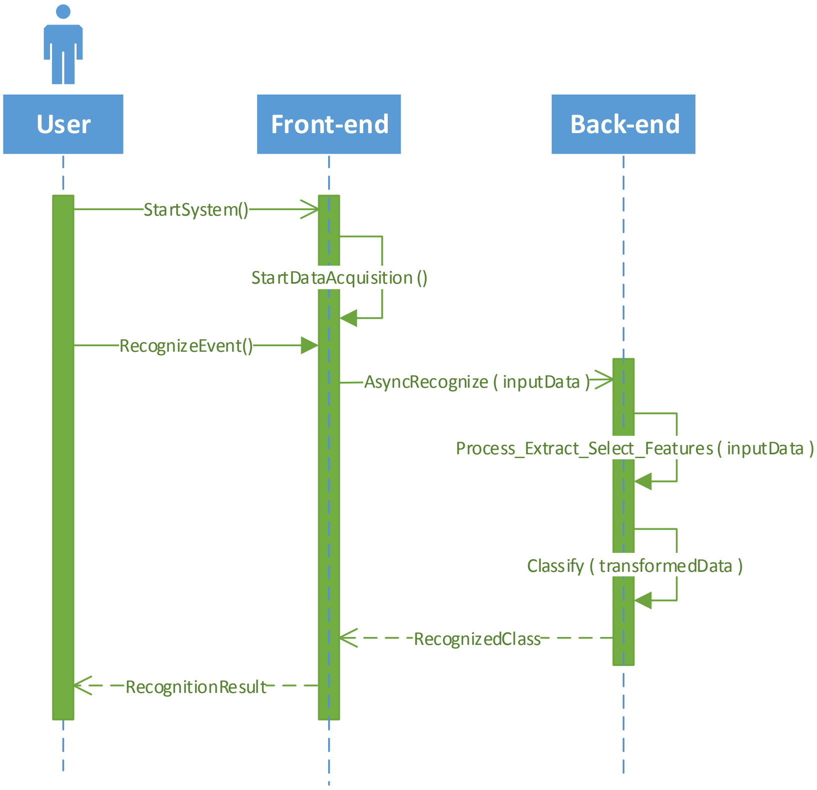


Figure 49. UML sequence diagram with a high-level example of an HCI scenario: the recognition of an isolated word used as a command to control an application.

5.4.2. Handling Live Data

The diagram of Figure 49, describes a command and control type of application, using our multimodal SSI framework, which can be used as a test-bed capable of evaluating different feature extraction, feature selection and classification methods, from live input data. We chose to implement each part in a different computer language to take advantage of the characteristics of different development environments that better matches our purposes in each module. Thus, the front-end part was implemented using .NET technologies in C#, enabling access to consistent programming models and available libraries suitable for rendering user interfaces. Additionally, it also allows using well-known 3rd party Application Programming Interfaces (API), which provide easy access to low-level data from multiple devices, as depicted in Figure 50.

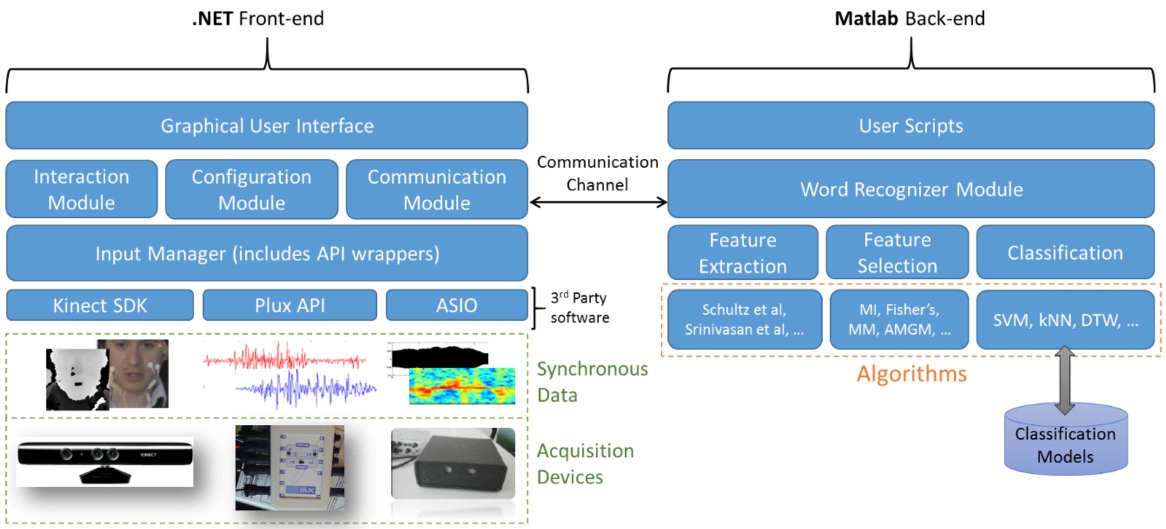


Figure 50. Architecture diagram of a multimodal SSI prototype for isolated word recognition. On the left the front-end part of the system (including devices and the corresponding data) is represented. On the right, the back-end part, including its modules and algorithms, is depicted.

The front-end part of the prototype includes the following modules:

- **Graphical User Interface** – The graphical user interface, for operator use, was developed using Windows Presentation Foundation (WPF) and shows the input raw signals, as well as the recognition results and respective confidence threshold when applicable.
- **Interaction Module** – This module contains the application logic related with the application state (e.g. acquiring data) and user events (e.g. button click).
- **Configuration Module –** This module handles the logic concerning the configuration of the system. The system configuration can be changed via application or via an XML file and

enables selecting all kind of mutable options, such as algorithms to be used or backend address.

- **Communication module –** The communication module is responsible for communicating with the back-end and passing all the necessary information. This module can be seen as an interface to the back-end features.
- **Input Manager –** The input manager handles the input data from the input SSI modalities and respective buffering strategy.

Additionally, the front-end also includes 3rd party system software that allows low-level access to the hardware devices, such as the Kinect SDK, Audio Stream Input/Output (ASIO) drivers or the API provided by Plux to access the EMG recording device.The back-end part of the system was developed using Matlab, a well-known platform from the research and academic communities, due to its numerical computation capabilities, range of applications in science (mainly mathematics and engineering), and built-in functions for faster prototyping.

The back-end code foundation, initially used for analysis, was refactored, keeping its functionalities and maintaining a set of user scripts that allow to test the features provided by the back-end. Both user scripts and communication module (described above), access the interface provided by the module referred to as the word recognizer. This module receives the configuration information (i.e. which algorithms to use, vocabulary, etc.) and the data to be classified. The module output will be the class label of the recognized word. In more detail, the word recognizer calls a set of scripts, given by the configuration passed as a parameter and follows a conventional machine learning pipeline, composed by feature extraction and/or selection and/or reduction and, finally, classification. The classification algorithms rely (if applicable) on a model previously built. The current implementation can be easily extended with different techniques, namely, other classification methods or feature reduction algorithms, by simply adding a new module that respects the established contracts. For example, the process for adding a new classification algorithm consists in adding a new scripted function to the back-end.

5.5. Discussion

The reported experiments gives us a direct answer to the challenges brought by the joint exploration of SSI modalities for European Portuguese – EP. These experiments create the conditions for in-depth studies with non-invasive input modalities for EP, and enable access to “ground truth” data for such studies. The experiments, considering multiple modalities, provided a glimpse on how the multitude of data and the knowledge acquired through our framework, can be used to serve the more complex goal of developing SSI for European Portuguese using multimodal, non-invasive, data.

The first experiment provided insights regarding the importance of some of the extracted signals features, made available by the proposed framework. By allowing the synchronous acquisition of SEMG and US, direct data measurements regarding tongue movements (obtained after US annotation), have been made available. This allowed a more informed and novel exploration of the EMG data, namely concerning its suitability to detect tongue movements. The processing approach presented, although simple, shows how the acquired multimodal data can help deepen the knowledge regarding the different SEMG channels and can inform the development of more complex processing methods, or the design of additional data acquisitions.

The second experiment illustrates the use of multiple modalities for Visual Speech Recognition – VSR, using articulatory measures. We have shown how redundancy between input modalities (in the context the speech production stage, since they both target the articulators' movement), can be useful if used for validation of the method, spurious movement detection or to obtain additional details, which are not captured by only one of the modalities.

The classification results of this approach, when compared with a previous study (Freitas et al., 2013), show an average absolute improvement of 7.1% of the word error rate, across all datasets, with the best dataset (vocabulary mix) achieving an absolute improvement of 10.3%. However, it should be noticed that the previous published results, although using the same vocabulary, used a dataset with 8 speakers instead of 3 and considered silent speech instead of normal speech, thus, an improvement was somewhat expected. Better results were also found for vocabularies with higher amount of lip articulation.

In short, considering the case of a small dataset and small vocabularies, our results point to the fact that articulatory measures achieve a performance at least as good as appearance-based methods.

The results of the third study point to a performance improvement (for most cases) using single and multiple modalities supported by Feature Selection (FS) methods. For single modalities, a noticeable improvement could be found in the Video modality, whereas Depth produced the worst results, in accordance to what was found previously in the literature for the Audio-Visual Speech Recognition scenario (Gurban and Thiran, 2009). For the combination of multiple modalities, we have noticed the following: (1) if no FS is applied, small improvements could still be noticed, particularly for the worst individual modalities considered in the fusion (e.g. Video combined with SEMG improved the results of Video by 7.8%); (2) after applying FS, when compared with the ones obtained individually, similar levels of word error rate could be achieved, with noticeable improvements for the case of Depth input. Thus, feature selection techniques appear to be useful, but for some cases, such as SEMG, no improvement could be noticed.

Summarizing, using multiple modalities seems to be a useful source of “ground truth” data for better understanding single modalities. On the other hand, combining multiple modalities for classification with small datasets and few observations does not improve the best single modalities (e.g. Video), but does improve the performance of worst performing modalities.

When comparing FR and FS techniques, FS presents better results. The most probable cause is the small amount of training data relative to the sample dimensionality. Previous studies (Gurban and Thiran, 2009; Qiao et al., 2009) have shown that when this situation occurs, the LDA within-scatter matrix becomes sparse, reducing the efficiency of the LDA transform. Regarding the considered FS techniques, we have observed that, on average, supervised techniques performed better than unsupervised ones. Selecting the best features can also be used as a way to compress data, removing the most redundant information and making it appropriate for scenarios where data storage or communication bandwidths are reduced (e.g. mobile communication). Finally, we have demonstrated how the multimodal framework can used to build and test current and future multimodal SSI applications. Also, the software engineering solution here presented allows to easily evaluate new parts of the processing pipeline with live data input.

5.6. Summary

This chapter concludes the description of the multimodal SSI framework proposed in this thesis. Here, we have presented several techniques for processing the modalities, providing four concrete examples of its use and analyzing different speech articulators, such as the tongue and the lips.

The first experiment presented in this chapter, consisted in using a novel approach to assess the capability of SEMG in detecting tongue movements for SSI. The approach uses synchronized US imaging and surface EMG signals to analyze signals of the tongue. Results shown that tongue movement can be detected using SEMG with some accuracy but with variation across speakers and possibility of high FP rates, suggesting the need for a solution adapted to each user.

The second experiment here presented, reported a high-level analysis using Video and UDS in a combined way. The results of this analysis included the detection of articulatory events concerning the lips and facial movement, as well as a classification experiment based on articulatory features.

In the third experiment, a classification experiment based on single and multiple modalities was presented. This experiment also assessed the impact of feature selection filters on silent speech data. Results shown that feature selection leads to word error rate accuracy improvements, which were achieved either using only a single input modality or a combination of several modalities. Looking at the fusing of modalities, the highest improvement was found for Video alone and for

Video combined with UDS using an SVM classifier. In terms of feature selection techniques, the supervised methods based on Fisher’s ratio, attained the best results.

The overall conclusions of the experiments are that input multimodal approaches for SSI can provide interesting benefits either by combining “ground truth” data from more invasive/obtrusive modalities, by using these for method validation and artifact removal, or for classification experiments, when fusing a more accurate (in terms of WER) modality, with others that do not perform as well.

In the last section of the chapter, we provided an example of how our framework can be used to create a multimodal SSI for EP system, for a command and control HCI scenario.

CHAPTER VI

The Challenge of Nasality Detection in Silent Speech

"...hardly ever can a youth transferred to the society of his betters unlearn the nasality and other vices of speech bred in him by the associations of his growing years."

William James, The principles of Psychology, Vol. 1, 1890

Contents

A known challenge in SSIs, far from solved, is the detection of the nasality in speech production, it being unknown if there is enough information to perform said detection in existing SSI modalities, such as SEMG. Nasality is an important characteristic of several languages, including EP (Almeida, 1976; Teixeira and Vaz, 2000), which is the target language for this thesis. Additionally, no SSI exists for EP and, as shown before in Freitas et al. (2012a) and Chapter III, nasality can negatively impact accuracy for this language. Given the particular relevance of nasality for EP (Lacerda and Head, 1966; Sampson, 1999), we have conducted an experiment that, using the framework described in Chapter IV and V, aimed at expanding the current state-of-the-art in this area, determining the possibility of detecting nasal vowels with non-invasive modalities such as SEMG and UDS, consequently improving SSI interaction systems. Surface Electromyography (EMG) is one of the approaches reported in literature that is suitable for implementing an SSI, having achieved promising results (Schultz and Wand, 2010).

The main idea behind this experiment consisted in using ground truth information from RT-MRI image sequences containing information about the velum movement, with the myoelectric signal collected using surface EMG sensors and the Doppler shifts captured by the ultrasonic sensor. By combining these sources, ensuring compatible scenario conditions and time alignment, we were able to estimate both the time when the velum moves and the type of movement (i.e. ascending or descending) under a nasality phenomenon, and to establish the differences between nasal and oral vowels using surface EMG or even UDS.

The remainder of this chapter is structured into two experiments: in the first experiment we present an exploratory analysis and classification experiment with SEMG and in the second, we conduct a correlational study using UDS.

6.1. Nasality detection based on Surface Electromyography

The use of surface EMG electrodes to target deeper muscles presented several difficulties. It was not clear to what depth a surface electrode can detect the myoelectric signal, not only because of the distance, but also due to signal propagation conditions in different tissues and the respective noise associated with them. Also, the signal output of surface electrodes in the region of the face and neck was likely to reflect a high level of cross talk, i.e. interference from signals from muscles that lie in the vicinity of the muscle fibers of interest, due to the superposition of numerous muscle fibers. Therefore, this experiment did not only analyzed the possibility of nasal vowel detection using SEMG but also assessed if deeper muscles could be sensed using surface electrodes in the regions of the face and neck and the best electrode location to do so.

Our work differed from previous studies (Bell-Berti, 1976; Fritzell, 1969; Kuehn et al., 1982; Lubker, 1968), which used intramuscular electrodes or surface electrodes placed directly in the velum, as none of them used surface electrodes placed in the face and neck regions, a significantly less invasive approach and quite more realistic and representative of the SSIs case scenarios. Also, although intramuscular electrodes may offer more reliable myoelectric signals, they also require considerable medical skills to implement and, for both these reasons, intramuscular electrodes were discarded for this study.

The problems described led to challenging tasks that have the potential to impact speech, health and accessibility technologies, by improving nasal vowel recognition in speech and in silent speech interfaces.

6.1.1. Method

To determine the possibility of detecting nasal vowels using surface EMG we needed to know when the velum is moving to avoid misinterpreting artifacts and noise as signals coming from the target muscles. To overcome this problem we took advantage of an earlier data collection based on RT-MRI (A. Teixeira et al., 2012), which provided an excellent method to estimate when the velum is moving and to interpret EMG data.

Recent advances in MRI technology allow for real-time visualization of the vocal tract with an acceptable spatial and temporal resolution. This technology enabled us to have access to real time images with relevant articulatory information for our study, including velum raising and lowering. In order to make the correlation between the two signals, audio recordings were performed in both data collections by the same set of speakers. Note that EMG and RT-MRI data cannot be collected together, so the best option was to collect the same corpus for the same set of speakers, at different times, reading the same prompts in EMG and RT-MRI, taking advantage of the offline synchronization method previously described in section 4.3.1.

6.1.1.1. Surface EMG setup

For this study we have used the same EMG acquisition system from Plux (“Plux Wireless Biosignals,” n.d.), already used in previous chapters (sections 3.1 and 4.2.1). For this experiment we used 5 pairs of EMG surface electrodes attached to the skin, using single-use 2.5cm diameter clear plastic self-adhesive surfaces. One of the difficulties found while preparing this particular study was that no specific background literature in speech science existed on ideal sensor position to detect the velum muscles referred to in section 2.1.2.1. Hence, based on anatomy and physiology literature (for example (Hardcastle, 1976)) and preliminary trials, we determined a set of positions that cover as much as possible the most likely positions that are best for detecting the targeted muscles. To measure the myoelectric activity we used both a bipolar and monopolar surface electrode configuration. In the monopolar configuration, instead of having both electrodes placed directly on the articulatory muscles (as in the bipolar configuration), one of the electrodes was used as a reference (i.e. located in a place with low or negligible muscle activity). In both configurations the result was the amplified difference between the pair of electrodes.

As depicted in Figure 51, the 5 sensors pairs were positioned in the following locations:

- EMG 1 was placed in the area superior to the mandibular notch, superficial to the mandibular fossa;
- EMG 2 was placed in the area inferior to the ear between the Mastoid process structure and the mandible angle, on the right side of the face using a monopolar configuration;
- EMG 3 was placed in the same position as EMG2 but used a bipolar configuration and was placed on the left side of the face;
- EMG 4 was placed in the superior neck area, beneath the mandibular corpus, at an equal distance from the mandible angle (EMG 2) and the mandible mental prominences (EMG 5);
- EMG 5 was placed in the superior neck area, beneath the mandible mental prominences.

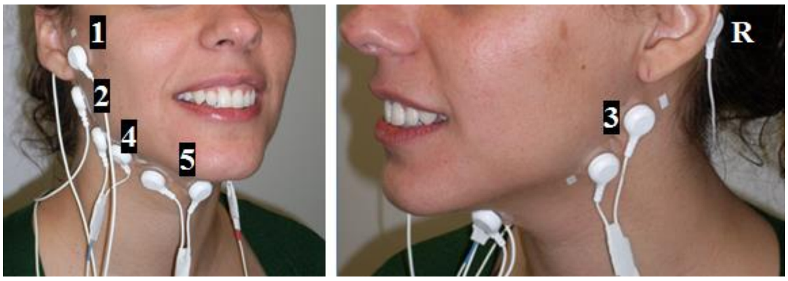


Figure 51. EMG electrode positioning and the respective channels (1 to 5) plus the reference electrode (R). EMG 1 and 2 use unipolar configurations whereas EMG 3, 4 and 5 use bipolar configurations.

The reference electrodes (EMG R) were placed on the mastoid portion of the temporal bone and on the cervical vertebrae. Considering the sensors' location, we expected them to acquire unwanted myoelectric signals due to the superposition of the muscles in these areas, such as the jaw muscles. However, in spite of the muscles of the velum being remote from this peripheral region, we expected to be able to select a sensor location that enabled us to identify and classify the targeted muscle signal with success.

We sampled the recording signal at 600Hz and used 12 bit samples. For system validation, we conducted several preliminary tests on larger superficial muscles.

We recorded the audio using a laptop integrated dual-microphone array, using a sample rate of 8000Hz, 16 bits per sample and a single audio channel. As the audio quality was not an issue in this collection, we opted for this solution instead of a headset microphone, which could have caused interference with the EMG signal.

6.1.1.2. RT-RMI Setup

The RT-MRI data collection was previously conducted at IBILI/Coimbra for nasal production studies. Images were acquired in the midsagittal and coronal oblique planes of the vocal tract (see Figure 52) using an Ultra-Fast RF-spoiled Gradient Echo (GE) pulse sequence which yielded a frame rate of 14 frames per second. Each recorded sequence contained 75 images. Additional information concerning the image acquisition protocol can be found in (Silva et al., 2012).

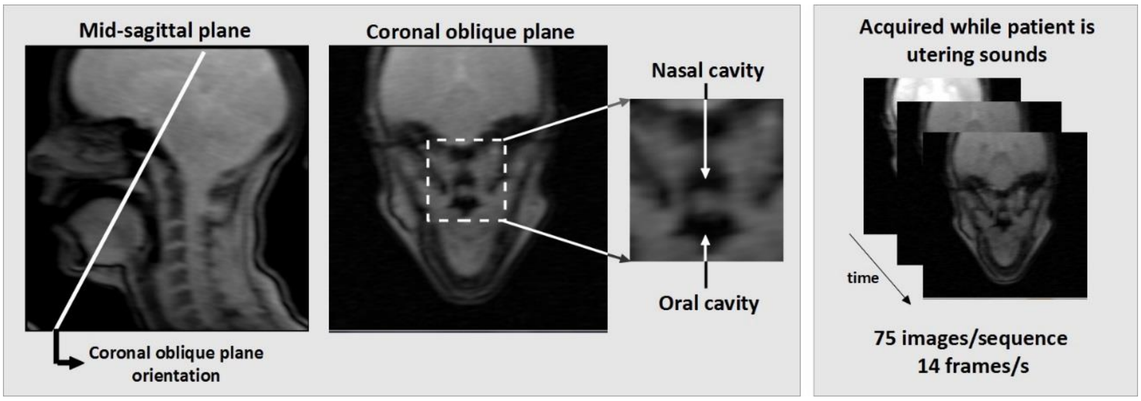


Figure 52. From left to right: mid-sagittal plane depicting orientation of the oblique plane used during acquisition, sample oblique plane showing the oral and nasal cavities and image sequence details (A. Teixeira et al., 2012).

The audio was recorded simultaneously with the real-time images, inside the scanner, at a sampling rate of 16000Hz, and using a fiber optic microphone. For synchronization purposes a TTL pulse was generated from the RT-MRI scanner (A. Teixeira et al., 2012). Currently, the corpus contains only three speakers due to the elevated costs per recording session and the availability of the technology involved.

6.1.1.3. Analysis

In the signal analysis, to measure the dependence between the MRI information and the EMG signal we used the mutual information concept. The mutual information concept (Pereda et al., 2005), also known as trans-information, derived from the bases of information theory, is given by:

$$I(X;Y) = \sum_{y \, € \, Y} \sum_{x \, € \, X} p(x,y) \, log\left(\frac{p(x,y)}{p(x) \, p(y)}\right) \tag{27}$$

where $p(x,y)$ is the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distribution functions of X and Y, respectively. This measures the common dependence of the two variables, providing the difference of information of one signal given the other. In this case we used normalized mutual information (Yao, 2003) to investigate the relation between the RT-MRI signal and the nasal zones of the EMG signal, allowing us to further investigate the relation between the nasality information extracted from the RT-MRI and the EMG signal from another viewpoint.

To confirm the existence of significant differences among the results of the EMG channels, we used Analysis of Variance (ANOVA) of the error rate, using SPSS (SPSS 19.0 – SPSS Inc., Chicago, IL, USA) and R (Everitt and Hothorn, 2009; R Development Core Team, 2011).

6.1.2. Corpora

The two corpora collected in this study (RT-MRI and EMG) share the same prompts. The set of prompts is composed by several nonsense words that contain five EP nasal vowels ([6~, e~, i~, o~, u~]) in isolated, word-initial, word-internal and word-final context (e.g. ampa [6~p6], pampa [p6~p6], pam [p6]). The nasal vowels were flanked by the bilabial stop or the labiodental fricative. For comparison purposes the set of prompts also included isolated oral vowels, as well as oral vowels in context. In the EMG data collection a total of 90 utterances per speaker were recorded. A detailed description of the RT-MRI corpus can be found in (A. Teixeira et al., 2012). In the EMG data collection, we recorded three silence prompts as well, (i.e. prompts where the speaker does not speak or makes any kind of movement) to further validate the system and the acquired EMG signal.

The 3 speakers used in this study were all female native speakers of EP, with the following ages: 33, 22 and 22 years. No history of hearing or speech disorders was known for any of them. The first speaker was an expert in the area of Phonetics and the remaining speakers were students in the area of Speech Therapy.

6.1.2.1. Surface EMG Data Collection

For this data collection we used the same speakers from the RT-MRI recordings. At first each speaker recorded a single session, meaning that the sensors were never removed during the recording. Later, for validation purposes and to support the assessment of aspects such as reproducibility, we recorded four additional sessions for one random speaker (Speaker 1) with the same prompts. Between sessions of the same speaker all sensors were removed. Before placing the sensors, the sensor location was first cleaned with alcohol. While uttering the prompts no other movement, besides the one associated with speech production, was made, including any kind of neck movement. The recordings took place in an isolated quiet room. An assistant was responsible for pushing the record button and also for stopping the recording in order to avoid unwanted muscle activity. The prompts were presented to the speaker in a random order. In this data collection two signals were acquired:

myoelectric and audio signals. For synchronization purposes, after starting the recording, a marker was generated in both signals.

6.1.3. Processing

The chapter on the processing stage includes the following: (1) a description of the methods for extraction of the velum information from the RT-MRI data; (2) a section describing how the signals from these distinct datasets were synchronized by applying the offline synchronization method of the multimodal framework; (3) velum movement information segmentation into nasal and non-nasal zones; (4) the EMG signal pre-processing and feature extraction steps; and (5) the classification model used for this experiment.

6.1.3.1. The extraction of velum movement information from RT-MRI data

The method for extracting velum movement information was previously developed, in the context of the projects Heron II (PTDC/EEA-PLP/098298/2008) and IRIS (FP7-PEOPLE-2013-IAPP, ref. 610986), by Dr. Samuel Silva from the University of Aveiro (Silva et al., 2013, 2012). In this section we present a brief description of how this was achieved and forward the reader to Silva et al. (2013) for more details.

The main interest for this study was to interpret velum position/movement using the mid-sagittal RT-MRI sequences of the vocal tract. The method used consists in measuring the area variation between the velum and the pharynx, which has a direct relation with the velum position. After using an image with the velum fully lowered to define the ROI, a region growing algorithm was applied, using a seed defined in a dark (hypo intense) pixel inside the ROI. This ROI was roughly positioned between the open velum and the back of the vocal tract. The main purpose was for the velum to move over that region when closing.

Figure 53 presents the contours of the segmented region over different image frames encompassing velum lowering and rising. For representation purposes, in order not to occlude the image beneath, only the contour of the segmented region is presented. Processing was always performed over the pixels enclosed in the depicted region. Notice that the blue boundaries presented in the images depict the result of the region growing inside the defined ROI (which just limits the growth) and not the ROI itself. The number of hypo intense pixels (corresponding to an area) inside the ROI decreases when the velum closes and increases when the velum opens. Therefore, a closed velum corresponds to area minimum while an open velum corresponds to local area maximum, which allows the detection of the frames where the velum is open. Since for all image sequences there was no informant movement, the ROI had to be set only once for each informant, and could then be reused throughout all the processed sagittal real-time sequences.

These images allowed for the derivation of a signal over time that described the velum movement (also shown in Figure 53 and depicted as dashed line in Figure 54). As can be observed, minima correspond to a closed velopharingeal port (oral sound) and maxima to an open port (nasal sound).

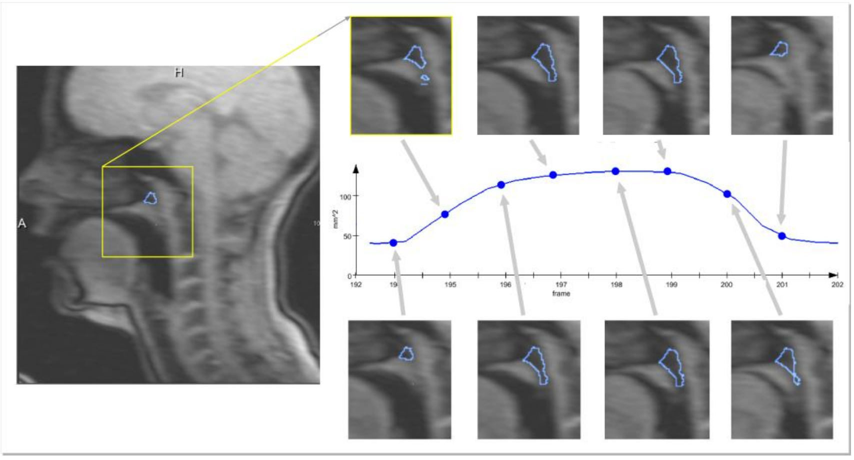


Figure 53. Mid-sagittal RT-MRI images of the vocal tract for several velum positions, over time, showing evolution from a raised velum to a lowered velum and back to initial conditions. The presented curve, used for analysis, was derived from the images (Freitas et al., 2014c).

6.1.3.2. Signal Synchronization

In order to address the nasal vowel detection problem we needed to synchronize the EMG and RT-MRI signals. For the synchronization of EMG and RT-MRI signals, we started by aligning both the EMG signal and the information extracted from the RT-MRI with the corresponding audio recordings. Next, we resampled the audio recordings to 12000Hz and applied DTW to the signals to find the optimal match between the two sequences. Based on the DTW result we mapped the information extracted from RT-MRI from the original production to the EMG time axis, establishing the required correspondence between the EMG and the RT-MRI information, as depicted in Figure 54. In order to validate the alignment, we annotated the audio data of Speaker 1 and compared the beginning of each word in the warped RT-MRI audio signal with the EMG audio signal. The relative position of the nasality signal of the EMG audio signal was found to be very similar to the one observed in the RT-MRI audio signal, which was an indication of good synchronization.

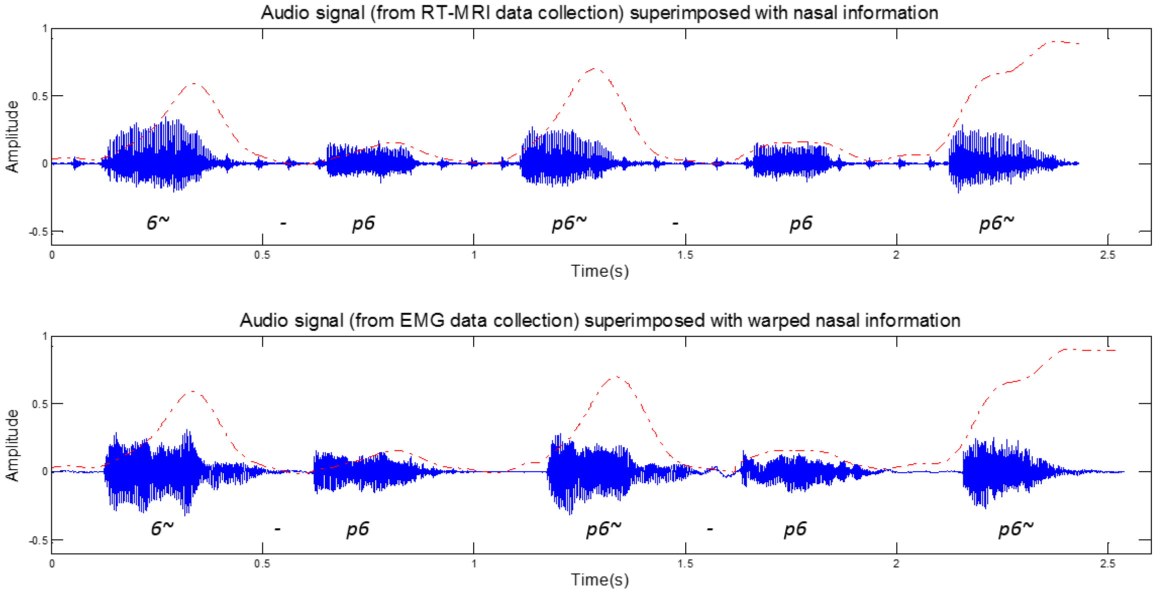


Figure 54. In the top chart we see the audio signal collected at the same time as the RT-MRI and the respective velum movement information (dashed line). In the bottom chart we see the audio signal collected at the same time as the EMG and the warped signal representing the velum movement information extracted from RT-MRI. Both charts represent the sentence [6~pɐ, p6~pɐ, p6~].

6.1.3.3. Signal segmentation into nasal and non-nasal zones

Based on the information extracted from the RT-MRI signal (after signal alignment), we were able to segment the EMG signal into nasal and non-nasal zones, using the zone boundary information to know which parts of the EMG signal are nasal or non-nasal, as depicted in Figure 55. If we consider a normalized RT-MRI signal x, we determined that $x(n) \geq \bar{x} + \left(\frac{\sigma}{2}\right)$ is nasal and $x(n) < \bar{x} + \left(\frac{\sigma}{2}\right)$ is non-nasal, based on an empirical analysis of the signals of all users. However, in order to include the whole transitional part of the signal (i.e. lowering and raising of the velum) in the nasal zones, we used the angle between the nearest peak and the points where the $x(n) = \bar{x}$ to calculate the nasal zone boundaries. This was done after testing different methods and their ability to cope with signal variability among speakers.

As detailed in Figure 55, by picturing two right triangles formed between the peak and $x(n) = 0$, by knowing angle θ1 (between the peak and the yy axis), the opposing *cathetus* of θ2 and assuming that θ1 = θ2, we were able to determine the magnitude of $v2$ and set a zone boundary that included either the lowering or the raising part of the signal. In one of the speakers there were a few cases where the velum remained open for longer time intervals. For these situations we used different peaks for the initial and end boundary of a nasal zone.

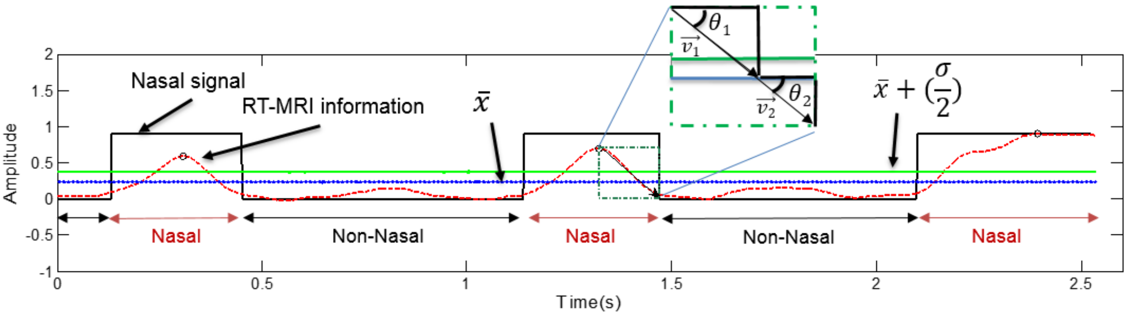


Figure 55. Exemplification of the EMG signal segmentation into nasal and non-nasal zones based on the information extracted from the RT-RMI (dashed red line). The square wave depicted with a black line represents the velum information split into two classes where 0 stands for non-nasal and 1 for nasal. The blue line is the average of the RT-MRI information (after normalization) and the green line is the average plus half of the standard deviation.

6.1.3.4. Pre-processing and Feature Extraction

In order to facilitate the analysis we pre-processed the EMG signal by normalizing it and applying a 12-point moving average filter with zero-phase distortion to the absolute value of the normalized EMG signal. An example of this pre-processing is depicted in Figure 56.

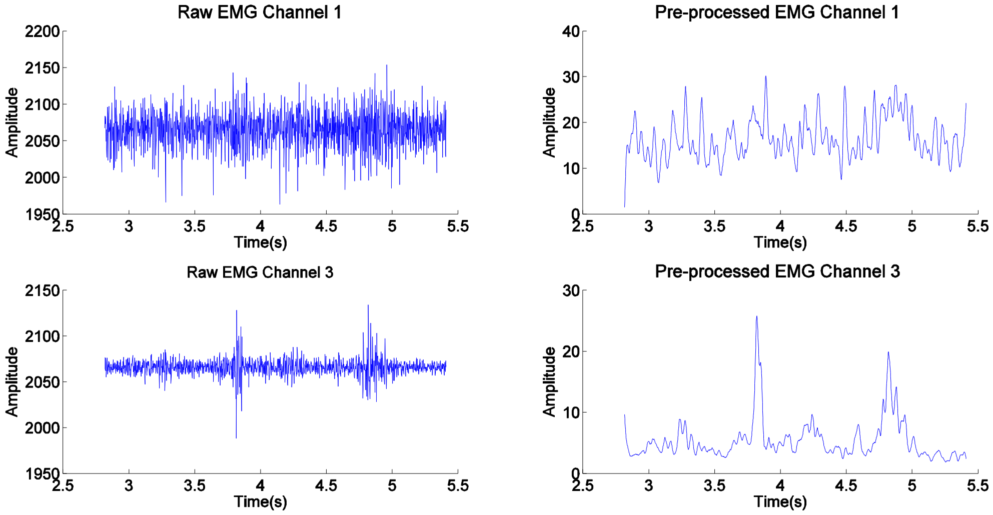


Figure 56. Raw EMG signal and pre-processed EMG signal of channel 1 (top) and 3 (bottom) for the sentence [6~p6, p6~p6, p6~] from speaker 1. The pre-processed signal was normalized and filtered using a 12-point moving average filter.

From each signal frame, we extracted 9 first order temporal features similar to the ones used by Hudgins et al. (1993). We then constructed our feature vector using the following data: mean, absolute mean, standard deviation, maximum, minimum, kurtosis, energy, zero-crossing rate and mean absolute slope. We considered 100ms frames and a frame shift of 20ms. Both feature set and frame sizes were determined empirically after several experiments.

6.1.3.5. Classification

For classification we used SVMs with a Gaussian Radial Basis Function. For estimating classifier performance we applied a 10-fold cross-validation technique to the whole set of frames from the 3 speakers to split the data into train and test sets. Relevant statistics and class distribution of this dataset are described in Table 20.

Table 20. Class distribution for all speakers for each EMG channel by zones (nasal and non-nasal) and frames

	Speaker 1		**Speaker 2**		**Speaker 3**		**All Speakers**
Utterances	15		14		15		44
Total Frames	836	53.2%	283	18.0%	453	28.8%	1572
Nasal frames	357	42.7%	195	68.9%	249	55.0%	801
Non-nasal frames	479	57.3%	88	31.1%	204	45.0%	771
Total Zones (nasal and non-nasal)	76	35.3%	65	30.2%	74	34.4%	215
Nasal Zones	45	59.2%	45	69.2%	45	60.8%	135
Non-nasal Zones	31	40.8%	20	30.8%	29	39.2%	80

6.1.4. Experimental Results

In this section we present the results of the analysis combining the EMG signal with the information extracted from the RT-MRI signal, two classification experiments, and a reproducibility assessment. In the first classification experiment, the EMG signal was divided into frames and each frame was classified as being nasal or non-nasal. The second experiment also divided the EMG signal into frames, but the classification was made by nasal and non-nasal zones, whose limits were known *a priori* based on the information extracted from the RT-MRI. The final part of this section addresses the problem of EMG signal variability across recording sessions.

For the signal and statistical analysis of the EMG signal we used 43 observations per speaker covering all EP nasal vowels, each containing the nasal vowel in several word positions (initial, internal and final) and flanked by [p]. We have also used, for visual analysis only, 4 observations per speaker containing isolated Portuguese nasal vowels [6~, e~, i~, o~, u~].

6.1.4.1. Exploratory Analysis

After we extracted the required information from the RT-MRI images and aligned it with the EMG signal, we started visually exploring possible relations between the signals. After pre-processing the EMG data, the resulting signal for all channels, along with the data derived from the RT-MRI and aligned as described in the previous section, is depicted in Figure 57. Based on a visual analysis, it was worth noticing that several peaks anticipated the nasal sound, especially in channels 2, 3 and 4. These peaks were most accentuated for the middle and final word position.

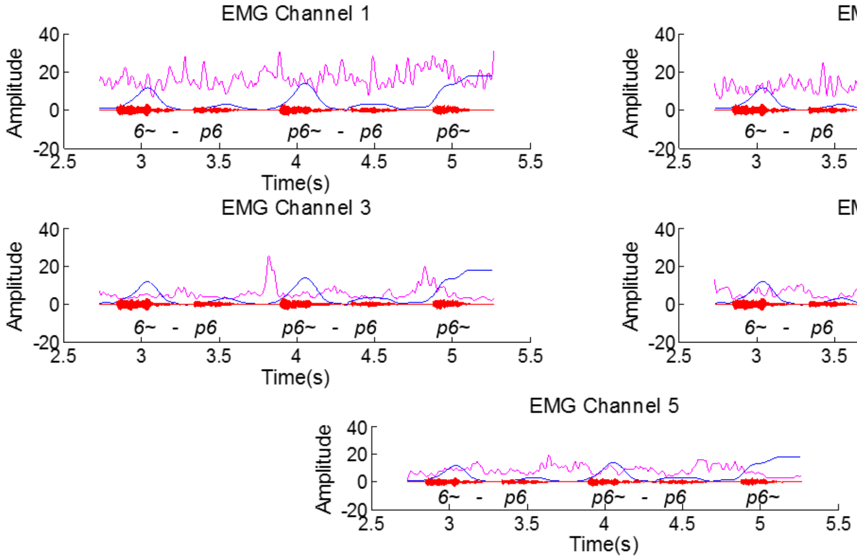


Figure 57. Filtered EMG signal (12-point moving average filter) for the several channels (pink), the aligned RT-MRI information (blue) and the respective audio signal for the sentence [6~p6, p6~p6, p6~] from speaker 1. An amplitude gain was applied to the RT-MRI information and to the EMG for better visualization of the superimposed signals.

By using surface electrodes the risk of acquiring myoelectric signal superposition was relatively high, particularly caused by muscles related with the movement of the lower jaw and the tongue, given the position of the electrodes. However, when we analyzed an example of a close vowel such as [i~], where the movement of the jaw is less prominent, the peaks found in the signal still anticipated the RT-MRI velar information for channels 3 and 4, as depicted in Figure 58. Channel 5 also exhibited a more active behavior in this case, which might be caused by its position near the tongue muscles and the tongue movement associated with the articulation of the [i~] vowel.

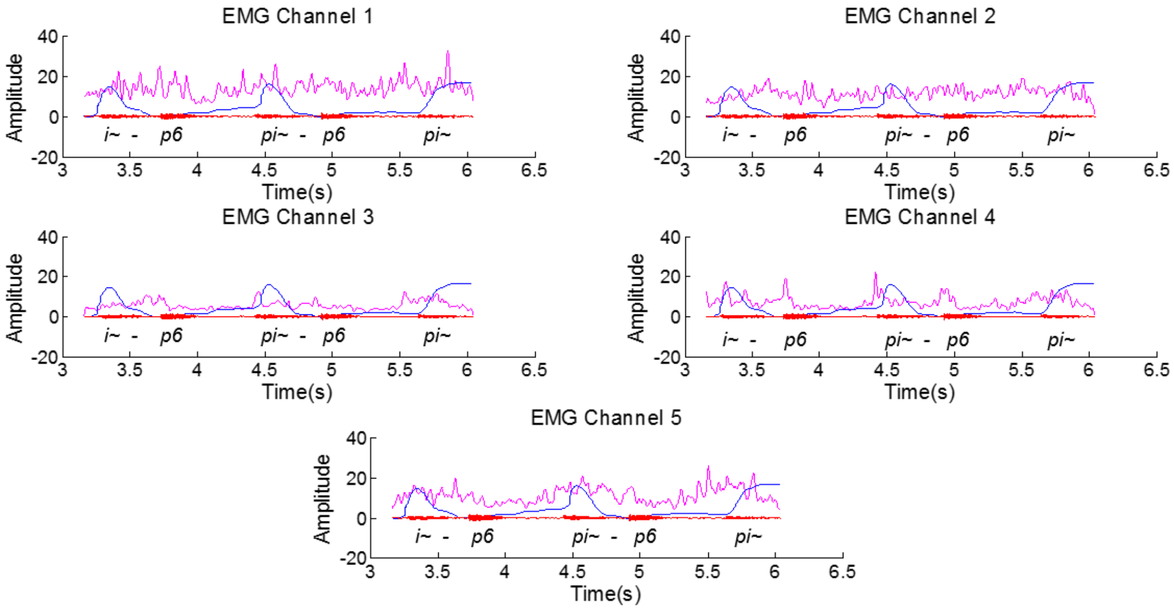


Figure 58. Filtered EMG signal (12-point moving average filter) for several channels (pink), the aligned RT-MRI information (blue) and the respective audio signal for the sentence [i~p6, i~p6, pi~] from speaker 1. An amplitude gain was applied to the RT-MRI information and to the EMG for better visualization of the superimposed signals.

Figure 59 shows the audio, RT-MRI and the EMG signal for an utterance that contains isolated EP nasal vowels in the following order: [6~, e~, i~, o~, u~]. These particular utterances were relevant since, in this case, minimal movement of the external articulators such as the lower jaw were required. If we consider the same analysis for isolated nasal vowels of the same speaker, EMG Channel 1 signal exihibted a clearer signal, apparently with less muscle crosstalk. Peaks could be noticed before the nasal vowels. For the remaining channels a clear relation with all the vowels did not manifest itself, although signal amplitude variations could be noticed in the last three vowels for Channel 3 and 5.

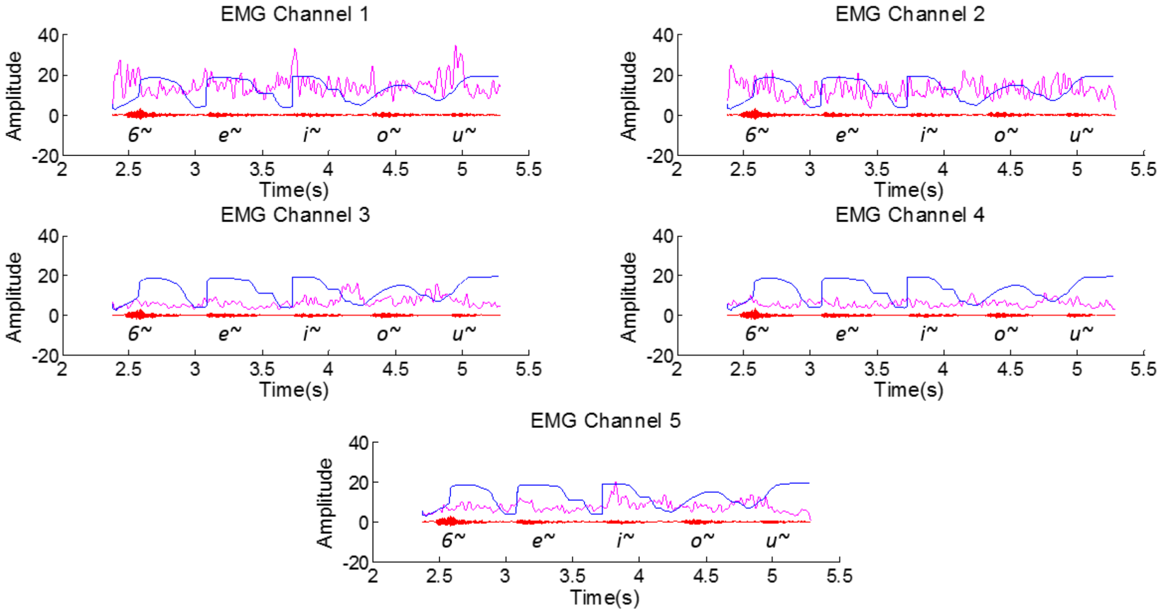


Figure 59. Portuguese vowels in an isolated context (Pre-processed EMG signal for all EMG channels (pink), the aligned RT-MRI information (blue) and the respective audio signal (red) for [6~, e~, i~, o~, u~]). An amplitude gain was applied to the RT-MRI information and to the EMG for better visualization of the superimposed signals.

To confirm our analysis presented above we conducted a quantitative analysis, investigating the existence of mutual information between the EMG signal and the information extracted from the RT-MRI signal. When conducting this analysis for the nasal zones of all speakers simultaneously, the mutual information values for channels 3, 4 and 5 were sligthly higher, as depicted by the boxplots presented in Figure 60. When considering each speaker individually, we found that at least for one speaker it showed that the best results could be found for channels 3, 4 and 5 as well. Also, for the same speaker we found that the amount of mutual information for non-nasal zones was close to zero.

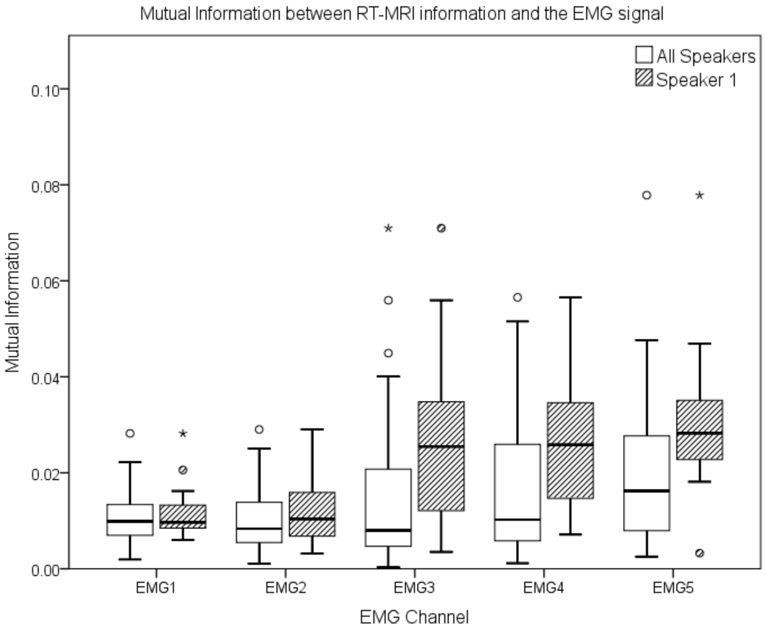


Figure 60. Boxplot of the mutual information in the nasal zones between the RT-MRI information and the EMG signal of all speakers and for a single speaker.

We analyzed other situations as well, such as the relation between the RT-MRI signal and the non-nasal zones of the EMG signal on the one hand, and the relation between the RT-MRI signal and the whole EMG signal on the other. No relevant information was found. The results we obtained did not allow us to draw a clear conclusion. We also applied other measures such as the Pearson's product-moment correlation coefficient, which measures the degree of linear dependence between two signals. However, we found magnitudes below 0.1, indicating a very weak linear relationship between the signals.

The fact that all seemed to point to the presence of differences between the two classes (nasal and non-nasal) motivated an exploratory classification experiment based on SVM, which has shown to yield acceptable performance in other applications, even when trained with small data sets. The results of this experiment are presented in what follows.

6.1.4.2. Frame-based Nasality Classification

In a real use situation the information about the nasal and non-nasal zones extracted from the RT-MRI signal is not available. Thus, in order to complement our study and because we want to have a nasality feature detector, we conducted an experiment where we classified the EMG signal frames as belonging to one of two classes: nasal or non-nasal.

The results of three relevant measures (error rate, sensitivity and specificity) are depicted in Figure 61. Besides the mean value of the 10-fold, 95% confidence intervals are also included.

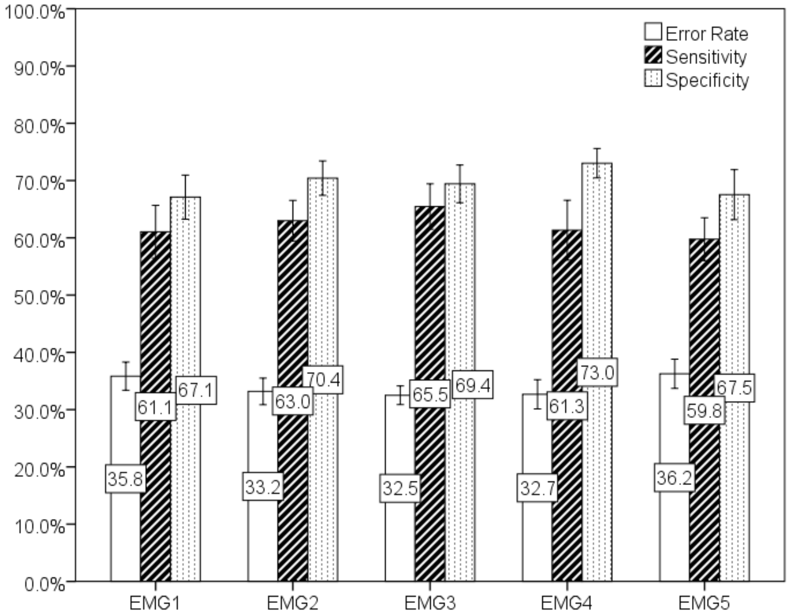


Figure 61. Classification results (mean value of the 10-fold for error rate, sensitivity and specificity) for all channels and all speakers. Error bars show a 95% confidence interval.

Results showed the best result for EMG Channel 3 with a 32.5% mean error rate, with a mean sensitivity and mean specificity of 65.5% and 69.4%, respectively. Channels 4 and 2 showed similar results and achieved second and third best results with mean error rates of 32.7% and 33.2%, with slightly lower sensitivity values of 61.3% and 63.0% and higher specificity values of 73.0% and 70.4%.

We then performed classification for each individual speaker. Error rate results are shown in Figure 62 for each speaker (left) and overall per channel (right), along with the corresponding 95% confidence interval. The mean error rate attained the best overall result of 24.3% for EMG channel 3. The best results for each individual speaker were found for Speaker 3 with 23.4% and 23.6% mean error rate in EMG channels 4 and 3. For Speaker 1 and 2, EMG channel 3 presented the best results with 25.7% and 23.7% mean error rate. On average Speaker 1 presented the least variability of results as shown by the confidence intervals. It is also interesting to note the difference of results in EMG channel 1, where speaker 2 attained a mean error rate of 24.1%. However, the class distribution slightly changed for Speaker 2 with 68.9% nasal frames, compared with 42.7% and 55.0% frames for Speaker 1 and 3. A closer look into the data of speaker 2 revealed that the higher amount of nasal frames could be explained by common breaths between words, implying an open velum.

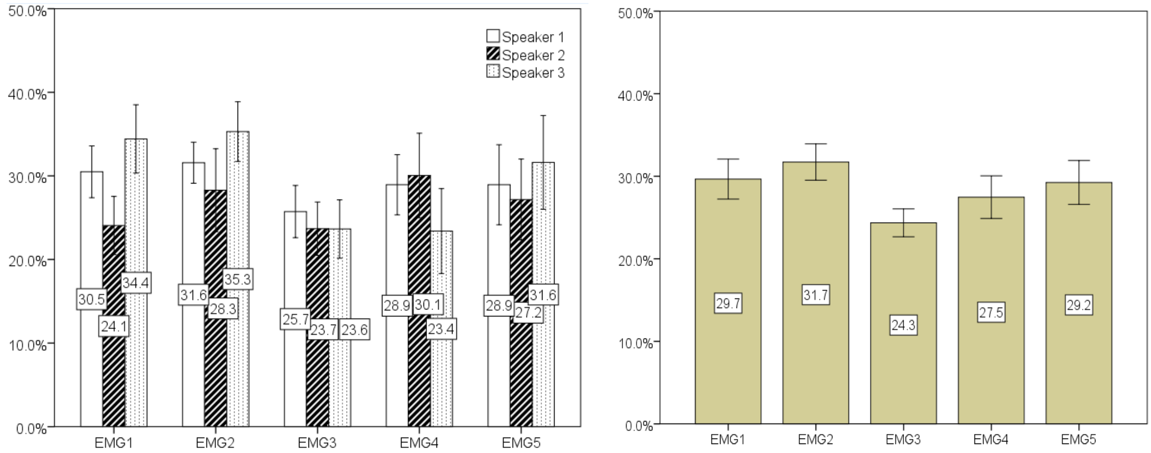


Figure 62. The graph on the left shows the mean error rate for each speaker clustered by EMG channel. The graph on the right shows the mean of the error rates from each speaker clustered by EMG channel. Error bars show a 95% confidence interval.

From a different perspective, when we subtracted the global mean error rate of all channels, then, as seen in Figure 63, Channel 3 exhibited a mean error rate 4.1% below this mean. Analyzing it by speaker, the best result was achieved for EMG channel 4 with 5.1% below the mean, and a noticeable result was found for EMG channel 1, where Speaker 2 obtained results below the mean error rate. However, for these speakers, the 95% confidence interval was considerably higher, showing some instability in the results.

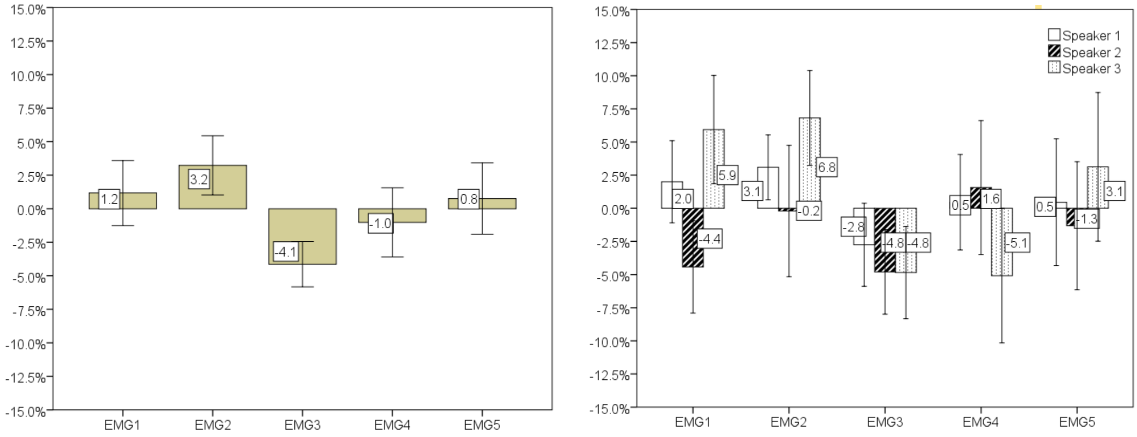


Figure 63. The difference between the mean error rate of all channels and the respective result of each channel for all (left) and each (right) speaker. Error bars show a 95% confidence interval.

When looking at the results grouped by nasal vowel, as shown in Table 21, an improvement could be noticed, particularly for the [u~] case with a 27.5% mean error rate using EMG channel 3.

Table 21. Mean error rate grouped by nasal vowel

EMG Channel	[6~p6,p6~p6,p6~]	[e~p6,pe~p6,pe~]	[i~p6,pi~p6,pi~]	[o~p6,po~p6,po~]	[u~p6,pu~p6,pu~]
1	36.2%	33.6%	38.7%	32.9%	35.6%
2	**34.2%**	33.9%	34.6%	33.5%	29.9%
3	39.8%	31.4%	35.8%	**29.4%**	**27.5%**
4	38.8%	**28.6%**	**32.8%**	35.1%	28.1%
5	39.5%	36.8%	36.1%	33.5%	35.0%
Mean	37.7%	32.9%	35.6%	32.9%	31.2%

To assess if any advantage could be extracted by using channel combination to improve classification, we experimented classification with multiple EMG channels. The most relevant combinations are shown in Table 22. The best results for all speakers and for each speaker individually were worse than the ones obtained previously.

Table 22. Mean error rate using multiple channels combinations

EMG Channel	**All Speakers**	**Speaker 1**	**Speaker 2**	**Speaker 3**
1 + 3	35.0%	30.1%	**24.7%**	29.5%
2 + 3	36.3%	**28.0%**	31.3%	33.9%
2 + 4	34.9%	31.4%	30.2%	33.0%
3 + 4	**32.9%**	28.7%	33.0%	**27.6%**
2 + 3 + 4	35.7%	29.0%	33.5%	32.1%
1 + 3 + 4 + 5	39.1%	36.3%	32.5%	34.9%

6.1.4.3. Statistical Analysis

In order to understand if the error rate results of each channel, described in the previous section, had significant differences among them, we decided to perform a statistical analysis of the results. Thus, assuming that the EMG channels are independent, we performed a one-way ANOVA between the error rate results of all speakers, as the reduced number prevented repeated-measures analysis. For all five channels, considered independently, the hypothesis of normal distribution for the EMG channels was assessed using the Shapiro-Wilk test. All tests gave p-values above 0.1, supporting the assumption of normality. The required homogeneity of variance was validated using Levene tests, resulting in p-values above 0.05. There was a significant difference between channel results at the p<.05 level for all five channels [F(4, 145) = 5.848, p < 0.001].

To assess the significance of the differences among the error rate of the channels, we performed a post-hoc analysis using the Tukey Honest Significant Differences (Tukey HSD) test. The pairs of EMG channels showing statistical significant differences, considering a confidence level of 0.05, were the following: EMG1 and EMG3 (p = 0.011); EMG2 and EMG3 (p < 0.001); and EMG3 and EMG5 (p = 0.024). Taking into consideration the results, the error rate for EMG channel

3 appeared several times as having significantly better error rates (lower values). The results for channel 4 did not showed significant differences in post-hoc comparisons. The only pair approaching significance (p=0.070) was EMG4, EMG2.

6.1.4.4. Nasal and Non-nasal Zone Classification

In this section, instead of classifying each frame, we classified a zone (i.e. region) of the EMG signal. As described in section 6.1.3.3, the RT-MRI information permitted us to split the EMG signal into nasal and non-nasal zones. Thus, because we knew the zone boundaries in the EMG signal, we could conduct a classification experiment based on the majority of nasal/non-nasal frames of a certain zone. This zone boundary information extracted from the RT-MRI signal, although not possible to use in a real classification scenario, allowed us to further explore and validate our methodology.

Assuming a decision by majority, a zone was condered nasal if the number of nasal frames was equal or higher than the number of non-nasal frames. Results using this technique are depicted in Table 23 and showed an absolute improvement of 11.0% when compared to what was achieved earlier using a single frame of the signal. When looking at a specific part of the zone, the results were still better than the ones achieved previously. The most important information seemed to be located in the initial half of each zone, since an accuracy degradation trend was observed for all channels when the initial part of the zone was not considered. When only the nasal zones were considered, the error rate reached 12.6% for EMG channel 4. The same trend of better results in the initial part of the zone could be verified as well, as depicted in Table 24. If only the non-nasal zones were taken into account, we observed the best results in EMG channel 3 with 25% mean error rate when using only the central part of the non-nasal zones, i.e. the 25%-75% interval.

Table 23. Mean error rates using a classification technique based on the majority of nasal/non-nasal frames for each zone. Each column of the table shows which part of the zone is being considered, where 0% represents the zone start and 100% the zone end (e.g. in the 50%-100% interval only the samples in the last half of each zone are being considered).

	Part of the zone considered			
EMG Channel	*[0-100%]*	*[0%-50%]*	*[25%-75%]*	*[50%-100%]*
1	25.2%	28.0%	24.3%	**27.7%**
2	**21.5%**	25.2%	28.7%	32.2%
3	25.7%	27.1%	28.7%	33.7%
4	24.3%	**22.9%**	**23.3%**	31.7%
5	32.2%	29.0%	30.7%	36.1%
Mean	**25.8%**	26.5%	27.1%	32.3%

Table 24. Mean error rates using a classification technique based on the majority of nasal/non-nasal frames for each nasal zone.

	Part of the NASAL zone considered			
EMG Channel	*[0-100%]*	*[0%-50%]*	*[25%-75%]*	*[50%-100%]*
1	18.5%	18.5%	20.5%	22.8%
2	14.1%	19.3%	26.0%	22.8%
3	21.5%	23.7%	29.9%	26.0%
4	**12.6%**	**15.6%**	**13.4%**	**16.5%**
5	23.7%	24.4%	22.1%	27.6%
Mean	**18.1%**	20.3%	22.4%	23.2%

6.1.4.5. Reproducibility Assessment

The EMG signal varied across recording sessions, even for the same speaker. For that reason, we assessed the effect of such variability in our study. Using 4 additional sessions from speaker 1, recorded at different times, we conducted the same analysis as presented in the previous sections. The results were very similar for all sessions, as can be seen in Figure 64, showing evidence that the existing variability among sessions had no major influence on the results. Considering a distribution of 2448 total frames, where 47.0% are nasal and 53.0% are non-nasal, we observed a best error rate of 26.9% for EMG channel 3, as depicted in Figure 64.

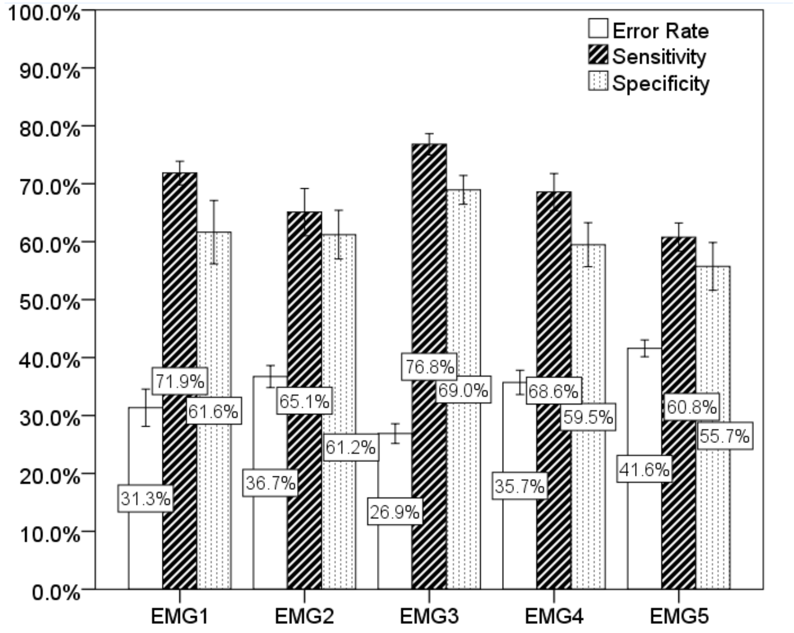


Figure 64. Classification results (mean value of the 10-fold for error rate, sensitivity and specificity) for all channels of Speaker 1. These results are based on four additional sessions from this speaker recorded a posteriori. Error bars show a 95% confidence interval.

Taking into consideration that the dataset distribution presented a higher amount of nasal frames, when we compared the results with the ones reported in the previous sections, we noticed a slightly higher error rate, particularly for channels 4 and 5. For EMG channel 3, the best in both conditions, the difference was very small (1.2% absolute value).

6.2. Nasality detection based on Ultrasonic Doppler sensing

As shown before in Chapter III, nasality can cause word recognition degradation in UDS based interfaces for EP (Freitas et al., 2012b). However, if benefits were to be found, an UDS-based SSI could eventually be included in a multimodal interface as one of the core input modalities.

This section describes an exploratory analysis on the existence of velum movement information detected in the Ultrasonic Doppler signal. The reflected signal contains information about the articulators and the moving parts of the face of the speaker, however, it is yet unclear how to distinguish between articulators and if velum movement information is actually being captured. Therefore, considering our aim of detecting velum movement and to provide a ground truth for our research, we used images collected from RT-MRI and extracted the velum aperture information during the nasal vowels of EP, following a similar strategy as the one described in the previous experiment with SEMG (see section 6.1). Then, by combining and registering these two sources, ensuring compatible conditions and proper time alignment, we were able to estimate the time when the velum moves and the type of movement (i.e. ascending or descending) during a nasal vowel production phenomenon. Using this method we were able to correlate the features extracted from the UDS signal with the signal that represents the velum movement and analyse if velum information was captured in our UDS signal analysis, for all nasal vowels.

6.2.1. Method

In order to understand if velum movement information could be found in the Doppler shifts of the echo signal, we used a signal that describes the velum movement as a reference. This signal was extracted from RT-MRI images, as described in section 6.1.3.1.

The exploratory analysis consisted of measuring the correlation between the extracted information from both datasets (after processing them) and finding at which frequency nasal information is more likely to be found, and for which vowels.

6.2.2. Corpora

For this study, we used part of the corpus collected in the UDS experiment described in Chapter III, section 3.3.2. This UDS dataset shares a set of prompts with the RT-MRI dataset. This subset of prompts is composed by several nonsense words that contain five EP nasal vowels ([6~, e~, i~, o~, u~]) isolated and in word-initial, word-internal and word-final context (e.g. ampa [6~p6], pampa [p6~p6], pam [p6~]). The nasal vowels are flanked by the bilabial stop or the labiodental fricative.

This set contains 3 utterances per nasal vowel and data from a single speaker. The UDS data was recorded at a distance of 12 cm from the speaker.

6.2.3. Processing

We divided the processing part into signal synchronization and feature extraction. In the former, we applied a similar process to the experience of section 6.1. In the latter, we extracted the signal features based on a method also described in Livescu et al. (2009) and Zhu (2008)

6.2.3.1. Signal synchronization

In order to be able to take advantage of the RT-MRI velum information we needed to synchronize the UDS and RT-MRI signals. We started by aligning both UDS and the information extracted from the RT-MRI with the corresponding audio recordings. We resampled the audio recordings to 12 kHz and applied Dynamic Time Warping (DTW) to the signals, in order to find the optimal match between the two sequences. Based on the DTW result we mapped the information extracted from RT-MRI from the original production to the UDS time axis, establishing the correspondence we required between the UDS and the RT-MRI information, as depicted in Figure 65.

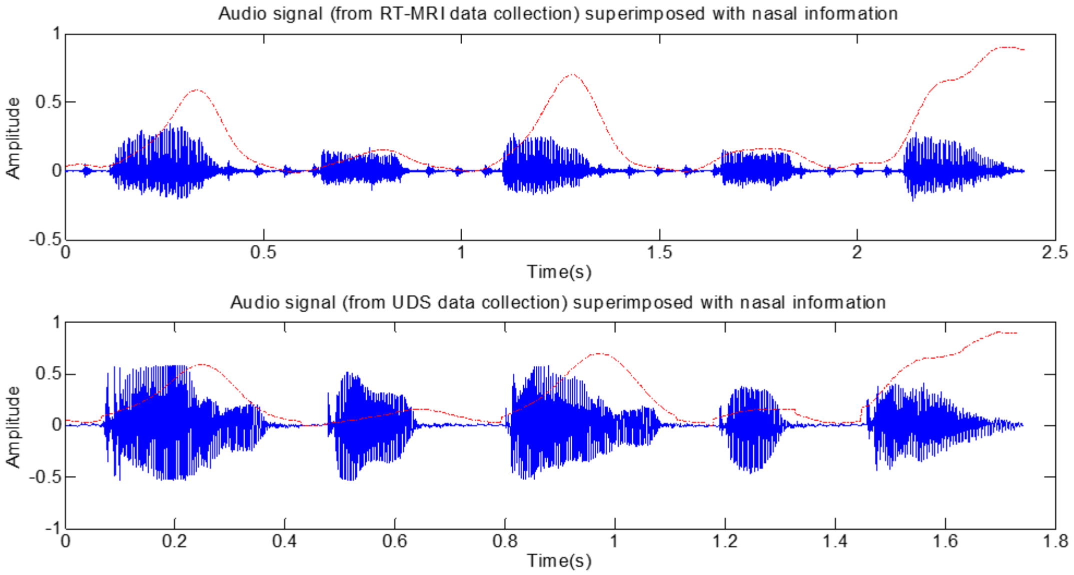


Figure 65. In the top chart we see the audio signal collected at the same time as the RT-MRI and the respective velum movement information (dashed line). In the bottom chart we see the audio signal collected at the same time as the UDS data and the warped signal representing the velum movement information extracted from RT-MRI. Both charts represent the sentence [6~p6, p6~p6, p6~].

6.2.3.2. Feature Extraction

For this experiment we selected two types of features - frequency-band energy averages and energy-band frequency averages (Livescu et al., 2009; Zhu, 2008). To obtain the frequency-band energy averages, we split the signal spectrum into several non-linearly divided bands centred on the carrier. After that, we computed the mean energy for each band. The frequency interval for each band n is given by:

$$Interval_n = [fmin_n, fmax_n], -5 \leq n \leq 4 \tag{28}$$

where $fmin_0 = 4000\,Hz$ (carrier frequency), $fmin_n = fmax_{n-1}$, $fmax_n = fmin_n + \alpha\,(|n| + 1)$, and $\alpha = 40\,Hz$. As such, the bandwidth slowly increased from 40 Hz to 280 Hz, with higher frequency resolution near the carrier.

In order to compute the energy-band frequency averages we split the spectrum into several energy bands and computed the frequency centroid for each band. We extracted values from 14 bands (7 below and 7 above the carrier frequency) using 10 dB energy intervals that ranged from 0 dB to -70 dB. For example, the interval from 0dB to -10dB would give origin to two frequency centroids, one corresponding to the frequencies before the carrier and other corresponding to the frequencies after the carrier.

6.2.4. Experimental Results

In order to achieve our aim of finding if velum movement information is present in the ultrasonic signal, we decided to measure the strength of association between the features of both signals (UDS and RT-MRI velum information). Below, we present several results based on Pearson's product-moment correlation coefficient, which measures how well the two signals are related as well as the results of the application of Independent Component Analysis to the extracted features. The correlation values range between -1 and 1, thus the greater the absolute value of a correlation coefficient, the stronger the linear relationship is. The weakest relationship is indicated by a correlation coefficient equal to 0.

6.2.4.1. Exploratory Analysis

When comparing the RT-MRI velum information with the features obtained along each frequency band, based on correlation magnitude presented in Figure 66, it was not clear which band presented the higher correlation, although the values near the carrier were slightly higher. However, when we split our analysis by vowel, more interesting results became visible. Figure 67 shows the correlation results for utterances where only the nasal vowel [6~] occurred (e.g. ampa [6~p6], pampa [p6~p6], pan [p6~]). A more distinct group of correlation values became visible at the frequency interval [4040..4120] Hz. When looking at the nasal vowel [e~], a stronger correlation became apparent as well in that interval. However, in the case of the nasal vowel [o~] and [u~], higher correlation values were found in the [3880..4040] Hz range, with an average correlation magnitude of 0.42 for [o~] and 0.44 for [u~] (depicted in Figure 68). For the nasal vowel [i~], we found much lower correlation values when compared with the remaining vowels such as [6~], [o~] or [u~]. The best interval could be found in the [4240..4400] Hz range with an average correlation magnitude of 0.25.

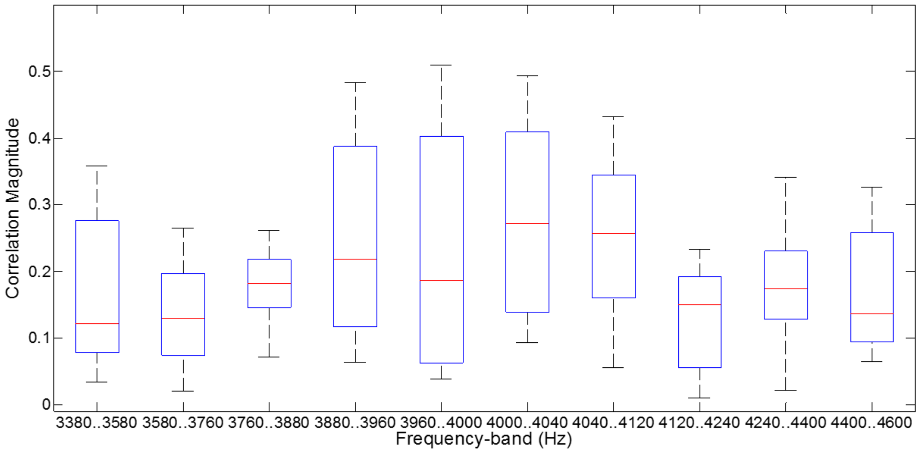


Figure 66. Boxplot for all utterances. The x-axis lists the frequency-band features and the y-axis corresponds to the absolute Pearson's correlation value. The central mark is the median and the edges of the box are the 25th and 75th percentiles.

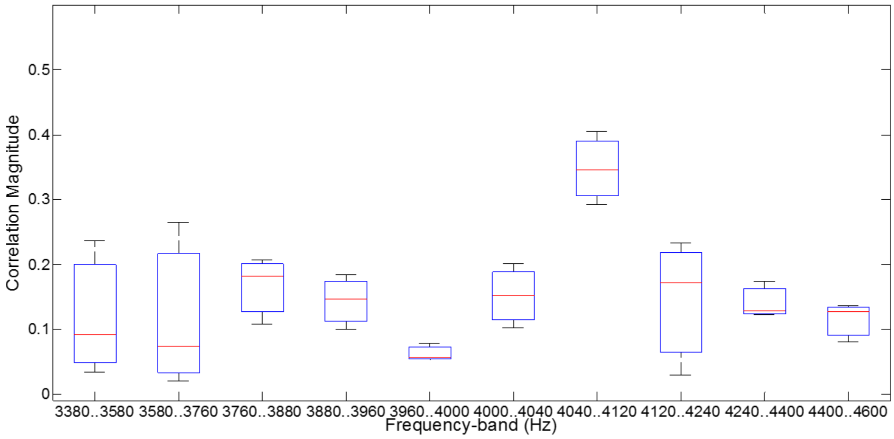


Figure 67. Boxplot for utterances with [6~]. The x-axis lists the frequency-band features and the y-axis corresponds to the absolute Pearson's correlation value.

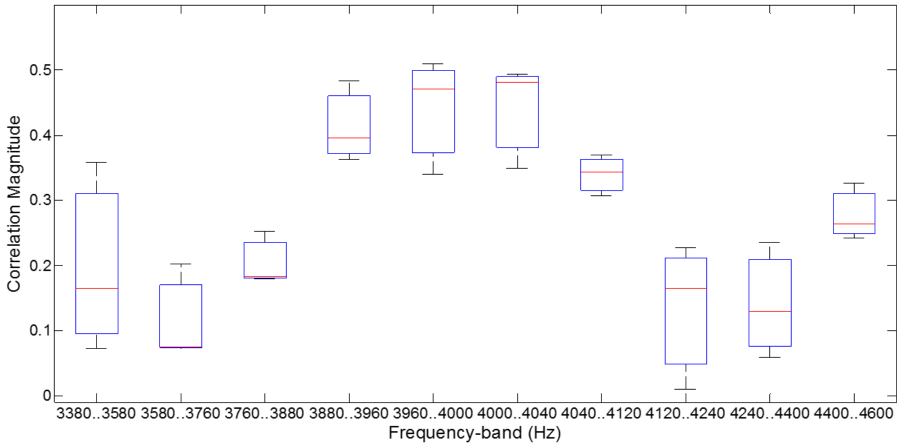


Figure 68. Boxplot for utterances with [u~]. The x-axis lists the frequency-band features and the y-axis corresponds to the absolute Pearson's correlation value.

When looking at the energy-band features for all vowels, we found similar values for the energy bands below -30dB, where the highest average correlation value was achieved by the [-30..-40] dB range above and below the carrier with 0.23. If we split out analysis by vowel, the highest value was achieved by the nasal vowel [o~] with an average correlation of 0.43 for the [-40..-50] dB interval above the carrier. The second best result using energy-band features was obtained by the nasal vowel [u~] in the [-30..-40] dB range with values of 0.40 above the carrier and 0.39 below the carrier.

6.2.4.2. Applying Independent Component Analysis

As mentioned earlier, the Ultrasonic Doppler signal can be seen as the sum for all articulators and moving parts of the face of the speaker. Thus, the signal can be interpreted as a mix of multiple signals. Considering our goal, an ideal solution would be to find a process to isolate the signal created by the velum. Independent Component Analysis (ICA) is a method used for separating a multivariate signal with independent sources linearly mixed. We used this method to understand if by applying blind source separation we could obtain independent components that relate to each articulator movement, including the velum.

For that purpose we applied the FastICA algorithm (Hyvarinen, 1999) using the RT-MRI information as *a priori* to build the separating matrix. This allowed us to obtain independent components with a higher correlation value than when compared to the extracted features without any transformation, as shown in Table 25. Also, in the process of transformation, due to the singularity of the covariance matrix, we observed a dimensionality reduction of 4 to 8 components depending on the utterance when using frequency-band features. When using energy-band features we observed a dimensionality reduction of 6 to 12 components.

Table 25. Average correlation magnitude values with 95% confidence interval for frequency-band and energy band features for the best independent components of each utterance.

Average correlation magnitude		
	Frequency	*Energy*
All vowels	0.42 ± 0.05	0.41 ± 0.04
[6~]	0.44 ± 0.04	0.33 ± 0.05
[e~]	0.41 ± 0.09	0.41 ± 0.10
[i~]	0.30 ± 0.05	0.42 ± 0.03
[o~]	0.47 ± 0.08	0.41 ± 0.05
[u~]	0.48 ± 0.14	0.50 ± 0.07

6.3. Discussion

The global results of this case study pointed to the fact that SEMG can be used to reduce the error rate caused by nasality in languages where this characteristic is particularly relevant, such as EP. For UDS, the results were not conclusive, but a moderate correlation between velum movement information and the UDS signal features was found.

The applied methodology used the audio signal to synchronize two distinct signals that otherwise were very hard to align. Although the two sources of information were recorded at different times, it is our belief that by reproducing the articulation movements we were able to obtain a very good indication of how the velum behaves.

In the first experiment, there was a noticeable trend that points to the EMG electrode pairs placed below the ear between the mastoid process and the mandible in the upper neck area as being the most promising sensors for detecting the myoelectric signal generated by the velum movement. To the best of our knoweledge, this is the first study that uses SEMG sensors in the regions of the face and neck to detect velum movement and as such no direct comparison with the literature could be made.

In a first stage, when overlaying the aligned RT-MRI and EMG signals the resulting match between them was more prominent for channels 2, 3 and 4, particularly for nasal vowels in medial and final word positions. However, when looking at the close vowel case and at the vowels in an isolated context, EMG channels 3 and 4 emerged as the ones with the closest match with the RT-MRI information. In the visual analysis, the fact that we obtained a more evident match in the case of vowels in medial and final word position suggested that it could be interesting to further analyse the differences between the position of nasal vowels. The word-final context only required an opening movement, while a word-initial position could conceivably in some cases require only a closing movement. We pose that at least one could say that any opening movement is under weak temporal constraints. The word-internal context required a clear opening-closing movement under strict temporal constraints given by the linguistically appropriate duration between the flanking plosives. Thus it is perhaps not surprising that the latter context gave clear results. However, different contexts, ideally where no jaw movement is required, should be considered to discard or minimize possible muscle crosstalk.

In our data analysis, using a different approach, we also compared the information present in the nasal zones of the EMG signal with the velar information of the RT-MRI signal using mutual information to measure the relation between the signals. The better results were found in EMG channels 3, 4 and 5, which matches up with our visual analysis of the signal.

Following our data analysis, we investigated if it was possible to develop a classifier that enabled us to distinguish between nasal and non-nasal frames of an EMG signal in a realistic frame independent scenario. The results for this scenario showed that it is possible (above chance level) to distinguish between the two frame classes, following the results from our previous analysis, particularly in EMG channels 2, 3 and 4. Specificity and sensitivity measures found for these channels were also slightly higher when compared with the remaining channels, showing that more true positive/negative results were accurately found. When looking at the results of the classifier per speaker, an improvement in the results could be noticed, showing a better modelling of our classification task. In this case, EMG channel 3 presented the best results for all speakers. EMG channel 1 and 4 presented substantial differences between speakers and recording sessions, showing that further exploration is required, particularly for channel 1, which, due to its position near several bone structures, may present high sensitivity with respect to sensor position or anatomic structure in that area when it comes to tracking this set of muscles.

The statistical analysis also showed that significant differences could be found between the results of EMG channel 3 and the EMG channels 1, 2 and 5, in line with what was discussed above, where EMG channel 3 emerged as the channel with the best results.

In our study, we also attempted to combine multiple channels, however, this did not improve the results we obtained. This could be an indication of the following conjectures: (1) the muscles' signals captured by the EMG channels overlap, thus, combining multiple channels does not improve the nasal/non-nasal class separation; (2) due to the superimposition of muscles in that area, adding a new channel with information of other muscles will create a less accurate model of the classes; (3) the muscles related with velum movement are not being captured by the added channel(s).

In a scenario where each zone was classified based on the majority frame type (nasal or non-nasal), we found a noteworthy improvement in the accuracy rates, reaching 12.6% for all users. Although this was not a realistic scenario, in the sense that an interface using surface EMG to detect velum movement does not know *a priori* the nasal zone boundary, these results allowed us to understand that the use of neighboring frames and introducing frame context might help to improve the accuracy of our methodology. It is also interesting to note the higher error rate in non-nasal frames and the fact that better results were obtained not using the whole zone, but only the central part of each zone.

An alternative approach to the classifier we developed was to detect velum movement events, since EMG activity was most likely to be found when the position of the velum needed to change. However, based on the results obtained in this study and the physiology of the area in question, it did not seem likely that a position without other muscle crosstalk could be found. As such, an event-based approach would need to find a way to avoid false positives originating from neighboring muscles.

Analyzing the reproducibility of the results, additional sessions from a random speaker recorded a posteriori, supported the previously achieved results, further evidencing that, for the speaker in question, the position selected for EMG channel 3 attained the best results.

The overall results of this experiment seemed to indicate that information regarding the velum movement was actually captured by the SEMG sensors, however, it was not clear which muscles were being measured.

Additionaly, although the methodology used in this study partially relied on RT-MRI information for scientific substantiation, a technology which requires a complex and expensive setup, the proposed solution to detect nasality was solely based on a single, non-invasive sensor of surface EMG. Thus, the development of a multimodal SSI based on SEMG for EP, with language adapted sensor positioning, now seems to be a possibility.

The second experiment, using UDS, altough not conclusive, showed a promising relation between some information extracted from the UDS signal and the velum movement. Knowing that the velum is a slow articulator, as shown by the RT-MRI velum movement information in Figure 65, and considering Eq. 3, we expected that velum movement, if detected by UDS, would be found in the regions near the carrier, which is where the results for [6~, e~, o~ and u~] present higher correlation. However, the velum is not the only slowly moving articulator and a different corpus which allows, for example, to discard jaw movements should be considered for future studies.

Another point of discussion is the differences found between nasal vowels. When looking at the correlation results of frequency-band features, we noticed a difference between [i~] and the remaining vowels. One possible explanation for this difference could be the articulation variances of each nasal vowel previously reported in literature (Schwartz, 1968). Since our technique was based on the reflection of the signal it is plausible that the tongue position influenced the detection of the velum, particularly for the case of [i~], in which the tongue posture may block the UDS signal.

According to Livescu et al. (2009), a clear difference between close and open vowels is to be expected. Although this is true for the nasal vowel [i~], we could not verify this in the case of [u~], which presented the highest correlation values, along with [o~]. Further investigation is required to understand if for example the rounding of the lips during the articulation of these two vowels is influencing the signal reflection and in which way.

For the utterances containing the [u~] nasal vowels, we noticed some alignment inaccuracies mainly at the end of the first phoneme. Thus, further improvements need to be considered for this particular case.

In this study we also applied blind source separation as an attempt to split the signal into independent components. This technique gave slightly better results for both sets of features, showing that some isolation of the velum movement in the Doppler shifts could be possible. Also noteworthy is the fact that this process has led to a dimensionality reduction (in the number of components) depending on the utterance, which may have a relation with the number of mobile articulators that can cause Doppler shifts in the signal (i.e. tongue, lower jaw, velum, lips, cheeks, oral cavity).

6.3.1. Speaker Postures

Regarding the different speaker postures during acquisition, for both RT-MRI, UDS and EMG data, and how this can influence the matching of both datasets, several studies have been presented in the literature concerning the differences in vocal tract configuration between sitting/upright and supine acquisitions of articulatory data. In general, mild differences in vocal tract shape and articulator position due to body posture were reported by several authors (Kitamura et al., 2005; Stone et al., 2007; Tiede et al., 2000; Wrench et al., 2011). These mostly refer to an overall tendency of the articulators to deform according to gravity, resulting in more retracted positions of the tongue and, to a lesser extent, of the lips and lower end of the uvula in the supine acquisitions. To which extent this effect is observed varies between speakers (Kitamura et al., 2005). None of these studies specifically addressed the velum or nasality, but an analysis of vocal tract outlines, when they contemplate the velum (Kitamura et al., 2005), does not seem to show any major difference. Considering the acoustic properties of speech, one important finding is that no significant differences have been found between the different body postures during acquisition (Engwall, 2006; Stone et al., 2007). Furthermore, the acquisition of running speech, as shown by Tiede et al. (2000) and Engwall (2006), minimizes vocal tract shape and articulator position differences occurring between upright and supine positions.

Concentrating on the velopharyngeal mechanism, Perry (2011) studied its configuration for upright and supine positions during speech production, and concluded that no significant difference was present regarding velar length, velar height and levator muscle length.

Regarding muscle activity, Moon et al. (Moon and Canady, 1995) presented a study where the activity of the *levator veli palatini* and *palatoglossus* muscles for upright and supine speaker postures were assessed using EMG. The activation levels were smaller for the supine position during the closing movement (with gravity working in the same direction), but no timing differences were reported.

Therefore, considering the methodology used and the knowledge available in the literature, the different postures do not seem to be a major differencing factor between the datasets that would preclude their matching with the acoustic signal. Furthermore, no evidence exists that muscle activity is relevantly affected by speaker posture. For both postures, apart from different activation levels, similar muscular configurations and activity seem to be observed, supporting the assumption that, after aligning the datasets, muscle activity relating to the velopharyngeal mechanism should also be aligned between both signals. This allowed us to infer muscle activity intervals from the velum movement information extracted from the MRI images.

6.4. Summary

This chapter described two experiments, which can be seen as case studies of the use of a framework for addressing the challenge of nasality detection. The first case study investigated the potential of SEMG for nasality detection (i.e. detecting the myoelectric signal associated with the velum movement). The second case study had a similar goal but explored a non-obtrusive technology – Ultrasonic Doppler. In both studies RT-MRI was used to extract ground information to help address the challenge of nasality detection. Simply stated, the information extracted from the RT-MRI images allowed us to know when to expect nasal information. Thus, by synchronizing both signals, based on simultaneously recorded audio signals from the same speaker, we were able to explore the existence of useful information in the EMG/UDS signal about velum movement.

The results of the first study showed that in a real use situation (in a frame-based classification experiment) error rates of 23.7% can be achieved for sensors positioned below the ear, between the mastoid process and the mandible in the upper neck region. They also showed that careful articulation and positioning of the sensors can influence nasal vowel detection results. These outcomes indicate that the described approach can be a valuable contribution to the state-of-the-art in the area of EMG speech interfaces, and that this approach can be applied in parallel with other techniques.

In the second study we measured the strength of association between the features that describe the ultrasonic signal data and RT-MRI data. The results we obtained showed that for features based on the energy of pre-determined frequency bands, we were able find moderate correlation values, for the case of the vowels [6~], [o~] and [u~] and weaker correlation values in the [i~] case. Moderate correlation values were also found using energy based features for bands below -30dB.

CHAPTER VII

Conclusions and Future work

"O que dá o verdadeiro sentido ao encontro é a busca e que é preciso andar muito para alcançar o que está perto."

José Saramago, Todos os Nomes

Contents

In this final chapter we start by presenting the major milestones over the course of the research work that supported this thesis. Afterwards, we present the main results and confront them with the stated hypotheses, arguing about its corroboration based on the achieved evidences and observations. To close, we conclude our work with considerations about future research and some final remarks.

7.1. Work Overview

The work presented in this thesis followed an empirical research methodology, where evidence is reached through experimental studies to support the testing of the formulated hypotheses. Most of the experiments relied on quantitative measurements.

The overall approach to achieve the established goals can be divided into the following stages: problem identification; problem assessment; establishment of thesis hypotheses; setting of thesis objectives to satisfy and demonstrate the hypotheses; state-of-the-art analysis; development of experiments; analyze and compare results; validate hypotheses and formulate conclusions.

In our methodology, we have proposed four hypotheses and came up with four matching objectives. Following this strategy, we have divided the work into the following stages, aligned with the four objectives (O1, O2, O3 and O4) presented in Chapter I and depicted in Figure 69.

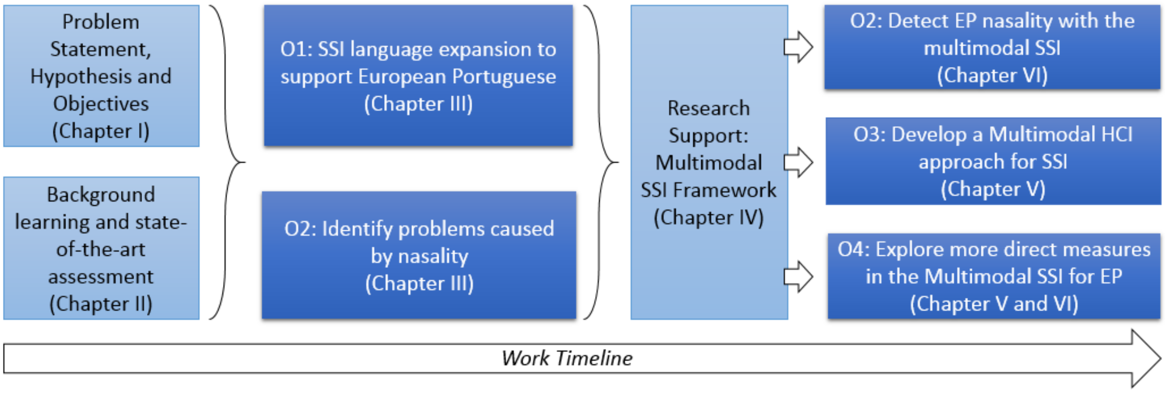

Problem Statement, Hypothesis and Objectives (Chapter I)
Background learning and state-of-the-art assessment (Chapter II)
O1: SSI language expansion to support European Portuguese (Chapter III)
O2: Identify problems caused by nasality (Chapter III)
Research Support: Multimodal SSI Framework (Chapter IV)
O2: Detect EP nasality with the multimodal SSI (Chapter VI)
O3: Develop a Multimodal HCI approach for SSI (Chapter V)
O4: Explore more direct measures in the Multimodal SSI for EP (Chapter V and VI)
Work Timeline

Figure 69. Overview of work stages chronologically ordered from bottom to top and its relation with the thesis structure and the defined objectives.

For **Objective 1 – SSI language expansion to support European Portuguese –** we started by conducting an in depth study of related work, including learning the necessary background knowledge in areas such as anatomy, physiology, phonetics and silent speech. Then, after attaining a general idea of the state-of-the-art in SSI, a preliminary evaluation of different types of SSIs was made, the major problems were identified and the aim of this thesis was defined, by enumerating the four hypotheses. These initial studies contributed to determine which SSI modalities were more suited for the problem and what were the available resources in the scientific community. Results of this work comprised the state-of-the-art assessment, a successful introduction of a new language (European Portuguese) in SSI and findings on the main issues to be solved by our research, as reported in (Freitas et al., 2012a, 2011).

As a result of the preliminary studies conducted under O1, we were able to confirm nasality as one of the problems in the language adoption of EP, in line with our stated **Objective 2 - Detect**

EP nasality with the multimodal SSI, by overcoming identified limitations of current techniques. To overcome the challenge of O2, we have decided to work on a way to capture the missing information.

Motivated by the good results achieved on O1 and, in parallel with the work described to fulfill O2, we have started to build the foundations for supporting multimodal research in SSI. This work entailed the design and development of a multimodal framework for the joint exploitation of input SSI modalities.

By applying the framework into different scenarios, we have investigated ways to achieve what had been proposed in O2, O3 and O4, as follows. For addressing the problems caused by nasality in European Portuguese, we have chosen SEMG and UDS as two potential input modalities capable of tackling this issue. As such, by leveraging the multimodal framework architecture in an offline synchronization scenario, we took advantage of a multimodal approach and extracted more direct measures from RT-MRI data that could be exploited as ground truth for our experiments, also in line with **Objective 4 - Explore more direct measures in the Multimodal SSI for EP**. With accurate information about the velum, we were able to confront EMG and UDS data with an estimation of velum movement and analyze the both signals enlightened by the RT-MRI velum information, which worked as ground truth data. Results from this study were reported in (Freitas et al., 2014c) and allowed to complete with success O2.

With this study at hand and using our computing framework, we have also created the conditions to address **Objective 3 - Develop a Multimodal HCI approach for SSI**. In this context, we have analyzed if combining (or fusing) non-invasive modalities could eventually create an SSI that would be beneficial for our scenario, in terms of word error rate. However, when multiple input modalities are considered, the large dimensionality of the feature space augments the complexity of the classification techniques and thus, of the word recognition task. This issue motivated the evaluation of several feature selection and feature reduction techniques, used to minimize redundancy and maximize relevancy of our data and consequently, improve our recognition results. Results of this study were reported and analyzed in (Freitas et al., 2014a).

In the scope of **Objective 4 - Explore more direct measures in the Multimodal SSI for EP,** studies with other modalities were also performed. An example is the use of US imaging to assess the potential of EMG in the detection of tongue gestures, using a dataset with up to seven streams of synchronous data. In parallel with the previous studies, we have also explored the use of articulatory information that can potentially be combined amongst different modalities to achieve a more complete representation of the human speech production process. Results related with the satisfaction of O4 were reported and discussed in (Freitas et al., 2014d).

The dissemination of intermediate results was performed through the publication of papers in international conferences and top journals of the area and the presentation of the work in seminars and workshops, as listed in detail section 1.5 of Chapter I. The activity of writing the papers and this thesis was transversal to the whole thesis, along with a constant update to the state-of-the-art. During the final stage of research, the full thesis document was written, by simply integrating and further developing the published peer-reviewed papers.

7.2. Key Results and Outcomes

Overall, from a more generic perspective, the main contribution of this thesis are a set of studies related with EP adoption in Silent Speech Interfaces that, in some cases, takes advantage from more direct measures, supported by a multimodal research framework that provides the means for a joint exploration of modalities.

In more concrete terms, the key results of this work provide a response for the four stated hypotheses as follows.

The first contribution of this thesis is applying for the first time existing methods in the area of SSI to EP. Before the work here described only contributions in related fields were found for this language, such as in speech production studies for EP (Martins et al., 2008), not knowing if differences in SSI performance could be noticed by the adoption of EP. After several experiments, we have demonstrated that is possible to adapt and even extend existing international work in SSI for EP, when considering an isolated word recognition task. We have found that recognition of nasal sounds may cause an inferior performance for several input modalities, such as SEMG, UDS and Video.

These first results, led us, in a certain way, to our second main contribution, related with the demonstration of the second hypothesis, based on the study of the problems caused by nasality in EP and how specific SSI modalities could detect or not such phenomena, which, along with prosody, is an evident and agreed challenge for SSI. Thus, our second contribution, supported by our prior experimental confirmation that adopting a language with strong nasal characteristics causes a decrease in the performance of an SSI, consists of a non-invasive solution for the detection of nasality based on a single SEMG sensor positioned below the ear, between the mastoid process and the mandible in the upper neck region. We have successfully included this nasality detection sensing capability via SEMG, in our multimodal SSI framework. To the best of our knowledge, prior to this conclusion, it was unknown in the Speech scientific community, if such muscles could actually be detected using SEMG. We believe that our work on this specific topic, might provide further insights for other areas related with facial muscles physiology.

Our fourth contribution of this thesis is the developed multimodal SSI computing framework, supporting the majority of the experimental studies here presented. This state-of-the-art framework was used with different goals and in different settings that go from, data collection with synchronous acquisition of multimodal data; processing, extracting and selecting and/or reducing the best features of each input modality; to analysis and classification, via machine learning techniques, of the processed information. With this computing framework, we were able to create the conditions to provide an adequate evidence for the satisfaction of the third (a multimodal HCI approach, based on less invasive modalities, has the potential to improve recognition results) and fourth hypothesis (supplementary measures acquired from more invasive modalities can be use as ground truth). Hence, analyzing the outcome of multimodal studies with less invasive modalities, we have noticed word recognition performance benefits, when combining modalities, in some identified cases. However, these results should be looked at with caution and can hardly be generalized, since other factors such as dataset size, vocabulary, speakers, language, applied techniques, etc., influence the performance of an SSI. Nonetheless, we believe to have taken an important step in using multimodal approaches like the ones described in this document, towards a complete or significant representation of the human speech production process, and with the possibility of more comprehensive articulatory approach.

A fifth and relevant contribution of this thesis, related with our fourth hypothesis is a method for taking advantage of supplementary direct measures acquired from more invasive modalities in multimodal setups. We believe that this method, also applied in O2, is of the uttermost importance for research in this area since it allows for accurate collection of information about the articulators to be used as ground truth data, leading us to an enhanced comprehension of other modalities. We have shown examples of the application of such technique, to the precise study of tongue and velum gestures, using US imaging and RT-MRI, respectively, in order to assess the potential of input SSI modalities such as SEMG and UDS.

7.3. Future work

This section describes some of the future work to continue and enhance the research here presented. We first describe the vision of the author for leveraging the developed framework beyond of what was presented, and then we discuss some future ideas for related areas.

7.3.1. Leveraging the framework

We can think of ways of improving and evolving the proposed framework, without the need to redefine the framework itself.

Firstly, the presented framework allows easily adding new input SSI modalities, by increasing the support to the number of stages in the speech production process, from which we extract information. This means to add modalities such as EEG, either to aid in better understanding other modalities, or to see if word recognition can be improved when looking at the brain signals. Another possibility is to add vibration (such as NAM - Non-Audible Murmur microphone), or electromagnetic sensors to obtain further information from the vocal tract. However, as mentioned before, with this method, human glottal activity is required and some scenarios involving speech impaired users would no longer be supported.

Secondly, baseline processing for each modality could be further improved and other signal processing methods could be explored for each modality (e.g. different feature extraction techniques). Nonetheless, our short-term goal continues to be the understanding of the best way to fuse the data coming from multiple sources, without falling into a high-dimensionality problem, and maximizing relevance and minimizing redundancy of extracted features. For that reason, we would like to explore the combination of feature selection with feature discretization techniques, in order to improve the classification performance, particularly in scenarios where there are only a few examples (instances) available.

We can also take advantage of articulatory information to improve the baseline results. An option would be to use the obtained information about the lip movement, refine it, and tackle one of the RGB Video limitations by combining it with information from the tongue using SEMG. In the long-term, we expect to design an articulatory SSI for EP and other languages, which will be able to interpret articulators' movement using multiple modalities. Besides complementing themselves, the redundancy of information found across some modalities can also be used for confirmation of some gestures. However, we would still need to carry a more detailed analysis of the characteristics of individual articulatory movements, to explore combinations of several sensors and to develop a classifier or classifiers, for integration in a multimodal SSI with an articulatory basis.

With our current framework, we will also be able to easily include new and more accurate state-of-the-art sensors, such as time-of-flight depth cameras, which provide more precise depth information when compared to the ones we used. With such technology, we might be able to extract movements such as lip protrusion with just one sensor facing the user in a frontal posture, which usually requires a lateral view of the speaker.

Future use of the framework includes obtaining more extensive vocabularies, other languages with distinctive characteristics and continuous speech scenarios. However, it is important to have a more comprehensive understanding of the capabilities of each modality before tackling more cumbersome scenarios that require a solid and stable knowledge basis.

To support our studies, it is also foreseen that further RT-MRI sessions be conducted in order to collect new corpora, targeting the exploration of other phonetic phenomena of EP, such as nasal consonants, which was not possible in our case due to the lack of RT-MRI information.

Last but not least, we plan to continue our work on multimodal SSI using non-invasive modalities, to provide an adequate response to specific user scenarios, striving for additional robustness in situations, such as noisy environments and where existing disabilities might hinder single modality interaction.

7.3.2. Expanding to other areas of research

The work here presented has been focused in SSI, a system capable of supporting human-computer communication even when no acoustic signal is available. However, the techniques employed in such systems, including the ones presented in this thesis, could be used for other purposes in the scope of communication in general (e.g. human-human communication, human-machine communication, speech therapy, etc.).

To take SSI systems beyond the research stage, and integrating them in real applications, would require high usability and user acceptance level standards, particularly for SSI systems that target HCI scenarios. As part of the future work, usability evaluation is a must for this type of interfaces and should always be taken into consideration in a future system design. Evaluating the interface usability allows to identify shortcomings, refine previously established user requirements and improve the existent prototypes. Additionally, one could assess the system efficiency, task effectiveness and, very important, if the user enjoyed and presented a positive attitude towards the evaluated system.

Another area of application, relates to users with speech impairments. We believe that the concepts here presented could also be extended for speech therapy. A speech therapist often needs to assess the capabilities of an individual and, besides relying mostly on the acoustical signal to do so, the technological means to aid such tasks are scarce and frequently require a significant amount of manual input. Hence, we believe that a comprehensive assessment of speech capabilities in terms of articulation, like the ones we obtained using our multimodal framework, could provide added value and thus be beneficial for this scenario. In this context, an extended version of the framework, including for example fMRI and EEG used as ground truth modalities, could help improving and automating current speech therapy methods.

7.4. Final Remarks

This thesis unveils numerous research paths for future research and, for that reason, we believe it represents a large stepping-stone for achieving higher ends in Silent Speech and related fields. In a certain way, it lays the foundations for broader ideas that encompass and combine important areas, such as speech therapy, phonetics, speech production and HCI.

Our strategy of never losing sight of more user-friendly modalities and aiming towards a comprehensive approach, in what speech production is concerned, removes several barriers usually found in research-stage solutions that include, for example, the use of invasive modalities. This focus on non-invasive modalities eventually allows us to reach the end-user at a faster pace, since there is a smaller or inexistent necessity of adapting solutions that have known usability issues. Additionally, we believe that the search for complimentary knowledge also opens the door for the design and development of improved HCI systems, especially in the therapy sector. For example, an enhanced representation of speech production can be leveraged for future speech therapy systems, with knowledge benefits for the therapist and more productive sessions for the patient.

We expect to take this work to further lengths, but more than that, we expect that the developments here described contribute not only for further scientific advances in this area, but also lead to useful applications that help and benefit society, in particular speech-impaired users.

"The rest is silence."

William Shakespeare, Hamlet (1600-02), Act V, scene 2, line 368

REFERENCES

Aha, D.W., Kibler, D., Albert, M.K., 1991. Instance-based learning algorithms. Mach. Learn. 6, 37–66.

Albuquerque, L., Oliveira, O., Teixeira, T., Sá-Couto, P., Freitas, J., Dias, M.S., 2014. Impact of age in the production of European Portuguese vowels, in: 15th Annual Conf. of the Int. Speech Communication Association (Interspeech 2014). Singapore, pp. 940–944.

Alfonso, P.J., Baer, T., 1982. Dynamics of vowel articulation. Lang. Speech 25, 151–173.

Alghowinem, S., Wagner, M., Goecke, R., 2013. AusTalk—The Australian speech database: Design framework, recording experience and localisation, in: 8th Int. Conf. on Information Technology in Asia (CITA 2013). IEEE, pp. 1–7.

Almeida, A., 1976. The Portuguese nasal vowels: Phonetics and phonemics, in: Schmidt-Radefelt, J. (Ed.), Readings in Portuguese Linguistics. Amsterdam, pp. 348–396.

Anderson, S., Liberman, N., Bernstein, E., Foster, S., Cate, E., Levin, B., Hudson, R., 1999. Recognition of elderly speech and voice-driven document retrieval, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1999). IEEE, pp. 145–148.

Articulate Assistant Advanced Ultrasound Module User Manual, Revision 212, [WWW Document], n.d. . Articul. Instruments, Ltd. URL http://www.articulateinstruments.com/aaa/ (accessed 5.24.14).

Baba, A., Yoshizawa, S., Yamada, M., Lee, A., Shikano, K., 2002. Elderly acoustic model for large vocabulary continuous speech recognition. IEICE Trans. Inf. Syst. J85-D-2, 390–397.

Babani, D., Toda, T., Saruwatari, H., Shikano, K., 2011. Acoustic model training for non-audible murmur recognition using transformed normal speech data. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2011) 5224–5227. doi:10.1109/ICASSP.2011.5947535

Bastos, R., Dias, M.S., 2008. Automatic camera pose initialization, using scale, rotation and luminance invariant natural feature tracking. J. WSCG 16, 34–47.

Bastos, R., Dias., M.S., 2009. FIRST - Fast Invariant to Rotation and Scale Transform: Invariant Image Features for Augmented Reality and Computer Vision. VDM Verlag.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: European Conference on Computer Vision (ECCV 2006). Springer, pp. 404–417.

Beddor, P.S., 1993. The perception of nasal vowels. Nasals, nasalization, and the velum 5, 171–196.

Bell-Berti, F., 1976. An electromyographic study of velopharyngeal function in speech. J. Speech, Lang. Hear. Res. 19, 225–240.

Bell-Berti, F., Krakow, R.A., Ross, D., Horiguchi, S., 1993. The rise and fall of the soft palate: The Velotrace. J. Acoust. Soc. Am. 93, 2416.

Betts, B.J., Jorgensen, C., Field, M., 2006. Small Vocabulary Recognition Using Surface Electromyography in an Acoustically Harsh Environment. J. Human-Computer Interact. 18, 1242–1259. doi:10.1.1.101.7060

Birkholz, P., Schutte, M., Preuß, S., Neuschaefer-Rube, C., 2014. Towards non-invasive velum state detection during speaking using high-frequency acoustic chirps, in: Hoffmann, R. (Ed.), Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014. TUDPress, Dresden, pp. 126–133.

Bresch, E., Kim, Y.-C., Nayak, K., Byrd, D., Narayanan, S., 2008. Seeing speech: Capturing vocal tract shaping using real-time magnetic resonance imaging. IEEE Signal Process. Mag. 25, 123–132. doi:10.1109/MSP.2008.918034

Bressmann, T., 2005. Comparison of nasalance scores obtained with the nasometer, the nasalview, and the oronasal system. Cleft palate-craniofacial J. 42, 423–433.

Brumberg, J.S., Guenther, F.H., Kennedy, P.R., 2013. An Auditory Output Brain–Computer Interface for Speech Communication, in: Brain-Computer Interface Research. Springer, pp. 7–14.

Brumberg, J.S., Kennedy, P.R., Guenther, F.H., 2009. Artificial speech synthesizer control by brain-computer interface., in: Proceedings of Interspeech 2009. pp. 636–639.

Brumberg, J.S., Nieto-Castanon, A., Kennedy, P.R., Guenther, F.H., 2010. Brain-Computer Interfaces for Speech Communication. Speech Commun. 52, 367–379. doi:10.1016/j.specom.2010.01.001

Brumberg, J.S., Wright, E.J., Andreasen, D.S., Guenther, F.H., Kennedy, P.R., 2011. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front. Neurosci. 5.

Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. 2, 121–167.

Burnham, D., Estival, D., Fazio, S., Viethen, J., Cox, F., Dale, R., Cassidy, S., Epps, J., Togneri, R., Wagner, M., 2011. Building an Audio-Visual Corpus of Australian English: Large Corpus Collection with an Economical Portable and Replicable Black Box., in: Proceedings of Interspeech 2011. pp. 841–844.

Calliess, J.-P., Schultz, T., 2006. Further investigations on unspoken speech. Universitat Karlsruhe (TH), Karlsruhe, Germany.

Carstens, n.d. . URL http//www.articulograph.de/ (accessed 10-30-2014).

Carvalho, P., Oliveira, T., Ciobanu, L., Gaspar, F., Teixeira, L., Bastos, R., Cardoso, J., Dias, M., Côrte-Real, L., 2013. Analysis of object description methods in a video object tracking environment. Mach. Vis. Appl. 24, 1149–1165. doi:10.1007/s00138-013-0523-z

Chan, A.D.C., 2003. Multi-expert automatic speech recognition system using myoelectric signals. The University of New Brunswick (Canada).

Chan, A.D.C., Englehart, K., Hudgins, B., Lovely, D.F., 2001. Hidden Markov model classification of myoelectric signals in speech, in: Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 1727–1730.

Chan, A.D.C., Englehart, K., Hudgins, B., Lovely, D.F., 2002. Hidden Markov model classification of myoelectric signals in speech. Eng. Med. Biol. Mag. IEEE 21, 143–146.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A Library for Support Vector Machines. ACM Trans. Intell. Syst. Technol. 2, 27:1–27:27. doi:10.1145/1961189.1961199

Chavuenet, W., 1871. A manual of spherical and practical astronomy.

Cisek, E., Triche, K., 2005. Depression and Social Support Among Older Adult Computer Users, in: 113th Annual Convention of the American Psychological Association.

Cootes, T.F., Edwards, G.J., Taylor, C.J., 2001. Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23, 681–685. doi:10.1109/34.927467

Cover, T.M., Thomas, J.A., 2005. Elements of Information Theory, Elements of Information Theory. doi:10.1002/047174882X

Dalston, R., 1982. Photodetector assessment of velopharyngeal activity. Cleft Palate J. 19, 1–8.

Dalston, R.M., 1989. Using simultaneous photodetection and nasometry to monitor velopharyngeal behavior during speech. J. Speech, Lang. Hear. Res. 32, 195–202.

Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection, in: International Conference on Machine Learning. pp. 74–81.

DaSalla, C.S., Kambara, H., Koike, Y., Sato, M., 2009. Spatial filtering and single-trial classification of EEG during vowel speech imagery, in: Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology. ACM, p. 27.

Davidson, W., Abramowitz, M., 2006. Molecular expressions microscopy primer: Digital image processing-difference of gaussians edge enhancement algorithm, Olympus America Inc., and Florida State University.

De Luca, C.J., 1979. Physiology and mathematics of myoelectric signals. IEEE Trans. Biomed. Eng. 26, 313–25.

De Wachter, M., Matton, M., Demuynck, K., Wambacq, P., Cools, R., Van Compernolle, D., 2007. Template-Based Continuous Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15, 1377–1390. doi:10.1109/TASL.2007.894524

Debruyne, F., Delaere, P., Wouters, J., Uwents, P., 1994. Acoustic analysis of tracheo-oesophageal versus oesophageal speech. J. Laryngol. Otol. 108, 325–328.

Denby, B., 2013. Down with Sound, the Story of Silent Speech, in: Workshop on Speech Production in Automatic Speech Recognition.

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J.M., Brumberg, J.S., 2010. Silent speech interfaces. Speech Commun. 52, 270–287. doi:10.1016/j.specom.2009.08.002

Denby, B., Stone, M., 2004. Speech synthesis from real time ultrasound images of the tongue. 2004 IEEE Int. Conf. Acoust. Speech, Signal Process. 1. doi:10.1109/ICASSP.2004.1326078

Deng, Y., Heaton, J.T., Meltzner, G.S., 2014. Towards a Practical Silent Speech Recognition System, in: Proceedings of Interspeech 2014. pp. 1164–1168.

Deng, Y., Patel, R., Heaton, J.T., Colby, G., Gilmore, L.D., Cabrera, J., Roy, S.H., Luca, C.J. De, Meltzner, G.S., 2009. Disordered speech recognition using acoustic and sEMG signals, in: Proceedings of Interspeech 2009. ISCA, pp. 644–647.

Dias, M.S., Bastos, R., Fernandes, J., Tavares, J., Santos, P., 2009. Using hand gesture and speech in a multimodal augmented reality environment, in: Gesture-Based Human-Computer Interaction and Simulation. Springer, pp. 175–180.

Dias, M.S., Pires, C.G., Pinto, F.M., Teixeira, V.D., Freitas, J., 2012. Multimodal user interfaces to improve social integration of elderly and mobility impaired. Stud. Heal. Technol. Informatics 177, 14–25. doi:10.3233/978-1-61499-069-7-14

Dubois, C., Otzenberger, H., Gounot, D., Sock, R., Metz-Lutz, M.-N., 2012. Visemic processing in audiovisual discrimination of natural speech: a simultaneous fMRI–EEG study. Neuropsychologia 50, 1316–1326.

Ekman, P., Rosenberg, E.L., 1997. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press.

Engwall, O., 2006. Assessing MRI measurements: Effects of sustenation, gravitation and coarticulation, in: Harrington, J., Tabain, M. (Eds.), Speech Production: Models, Phonetic Processes and Techniques. Psychology Press, New York, pp. 301–314.

Everitt, B.S., Hothorn, T., 2009. A Handbook of Statistical Analyses Using R, Water.

Fabre, D., Hueber, T., Badin, P., 2014. Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression, in: Proceedings of Interspeech 2014. pp. 2293–2297.

Fagan, M.J., Ell, S.R., Gilbert, J.M., Sarrazin, E., Chapman, P.M., 2008. Development of a (silent) speech recognition system for patients following laryngectomy. Med. Eng. Phys. 30, 419–425. doi:10.1016/j.medengphy.2007.05.003

Ferreira, A., 2014. Feature Selection and Discretization for High-Dimensional Data. PhD Thesis, Instituto Superior Técnico, Lisboa, Portugal.

Ferreira, A., Figueiredo, M., 2012. Efficient feature selection filters for high-dimensional data. Pattern Recognit. Lett. 33, 1794–1804. doi:10.1016/j.patrec.2012.05.019

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Hum. Genet. 7, 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Fitzpatrick, M., 2002. Lip-reading cellphone silences loudmouths. New Sci. Ed. 2002 3.

Florescu, V.M., Crevier-Buchman, L., Denby, B., Hueber, T., Colazo-Simon, A., Pillot-Loiseau, C., Roussel-Ragot, P., Gendrot, C., Quattrocchi, S., 2010. Silent vs vocalized articulation for a portable ultrasound-based silent speech interface., in: Proceedings of Interspeech 2010. pp. 450–453.

Fox, S., 2006. Are “wired Seniors” Sitting Ducks? Pew Internet & American Life Project.

Fraiwan, L., Lweesy, K., Al-Nemrawi, A., Addabass, S., Saifan, R., 2011. Voiceless Arabic vowels recognition using facial EMG. Med. Biol. Eng. Comput. 49, 811–818. doi:10.1007/s11517-011-0751-1

Francisco, A.A., Jesse, A., Groen, M.A., McQueen, J.M., 2014. Audiovisual temporal sensitivity in typical and dyslexic adult readers, in: Proceedings of Interspeech 2014.

Freitas, J., Calado, A., Barros, M.J., Dias, M.S., 2009. Spoken Language Interface for Mobile Devices, in: Human Language Technology: Challenges of the Information Society. pp. 24–35. doi:10.1007/978-3-642-04235-5_3

Freitas, J., Ferreira, A., Figueiredo, M., Teixeira, A., Dias, M.S., 2014a. Enhancing Multimodal Silent Speech Interfaces with Feature Selection, in: 15th Annual Conf. of the Int. Speech Communication Association (Interspeech 2014). Singapore, pp. 1169–1173.

Freitas, J., Teixeira, A., Dias, M.S., 2012a. Towards a Silent Speech Interface for Portuguese: Surface Electromyography and the nasality challenge, in: International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2012). pp. 91–100.

Freitas, J., Teixeira, A., Dias, M.S., 2013. Multimodal Silent Speech Interface based on Video, Depth, Surface Electromyography and Ultrasonic Doppler: Data Collection and First Recognition Results, in: Int. Workshop on Speech Production in Automatic Speech Recognition (SPASR 2013). Lyon.

Freitas, J., Teixeira, A., Dias, M.S., 2014b. Can Ultrasonic Doppler Help Detecting Nasality for Silent Speech Interfaces? - An Exploratory Analysis based on Alignement of the Doppler Signal with Velum Aperture Information from Real-Time MRI, in: International Conference on Physiological Computing Systems (PhyCS 2014). pp. 232 – 239.

Freitas, J., Teixeira, A., Dias, M.S., Bastos, C., 2011. Towards a Multimodal Silent Speech Interface for European Portuguese, in: Speech Technologies. InTech, Ivo Ipsic (Ed.), pp. 125–149. doi:10.5772/16935

Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014c. Velum Movement Detection based on Surface Electromyography for Speech Interface, in: International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2014). pp. 13–20.

Freitas, J., Teixeira, A., Silva, S., Oliveira, C., Dias, M.S., 2014d. Assessing the Applicability of Surface EMG to Tongue Gesture Detection, in: Proceedings of IberSPEECH 2014, Lecture Notes in Artificial Intelligence (LNAI). Springer, pp. 189–198.

Freitas, J., Teixeira, A., Vaz, F., Dias, M.S., 2012b. Automatic Speech Recognition Based on Ultrasonic Doppler Sensing for European Portuguese, in: Advances in Speech and Language Technologies for Iberian Languages, Communications in Computer and Information Science. Springer Berlin Heidelberg, pp. 227–236. doi:10.1007/978-3-642-35292-8_24

Fritzell, B., 1969. The velopharyngeal muscles in speech: An electromyographic and cineradiographic study. Acta Otolaryngolica 50.

Galatas, G., Potamianos, G., Makedon, F., 2012a. Audio-Visual Speech Recognition Incorporating Facial Depth Information Captured by the Kinect, in: 20th European Signal Processing Conference. pp. 2714–2717.

Galatas, G., Potamianos, G., Makedon, F., 2012b. Audio-Visual Speech Recognition Using Depth Information From The Kinect in Noisy Video Condition, in: Proceedings of the 5th International Conference on PErvasive Technologies Related to Assistive Environments - PETRA '12. pp. 1–4. doi:10.1145/2413097.2413100

Gan, T., Menzel, W., Yang, S., 2007. An Audio-Visual Speech Recognition Framework Based on Articulatory Features, in: Auditory-Visual Speech Processing. pp. 1–5.

Gerdle, B., Karlsson, S., Day, S., Djupsjöbacka, M., 1999. Acquisition, processing and analysis of the surface electromyogram, in: Modern Techniques in Neuroscience Research. Springer, pp. 705–755.

Ghosh, P.K., Narayanan, S., 2011. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. J. Acoust. Soc. Am. doi:10.1121/1.3634122

Gilbert, J.M., Rybchenko, S.I., Hofe, R., Ell, S.R., Fagan, M.J., Moore, R.K., Green, P., 2010. Isolated word recognition of silent speech using magnetic implants and sensors. Med. Eng. Phys. 32, 1189–1197. doi:10.1016/j.medengphy.2010.08.011

Gonzalez, J.A., Cheah, L.A., Bai, J., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D., 2014. Analysis of Phonetic Similarity in a Silent Speech Interface based on Permanent Magnetic Articulography, in: Proceedings of Interspeech 2014. pp. 1018–1022.

Gurban, M., Thiran, J.-P., 2009. Information Theoretic Feature Extraction for Audio-Visual Speech Recognition. IEEE Trans. Signal Process. 57, 4765–4776. doi:10.1109/TSP.2009.2026513

Gurbuz, S., Tufekci, Z., Patterson, E., Gowdy, J.N., 2001. Application of affine-invariant Fourier descriptors to lipreading for audio-visual speech recognition, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2001). IEEE, pp. 177–180.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182. doi:10.1162/153244303322753616

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L., 2006. Feature Extraction, Foundations and Applications, Soft Computing.

Hämäläinen, A., Pinto, F., Dias, M., Júdice, A., Freitas, J., Pires, C., Teixeira, V., Calado, A., Braga, D., 2012. The first European Portuguese elderly speech corpus, in: Proceedings of IberSPEECH. Madrid, Spain.

Hardcastle, W.J., 1976. Physiology of speech production: an introduction for speech scientists. Academic Press New York.

Hardcastle, W.J., Laver, J., Gibbon, F.E., 2012. The handbook of phonetic sciences. John Wiley & Sons.

Harris, C., Stephens, M., 1988. A combined corner and edge detector, in: Alvey Vision Conference. Manchester, UK, p. 50.

Hasegawa, T., Ohtani, K., 1992. Oral image to voice converter-image input microphone, in: Singapore ICCS/ISITA'92.'Communications on the Move'. IEEE, pp. 617–620.

Heaton, J.T., Robertson, M., Griffin, C., 2011. Development of a wireless electromyographically controlled electrolarynx voice prosthesis, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, (EMBC 2011). pp. 5352–5355. doi:10.1109/IEMBS.2011.6091324

Heistermann, T., Janke, M., Wand, M., Schultz, T., 2014. Spatial Artifact Detection for Multi-Channel EMG-Based Speech Recognition. Int. Conf. Bio-Inspired Syst. Signal Process. 189–196.

Helfrich, H., 1979. Age markers in speech, Social markers in speech. Cambridge University Press Cambridge, England.

Heracleous, P., Badin, P., Bailly, G., Hagita, N., 2011. A pilot study on augmented speech communication based on Electro-Magnetic Articulography. Pattern Recognit. Lett. 32, 1119–1125.

Heracleous, P., Hagita, N., 2010. Non-audible murmur recognition based on fusion of audio and visual streams., in: Proceedings of Interspeech 2010. pp. 2706–2709.

Heracleous, P., Nakajima, Y., Lee, A., Saruwatari, H., Shikano, K., 2003. Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation. IEEE Work. Autom. Speech Recognit. Underst. (ASRU 2003). doi:10.1109/ASRU.2003.1318406

Herff, C., Janke, M., Wand, M., Schultz, T., 2011. Impact of Different Feedback Mechanisms in EMG-based Speech Recognition 2213–2216.

Hermens, H.J., Freriks, B., Disselhorst-Klug, C., Rau, G., 2000. Development of recommendations for SEMG sensors and sensor placement procedures. J. Electromyogr. Kinesiol. 10, 361–374.

Hilberg, O., Jackson, A., Swift, D., Pedersen, O., 1989. Acoustic rhinometry: evaluation of nasal cavity geometry by acoustic reflection. J. Appl. Physiol. 66, 295–303.

Hofe, R., Bai, J., Cheah, L.A., Ell, S.R., Gilbert, J.M., Moore, R.K., Green, P.D., 2013a. Performance of the MVOCA silent speech interface across multiple speakers, in: Proc. of Interspeech 2013. pp. 1140–1143.

Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K., Rybchenko, S.I., 2010. Evaluation of a silent speech interface based on magnetic sensing., in: Proceedings of Interspeech 2010. pp. 246–249.

Hofe, R., Ell, S.R., Fagan, M.J., Gilbert, J.M., Green, P.D., Moore, R.K., Rybchenko, S.I., 2013b. Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing. Speech Commun. 55, 22–32. doi:10.1016/j.specom.2012.02.001

Holzrichter, J.F., Burnett, G.C., Ng, L.C., Lea, W.A., 1998. Speech articulator measurements using low power EM-wave sensors. J. Acoust. Soc. Am. doi:10.1121/1.421133

Holzrichter, J.F., Foundation, J.H., Davis, C., 2009. Characterizing Silent and Pseudo-Silent Speech using Radar-like Sensors, in: Interspeech 2009. pp. 656–659.

Horiguchi, S., Bell-Berti, F., 1987. The Velotrace: A device for monitoring velar position. Cleft Palate J. 24, 104–111.

Hu, R., Raj, B., 2005. A robust voice activity detector using an acoustic Doppler radar, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2005). IEEE, pp. 319–324.

Hudgins, B., Parker, P., Scott, R.N., 1993. A new strategy for multifunction myoelectric control. IEEE Trans. Biomed. Eng. 40, 82–94.

Hueber, T., Bailly, G., Denby, B., 2012. Continuous Articulatory-to-Acoustic Mapping using Phone-based Trajectory HMM for a Silent Speech Interface, in: Proceedings of Interspeech 2012. pp. 723–726.

Hueber, T., Benaroya, E.L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2010. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. 52, 288–300. doi:10.1016/j.specom.2009.11.004

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2009. Visuo-phonetic decoding using multi-stream and context-dependent models for an ultrasound-based silent speech interface., in: Proceedings of Interspeech 2009. pp. 640–643.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., Stone, M., 2008. An ultrasound-based silent speech interface. J. Acoust. Soc. Am. doi:10.1121/1.2936013

Hueber, T., Chollet, G., Denby, B., Stone, M., Zouari, L., 2007. Ouisper: Corpus Based Synthesis Driven by Articulatory Data, in: International Congress of Phonetic Sciences. Saarbrücken, pp. 2193–2196.

Human Brain Project [WWW Document], n.d. URL https://www.humanbrainproject.eu/ (accessed 8.8.14).

Hyvarinen, A., 1999. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Networks 10, 626–634.

Ishii, S., Toda, T., Saruwatari, H., Sakti, S., Nakamura, S., 2011. Blind noise suppression for Non-Audible Murmur recognition with stereo signal processing. IEEE Work. Autom. Speech Recognit. Underst. 494–499. doi:10.1109/ASRU.2011.6163981

Itoi, M., Miyazaki, R., Toda, T., Saruwatari, H., Shikano, K., 2012. Blind speech extraction for Non-Audible Murmur speech with speaker's movement noise, in: IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2012). IEEE, pp. 320–325.

Jawbone, n.d. Jawbone Headset [WWW Document]. URL https://jawbone.com (accessed 8.18.14).

Jennings, D.L., Ruck, D.W., 1995. Enhancing automatic speech recognition with an ultrasonic lip motion detector, in: IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 1995). IEEE, pp. 868–871.

Jessen, M., 2007. Speaker classification in forensic phonetics and acoustics, in: Müller, C. (Ed.), Speaker Classification I. Springer, pp. 180–204.

Johnson, A.E., Hebert, M., 1999. Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans. Pattern Anal. Mach. Intell. 21, 433–449.

Jorgensen, C., Binsted, K., 2005. Web Browser Control Using EMG Based Sub Vocal Speech Recognition. Proc. 38th Annu. Hawaii Int. Conf. Syst. Sci. 294c–294c. doi:10.1109/HICSS.2005.683

Jorgensen, C., Dusan, S., 2010. Speech interfaces based upon surface electromyography. Speech Commun. 52, 354–366. doi:10.1016/j.specom.2009.11.003

Jorgensen, C., Lee, D.D., Agabont, S., 2003. Sub auditory speech recognition based on EMG signals, in: Proceedings of the International Joint Conference on Neural Networks, 2003. IEEE, pp. 3128–3133.

Jou, S.-C., 2008. Automatic speech recognition on vibrocervigraphic and electromyographic signals. PhD Thesis, Language Technologies Institute and Carnegie Mellon University.

Jou, S.-C., Maier-Hein, L., Schultz, T., Waibel, A., 2006a. Articulatory feature classification using surface electromyography, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2006). IEEE, pp. I–I.

Jou, S.-C., Schultz, T., Waibel, A., 2004. Adaptation for soft whisper recognition using a throat microphone., in: Proceedings of Interspeech 2004.

Jou, S.-C., Schultz, T., Waibel, A., 2007. Continuous electromyographic speech recognition with a multi-stream decoding architecture, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007). IEEE, pp. IV–401.

Jou, S.-C., Schultz, T., Walliczek, M., Kraft, F., Waibel, A., 2006b. Towards continuous speech recognition using surface electromyography., in: Proceedings of Interspeech 2006.

Júdice, A., Freitas, J., Braga, D., Calado, A., Dias, M.S., Teixeira, A., Oliveira, C., 2010. Elderly Speech Collection for Speech Recognition Based on Crowd Sourcing, in: Int. Conf. on Software Development for Enhancing Accessibility and Fighting Info-Exclusion (DSAI 2010). SAE, Oxford, UK.

Junqua, J.-C., Fincke, S., Field, K., 1999. The Lombard effect: A reflex to better communicate with others in noise, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1999). IEEE, pp. 2083–2086.

Kalaiselvi, K., Vishnupriya, M.S., 2014. Non-Audible Murmur (NAM) Voice Conversion by Wavelet Transform. Int. J.

Kalgaonkar, K., Hu, R.H.R., Raj, B., 2007. Ultrasonic Doppler Sensor for Voice Activity Detection. IEEE Signal Process. Lett. 14, 754–757. doi:10.1109/LSP.2007.896450

Kalgaonkar, K., Raj, B., 2007. Acoustic Doppler sonar for gait recogination, in: IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS 2007). Ieee, pp. 27–32. doi:10.1109/AVSS.2007.4425281

Kalgaonkar, K., Raj, B., 2008. Ultrasonic Doppler sensor for speaker recognition, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2008). Ieee, pp. 4865–4868. doi:10.1109/ICASSP.2008.4518747

Kalgaonkar, K., Raj, B., 2009. One-handed gesture recognition using ultrasonic Doppler sonar, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2009). Ieee, pp. 1889–1892. doi:10.1109/ICASSP.2009.4959977

Ke, Y., Sukthankar, R., 2004. PCA-SIFT: A more distinctive representation for local image descriptors, in: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004). IEEE, pp. II–506.

Kent, R.D., 1972. Some considerations in the cinefluorographic analysis of tongue movements during speech. Phonetica 26, 16–32. doi:10.1159/000259387

Kirchhoff, K., Fink, G.A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Commun. 37, 303–319. doi:10.1016/S0167-6393(01)00020-6

Kitamura, T., Takemoto, H., Honda, K., Shimada, Y., Fujimoto, I., Syakudo, Y., Masaki, S., Kuroda, K., Oku-Uchi, N., Senda, M., 2005. Difference in vocal tract shape between upright and supine postures: Observations by an open-type MRI scanner. Acoust. Sci. Technol. 26, 465–468.

Kober, H., Möller, M., Nimsky, C., Vieth, J., Fahlbusch, R., Ganslandt, O., 2001. New approach to localize speech relevant brain areas and hemispheric dominance using spatially filtered magnetoencephalography. Hum. Brain Mapp. 14, 236–250.

Kroos, C., 2012. Evaluation of the measurement precision in three-dimensional Electromagnetic Articulography (Carstens AG500). J. Phon. 40, 453–465.

Kuehn, D.P., Folkins, J.W., Cutting, C.B., 1982. Relationships between muscle activity and velar position. Cleft Palate J 19, 25–35.

Kuehn, D.P., Folkins, J.W., Linville, R.N., 1988. An electromyographic study of the musculus uvulae. Cleft Palate J 25, 348–355.

Lacerda, A., Head, B., 1966. Análise de sons nasais e sons nasalizados do português. Rev. do Laboratório Fonética Exp. Coimbra 6, 5–70.

Lahr, R.J., 2006. Head-worn, trimodal device to increase transcription accuracy in a voice recognition system and to process unvocalized speech. US 7082393 B2.

Lee, J.A., Verleysen, M., 2007. Nonlinear dimensionality reduction. Springer.

Levelt, W.J.M., 1995. The ability to speak: from intentions to spoken words. Eur. Rev. doi:10.1017/S1062798700001290

Linville, S.E., 2001. Vocal aging. Singular Thomson Learning, San Diego, CA;

Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., Dawson-Haggerty, S., Woods, B., Frankel, J., Magami-Doss, M., Saenko, K., 2007. Articulatory Feature-Based Methods for Acoustic and Audio-Visual Speech Recognition: Summary from the 2006 JHU Summer workshop, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2007). doi:10.1109/ICASSP.2007.366989

Livescu, K., Zhu, B., Glass, J., 2009. On the phonetic information in ultrasonic microphone signals, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2009). IEEE, pp. 4621–4624.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. J. Neural Eng. 4.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110.

Lubker, J.F., 1968. An electromyographic-cinefluorographic investigation of velar function during normal speech production. Cleft Palate J 5, 1–18.

Magen, H.S., 1997. The extent of vowel-to-vowel coarticulation in English. J. Phon. 25, 187–205.

Maier-Hein, L., Metze, F., Schultz, T., Waibel, A., 2005. Session independent non-audible speech recognition using surface electromyography, in: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2005). pp. 331–336.

Manabe, H., 2003. Unvoiced Speech Recognition using EMG - Mime Speech Recognition -, in: CHI’03 Extended Abstracts on Human Factors in Computing Systems. ACM, pp. 794–795. doi:10.1145/765891.765996

Manabe, H., Zhang, Z., 2004. Multi-stream HMM for EMG-based speech recognition., in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4389–4392. doi:10.1109/IEMBS.2004.1404221

Martins, P., 2014. Ressonância Magnética em Estudos de Produção de Fala. PhD Thesis, University of Aveiro.

Martins, P., Carbone, I., Pinto, A., Silva, A., Teixeira, A., 2008. European Portuguese MRI based speech production studies. Speech Commun. 50, 925–952. doi:10.1016/j.specom.2008.05.019

MATLAB, Statistics Toolbox, n.d. . Release 2013a, The MathWorks, Inc., Natick, Massachusetts, United States.

McGill, S., Juker, D., Kropf, P., 1996. Appropriately placed surface EMG electrodes reflect deep muscle activity (psoas, quadratus lumborum, abdominal wall) in the lumbar spine. J. Biomech. 29, 1503–1507.

McGurk, H., MacDonald, J., 1976. Hearing lips and seeing voices. Nature 264, 746–748.

McLoughlin, I.V., 2014. The Use of Low-Frequency Ultrasound for Voice Activity, in: Proceedings of Interspeech 2014. pp. 1553–1557.

Meltzner, G.S., Colby, G., Deng, Y., Heaton, J.T., 2010. Signal acquisition and processing techniques for sEMG based silent speech recognition, in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 4848–4851.

Meltzner, G.S., Sroka, J., Heaton, J.T., Gilmore, L.D., Colby, G., Roy, S., Chen, N., Luca, C.J. De, 2008. Speech Recognition for Vocalized and Subvocal Modes of Production using Surface EMG Signals from the Neck and Face, in: Proceedings of Interspeech 2008.

Moller, K., Martin, R., Christiansen, R., 1971. A technique for recording velar movement. Cleft Palate J. 8, 263–276.

Moon, J.B., Canady, J.W., 1995. Effects of gravity on velopharyngeal muscle activity during speech. Cleft palate-craniofacial J. 32, 371–375.

Morse, M.S., Gopalan, Y.N., Wright, M., 1991. Speech recognition using myoelectric signals with neural networks, in: Proceedings of the Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 1877–1878.

Morse, M.S., O'Brien, E.M., 1986. Research summary of a scheme to ascertain the availability of speech information in the myoelectric signals of neck and head muscles using surface electrodes. Comput. Biol. Med. 16, 399–410.

Müller, M., 2007. Dynamic time warping. Inf. Retr. Music motion 69–84.

Nakajima, Y., 2005. Development and evaluation of soft silicone NAM microphone. Technical Report IEICE, SP2005-7.

Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003a. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP 2003) 5. doi:10.1109/ICASSP.2003.1200069

Nakajima, Y., Kashioka, H., Shikano, K., Campbell, N., 2003b. Non-audible murmur recognition. Eurospeech 2601–2604.

Nakamura, H., 1988. Method of recognizing speech using a lip image. 4,769,845.

Narayanan, S., Bresch, E., Ghosh, P.K., Goldstein, L., Katsamanis, A., Kim, Y., Lammert, A.C., Proctor, M.I., Ramanarayanan, V., Zhu, Y., 2011. A Multimodal Real-Time MRI Articulatory Corpus for Speech Research., in: Proceedings of Interspeech 2011. pp. 837–840.

Nijholt, A., Tan, D., 2008. Brain-computer interfacing for intelligent systems. Intell. Syst. IEEE 23, 72–79.

Noda, K., Yamaguchi, Y., Nakadai, K., Okuno, H.G., Ogata, T., 2014. Lipreading using Convolutional Neural Network, in: Proceedings of Interspeech 2014.

Oliveira, C., Albuquerque, L., Hämäläinen, A., Pinto, F.M., Dias, M.S., Júdice, A., Freitas, J., Pires, C., Teixeira, V., Calado, A., Braga, D., Teixeira, A., 2013. Tecnologias de Fala para Pessoas Idosas, in: Laboratório Vivo de Usabilidade (Living Usability Lab). ARC Publishing, pp. 167–181.

Oppenheim, A. V, Schafer, R.W., Buck, J.R., 1999. Discrete Time Signal Processing, Book. Prentice-Hall.

Otani, M., Shimizu, S., Hirahara, T., 2008. Vocal tract shapes of non-audible murmur production. Acoust. Sci. Technol. 29, 195–198.

Papcun, G., Hochberg, J., Thomas, T.R., Laroche, F., Zacks, J., Levy, S., 1992. Inferring articulation and recognizing gestures from acoustics with a neural network trained on x-ray microbeam data. J. Acoust. Soc. Am. 92, 688–700.

Patil, S.A., Hansen, J.H.L., 2010. The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification. Speech Commun. 52, 327–340. doi:10.1016/j.specom.2009.11.006

Pellegrini, T., Hämäläinen, A., de Mareüil, P.B., Tjalve, M., Trancoso, I., Candeias, S., Dias, M.S., Braga, D., 2013. A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance., in: Proceedings of Interspeech 2013. pp. 852–856.

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M.S., Braga, D., 2012. Impact of age in ASR for the elderly: preliminary experiments in European Portuguese, in: Advances in Speech and Language Technologies for Iberian Languages. Springer, pp. 139–147.

Pera, V., Moura, A., Freitas, D., 2004. Lpfav2: a new multi-modal database for developing speech recognition systems for an assistive technology application, in: 9th Conference Speech and Computer (SPECOM 2004).

Pereda, E., Quiroga, R.Q., Bhattacharya, J., 2005. Nonlinear multivariate analysis of neurophysiological signals. Prog. Neurobiol. 77, 1–37.

Perkell, J.S., Cohen, M.H., Svirsky, M.A., Matthies, M.L., Garabieta, I., Jackson, M.T.T., 1992. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J. Acoust. Soc. Am. 92, 3078–3096.

Perry, J.L., 2011. Variations in velopharyngeal structures between upright and supine positions using upright magnetic resonance imaging. Cleft Palate-Craniofacial J. 48, 123–133.

Petajan, E., 1984. Automatic lipreading to enhance speech recognition. University of Illinois.

Phang, C.W., Sutanto, J., Kankanhalli, A., Li, Y., Tan, B.C.Y., Teo, H.-H., 2006. Senior citizens' acceptance of information systems: A study in the context of e-government services. IEEE Trans. Eng. Manag. 53, 555–569.

Plux Wireless Biosignals, n.d. . URL http//www.plux.info/ (accessed 10-30-2014).

Porbadnigk, A., Wester, M., Calliess, J., Schultz, T., 2009. EEG-based speech recognition impact of temporal effects, in: Int. Conf. on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2009). doi:10.1.1.157.8486

Potamianos, G., Neti, C., Gravier, G., Garg, A., Senior, A.W., 2003. Recent advances in the automatic recognition of audiovisual speech. Proc. IEEE 91, 1306–1326.

Qiao, Z., Zhou, L., Huang, J.., 2009. Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data. Int. J. Appl. Math. 39, 48–60.

QREN 13852 AAL4ALL, n.d. . R&D Proj. URL http//www.aal4all.org/, last accessed 14th Oct. 2014.

QREN 7900 LUL - Living Usability Lab, n.d. . R&D Proj. URL http//www.microsoft.com/pt-pt/mldc/lul/en/default.aspx last acessed 14th Oct. 2014.

Quatieri, T.F., Brady, K., Messing, D., Campbell, J.P., Campbell, W.M., Brandstein, M.S., Weinstein, C.J., Tardelli, J.D., Gatewood, P.D., 2006. Exploiting nonacoustic sensors for speech encoding. IEEE Trans. Audio. Speech. Lang. Processing 14. doi:10.1109/TSA.2005.855838

R Development Core Team, R., 2011. R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. doi:10.1007/978-3-540-74686-7

Rabiner, L.R., Juang, B.-H., 1993. Fundamentals of speech recognition. PTR Prentice Hall Englewood Cliffs.

Rabiner, L.R., Rosenberg, A.E., Levinson, S.E., 1978. Considerations in dynamic time warping algorithms for discrete word recognition. J. Acoust. Soc. Am. 63, S79–S79.

Raj, B., Kalgaonkar, K., Harrison, C., Dietz, P., 2012. Ultrasonic Doppler Sensing in HCI. IEEE Pervasive Comput. 11, 24–29. doi:10.1109/MPRV.2012.17

Rossato, S., Teixeira, A., Ferreira, L., 2006. Les nasales du portugais et du français: une étude comparative sur les données EMMA. XXVI Journées d'Études la Parol.

Sá, F., Afonso, P., Ferreira, R., Pêra, V., 2003. Reconhecimento Automático de Fala Contínua em Português Europeu Recorrendo a Streams Audio-Visuais, in: The Proceedings of COOPMEDIA.

Saenko, K., Darrell, T., Glass, J., 2004. Articulatory Features for Robust Visual Speech Recognition, in: 6th International Conference on Multimodal Interfaces. pp. 1–7.

Saenko, K., Livescu, K., Glass, J., Darrell, T., 2009. Multistream articulatory feature-based models for visual speech recognition. IEEE Trans. Pattern Anal. Mach. Intell. 31, 1700–1707. doi:10.1109/TPAMI.2008.303

Sampson, R., 1999. Nasal Vowel Evolution in Romance. Oxford University Press, Oxford.

Schultz, T., 2007. Speaker characteristics, in: Speaker Classification I. Springer, pp. 47–74.

Schultz, T., Wand, M., 2010. Modeling coarticulation in EMG-based continuous speech recognition. Speech Commun. 52, 341–353. doi:10.1016/j.specom.2009.12.002

Schwartz, M.F., 1968. The acoustics of normal and nasal vowel production. Cleft palate J 5, 125–140.

Scobbie, J.M., Wrench, A.A., van der Linden, M., 2008. Head-Probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement., in: Proceedings of the 8th International Seminar on Speech Production. pp. 373–376.

Seaver, E.J., Karnell, M.P., Gasparaitis, A., Corey, J., 1995. Acoustic rhinometric measurements of changes in velar positioning. Cleft palate-craniofacial J. 32, 49–54.

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

Shaikh, A.A., Kumar, D.K., Yau, W.C., Che Azemin, M.Z., Gubbi, J., 2010. Lip reading using optical flow and support vector machines, in: 3rd International Congress on Image and Signal Processing (CISP 2010). IEEE, pp. 327–330.

Shankweiler, D., Harris, K.S., Taylor, M.L., 1968. Electromyographic studies of articulation in aphasia. Arch. Phys. Med. Rehabil. 49, 1–8.

Shi, J., Tomasi, C., 1994. Good features to track, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1994). IEEE, pp. 593–600.

Shin, J., Lee, J., Kim, D., 2011. Real-time lip reading system for isolated Korean word recognition. Pattern Recognit. 44, 559–571.

Silva, S., Martins, P., Oliveira, C., Silva, A., Teixeira, A., 2012. Segmentation and Analysis of the Oral and Nasal Cavities from MR Time Sequences, Image Analysis and Recognition, in: Proc. ICIAR, LNCS, 7325, Springer. pp.214-221.

Silva, S., Teixeira, A., 2014. Automatic Annotation of an Ultrasound Corpus for Studying Tongue Movement, in: Proc. ICIAR, LNCS 8814. Springer, Vilamoura, Portugal, pp. 469–476.

Silva, S., Teixeira, A., Oliveira, C., Martins, P., 2013. Segmentation and Analysis of Vocal Tract from MidSagittal Real-Time MRI, in: Image Analysis and Recognition. Springer, pp. 459–466.

Srinivasan, S., Raj, B., Ezzat, T., 2010. Ultrasonic sensing for robust speech recognition, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010). doi:10.1109/ICASSP.2010.5495039

Stephanidis, C., Akoumianakis, D., Sfyrakis, M., Paramythis, A., 1998. Universal accessibility in HCI: Process-oriented design guidelines and tool requirements, in: Proceedings of the 4th ERCIM Workshop on User Interfaces for All. Stockholm, pp. 19–21.

Stone, M., Lundberg, A., 1996. Three-dimensional tongue surface shapes of English consonants and vowels. J. Acoust. Soc. Am. 99, 3728–3737. doi:10.1121/1.414969

Stone, M., Stock, G., Bunin, K., Kumar, K., Epstein, M., Kambhamettu, C., Li, M., Parthasarathy, V., Prince, J., 2007. Comparison of speech production in upright and supine position. J. Acoust. Soc. Am. 122, 532–541.

Stork, D.G., Hennecke, M.E., 1996. Speechreading by humans and machines: models, systems, and applications. Springer.

Stover, S.E., Haynes, W.O., 1989. Topic manipulation and cohesive adequacy in conversations of normal adults between the ages of 30 and 90. Clin. Linguist. Phon. 3, 137–149.

Strevens, P., 1954. Some observations on the phonetics and pronunciation of modern Portuguese. Rev. Laboratório Fonética Exp. 5–29.

Sugie, N., Tsunoda, K., 1985. A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production. IEEE Trans. Biomed. Eng. 32, 485–490.

Suppes, P., Lu, Z.-L., Han, B., 1997. Brain wave recognition of words. Proc. Natl. Acad. Sci. 94, 14965–14969.

Tao, F., Busso, C., 2014. Lipreading Approach for Isolated Digits Recognition under Whisper and Neutral Speech, in: Proceedings of Interspeech 2014.

Teixeira, A., 2000. Síntese Articulatória das Vogais Nasais do Português Europeu [Articulatory Synthesis of Nasal Vowels for European Portuguese]. PhD Thesis, Universidade de Aveiro.

Teixeira, A., Martinez, R., Silva, L.N., Jesus, L.M.T., Príncipe, J.C., Vaz, F.A.C., 2005. Simulation of human speech production applied to the study and synthesis of European Portuguese. EURASIP J. Appl. Signal Processing 1435–1448.

Teixeira, A., Martins, P., Oliveira, C., Ferreira, C., Silva, A., Shosted, R., 2012. Real-time MRI for Portuguese: Database, methods and applications, in: Lecture Notes in Computer Science (including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 306–317. doi:10.1007/978-3-642-28885-2_35

Teixeira, A., Vaz, F., 2000. Síntese Articulatória dos Sons Nasais do Português, in: Anais Do V Encontro Para O Processamento Computacional Da Língua Portuguesa Escrita E Falada (PROPOR). pp. 183–193.

Teixeira, V., Pires, C., Pinto, F., Freitas, J., Dias, M.S., Rodrigues, E.M., 2012. Towards elderly social integration using a multimodal human-computer interface, in: Workshop on AAL Latest Solutions, Trends and Applications (AAL 2012). pp. 3–13.

The UCLA Phonetics Laboratory, 2002. Dissection of the Speech Production Mechanism.

Tiede, M.K., Masaki, S., Vatikiotis-Bateson, E., 2000. Contrasts in speech articulation observed in sitting and supine conditions, in: Proceedings of the 5th Seminar on Speech Production, Kloster Seeon, Bavaria. pp. 25–28.

Toda, T., 2010. Voice conversion for enhancing various types of body-conducted speech detected with non-audible murmur microphone. J. Acoust. Soc. Am. 127, 1815. doi:10.1121/1.3384185

Toda, T., 2012. Statistical approaches to enhancement of body-conducted speech detected with non-audible murmur microphone, in: ICME International Conference on Complex Medical Engineering (CME 2012). IEEE, pp. 623–628.

Toda, T., Nakamura, K., Nagai, T., Kaino, T., Nakajima, Y., Shikano, K., 2009. Technologies for processing body-conducted speech detected with non-audible murmur microphone, in: Proceedings of Interspeech 2009.

Tombari, F., Salti, S., Di Stefano, L., 2010. Unique signatures of histograms for local surface description, in: Computer Vision–ECCV 2010. Springer, pp. 356–369.

Toth, A.R., Kalgaonkar, K., Raj, B., Ezzat, T., 2010. Synthesizing speech from Doppler signals., in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2010). pp. 4638–4641.

Tran, T., Mariooryad, S., Busso, C., 2013. Audiovisual corpus to analyze whisper speech, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013). pp. 8101–8105. doi:10.1109/ICASSP.2013.6639243

Tran, V.A., Bailly, G., Loevenbruck, H., Toda, T., 2010. Improvement to a NAM-captured whisper-to-speech system. Speech Commun. 52, 314–326. doi:10.1016/j.specom.2009.11.005

Tran, V.-A., Bailly, G., Lœvenbruck, H., Toda, T., 2009. Multimodal HMM-based NAM-to-speech conversion, in: Interspeech 2009. pp. 656–659.

Trigo, L., 1993. The inherent structure of nasal segments. Nasals, nasalization, and the velum 5, 369–400.

Tronnier, M., 1998. Nasals and Nasalisation in Speech Production with Special Emphasis on Methodology and Osaka Japanese. Lund University.

Vipperla, R., Wolters, M., Georgila, K., Renals, S., 2009. Speech input from older users in smart environments: Challenges and perspectives, in: Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments. Springer, pp. 117–126.

Wand, M., Himmelsbach, A., Heistermann, T., Janke, M., Schultz, T., 2013a. Artifact removal algorithm for an EMG-based Silent Speech Interface., in: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. pp. 5750–3. doi:10.1109/EMBC.2013.6610857

Wand, M., Janke, M., Schultz, T., 2011. Investigations on Speaking Mode Discrepancies in EMG-Based Speech Recognition., in: Interspeech 2011. pp. 601–604.

Wand, M., Janke, M., Schultz, T., 2012. Decision-tree based Analysis of Speaking Mode Discrepancies in EMG-based Speech Recognition., in: International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2012). pp. 101–109.

Wand, M., Janke, M., Schultz, T., 2014. The EMG-UKA Corpus for Electromyographic Speech Processing, in: Proceedings of Interspeech 2014.

Wand, M., Schulte, C., Janke, M., Schultz, T., 2013b. Array-based Electromyographic Silent Speech Interface, in: International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2013).

Wand, M., Schultz, T., 2011a. Session-independent EMG-based Speech Recognition., in: International Conference on Bio-Inspired Systems and Signal Processing (BIOSIGNALS 2011). pp. 295–300.

Wand, M., Schultz, T., 2011b. Analysis of phone confusion in EMG-based speech recognition, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 2011). pp. 757–760. doi:10.1109/ICASSP.2011.5946514

Wand, M., Schultz, T., 2014. Towards Real-life Application of EMG-based Speech Recognition by using Unsupervised Adaptation, in: Proceedings of Interspeech 2014. pp. 1189–1193.

Wang, J., Balasubramanian, A., Mojica de la Vega, L., Green, J.R., Samal, A., Prabhakaran, B., 2013. Word Recognition from Continuous Articulatory Movement Time-Series Data using Symbolic Representations, in: ACL/ISCA Workshop on Speech and Language Processing for Assistive Technologies. Grenoble, France, pp. 119–127.

Wang, J., Samal, A., Green, J.R., 2014. Across-speaker Articulatory Normalization for Speaker-independent Silent Speech Recognition Contribution of Tongue Lateral to Consonant Production, in: Proceedings of Interspeech 2014. pp. 1179–1183.

Wang, J., Samal, A., Green, J.R., Rudzicz, F., 2012a. Sentence recognition from articulatory movements for silent speech interfaces, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012). IEEE, pp. 4985–4988.

Wang, J., Samal, A., Green, J.R., Rudzicz, F., 2012b. Whole-Word Recognition from Articulatory Movements for Silent Speech Interfaces, in: Proceedings of Interspeech 2012.

Warren, D., DuBois, A., 1964. A pressure-flow technique for measuring velopharyngeal orifice area during continuous speech. Cleft Palate J 1, 52–71.

Watterson, T., Lewis, K., Brancamp, T., 2005. Comparison of nasalance scores obtained with the Nasometer 6200 and the Nasometer II 6400. Cleft palate-craniofacial J. 42, 574–579.

Weiss, M.S., Yeni-Komshian, G.H., Heinz, J.M., 1979. Acoustical and perceptual characteristics of speech produced with an electronic artificial larynx. J. Acoust. Soc. Am. 65, 1298–1308.

Wells, J., Barry, W., Grice, M., Fourcin, A., Gibbon, D., 1992. Standard computer-compatible transcription, Esprit project 2589 (SAM), Doc. no. SAM-UCL-037. London.

Wester, M., Schultz, T., 2006. Unspoken speech – speech recognition based on electroencephalography. Universitat Karlsruhe (TH), Karlsruhe, Germany.

Wilpon, J.G., Jacobsen, C.N., 1996. A study of speech recognition for children and the elderly, in: IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP 1996). IEEE, pp. 349–352.

Wrench, A.A., Cleland, J., Scobbie, J.M., 2011. An ultrasound protocol for comparing tongue contours: upright vs. supine, in: Proceedings of 17th ICPhS, Hong Kong. pp. 2161–2164.

Xue, S.A., Hao, G.J., 2003. Changes in the Human Vocal Tract Due to Aging and the Acoustic Correlates of Speech ProductionA Pilot Study. J. Speech, Lang. Hear. Res. 46, 689–701.

Yaling, L., Wenjuan, Y., Minghui, D., 2010. Feature extraction based on lsda for lipreading, in: International Conference on Multimedia Technology (ICMT), 2010. IEEE, pp. 1–4.

Yao, Y.Y., 2003. Information-theoretic measures for knowledge discovery and data mining, in: Entropy Measures, Maximum Entropy Principle and Emerging Applications. Springer, pp. 115–136.

Yau, W.C., Arjunan, S.P., Kumar, D.K., 2008. Classification of voiceless speech using facial muscle activity and vision based techniques. TENCON 2008 - 2008 IEEE Reg. 10 Conf. doi:10.1109/TENCON.2008.4766822

Yu, L., Liu, H., 2003. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. Int. Conf. Mach. Learn. 1–8. doi:citeulike-article-id:3398512

Zahner, M., Janke, M., Wand, M., Schultz, T., 2014. Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach, in: Proceedings of Interspeech 2014.

Zhao, G., Barnard, M., Pietikainen, M., 2009. Lipreading with local spatiotemporal descriptors. IEEE Trans. Multimedia, 11, 1254–1265.

Zhu, B., 2008. Multimodal speech recognition with ultrasonic sensors. Msc Thesis, Massachusetts Institute of Technology.

Zhu, B., Hazen, T.J., Glass, J.R., 2007. Multimodal Speech Recognition with Ultrasonic Sensors, in: Interspeech 2007. pp. 662–665.