SPRINGER BRIEFS IN ELECTRICAL AND COMPUTER ENGINEERING · SPEECH TECHNOLOGY

João Freitas
António Teixeira
Miguel Sales Dias
Samuel Silva

An Introduction to Silent Speech Interfaces

EXTRAS ONLINE

Springer

SpringerBriefs in Speech technology

More information about this series at http://www.springer.com/series/10059

João Freitas • António Teixeira
Miguel Sales Dias • Samuel Silva

An Introduction to Silent Speech Interfaces

Springer

João Freitas
DefinedCrowd Corporation
Lisboa, Portugal

Microsoft Portugal
Microsoft Language Development Center
Lisboa, Portugal

Miguel Sales Dias
Instituto Universitário de Lisboa
(ISCTE-IUL)
ISTAR-IUL
Lisboa, Portugal

Microsoft Portugal
Microsoft Language Development Center
Lisboa, Portugal

António Teixeira
University of Aveiro
Department of Electronics,
Telecommunications and
Informatics/IEETA
Aveiro, Portugal

Samuel Silva
University of Aveiro
Department of Electronics,
Telecommunications and
Informatics/IEETA
Aveiro, Portugal

ISSN 2191-8112 ISSN 2191-8120 (electronic)
SpringerBriefs in Speech technology
ISBN 978-3-319-40173-7 ISBN 978-3-319-40174-4 (eBook)
DOI 10.1007/978-3-319-40174-4

Library of Congress Control Number: 2016943137

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Preface

Speech communication assumes a dominant role in how we communicate, and it is nowadays available to support interaction with machines in a wide range of scenarios, ranging from personal assistants for smartphones to home entertainment. While in many circumstances audible speech may suffice, there are a multitude of scenarios for which it is inadequate due to ambient noise, need for privacy, or as a result of existing speech impairments. Nevertheless, the audible speech signal, although it is the more significant outcome, is the end result of the human speech production process that includes multiple stages (e.g., message conception in the brain, pulmonary and muscular activity, and articulators' movement). As humans are able to interpret a set of non-audible features, such as lip movement, which become an integral part of the communication, researchers are starting to explore how the information from the different phases of the speech production process can be acquired and used in the development of enhanced speech input modalities for interaction of humans with machines. This area of research, commonly designated as silent speech interfaces (SSI), holds a potential solution for a more natural human–computer interaction (HCI) in the absence of audible speech and is theoretically able to address several issues inherent to automatic speech recognition (ASR) technology based on the acoustic signal, addressing a broader set of scenarios from speech recognition in noisy environments to communication with speech-impaired individuals. Commonly, these systems sense and collect data from key elements of the human speech production process—from glottal and articulators' activity, their neural pathways, or the brain itself—and create an alternative digital representation of speech.

While notable work exists, the literature on SSI still lacks an integrated view of key aspects, technologies, and findings that enable a systematic approach and a clear perception of what is at stake with SSI. In line with this current status, our goal was to address this gap collecting, in a single source, the grounds for further developments, hopefully motivating further interest of researchers for this multidisciplinary research field.

This book constitutes a broad and comprehensive overview of the existing technical approaches in the area of SSI. Each technique is described in the context of the human speech production process, allowing the reader to clearly understand the principles behind SSI in general and across different approaches. Additionally, the book explores the combined use of different data sources in order to tackle limitations of simpler SSI approaches and to address current challenges of this field. Another set of information deemed relevant concerns existing SSI applications, resources, and a simple tutorial on how to build an SSI.

All these contents aim to guide readers—even those not particularly acquainted with the different disciplines involved—from the application contexts in which SSI may play a key role to the more technical aspects of designing and developing such systems.

And our effort would not be completed without some thoughts on the future of SSI. After all, as important as understanding where we are, in this research field, is to set our minds to where we need to go from here. May readers find this book a source of knowledge and inspiration for the future of silent speech interfaces.

Lisboa, Portugal — João Freitas
Aveiro, Portugal — António Teixeira
Lisboa, Portugal — Miguel Sales Dias
Aveiro, Portugal — Samuel Silva
April 2016

Contents

Chapter 1
Introduction

Abstract The concept of silent speech, when applied to human–computer interaction (HCI), describes a system that allows for speech communication between humans and machines in the absence of an audible acoustic signal. This type of system can be used as an input HCI modality in high-background-noise environments such as in living rooms, or in aiding speech-impaired individuals. The audible acoustic signal is, in fact, just the end result of the complex process of speech production, which starts at the brain, triggers relevant muscular activity, and results in movements of the articulators. It is this information that silent speech interfaces (SSIs) strive to harness and, in this context, understanding the different stages of speech production is of major importance.

In this chapter, the reader finds a brief introduction into the historical context for the rising interest in silent speech, followed by an overview on the different stages involved in speech production. Along the way, we establish a correspondence between the natural speech production process and the technology, which will be further discussed in the following chapters, leading to the existing silent speech interface (SSI) systems. Additionally, we identify overall challenges in the development of SSI.

Keywords Silent speech interface (SSI) • Speech production • Speech-motor control • Articulators • SSI modalities • Challenges

1.1 Silent Speech

Speech is our most natural form of communication, fostering the flow of ideas from one human brain to another. Human communication is far more complex and richer than the resulting audible sound wave. Previous studies, notably the one that discovered the McGurk effect (McGurk and MacDonald 1976), showed that humans employ both hearing and visual senses in speech perception. Humans are in fact able to interpret related contextual information, such as lip motion, head and body movements, hand gesture, facial expressions, and specific characteristics of the speech signal such as prosody, which become an integral part of the communication process. While some of these information cues might provide some redundancy or be less important to communication, others can be essential to fully comprehend the message. Taking a broader understanding of speech communication, we realize that

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_1

speech can still be present even if it does not entail the actual production of sound. This is the rationale that directs us to the concept of silent speech, that is to say, viable and effective speech communication in the absence of an audible acoustic signal.

Furthermore, speech production goes beyond these external characteristics, encompassing different stages, starting from the conceptual idea, which then originates brain signals, subsequently transformed into pulmonary and muscular activity and then into sound waves. Therefore, the speech audio signal is just the end result of a larger set of events that we strive to understand and use for communication and interaction purposes.

Gaining insight over the different stages of speech production is, therefore, a relevant and challenging goal we need to embrace to fully model different approaches for the development of silent speech systems. Such insights will enable the continuous development of spoken language technology, particularly serving a wide range of applications, in both human–human and human–machine communication, via silent speech.

Spoken language technology has significantly evolved in recent years. Several professional and entertainment sectors use automatic speech recognition (ASR) systems in their daily practice, notably, with command and control, dictation, and speech analytics (keyword spotting, transcription) features, used in various scenarios, like call centers, security, defense, healthcare, automotive, law, finance, telecom, office and mobile productivity, and even in personal life (e.g., personal assistants). State-of-the-art speech recognition systems achieve very satisfactory performance rates in controlled conditions and some companies such as Google (Novet 2015) claim word error rates below 8 % in mobility scenarios. Nonetheless, these systems still rely solely on the audio signal, which is potentially affected by environmental noise and, additionally, can present strong variability across gender or age, such as the one observed for elderly and children speech, leading to performance degradation. Since it depends on the audio signal, automatic speech recognition is also inadequate for users without the ability to produce an audible acoustic signal or in scenarios where non-disturbance or privacy is required.

While the scientific community is actively working on novel methods to enhance speech recognition, which might improve performance in some of the alluded scenarios, researchers have also started to focus on the silent aspect of speech communication. *Silent speech interface (SSI) systems allow human–computer interaction (HCI) through speech in the absence of an acoustic signal*, addressing a broader set of scenarios, from speech recognition in noisy environments to communication involving speech-impaired individuals. Commonly, these systems sense and collect data from the human speech production process—from glottal and articulators' activity, their neural pathways or the brain itself—and create an alternative digital representation of speech that 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 by exploring biometric signals other than voice, measured by sensing devices such as ultrasound, electromyography, vision, and depth or other types of sensors.

A broader view of SSI, including not only interaction with machines but also between humans, is also possible (e.g., Gonzalez et al. (2016)). In this view, not adopted in this book, biometric signals, measured by some of the same devices used in HCI, are used to, for example, drive a speech synthesizer or even create a synthetic speech output by direct mapping from the measured signals. This kind of application is particularly useful for restoring the voice to those submitted to very drastic surgical interventions causing voice loss, such as laryngectomy. For additional information on this subject, the reader is forwarded to (Denby and Stone 2004; Gonzalez et al. 2016; Guenther et al. 2009; Toth et al. 2010)

Movies and books have been, and continue to be, the source of many ideas for research and commercial products. This also happened in the area of SSI. 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, although the notion of lipreading by humans already existed for even longer time (Clegg 1953), it took more than a decade for real solutions to appear[1]. It was only in 1984 that the first automatic visual lipreading system by Petajan (1984) was proposed, as an enhancement to ASR in noisy environments. A few years later, Nakamura (1988) also registered several patents with the same purpose. In the mid-1980s, researchers like Sugie and Tsunoda (1985), and Morse and O'Brien (1986), almost simultaneously, started exploring the muscular activity in the face to recognize, at first, isolated vowels and then words (Morse et al. 1991).

In the 1990s, 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. Also, with the evolution of video cameras, more studies on lipreading started to appear. A relevant example is Hasegawa and Ohtani (1992) that achieved a 91 % recognition rate using a video of the speaker's face from which lip and tongue features were extracted.

In the early 2000s, DARPA (Defense Advanced Research Projects Agency) focused on recovering glottal excitation cues from voiced speech in noisy environments, with the Advanced Speech Encoding Program. By 2002, in Japan, an NTT DoCoMo (Japan mobile service provider) press release announced a silent cellphone prototype using electromyography and optical capture of lip movement (Fitzpatrick 2002), specially targeting scenarios encompassing environmental noise and speech-impaired users.

In recent years, the SSI concept became more prominent in the speech research community and diverse modalities (i.e., ways users can consider to interact with a system) were used to drive SSI research. Among the chosen modalities, we can find more invasive ones such as intra-cortical electrodes and non-obtrusive modalities such as video or ultrasonic Doppler sensing (UDS).

[1] For a more detailed historical context the reader is pointed to Denby et al. (2010).

While notable work exists in the field, the literature on SSI still lacks an integrated view of key aspects, technologies and findings that enable a systematic approach and a clear perception of what is at stake with SSI, motivating further interest of researchers for this multidisciplinary research field. To provide such grounds is the main goal of this book.

To better understand the apparent paradox of silent speech, one needs first to understand speech production, a complex mechanism, and, to that purpose, knowledge is required from diverse fields such as Anatomy, Physiology, Phonetics and Linguistics. As such, in what follows, we provide a simplified view of the stages involved in speech production, showing a glimpse of how rich the human speech process is, and, at the same time, giving the reader the base knowledge to understand the topics discussed in the following chapters.

1.2 A Speech Production Primer for SSI

A good model of speech production—at least for the objectives of this book—is the one proposed by Levelt (1995) that divides the process into several stages (Fig. 1.1): the first stage, of conceptualization and formulation, occurring in the brain, converts communication intentions into messages, followed by the grammatical encoding of the preverbal message to surface structure. The next stage of speech production is the conversion from this 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 two phases: the electrical impulse fed into the articulators (articulatory control phase) and the actual process of articulation. The final stage consists on the consequent effects of the previous phases resulting in the acoustic speech signal and other effects (e.g., alterations in the face).

Understanding each of these stages entails extensive background knowledge (Hardcastle 1976; Seikel et al. 2009). Thus, in what follows, we present a selective description of the speech process, considering the different stages depicted in Fig. 1.1, focusing on the relevant and necessary topics to understand the work presented in the following chapters. We start by briefly explaining how motor control occurs for speech production. Then, we present some information regarding the

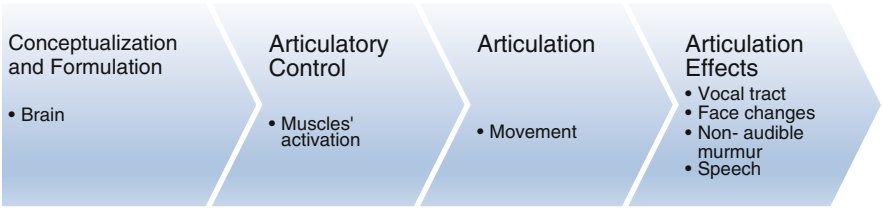

Fig. 1.1 Overview of the speech production process, following loosely the model proposed by Levelt (1995)

muscles involved in articulation control. Afterwards, we introduce the articulators, their position and their function. Providentially, essential and exhaustive descriptions can be found in the literature (Hardcastle 1976; Seikel et al. 2009; The UCLA Phonetics Laboratory 2002).

Speech production requires a particularly coordinated sequence of events to take place and is considered the most complex 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 event, our brain maps it into muscular 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 axons. The motor axons are then divided into several branches that end with a neuromuscular junction known as the motor endplate, meaning that a single motor neuron innervates several muscle fibers. The muscle fibers coalesce among several motor units.

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, e.g., when speaking, occur in the region of the face and neck. If these patterns are measured at the articulatory muscles, we are able to collect data potentially related to the resulting speech. This myoelectric activity occurs independently of the acoustic signal, i.e., it occurs if motor activity is present, whether the subject produces audible, murmured, or silent speech.

In the speech production process, the articulatory 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. A more detailed description of the muscles involved in articulation can be found in Chap. 2.

Articulation describes how humans produce speech sounds and which speech organs are involved in this process. Although all organs, surfaces, and cavities of the vocal tract are contributors to the production of speech, some, the articulators, are the most important and enable the production of the different sounds that languages use to convey information. The positioning of the articulators defines the articulatory and resonant characteristics of the vocal tract and they can be active or passive. Passive articulators (e.g., teeth, alveolar ridge on the upper jaw, hard palate (Seikel et al. 2009)) remain static during speech production. The active articulators move in

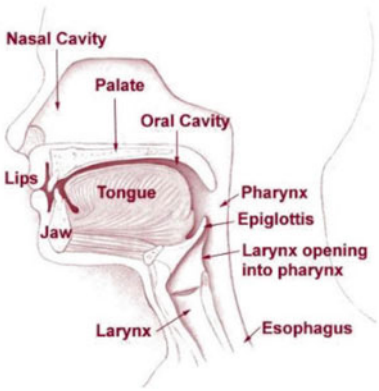

Fig. 1.2 Sagittal view of the vocal tract depicting its main regions (oral cavity, nasal cavity) and several articulators (e.g., tongue, lips, jaw). Obtained from https://upload.wikimedia.org/wikipedia/commons/d/d4/Illu01_head_neck.jpg

relation to passive articulators, through muscular action, to achieve different tract configurations and, in consequence, different sounds. Active articulators include the tongue, lower jaw, velum and lips, being the most important the tongue as it is involved in the production of the majority of sounds. A depiction of several articulators is presented in Fig. 1.2.

1.3 Current SSI Modalities: An Overview

The several stages of speech production, with their different measurable outputs, provide the means for the existence of a variety of modalities for transmission of information between the user and a machine, the main objective of a silent speech interface for HCI.

The modalities currently in use cover all stages of human speech production, as demonstrated by examples in Table 1.1. As shown in the table, SSI modalities go from the interpretation of signals from implants in the cerebral cortex to the measurement of visible effects in the face.

The organization presented above allows us to associate each type of modality found in the SSI literature to a stage of the human speech production process, providing a better understanding from where the information is extracted. The modalities mentioned are only examples.

After this very broad overview, showing the diversity of approaches, Tables 1.2, 1.3, 1.4, and 1.5 summarize the methods found in a systematic literature review performed by the authors. These tables, organized by the speech production stage, enable a comparison between the modalities based on their main advantages, limitations, invasiveness, and obtrusiveness. We define a modality as invasive, if it requires medical expertise or permanent attachment of sensors and, obtrusive, if it requires "wearing" or being equipped with a sensor in a non-ubiquitous way. The classification of a modality, as invasive or noninvasive, strongly depends on the perspective of the reader. For example, PMA may be considered a borderline case, since it requires the permanent attachment of sensors (typically using dental or chirurgical

Table 1.1 SSI modalities currently in use cover all stages of human speech production

Stage of speech production	Relevant examples
Conceptualization and formulation (Brain/central nervous 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 (SEMG) 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) (Florescu et al. 2010)
	• Video of the lips (Wand et al. 2016)
	• Measures of glottal area using low power radar (Holzrichter et al. 2009)
Articulation effects (vocal tract)	• Capturing the movements of a speaker's face through ultrasonic Doppler sensing (Freitas et al. 2012b; Srinivasan et al. 2010)
	• RGB and depth information (RBG-D) from 3D cameras regarding visible effects in the face (Freitas et al. 2014a)
	• Digital transformation of signals from a non-audible murmur (NAM) microphone (Nakajima et al. 2003a; Toda 2010)
	• Analysis of glottal activity effects using vibration sensors (Patil and Hansen 2010)

Table 1.2 Single SSI modalities related to the first stage of speech production (Conceptualization and formulation): overview of advantages, limitations, and whether or not they are considered to be invasive or 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 electroencephalographic sensors (Porbadnigk et al. 2009)	Recognizes unspoken speech; setup is far less complex when compared with other BCIs—brain–computer interfaces	Low recognition rates; reduced vocabularies; requires a learning process	No/Yes

glue), but, depending on the location (e.g., velum vs. lips), it may (not) require a high degree of medical expertise. Therefore, although we refer to PMA as an invasive modality, when compared with others that require a chirurgical intervention, it is acceptable to classify it as noninvasive.

Table 1.3 Single SSI modalities related to the second stage of speech production (Articulation Control): overview of advantages, limitations, and whether or not they are considered to be invasive or obtrusive

Modality	Main advantages	Limitations	Invasive/obtrusive
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 to 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

Table 1.4 Single SSI modalities related to the third stage of speech production (Articulation): overview of advantages, limitations, and whether or not they are considered to be invasive or obtrusive

Modality	Main advantages	Limitations	Invasive/obtrusive
Real-time characterization of the tongue using ultrasound (Hueber et al. 2008)	Achieved good results when combined with video	Probe stabilization; only shows the tongue (and surrounding structures)	No/Yes
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 (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 discomfort with magnetic beads in more hidden articulators such as the velum; complex setup	Yes/Yes
RGB-D information from 3D cameras of the visible articulators (Freitas et al. 2014a, b; Galatas et al. 2012)	Widely available; does not require glottal activity;	Only captures visible articulators (e.g., lips)	No/No
Analysis of glottal activity using electromagnetic wave sensors (Holzrichter et al. 2009; Quatieri et al. 2006)	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

Table 1.5 Single SSI modalities related to the fourth stage in speech production (Articulation Effects): overview on advantages, limitations, and whether or not they are considered to be invasive or obtrusive

Modality	Main advantages	Limitations	Invasive/obtrusive
Capture movements of a talker's face through ultrasonic Doppler 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
Digital transformation of signals from a non-audible murmur (NAM) 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

1.4 Best Systems

It is challenging to make a fair comparison between existing SSI systems for HCI because their results are dependent on the chosen modality, on the type of features extracted from the associated signal, and on the classification models selected to recognize the intended message, which are characteristics that strongly influence their performance. Furthermore, other factors need to be considered when comparing SSI results. In terms of accuracy rates, these present large variations, depending on the following factors:

- **Vocabulary and speech units' size**: The size of the vocabulary and the particular speech units considered greatly influence the performance of an SSI. At an early stage, SSI approaches tend to start by recognizing isolated words (usually digits) (Betts et al. 2006; Florescu et al. 2010; Freitas et al. 2012a; 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 2011a).
- **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 recent years (Denby et al. 2010; Wand and Schultz 2011b). The set of images in Fig. 1.3 demonstrates several SEMG setups. Similar sensor setups between speakers do not effortlessly guarantee the same outcomes.
- **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

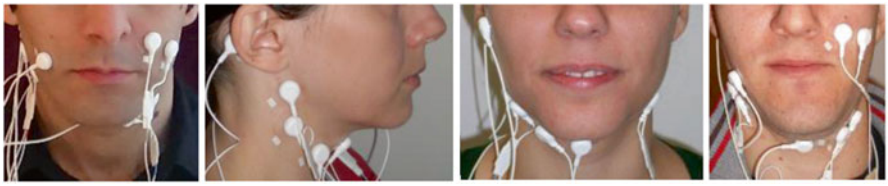

Fig. 1.3 SEMG setups for four different speakers, illustrating different possibilities of sensor positioning and highlighting the challenge of acquiring signals independent of speakers' anatomy

Hidden Markov Models (HMM), in order to obtain a more complete representation of each unit, many representations need to be collected. 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 advantage becomes particularly handy and of extreme importance for novel approaches using early prototypes, where many variables that need to be defined in data acquisition sessions are still unclear.

- **Acoustic feedback and user experience with SSI**: Word or phrase 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 evaluation, whether or not a user has prior 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 (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 pair of electrodes (Betts et al. 2006).
- **Language**: To the best of our knowledge, most of silent speech prototypes have been mainly developed and designed for English, with some exceptions for French (Tran et al. 2009), Japanese (Toda et al. 2009), Arabic (Fraiwan et al. 2011), and European Portuguese (Freitas et al. 2012b). However, language characteristics such as nasality, which is a prominent characteristic of European Portuguese, have been proven to influence performance (Freitas et al. 2012a).

1.5 Main Challenges in SSI

In the work presented by Denby and coworkers (Denby et al. 2010), several challenges to the development of SSI are enumerated, 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 to enable 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, SEMG, and EEG, have shown high sensitivity in sensor positioning, requiring previous training/adaptation or very accurate deployment.

- **Lombard and silent speech effects**: The effects resulting from silent speech articulation and the Lombard effect (different articulation when no auditory feedback is provided, as opposed to when the subject generates self-acknowledged audible sound), are not yet clear and require further investigation. This effect can be minimized depending on the user experience and proficiency with SSI use. However, relying on such prior experience would introduce a highly subjective requirement.
- **Prosody, emotions, and nasality**: The extraction of speech cues from prosody (intonation, tone, stress, and rhythm), that reflect the emotional state of the speaker, as well as irony or discourse focus is a major challenge in SSI, still unexplored in the research community. As for nasality, due to the modified or absent speech signal in SSI, the information needed to model this phenomenon must be obtained by other signal processing means.

Also very important and challenging is the language expansion of SSI technologies, that is, the extension of human languages with SSI support, one of the main drivers for the creation of this book.

1.6 Following Chapters

After the introduction to the area of SSI for HCI, in this first chapter, complemented with the presentation of information on speech production, the following chapters are organized into four parts:

- Chapters 2 and 3 provide an overview of the many different individual modalities that have been proposed in the area of SSI. In line with the discussed stages of speech production, these two chapters cover, on the one hand, the assessment of the brain and muscular activity that is on the genesis of speech production and, on the other, the consequences of such activity. In Chap. 2, readers will find recent work regarding the study of brain and muscular activities related to speech production, and the use of such knowledge to serve silent speech interfaces. Chapter 3 provides an overview of the technologies used to assess articulatory and visual aspects of speech production and how researchers have harnessed their capabilities for silent speech interfaces.
- Chapter 4 addresses how to combine modalities in order to tackle some of the challenges in SSI systems. Combination poses several challenges regarding the acquisition, synchronization, processing, and analysis of the required data. To answer these challenges, a framework to support research on multimodal SSI is presented.

- As this book aims at including real applications, the next part of the book, Chap. 5, is focused on providing application examples. It includes a tutorial on how to build a simple SSI system—that we challenge all readers to perform—and an example of a state-of-the-art multimodal SSI system. The tutorial aims at providing a first hands-on experience in the field with a minimum effort, by providing data and easy to run scripts.
- Finally, Chap. 6 presents some comments and considerations regarding the future of SSI.

References

Betts BJ, Binsted K, Jorgensen C (2006) Small vocabulary recognition using surface electromyography. J Human-Computer Interact. 18:1242–1259. http://dx.doi.org/10.1016/j.intcom.2006.08.012

Brumberg JS, Nieto-Castanon A, Kennedy PR, Guenther FH (2010) Brain-computer interfaces for speech communication. Speech Commun 52:367–379. doi:10.1016/j.specom.2010.01.001

Clegg DG (1953) The listening eye: a simple introduction to the art of lip-reading. Methuen, London

De Luca CJ (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

Denby B, Schultz T, Honda K, Hueber T, Gilbert JM, Brumberg JS (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

Fagan MJ, Ell SR, Gilbert JM, Sarrazin E, Chapman PM (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

Fitzpatrick M (2002) Lip-reading cellphone silences loudmouths. New Sci. Ed. 2002:3

Florescu VM, 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

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

Freitas J, Ferreira A, Figueiredo M, Teixeira A, Dias MS (2014a) Enhancing multimodal silent speech interfaces with feature selection. In: 15th Annual conference of the international speech communication association (Interspeech 2014). Singapore, pp 1169–1173

Freitas J, Teixeira A, Dias MS (2014b) Multimodal corpora for silent speech interaction. In: 9th Language resources and evaluation conference (LREC), pp 1–5

Freitas J, Teixeira A, Dias MS (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, Vaz F, Dias MS (2012a) Automatic speech recognition based on ultrasonic Doppler sensing for European Portuguese. In: Toledano DT, Ortega A, Teixeira A, Gonzalez-Rodriguez J, Hernandez-Gomez L, San-Segundo R, Ramos D (eds) Advances in speech and language technologies for Iberian languages, Communications in Computer and Information Science. Springer, Berlin, pp 227–236. doi:10.1007/978-3-642-35292-8_24

Galatas G, Potamianos G, Makedon F (2012) 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

Gonzalez JA, Cheah LA, Gilbert JM, Bai J, Ell SR, Green PD, Moore RK (2016) A silent speech system based on permanent magnet articulography and direct synthesis. Comput Speech Lang. doi:10.1016/j.csl.2016.02.002

Guenther FH, Brumberg JS, Joseph Wright E, Nieto-Castanon A, Tourville JA, Panko M, Law R, Siebert SA, Bartels JL, Andreasen DS, Ehirim P, Mao H, Kennedy PR (2009) A wireless brain-machine interface for real-time speech synthesis. PLoS One 4(12), e8218. doi:10.1371/journal.pone.0008218

Hardcastle WJ (1976) Physiology of speech production: an introduction for speech scientists. Academic, New York

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

Heistermann T, Janke M, Wand M, Schultz T (2014) Spatial artifact detection for multi-channel EMG-based speech recognition. In: International conference on bio-inspired systems and signal processing, pp 189–196

Hofe R, Bai J, Cheah LA, Ell SR, Gilbert JM, Moore RK, Green PD (2013a) Performance of the MVOCA silent speech interface across multiple speakers. In: Proc. of Interspeech 2013, pp 1140–1143

Hofe R, Ell SR, Fagan MJ, Gilbert JM, Green PD, Moore RK, Rybchenko SI (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 JF, Foundation JH, Davis C (2009) Characterizing Silent and Pseudo-Silent Speech using Radar-like Sensors. Interspeech 2009:656–659

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, Chollet G, Denby B, Dreyfus G, Stone M (2008) An ultrasound-based silent speech interface. J Acoust Soc Am. doi:10.1121/1.2936013

Jorgensen C, Dusan S (2010) Speech interfaces based upon surface electromyography. Speech Commun 52:354–366. doi:10.1016/j.specom.2009.11.003

Levelt WJM (1995) The ability to speak: from intentions to spoken words. Eur Rev. doi:10.1017/S1062798700001290

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

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

Morse MS, Gopalan YN, Wright M (1991) Speech recognition using myoelectric signals with neural networks, in: Proceedings of the Annual international conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp 1877–1878

Morse MS, O'Brien EM (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

Nakajima Y, Kashioka H, Shikano K, Campbell N (2003a) Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. IEEE International conference on acoustics, speech and signal processing (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. Patent No. 4769845
Novet J (2015) Google says its speech recognition technology now has only an 8% word error rate [WWW Document]. VentureBeat. http://venturebeat.com/2015/05/28/google-says-its-speech-recognition-technology-now-has-only-an-8-word-error-rate/ (accessed 1 January 2016)
Patil SA, Hansen JHL (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
Petajan E (1984) Automatic lipreading to enhance speech recognition. University of Illinois, Champaign
Porbadnigk A, Wester M, Calliess J, Schultz T (2009) EEG-based speech recognition impact of temporal effects. In: International conference on bio-inspired systems and signal processing (BIOSIGNALS 2009). doi:10.1.1.157.8486
Quatieri TF, Brady K, Messing D, Campbell JP, Campbell WM, Brandstein MS, Weinstein CJ, Tardelli JD, Gatewood PD (2006) Exploiting nonacoustic sensors for speech encoding. IEEE Trans. Audio. Speech. Lang. Processing 14. doi:10.1109/TSA.2005.855838.
Seikel JA, King DW, Drumright DG (2009) Anatomy and physiology for speech, language, and hearing, 4th edn. Delmar Learning, Clifton Park
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
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
The UCLA Phonetics Laboratory (2002) Dissection of the speech production mechanism
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, 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
Toth AR, 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 V-A, Bailly G, Lœvenbruck H, Toda T (2009) Multimodal HMM-based NAM-to-speech conversion. Interspeech 2009:656–659
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 5750–5753. 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, Koutník J, Schmidhuber J (2016) Lipreading with long short-term memory. arXiv Prepr. arXiv1601.08188.
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) 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 (2011b) Session-independent EMG-based Speech Recognition. In: International conference on bio-inspired systems and signal processing (BIOSIGNALS 2011). pp 295–300
Zhu B, Hazen TJ, Glass JR (2007) Multimodal Speech Recognition with Ultrasonic Sensors. Interspeech 2007:662–665

Chapter 2
SSI Modalities I: Behind the Scenes—From the Brain to the Muscles

Abstract Silent speech approaches can profit not only from an understanding of the brain and motor stages associated with speech production, but also from a direct use of the information coming from these stages. Therefore, in this chapter, readers can find a short overview of recent work regarding the study of brain and muscular activity related to speech production, and the use of such knowledge to serve silent speech interfaces (SSIs). In this context, the chapter takes in consideration the importance of understanding the sensorimotor cortex's role in speech production and the application of such knowledge in the context of brain–computer interfaces (BCI). Regarding muscular activity, the concept of myoelectric signal is introduced and the literature surveyed on the technologies used to measure it. For each of the mentioned speech production stages, recent accomplishments in the domain of silent speech interfaces are also covered.

Keywords Speech production • Silent speech interfaces • Brain activity • Myoelectric activity • Brain–computer interfaces • Electroencephalography • Electrocorticography • Magnetoencephalography • Muscular activity • Myoelectric signals • Surface electromyography

Human speech production starts in the brain, behind the scenes, in a remarkably complex hidden activity. Informally, one can say that it is also in the brain that the decision to move the articulators occurs. When this decision is taken, an electrical signal is sent to the muscles responsible for positioning the articulators, such that, when the air passes through the vocal tract, the correct sounds are uttered. This chapter links existing work in the area of silent speech interfaces (SSI) to the part of the natural process of speech production that occurs out of our sight.

The main goal of this chapter is not to provide an extensive literature review on the concepts and theories on the brain and muscular activities related to speech production processes. Instead, it provides an overall panorama of what the recent literature has to offer in these two topics, emphasizing the more notable accomplishments and the main technologies supporting them. While the focus and application areas for these technologies are vast, we restrict our coverage to those that are speech production related.

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_2

2.1 Brain Activity and Silent Speech Interfaces

With the recent evolution of cognitive and neurosciences, brain imaging and sensing technologies, we have grown our understanding and interpretation of the processes that happen in the brain and a new research area, referred to as brain–computer interface or BCI, has emerged. BCIs 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 (Akcakaya et al. 2014; Nijholt and Tan 2008). This evolution has motivated a constant advance on the methods to measure, understand, and harness brain activity (Brumberg et al. 2010; Nijholt and Tan 2008). A BCI can be based on several types of changes that occur during mental activity, such as electrical potentials (Chakrabarti et al. 2015; Conant et al. 2014), magnetic fields (Munding et al. 2015), or metabolic/hemodynamic recordings (Sorger et al. 2012). One such mental activity is the generation of "unspoken speech," which refers to the process where the subject imagines speaking a given word without moving any articulatory muscle or producing any sound. An SSI based on unspoken speech is particularly suited for subjects with physical disabilities such as the locked-in syndrome.

2.1.1 Mapping Speech-Related Brain Activity

A very important first step towards the application of BCIs for SSI is the exploration and understanding of the importance and function of multiple brain regions in the context of different stages of speech production. In many of the works and reviews in the literature, speech mapping is reported both for production and perception (Chakrabarti et al. 2015; Price 2012), with some authors arguing in favor of their intertwined nature (Pickering and Garrod 2013). Although understanding the mechanisms of speech perception (e.g., Chang et al. 2010; Mesgarani et al. 2014), can shed some light over the neural representation of phones and phonetic features, which may be relevant for silent speech interfaces, in the context of this book and for the sake of simplicity, we mainly emphasize speech production mapping. Furthermore, it is not our purpose to present a discussion of theoretical models of speech production neuroanatomy and the interested reader is forwarded to (Hickok 2012; Indefrey 2011; Price 2012).

Speech production involves a wide variety of brain structures (Andersen et al. 2014) and multiple technologies have been used to explore which regions on the sensorimotor cortex are important for speech production, covering from an overall assessment of the regions activated during speech, up to a more precise spatial and temporal description of such activity in relation to different articulators and phonetic settings.

Functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) play an important role in studying the anatomy of language. Price (Price 2012) reviews the literature concerning covert (i.e., silent/imagined) speech and overt (i.e., plain) speech considering the brain areas that control the lips, tongue,

jaw, larynx, and breathing. Price emphasizes the similarity among longitudinal studies' outcomes throughout the years, an important evidence of a consistent functional anatomy of speech across individuals. One of the main challenges identified is to understand how the identified regions interact with each other to produce speech. Although valuable in many aspects, these sensing technologies do not provide enough temporal resolution to allow sub-word study (Piai 2015), needed for a more detailed assessment of the underlying brain activity organization.

Conant et al. (2014) also discuss the current state-of-the-art of speech mapping in the human sensorimotor cortex. In a recent notable example, Bouchard et al. (2013) present a thorough study on the functional organization of the sensorimotor cortex for speech production. Considering multiple articulators (lips, jaw, tongue, and larynx) and through intracranial cortical recordings (electrocorticography, or ECoG), the authors analyze the spatial and temporal representations of cortical activity during the production of several consonant–vowel syllables and discuss the phonetic organization of spatial patterns. In this work, the temporal resolution of ECoG was crucial in disambiguating between the activity associated with consonants and vowels. Conant et al. (2014) highlight that, despite the knowledge already gathered concerning speech maps in the sensorimotor cortex, several aspects still need further exploration, namely regarding coordination and the influence of phonetic context in the activation pattern for particular phonemes. In this regard, the consideration of technologies supporting better spatial and temporal resolution to record cortical activity, combined with more detailed monitoring of speech articulators will enable further advances. On a detailed analysis of the use of ECoG for speech mapping (Chakrabarti et al. 2015) emphasizes that this technology is the one currently providing the best spatiotemporal resolution to allow the study of the dynamic nature of neural activity associated to speech production.

Munding et al. (2015) present an overview of the most notable works exploring speech mapping recurring to magnetoencephalography (MEG). Considering the characteristics of MEG, when it comes to speech mapping, this technology may be considered in-between electroencephalography (EEG) and ECoG recordings. When compared to EEG, it provides a better spatial resolution, although it does not reach the spatial and temporal resolutions provided by ECoG. Nevertheless, the latter requires a more complex setting due to its invasiveness and often provides only sparse anatomical coverage (Munding et al. 2015). Based on these characteristics of MEG which, as EEG, is able to cover the whole head, researchers have managed to perform a broad mapping of speech-related areas.

2.1.2 Measuring Brain Activity

In line with the technologies considered for speech mapping, several technologies have been developed to support BCI for SSI. These technologies provide different levels of spatial and temporal resolutions and require settings of different complexity and invasiveness, which eventually limit the application scenarios.

Current SSI approaches have been based on electrical potentials, more exactly on the sum of the postsynaptic potentials in the cortex. Two types of BCIs have been used for unspoken speech recognition: one noninvasive approach based on the interpretation of signals from EEG sensors and another, an invasive approach based on the interpretation of signals from intra-cortical microelectrodes in the speech-motor cortex (one exception can be found in Kellis et al. (2010), who explicitly consider the face-motor cortex). Approaches that use EEG and ECoG associated to other regions in the motor cortex, not directly related to speech production and resulting in a text/speech output, are not considered in this context (e.g., Guenther and Brumberg 2011; Oken et al. 2014).

Unspoken speech recognition tasks have also been tried based on magnetoencephalograms (MEG), by measuring 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).

2.1.3 Electroencephalographic Sensors

In this approach, EEG sensors are externally attached to the scalp, as depicted in Fig. 2.1. These sensors capture the potential in the respective area, which during brain activity can go up to 75 μV and during an epileptic seizure can reach 1 mV (Calliess and Schultz 2006). Results from this approach have achieved accuracies significantly above chance (Lotte et al. 2007) and point the Broca's and Wernicke's areas as the most relevant in terms of sensed information (Kober et al. 2001).

Fig. 2.1 EEG-based recognition system for unspoken speech. Adapted from Denby et al. (2010); Wester and Schultz (2006)

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.

Deng et al. (2010) were able to use EEG data to decode three different rhythms in a silent speech context. In a recent work, Iqbal et al. (2016) reported a very high classification accuracy for pairwise differentiation between imagined vowels /a/, /u/, and the rest position. Despite the notable results, the study context is still narrow, and it is yet to be shown how the proposed methods behave in a broader phonetic scope. In a similar experimental setting, considering the same vowels for Japanese, Matsumoto (2014) proposed an adaptive selection of data to improve classification.

2.1.4 Electrocorticographic Electrodes

Electrocorticography (ECoG), or intracranial electroencephalography (iEEG), is an invasive technique that consists in the implantation of an extracellular recording electrode and electrical hardware for amplification and transmission of electric brain activity. In comparison with EEG, ECoG sensing provides better spatial resolution and higher signal bandwidth (Leuthardt et al. 2011).

Relevant aspects of this procedure include: the location for implanting the electrodes; the type of electrodes; and the decoding technique. Due to the invasive nature of ECoG, increased risk and medical expertise required, this approach is only applied as a solution to restore speech communication in extreme cases, such as in the case of subjects with the locked-in syndrome which are medically stable and present normal cognition. When compared to EEG sensors, this type of systems presents a better performance since the recordings are not affected by motor potentials caused by unintentional movements (Brumberg et al. 2010).

Results for this approach, in the context of 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) can be achieved, in a 38 phonemes classification problem (Brumberg et al. 2011).

Pei et al. (2011a, b, 2012) present several experimental studies where they are able to predict vowels and consonants in overt and covert speech, with average accuracy around 40 % for overt speech and, 37 %, for covert speech (chance level of 25 %). These results reported for overt and covert speech were quite similar and therefore, these findings are an important evidence that cortex activity can also be used for silent speech decoding (Chakrabarti et al. 2015). One relevant hypothesis raised by the authors (Pei et al. 2012) concerns the possibility of leveraging a currently available speech recognition system, trained with spoken words data, and applying it directly to imagined speech decoding. In this regard, Martin et al. (2014) provides some positive evidence of this possibility, by showing that overt and covert speech share a common neural basis.

For a subject with locked-in syndrome, (Brumberg et al. 2013) reported the control of a speech synthesizer in real time, which eliminates the need of a typing process. Mugler et al. (2014) argued that most approaches in the literature consider whole word classification studies and, while this is an important evidence of the viability of imagined speech decoding, efforts should now move to the phoneme level in order to encompass the full complexity of speech. Using ECoG, the authors performed a classification of phonemes for American English with up to 36 % accuracy, when considering all phonemes, and up to 63 % for a single phoneme.

Herff et al. (2015) presented Brain-to-Text, a system that transforms brain activity resulting from overt speech production into the corresponding textual representation, presenting word error rates as low as 25 % and phone error rates below 50 %.

Chakrabarti et al. (2015), in the context of a review on the use of ECoG for speech decoding, discuss the challenges that still need to be addressed to reach a practical speech neuroprosthetic. Among them, it is important to note that existing studies have been performed in extremely controlled environments, typically in medical settings, with speakers lying still and data recorded over short periods of time considering very limited, discontinuous speech. The importance of such studies is indisputable to improve our knowledge regarding brain activity during speech, but a long way has yet to be made towards a silent speech BCI supporting natural communication in an everyday context.

2.2 Muscular Activity and Silent Speech Interfaces

During the human speech production process, in the phase before the actual articulation, myoelectric signals are sent to the multitude of muscles involved in speech production. The anatomical location of these muscles and their functions are well known, and sensing technology provides ways of measuring that activity, although not free of challenges, opening routes for its application in SSI.

2.2.1 Muscles in Speech Production

There are many muscles involved in the speech production process, which, according to (The UCLA Phonetics Laboratory 2002), can be grouped as follows: respiration, lip movement, mandibular movement, tongue, soft palate, pharynx, and larynx. In this book, we will focus on the muscles which have been previously identified as more relevant for SSI studies, particularly, some of the main muscles of the face and neck used in speech production, which are the following (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 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;
- *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 2.2 illustrates the muscles found in the areas of the face and neck, providing a glimpse of the complexity in terms of muscle physiology in this area.

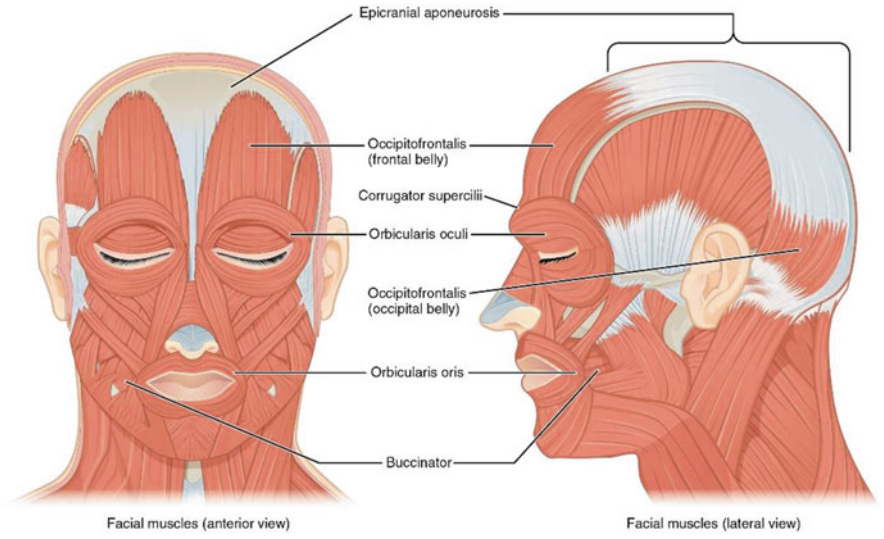

Fig. 2.2 Anterior and lateral view of the human facial muscles from (OpenStax_College 2013)

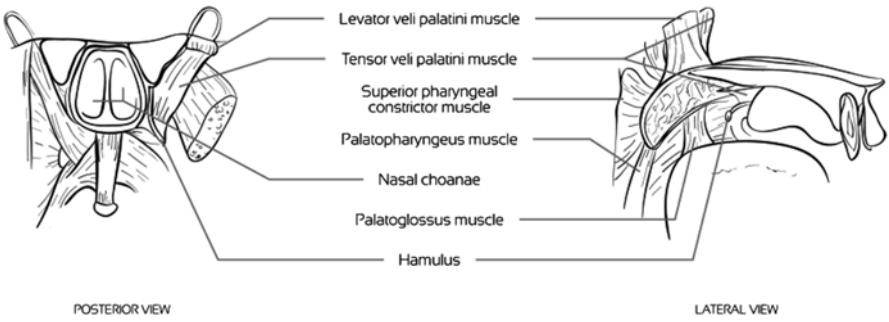

Fig. 2.3 Muscles of the soft palate from posterior (*left*) and the side (*right*) view

This set of muscles is not enough for all sounds. 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 Fig. 2.3:

- *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.

2.2.2 *Measuring Electrical Muscular Activity*

The articulators' muscles are activated through small electrical currents in the form of ion flows (muscular activity), 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. These patterns can be measured via a bioelectric technique based on different types of sensors. The process of recording and evaluating this electrical muscle activity is called electromyography (EMG).

Currently, there are two main sensing techniques to measure electromyography signals: intramuscular and surface electrodes. Both sensing techniques allow recording the electrical activity associated with the contraction of muscle fibers in the motor unit being analyzed.

The intramuscular electrodes are inserted close to the muscle fibers and can be divided into two groups: wire and needle. They are used in clinical contexts in order to detect muscle disorders, denervation and to diagnosis myopathies. Depending on the location and type of electrode the recorded, action potentials can include several muscle fibers. An overview about the types of intramuscular electrodes can be found in (Merletti and Farina 2009).

Surface electrodes are non-invasive; their attachment to the subject is usually done on some adhesive basis, which may obstruct movement, especially in facial muscles. By measuring facial muscles, the surface electrodes will measure the superposition of multiple fields (Gerdle et al. 1999). 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.

There are pros and cons in each sensing technique. The main difference between them is the distance and physiological structures that separates the electrode from the motor unit where the myoelectric signal is being generated. Since the intramuscular electrodes are inserted into the muscles, close to the muscles fibers, the noise caused by in between volumes is minimized. This advantage of being close to the muscle fibers and less noisy results can be seen as a disadvantage in terms of usability. Analyzing the particular case of an SSI based on EMG, intramuscular techniques introduce several usability challenges in real-world scenarios. The inherent invasive procedure associated with these sensors is a barrier for the development of a natural human–computer interface. For that reason, this book focuses on the second technique, which is when the muscular activity is measured by non-implanted electrodes and which is referred to in the literature as surface electromyography (SEMG). Surface electrodes also have the advantage of capturing a more representative signal of the fibers in the motor unit. In terms of positioning, although it is difficult to replicate exact positions with both electrode types, intramuscular electrodes due to their small detection capability are hard to position near the same muscle fiber(s).

2.2.3 Surface Electromyography

Surface EMG-based speech recognition has the potential to overcome 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

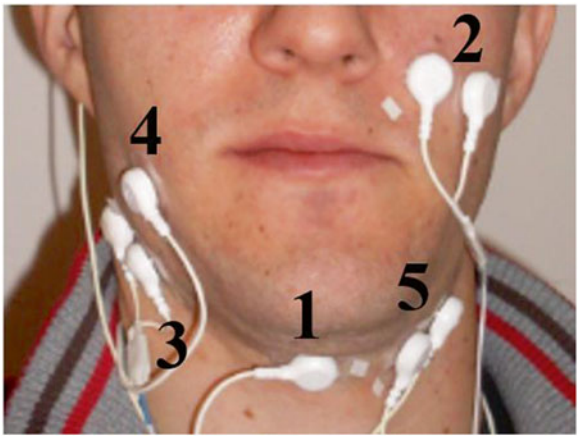

Fig. 2.4 Example of an SEMG setup for SSI with five channels

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). Figure 2.4 shows an example of SEMG sensors placed in the speaker's face and neck.

In the context of speech production, the sensor presence in relevant regions of the face and neck can cause the subject to alter his/her behavior, be distracted or restrained, subsequently altering the experiment result. However, pre-recorded instructions can be given to the speaker in order to minimize these effects. 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). The EMG signal is also not affected by noisy environments, however, in the presence of noise, differences may be found in the speech production process (Junqua et al. 1999).

In 1985, an SEMG-based speech prosthesis was developed by Sugie and Tsunoda (1985) and achieved an average correction rate of 64 % when recognizing five Japanese vowels. In this study, the authors used three EMG channels located in the o*rbicularis oris*, *zygomaticus major,* and the *digastricus*. 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).

Ten years later, in 2001, relevant results were reported by Chan et al. (2001) where five channels of surface Ag–AgCl sensors were used to recognize ten English digits. The author captured myoelectric signals from the *levator anguli oris*, the *zygomaticus major*, the *platysma*, the *depressor anguli oris*, and the *anterior belly of the digastric*, using electrodes embedded in a pilot's oxygen mask. In this study, accuracy rates as high as 93 % were achieved. The same author (Chan 2003) was the first to combine conventional automatic speech recognition (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 18 vowel and 23 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.

A slightly different approach was presented by Manabe in 2003 where instead of using electrodes positioned in the face, the speaker had ring-shaped electrodes in his/her fingers. When speaking, the fingers were pressed against the targeted facial muscles (Manabe and Zhang 2004; Manabe 2003). In these studies, accuracy as high as 64 % was achieved. This approach although interesting in terms of usability faced a problem related with the exact positioning of the electrodes. This issue was later studied by Maier-Hein et al. (2005). Maier-Hein et al. (2005) which presented a speech recognition system based on SEMG for audible and non-audible speech. Using a seven EMG channel system the authors achieved 87.1 % across different sessions on a ten-digit vocabulary. This study also draws interesting conclusions for ASR based on SEMG. They address important issues such as electrodes repositioning between recording sessions, temperature changes, and skin tissue variability by using model adaptation methods, showing the necessity of applying some kind of normalization across sessions and speakers. Additionally, the authors state that ideal number of channels lies between 2 and 5 and they point out that the differences between audible and non-audible speech articulation may have a relevant accuracy impact in this type of interfaces. In the follow-up of this work, in 2007, Jou et al. (2007a) reported an average accuracy of 70.1 % for a 101-word vocabulary in a speaker-dependent scenario. Later, 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 error rate of 21.9 % on a 108-word vocabulary task (Wand and Schultz 2011a), showing that this sort of interfaces could reach interesting error rates in the presence of bigger vocabularies.

In the last years, 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 acquisition devices (Manabe and Zhang 2004; Manabe 2003); recognition of syllables instead of phonemes or words (Lopez-Larraz et al. 2010); and even trying to recognize mentally rehearsed speech (Meltzner et al. 2008).

Since 2010, 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. 2011, 2012); 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); multimodal speech recognition systems

that include SEMG (Freitas et al. 2014); addressing new languages and their characteristics (Freitas et al. 2012, 2015); removing the need for initial training before actual use (Wand and Schultz, 2014); reducing the number of sensors and enhancing SEMG continuous speech modelling (Deng et al. 2014); EMG-based synthesis (Zahner et al. 2014); and the application of deep neural networks (Diener et al. 2015) to classify EMG signals.

In terms of electrodes setup, Wand et al., in 2013, presented a different method for collecting facial EMG signals based on multichannel electrode arrays. This type of setup, previously used in other EMG studies, was applied for the first time for speech recognition. The advantages of such configuration are to increase the number of available channels, to facilitate the positioning of a high number of electrodes and the detection of artifacts (Heistermann et al. 2014; Wand et al. 2013).

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

2.3 Conclusions

Part of the concept of silent speech is, in a way, to look beyond the visible or audible parts of the human speech process. In this chapter, we focused our attention in two parts of this process with these characteristics: the brain and the muscle activity related with speech. Following this order, we introduced related SSI modalities, which aim at capturing information about speech activity that happens behind the scenes, i.e., activity which is not visible or measured by human senses. In order to do so we explained which sensing technologies we currently have at our disposal to extract related information, how they work and some of the most relevant studies in these areas of research.

Our analysis of the SSI modalities related with the brain and the facial and neck muscles shows several lines of research with the overall goal of improving performance and usability of associated human–computer interfaces. However, because of the reasons pointed out in Chap. 1 (Sect. 1.4) establishing a direct comparison between them may not be fair. Additionally, depending on the HCI scenario, some sensing technologies may make more sense than others. For example, invasive approaches, like ECoG, have been mostly reported in extreme cases of paralysis (e.g., total locked-in syndrome) and do not make sense for simply tackling noisy environments. It all comes down to what the human–computer interface user is capable or is willing to use. As technology evolves, approaches like ECoG may no longer be necessary in the future, if similar results are obtained of modalities such as EEG.

In the next chapter, we will study ways to extract relevant information not only from the myoelectric signal but also from the resulting articulators' movement, bearing in mind that not all movements occur in the sight of the human eye.

References

Akcakaya M, Peters B, Moghadamfalahi M, Mooney AR, Orhan U, Oken B, Erdogmus D, Fried-Oken M (2014) Noninvasive brain-computer interfaces for augmentative and alternative communication. IEEE Rev Biomed Eng 7:31–49. doi:10.1109/RBME.2013.2295097

Andersen RA, Kellis S, Klaes C, Aflalo T 2014. Toward More Versatile and Intuitive Cortical Brain–Machine Interfaces. Curr. Biol. 24, R885–R897. doi:http://dx.doi.org/10.1016/j.cub.2014.07.068

Beddor PS (1993) The perception of nasal vowels. In: Huffman MK, Krakow RA (eds) Phonetics and phonology, vol 5, Nasals, nasalization and the velum. Academic, London

Betts BJ, 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

Bouchard KE, Mesgarani N, Johnson K, Chang EF (2013) Functional organization of human sensorimotor cortex for speech articulation. Nature 495:327–332

Brumberg JS, Kennedy PR, Guenther FH (2009) Artificial speech synthesizer control by brain-computer interface. Proc Interspeech 2009:636–639

Brumberg JS, Nieto-Castanon A, Kennedy PR, Guenther FH (2010) Brain-Computer Interfaces for Speech Communication. Speech Commun 52:367–379. doi:10.1016/j.specom.2010.01.001

Brumberg JS, Wright EJ, Andreasen DS, Guenther FH, Kennedy PR (2011) Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex. Front Neurosci 5

Brumberg JS, Guenther FH, Kennedy PR (2013) An auditory output brain--computer interface for speech communication. In: Guger C, Allison BZ, Edlinger G (eds) Brain-computer interface research, SpringerBriefs in Electrical and computer engineering. Springer, Heidelberg, pp 7–14. doi:10.1007/978-3-642-36083-1_2

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

Chakrabarti S, Sandberg H, Brumberg J, Krusienski D (2015) Progress in speech decoding from the electrocorticogram. Biomed Eng Lett 5:10–21. doi:10.1007/s13534-015-0175-1

Chan ADC (2003) Multi-expert automatic speech recognition system using myoelectric signals. The University of New Brunswick (Canada)

Chan ADC, Englehart K, Hudgins B, Lovely DF (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 ADC, Englehart K, Hudgins B, Lovely DF (2002) Hidden Markov model classification of myoelectric signals in speech. Eng Med Biol Mag IEEE 21:143–146

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428–1432

Conant D, Bouchard KE, Chang EF (2014) Speech map in the human ventral sensory-motor cortex. Curr. Opin. Neurobiol. 24, 63–67. doi:http://dx.doi.org/10.1016/j.conb.2013.08.015

DaSalla CS, 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

Denby B, Schultz T, Honda K, Hueber T, Gilbert JM, Brumberg JS (2010) Silent speech interfaces. Speech Commun 52:270–287. doi:10.1016/j.specom.2009.08.002

Deng S, Srinivasan R, Lappas T, D'Zmura M (2010) EEG classification of imagined syllable rhythm using Hilbert spectrum methods. J Neural Eng 7:46006

Deng Y, Heaton JT, Meltzner GS (2014) Towards a Practical Silent Speech Recognition System. Proceedings of Interspeech 2014:1164–1168

Diener L, Janke M, Schultz T (2015) Direct conversion from facial myoelectric signals to speech using deep neural networks. Neural Networks (IJCNN), 2015 Int. Jt. Conf. doi:10.1109/IJCNN.2015.7280404

Freitas J, Teixeira A, Dias MS (2012) 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, Ferreira A, Figueiredo M, Teixeira A, Dias MS (2014) 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, Silva S, Oliveira C, Dias MS (2015) Detecting nasal vowels in speech interfaces based on surface electromyography. PLoS One 10, e0127040. doi:10.1371/journal.pone.0127040

Fritzell B (1969) The velopharyngeal muscles in speech: an electromyographic and cineradiographic study. Acta Otolaryngol 50

Gerdle B, Karlsson S, Day S, Djupsjöbacka M (1999) Acquisition, processing and analysis of the surface electromyogram. In: Windhorst U, Johansson H (eds) Modern techniques in neuroscience research. Springer, Berlin, pp 705–755

Guenther FH, Brumberg JS (2011) Brain-machine interfaces for real-time speech synthesis. In: Engineering in Medicine and Biology Society, EMBC, 2011 Annual international conference of the IEEE, pp 5360–5363. doi:10.1109/IEMBS.2011.6091326

Hardcastle WJ (1976) Physiology of speech production: an introduction for speech scientists. Academic, New York

Heistermann T, Janke M, Wand M, Schultz T (2014) Spatial artifact detection for multi-channel EMG-based speech recognition. In: Proceedings of the International conference on bio-inspired systems and signal processing, pp. 189–196

Herff C, Janke M, Wand M, Schultz T (2011) Impact of different feedback mechanisms in EMG-based speech recognition. Interspeech 12:2213–2216

Herff C, Heger D, de Pesters A, Telaar D, Brunner P, Schalk G, Schultz T (2015) Brain-to-text: Decoding spoken phrases from phone representations in the brain. Front Neurosci 9:217. doi:10.3389/fnins.2015.00217

Hickok G (2012) Computational neuroanatomy of speech production. Nat Rev Neurosci 13:135–145

Indefrey P (2011) The spatial and temporal signatures of word production components: A critical update. Front Psychol 2:255. doi:10.3389/fpsyg.2011.00255

Iqbal S, Muhammed Shanir PP, Khan Y, Farooq O (2016) Time domain analysis of EEG to classify imagined Speech. In: Satapathy SC, Raju KS, Mandal JK, Bhateja V (eds) Proceedings of the Second international conference on computer and communication technologies, Advances in intelligent systems and computing. Springer, Delhi, pp 793–800. doi:10.1007/978-81-322-2523-2_77

Jorgensen C, Binsted K (2005) Web browser control using EMG based sub vocal speech recognition. In: Proceedings of the 38th Annual Hawaii international conference on system science, p 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, 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

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

Kellis S, Miller K, Thomson K, Brown R, House P, Greger B (2010) Decoding spoken words using local field potentials recorded from the cortical surface. J Neural Eng 7:56007

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

Kuehn DP, Folkins JW, Linville RN (1988) An electromyographic study of the musculus uvulae. Cleft Palate J 25:348–355

Leuthardt EC, Gaona C, Sharma M, Szrama N, Roland J, Freudenberg Z, Solis J, Breshears J, Schalk G (2011) Using the electrocorticographic speech network to control a brain–computer interface in humans. J Neural Eng 8:36004

Lopez-Larraz E, Mozos OM, Antelis JM, Minguez J (2010) Syllable-based speech recognition using EMG. Conf Proc IEEE Eng Med Biol Soc 2010:4699–4702. doi:10.1109/IEMBS.2010.5626426

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

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

Martin S, Brunner P, Holdgraf C, Heinze H-J, Crone NE, Rieger J, Schalk G, Knight RT, Pasley BN (2014) Decoding spectrotemporal features of overt and covert speech from the human cortex. Front Neuroeng 7

Matsumoto M (2014) Silent speech decoder using adaptive collection. In: Proceedings of the Companion publication of the 19th International conference on intelligent user interfaces, IUI Companion '14. ACM, New York, pp 73–76. doi:10.1145/2559184.2559190

Meltzner GS, Sroka J, Heaton JT, Gilmore LD, Colby G, Roy S, Chen N, Luca CJ. 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

Meltzner GS, Colby G, Deng Y, Heaton JT (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

Merletti R, Farina D (2009) Analysis of intramuscular electromyogram signals. Philos Trans A Math Phys Eng Sci 367:357–368

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science (80-.). 343, 1006–1010.

Morse MS, O'Brien EM (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

Morse MS, Gopalan YN, Wright M (1991) Speech recognition using myoelectric signals with neural networks. In: Proceedings ofthe Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society. IEEE, pp 1877–1878

Mugler EM, Patton JL, Flint RD, Wright ZA, Schuele SU, Rosenow J, Shih JJ, Krusienski DJ, Slutzky MW (2014) Direct classification of all American English phonemes using signals from functional speech motor cortex. J Neural Eng 11:035015. doi:10.1088/1741-2560/11/3/035015

Munding D, Dubarry A-S, Alario F-X (2015) On the cortical dynamics of word production: a review of the MEG evidence. Lang Cogn Neurosci 1:22. doi:10.1080/23273798.2015.1071857

Nijholt A, Tan D (2008) Brain-computer interfacing for intelligent systems. Intell Syst IEEE 23:72–79

Oken BS, Orhan U, Roark B, Erdogmus D, Fowler A, Mooney A, Peters B, Miller M, Fried-Oken MB (2014) Brain–computer interface with language model—electroencephalography fusion for locked-in syndrome. Neurorehabil. Neural Repair 28:387–394

OpenStax_College (2013) Front and side views of the muscles of facial expressions [WWW Document]. Anat. Physiol. Connexions Web site. URL http://cnx.org/content/col11496/1.6/ (accessed 4.3.16)

Pei X, Barbour DL, Leuthardt EC, Schalk G (2011a) Decoding vowels and consonants in spoken and imagined words using electrocorticographic signals in humans. J Neural Eng 8:046028. doi:10.1088/1741-2560/8/4/046028

Pei X, Leuthardt EC, Gaona CM, Brunner P, Wolpaw JR, Schalk G (2011b) Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. Neuroimage 54:2960–72. doi:10.1016/j.neuroimage.2010.10.029

Pei X, Hill J, Schalk G (2012) Silent communication: toward using brain signals. Pulse IEEE 3:43–46. doi:10.1109/MPUL.2011.2175637

Piai V (2015) The role of electrophysiology in informing theories of word production: a critical standpoint. Lang Cogn Neurosci 31(4):471–473. doi:10.1080/23273798.2015.1100749

Pickering MJ, Garrod S (2013) An integrated theory of language production and comprehension. Behav Brain Sci 36:329–347. doi:10.1017/S0140525X12001495

Price CJ (2012) A review and synthesis of the first 20 years of {PET} and fMRI studies of heard speech, spoken language and reading. Neuroimage 62:816–847. doi:http://dx.doi.org/10.1016/j.neuroimage.2012.04.062

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

Seikel JA, King DW, Drumright DG (2009) Anatomy and physiology for speech, language, and hearing, 4th edn. Delmar Learning, Clifton Park

Sorger B, Reithler J, Dahmen B, Goebel R (2012) A real-time fMRI-based spelling device immediately enabling robust motor-independent communication. Curr Biol 22:1333–1338. doi:10.1016/j.cub.2012.05.022

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

Telaar D, Wand M, Gehrig D, Putze F, Amma C, Heger D, Vu NT, Erhardt M, Schlippe T, Janke M (2014) BioKIT-Real-time decoder for biosignal processing. In: The 15th Annual conference of the international speech communication association (Interspeech 2014)

The UCLA Phonetics Laboratory (2002) Muscles of the speech production mechanism. In: Dissection manual for students of speech. p. Appendix B

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

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, Janke M, Schultz T (2011) Investigations on speaking mode discrepancies in EMG-based speech recognition. Interspeech 2011:601–604

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, Schulte C, Janke M, Schultz T (2013) Array-based electromyographic silent speech interface. In: International conference on bio-inspired systems and signal processing (BIOSIGNALS 2013), pp 89–96

Wand M, Janke M, Schultz T (2014) (2014) The EMG-UKA corpus for electromyographic speech processing. In: Proceedings of Interspeech 2014

Wester M, Schultz T (2006) Unspoken speech—speech recognition based on electroencephalography. Universitat Karlsruhe (TH), Karlsruhe

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

Chapter 3
SSI Modalities II: Articulation and Its Consequences

Abstract Brain and muscular activity originates the change in shape or position of articulators such as the tongue or lips and, as a consequence, the vocal tract assumes different configurations. Most of these changes, namely of articulators and tract, are internal and are not easy to measure but, in some cases, like the lips or the tongue tip, such changes are visible or have visible effects. Even without the production of speech sound, these different configurations of articulators provide valuable information that can be used in the context of silent speech interfaces (SSIs). In this chapter, the reader finds an overview of the technologies used to assess articulatory and visual aspects of speech production and how researchers have exploited their capabilities for the development of silent speech interfaces.

Keywords Visual speech • Video • Electromagnetic midsagittal articulography • Permanent magnetic articulography • Ultrasound • Ultrasonic Doppler • Non-audible murmur microphone

In the previous chapter, we discussed how brain and myoelectric activity can be harnessed using a variety of technologies to understand speech production and how to use this knowledge to enable silent speech interfaces (SSIs).

Due to brain and muscular activity, the vocal tract assumes particular configurations, including changes of externally visible articulators, such as the lips or the tongue tip. Even without the speech sound, these different changes and configurations provide valuable information regarding the speech content. For example, a backed tongue or rounded lips hint particular families of sounds, e.g., vowels [o] and [u], a relevant input for silent speech interfaces.

This chapter provides a brief overview of sensing technologies used to measure articulatory and visible changes of speech production and how researchers have used such capabilities for silent speech interfaces. Taking into consideration the state-of-the-art approaches and associated technologies, we have grouped them in the following manner: (1) approaches that are capable of measuring non-visible articulators and can track changes in the vocal track; (2) those that focus on visible articulators and visible effects of articulation; and lastly, (3) approaches that measure other effects which do not fall into the first two categories. This separation is based on the main capabilities of each sensing technology approach; however, it does not mean that

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_3

information from other groups cannot be collected. For example, electromagnetic articulography (EMA) can be used to collect lip movement information, but it also can give us information about the back of the tongue or the velum.

3.1 Measuring Non-visible Articulators and Changes in the Vocal Tract

Knowing the shape of the vocal track provides us with valuable information for what is being said. However, not all articulators are visible and tracking these changes in the vocal track can be a challenging task, especially if our aim is to build a human–computer interface (HCI).

In the state-of-the-art of silent speech interfaces (SSIs), we can find two solutions to capture information from hidden articulators: (1) measure the position of magnets placed in the articulators; (2) use image-based techniques, such as ultrasound.

In the following subsections, we explain in more detail each of these approaches and their respective results.

3.1.1 Electromagnetic and Permanent Magnetic Articulography

Using the principle of magnetic sensing, we can monitor the movement of fixed points in the articulators, collecting information from the articulation stage (explained on the first chapter) of the speech production process. A variant of this approach is the standard EMA system (Perkell et al. 1992), like the Carstens AG500 (Carstens Medizinelektronik 2016) that uses glued coils electrically connected to external equipment (Kroos 2012). However, although this approach enables accurate access to the Cartesian positions of the articulators in a given reference frame, the necessary electrical connections for this approach makes 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 showed interesting accuracy 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 English sentences (Wang et al. 2012a) and whole-words recognition (Wang et al. 2012b). The first achieved accuracy rates of 94.9 % across ten subjects and, the latter, 60.0 % in a speaker-dependent scenario. More recently, the same authors improved the results to an average speaker-dependent recognition accuracy of 80.0 % (Wang et al. 2013). The most recent studies explore techniques such as normalization methods for speaker-independent silent speech recognition (Wang et al. 2014).

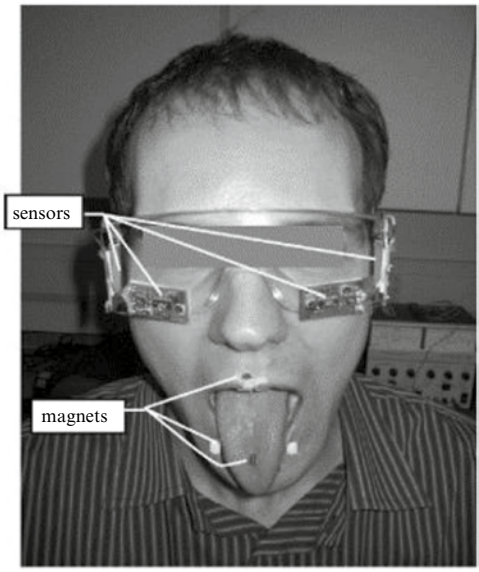

Fig. 3.1 Placement of magnets and magnetic sensors for a PMA-based SSI. Adapted from Denby et al. (2010)

A recent variant of this concept consists of using magnets attached to the vocal apparatus (see Fig. 3.1), 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, which are associated with specific articulatory gestures through statistical modelling. For example, in Fagan et al. (2008), magnets were placed on the lips, teeth, and tongue of a subject and were tracked by six dual axis magnetic sensors incorporated into a pair of glasses. Results from this laboratory experiment showed 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). Other authors obtained their best the best result (91.0 % word accuracy rate) considering an additional axis, five magnetic sensors and 71 words vocabulary (Hofe et al. 2013a). On a phonetic level, the same authors showed also the capability of this approach to detect voiced and unvoiced consonants (Hofe et al. 2013a, b). More recent studies 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, although PMA is capable of discriminating the place of articulation of consonants, it does not provide much information regarding the voicing and manner of articulation.

3.1.2 Vocal Tract Imaging

In recent years, imaging techniques have also played an important role in expanding our knowledge of articulatory features by providing an "inside view" of the vocal tract. Among these are technologies such as magnetic resonance imaging (MRI) both in 2D and 3D, and the novel protocols for real-time MRI (RT-MRI), which enable capturing dynamic aspects of speech production (Scott et al. 2014; Silva and Teixeira 2015). With these, the various articulators can be sensed, from the lips to the larynx, with a considerable frame rate. Nonetheless, one of the disadvantages of MRI is the effort and resources required to perform a study, due to heavy non-portability nature of the required equipment and its associated costs. As an alternative, ultrasound imaging technology has evolved to provide increasingly portable and cheaper devices that can be used in the field (Whalen and McDonough 2015). When compared to MRI, these are important advantages but it is also important to note that ultrasound sensing, featuring an ultrasonic probe positioned under the speaker's chin, only allows covering the tongue and, to some extent, the hard palate, which often means that additional technologies are used to cover the lips, for example.

Overall, the cost, portability, and ease of use of ultrasound, in comparison to other imaging approaches, turn it into a technology that can be more easily adapted for silent speech interfaces.

3.1.3 Ultrasound and Speech Articulation

Taking into account its advantages relatively to other imaging methods, ultrasound has been widely used to study speech production with an emphasis on the tongue, given its technical characteristics (Xu et al. 2016).

One of the main challenges that has been recurrently addressed by several researchers, concerns the stabilization of the ultrasound probe and its registration with other auxiliary technologies. Since the ultrasound probe is positioned under the chin, a change in the probe-chin angle or a translation of its position due to the speakers' movement would render the acquired data hard to compare. Therefore, methods are needed to serve two purposes: (1) ensure that the relative position between the probe and the speaker is kept, throughout the study; and if that is not ensured, (2) detect and measure any movement affecting the probe position and orientation and use it to correct the data. Examples of different approaches to these aspects can be found in the works of Miller and Finch (2011), Whalen et al. (2005), Mielke (2011), Zharkova and Hewlett (2009), and Hueber et al. (2010), along with efforts for their evaluation (Scobbie et al. 2008; Whalen et al. 2005), and the assessment of the impact of not using stabilization methods for clinical applications (e.g., Acher et al. 2014; Zharkova et al. 2015). Figure 3.2 presents an example of a US probe stabilization headset, ensuring that, even though the speaker can move his head, the relative position between the US probe and the speaker is kept throughout the study.

Fig. 3.2 Speaker using a probe stabilization headset. The US probe can be seen tightly secured below the chin

The multiplicity of works considering ultrasound to study speech articulation conveys its applicability in capturing the lingual features of different sounds such as rhothics (Lawson et al. 2015) and laterals (Turton 2015), in a variety of languages (e.g., Magloughlin 2016; Mielke 2011; Vietti et al. 2015), and aspects such as coarticulation (Zharkova et al. 2012).

In addition, ultrasound is also being used for therapy and rehabilitation to provide biofeedback (Bacsfalvi and Bernhardt 2011; Cleland et al. 2015), which brings this technology into scenarios where SSI can also provide benefits.

3.1.4 Ultrasound and Silent Speech Interfaces

In the silent speech context, an ultrasound transducer is placed beneath the chin, providing a partial view of the tongue surface in the midsagittal plane (Hueber et al. 2010). This type of approach is commonly combined with frontal and/or side video of the user's lips (as depicted in Fig. 3.3a). 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 to ensure 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).

More recent work using US and 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 Fig. 3.3b for an example of the resulting image data). Results for this technique show that, for an hour of

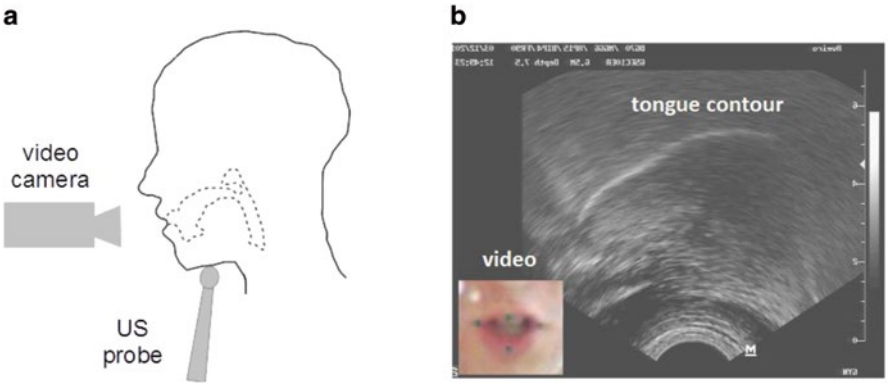

Fig. 3.3 Acquiring data for the tongue and lips: (**a**) an US probe is placed under the speaker's chin and a video camera captures the lips; (**b**) example of an US vocal tract image with embedded frontal lip view

continuous speech, 60 % of the phones are correctly identified in a sequence of tongue and lip images, demonstrating 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).

3.2 Measuring Visible Articulators and Visible Effects of Articulation

Obtaining information regarding visible articulators such as the lips or the jaw is generally done using image and video sensing techniques, but other technologies such as ultrasonic Doppler or depth cameras can also be used. These technologies are not only useful to obtain information on visible effects, but are also capable of obtaining information regarding other internal articulators such as the tip of the tongue (when visible), or subtle changes in the face. In the following subsections, we reveal some of the relevant methods and studies in the areas of visual speech recognition and ultrasonic Doppler sensing (UDS).

3.2.1 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). These authors discovered that vision affects the performance of the human speech perception

Fig. 3.4 Typical visual speech recognition system pipeline

because: (1) it permits to identify the source location; (2) it allows a better segmentation of the audio signal; and (3) 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 automatic speech recognition (AV-ASR) systems and visual speech recognition (VSR) (or visual-only ASR) systems, usually composed by the stages depicted in Fig. 3.4.

In VSR systems, a video composed of successive RGB frames is used as 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. Studies indicate that the majority of the systems that use multiple simultaneous input channels such as audio plus video have a better performance than 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, and degradation due to speech impairments (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; Wand et al. 2016), 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 accounting for information which could be extracted from jaws and cheeks.

3.2.2 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 appearance-based 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 (geometry-based) approaches perform 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, area of the mouth, and image moment descriptors of the lip contours, such as active shape models or lip-tracking models. Shape-based methods require accurate and reliable facial and lip feature detection and tracking, which have proven 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 methods 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 amount of pixels that is usually 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), or locality sensitive discriminant analysis (LSDA). Other nonlinear transforms, such as the discrete cosine transform (DCT), discrete wavelet transformation (DWT) or Haar transforms, or a combination of these methods (Potamianos et al. 2003; Yaling et al. 2010), have also been used. A comprehensive overview of these methodologies can be found in (Potamianos et al. 2003).

3.2.3 Local Feature Descriptors

An alternative approach to the methods already mentioned is to explore the use of local feature descriptors (Carvalho et al. 2013) following, for example, an appearance-based approach using feature extraction and tracking. This type of approach has proliferated in the areas of computer vision and augmented reality, due to its intrinsic low computational cost, allowing real-time solutions.

To fully address the problem, robust feature extraction and tracking mechanism are required and the computer vision community provides various alternatives, such as Harris and 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.

3.2.4 *Ultrasonic Doppler Sensing*

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 observer-source, c is the propagation velocity of the wave on the medium and f_0 the source frequency, the observed frequency will be:

$$f = \frac{c + v_o}{c + v_s} f_0 \tag{3.1}$$

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

$$f = \left(1 - \frac{v_s}{c}\right) f_0 \quad \text{or} \quad \Delta f = -\frac{v_s}{c} f_0 \tag{3.2}$$

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 whose frequency is measured by a receiver placed closely to the emitter. The observed Doppler shift will then be the double:

$$\Delta f = \frac{2v}{c} f_0 \tag{3.3}$$

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

UDS of speech is one of the approaches reported in the literature that is also suitable for implementing an SSI (Srinivasan et al. 2010; Zhu 2008). This technique is based on the emission of a pure tone in the ultrasound range towards the speaker's face. The reflected signal is received by an ultrasound sensor tuned to the transmitted frequency, which 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).

To put it simply, we could consider a scenario (much like the one depicted in Fig. 3.5), where a source T emits a wave with frequency f_0 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).

Below, an example of the spectrogram of the Doppler signal and the correspondent audio signal applied to a Portuguese word "canto" ([k6~tu], corner) is depicted on Fig. 3.6.

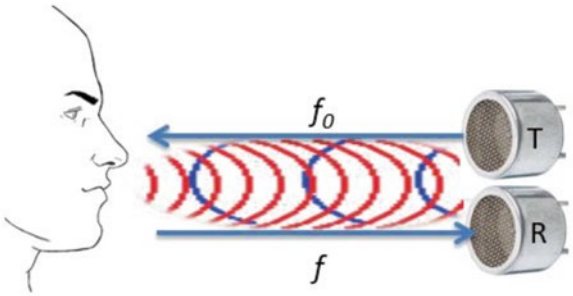

Fig. 3.5 Doppler effect representation (*T* transmitter, *R* receptor)

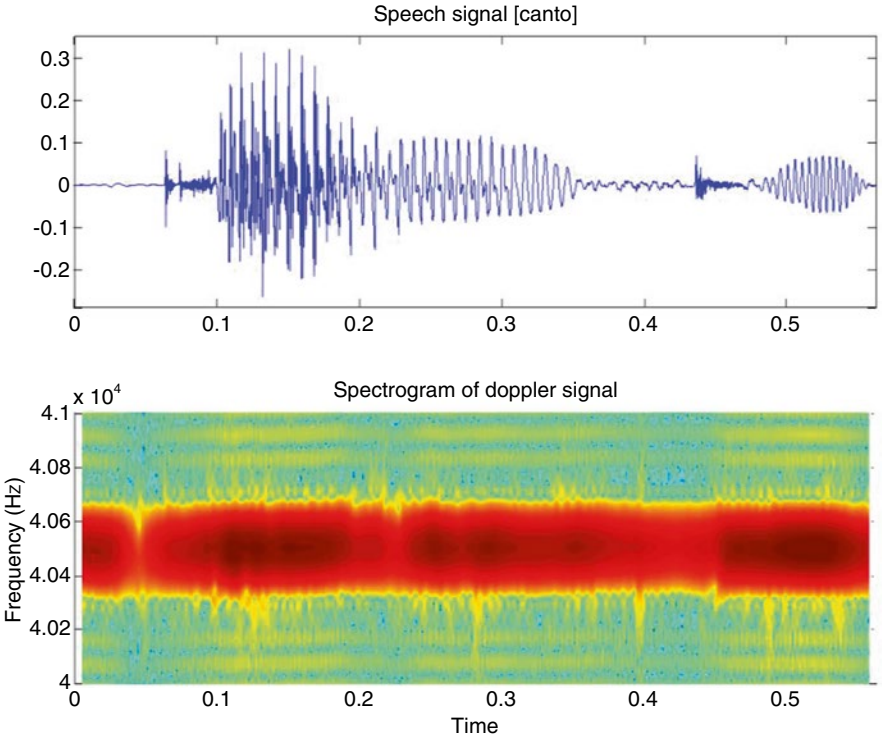

Fig. 3.6 Audio signal (*above*) and spectrogram of the Doppler signal (*below*) for the word "canto" (*corner*) (Freitas et al. 2011)

3.2.5 Ultrasonic Doppler Uses in Silent Speech Interfaces

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 human–computer interaction (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. 2012; 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, 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 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, ten linear predictive coding (LPC) spectral 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 four template utterances, in a speaker-dependent isolated digit recognition task, considering five test sessions and each session containing 100 utterances. For the cross-session scenario, an accuracy not higher than 12.6 % was achieved.

It was only a few years later 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 et al. 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 10 random digits of 20 speakers at a 15.2 cm distance in relation to the sensors. As far as feature extraction was concerned, the authors split the signal in frequency and magnitude sub-bands and extracted, for each frame, features based on energy-band frequency centroids and frequency sub-band energy averages. 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 obtained for the ultrasonic approach were very similar across multiple noise levels, with the 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 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 and coworkers used similar hardware to the setup previously described, adding however the synchronization of the two-channel (audio and ultrasound) output, locating the carrier at 8 kHz and the sensor, and positioning the sensor at 40.5 cm from the speaker. In terms of signal processing, the authors applied a fast Fourier transform (FFT) over the preprocessed signal and applied a DCT to the bins corresponding to the frequencies between 7 and 9.5 kHz, retaining the coefficients as features. For classification purposes, the authors adopted Hidden Markov

models (HMM) with 16 states and one Gaussian per state. The best results for fast speech showed an accuracy of 37.75 % and 18.17 % for slow speech.

In recent studies from the authors of this book, UDS was used to recognize European Portuguese (Freitas et al. 2012) with the best result of 27.8 % WER, in an isolated word recognition problem across several speakers. Additionally, the authors have also made exploratory studies for analyzing the capacity of UDS to detect nasality using RT-MRI as ground truth data (Freitas et al. 2014).

When compared with other secondary sensors (assuming the speech sensing as the primary source), ultrasonic Doppler sensors have the advantage of not requiring to be mounted on the speaker, and although their measurements are not as detailed as in physiological microphones (PMIC) or General Electromagnetic Motion System (GEMS) (Hu and Raj 2005), the reported results obtained with mutual information between UDS and acoustic speech signals were very similar to the ones reported for other secondary devices (Hu and Raj 2005). When compared with vision devices such as cameras, UDS sensors have a clear cost advantage, since an ultrasonic sensing setup can be bought 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 of this approach. Using the same criteria as adopted in Denby et al. (2010), we can conclude the following:

- **Works in noisy conditions**: 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 40.0 cm without requiring any additional hardware.
- **Ready for market**: Results for this approach are still preliminary.
- **Low cost**: The hardware used in this approach is commercially available and is very inexpensive.

UDS, as was stressed before, still has a margin for improvement, especially when applied to silent speech recognition. Additionally, 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 since both affect the ultrasound performance.

3.3 Measuring Other Effects

There are other, non-visible, effects of articulation that can be measured, such as non-acoustic signals propagated via human tissues or bones. In this section, we present the most representative methods in the area, focusing in the non-audible

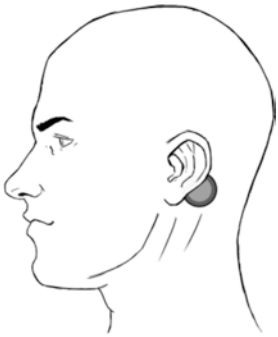

Fig. 3.7 Non-audible murmur microphone positioning based on Ishii et al. (2011)

murmur microphones due to their strong application as a silent speech modality. Other electromagnetic and vibration sensors which can be found in the literature are also briefly described.

3.3.1 Non-audible Murmur Microphones

Non-audible murmur is the term given by research community to speech actions, which a nearby person is not able hear or understand (Nakajima et al. 2003a). This type of speech, although not perceptible to nearby listeners, can be detected using the NAM microphone (see Fig. 3.7), introduced by Nakajima et al. (2003b). 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. 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 (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 recognition values of around 88 % using an iterative adaptation of normal speech to train an 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).

3.3.2 Other Electromagnetic and Vibration Sensors

The development of electromagnetic and vibration sensors was mainly motivated by several military programs in Canada, the United States, 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), which estimates the area between the vibrating vocal folds; GEMS, which measures the glottal tissue oscillations as well as other speech organs movement; and the tuned electromagnetic resonating collar (TERC), which measures small changes in the dielectric properties of the glottis during glottal activity (Brown et al. 2005). In terms of vibration microphones the following types can be found: Throat microphone, bone microphone, physiological microphone, and in-ear microphone. These are usually placed in the neck area to measure the vibration of the vocal cords.

These sensors have presented good results in terms of noise attenuation with gains up to 20 dB (Quatieri et al. 2006) and significant improvements in WER (Jou et al. 2004). Some authors (Quatieri et al. 2006) showed 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 was also considered by Advanced Speech Encoding Program of DARPA for non-acoustic communication (Denby et al. 2010).

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

The disadvantage found in these sensors is that they usually require at least a minimum of 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.

3.4 Conclusions

With this chapter, we have completed our review of existing SSI approaches, along with the associated technologies and their respective recognition accuracy results. This chapter focuses on articulation and its consequences. Here, we have presented multiple techniques with different capabilities. We have seen that one of these, namely ultrasound imaging, is capable of delivering valuable insights about the vocal tract changes. Other approach, such the one based in RGB sensing, can provide high-definition information of the face.

Considering the pros and cons of each modality, RGB data extracted from a video camera, emerge as an important modality capable of achieving the proposed objectives, since it gathers interesting characteristics in terms of silent speech interaction (i.e., low cost, noninvasive, works in noisy environments). However, it is limited to visible articulators such as the lips and therefore complementary information collected for other articulators, such as the tongue and the velum, could be used to improve the recognition performance.

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

UDS is an interesting approach mainly due to its non-obtrusive nature and low cost. Still, several issues remain unsolved: speaker dependence, sensor distance sensitivity, spurious movements made by the speaker, silent articulation, among others.

Permanent magnetic articulography has also been one the approaches addressed by 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, in turn, 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. They have been used with success as commercial applications, mainly in noise-cancellation scenarios. When compared with other SSI technologies, these sensors have the disadvantage of requiring glottal activity.

To summarize, all the mentioned techniques enable us to access information at all stages of the speech production process. However, each of the associated technologies shows, along with their strengths, their limitations too. We can then conclude that, in order to obtain a potentially more complete representation of the intent of the issuer of a silent speech utterance, we need an approach that combines a carefully selected mixture of these SSI modalities. This idea introduces several new and complex challenges that we will explore and tackle in the next chapter.

References

Acher A, Perrier P, Savariaux C, Fougeron C (2014) Speech production after glossectomy: methodological aspects. Clin Linguist Phon 28:241–256

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

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

Bacsfalvi P, Bernhardt BM (2011) Long-term outcomes of speech therapy for seven adolescents with visual feedback technologies: ultrasound and electropalatography. Clin Linguist Phon 25:1034–1043

Bastos R, Dias MS (2009) FIRST—fast invariant to rotation and scale transform: invariant image features for augmented reality and computer vision. VDM, Saarbrücken

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

Brown DR III, Keenaghan K, Desimini S (2005) Measuring glottal activity during voiced speech using a tuned electromagnetic resonating collar sensor. Meas Sci Technol 16:2381

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. Proc Interspeech 2011:841–844

Carstens Medizinelektronik (2016) 3D Electromagnetic Articulograph [WWW Document]. URL http://www.articulograph.de/. Accessed 4 April 2016

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

Cleland J, Scobbie JM, Wrench AA (2015) Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. Clin Linguist Phon 1–23

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 JM, Brumberg JS (2010) Silent speech interfaces. Speech Commun 52:270–287. doi:10.1016/j.specom.2009.08.002

Fabre D, Hueber T, Badin P (2014) Automatic animation of an articulatory tongue model from ultrasound images using Gaussian mixture regression. Proc Interspeech 2014:2293–2297

Fagan MJ, Ell SR, Gilbert JM, Sarrazin E, Chapman PM (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

Florescu VM, 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. Proc Interspeech 2010:450–453

Francisco AA, Jesse, A, Groen MA, McQueen JM (2014) Audiovisual temporal sensitivity in typical and dyslexic adult readers. Proc Interspeech 2014

Freitas J, Teixeira A, Dias MS, 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, Vaz F, Dias MS (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, pp 227–236. doi:10.1007/978-3-642-35292-8_24

Freitas J, Teixeira A, Dias MS (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: International conference on physiological computing systems (PhyCS 2014). pp 232–239

Gilbert JM, Rybchenko SI, Hofe R, Ell SR, Fagan MJ, Moore RK, 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 JA, Cheah LA, Bai J, Ell SR, Gilbert JM, Moore RK, Green PD (2014) Analysis of phonetic similarity in a silent speech interface based on permanent magnetic articulography. Proc Interspeech 2014:1018–1022

Gurbuz S, Tufekci Z, Patterson E, Gowdy JN (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.

Harris C, Stephens M (1988) A combined corner and edge detector. In: Alvey vision conference. Manchester, UK, p. 50.

Heracleous P, Hagita N (2010) Non-audible murmur recognition based on fusion of audio and visual streams. Proc Interspeech 2010:2706–2709

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

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

Hofe R, Ell SR, Fagan MJ, Gilbert JM, Green PD, Moore RK, Rybchenko SI (2010) Evaluation of a silent speech interface based on magnetic sensing. Proc Interspeech 2010:246–249

Hofe R, Bai J, Cheah LA, Ell SR, Gilbert JM, Moore RK, Green PD (2013a) Performance of the MVOCA silent speech interface across multiple speakers. In: Proc. of Interspeech 2013. pp 1140–1143

Hofe R, Ell SR, Fagan MJ, Gilbert JM, Green PD, Moore RK, Rybchenko SI (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 JF, Foundation JH, Davis C (2009) Characterizing silent and pseudo-silent speech using radar-like sensors. Interspeech 2009:656–659

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

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. Proc Interspeech 2009:640–643

Hueber T, Benaroya EL, 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, Bailly G, Denby B (2012) Continuous articulatory-to-acoustic mapping using phone-based trajectory HMM for a silent speech interface. Proc Interspeech 2012:723–726

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]. https://jawbone.com

Jennings DL, Ruck DW (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.

Jou S-C, Schultz T, Waibel A (2004) Adaptation for soft whisper recognition using a throat microphone. Proc Interspeech 2004

Kalaiselvi K, Vishnupriya MS (2014) Non-audible murmur (NAM) voice conversion by wavelet transform. Int. J.

Kalgaonkar K, Raj B (2007) Acoustic Doppler sonar for gait recognition. 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

Kalgaonkar K, Hu RHR, Raj B (2007) Ultrasonic Doppler sensor for voice activity detection. IEEE Signal Process Lett 14:754–757. doi:10.1109/LSP.2007.896450

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.

Kroos C (2012) Evaluation of the measurement precision in three-dimensional electromagnetic articulography (Carstens AG500). J Phon 40:453–465

Lawson E, Scobbie JM, Stuart-Smith J (2015) The role of anterior lingual gesture delay in coda/r/ lenition: an ultrasound tongue imaging study. Proc 18th ICPhS

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.

Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110

Magloughlin L (2016) Accounting for variability in North American English/?: Evidence from children's articulation. J Phon 54:51–67

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746–748

McLoughlin IV (2014) The use of low-frequency ultrasound for voice activity. Proc Interspeech 2014:1553–1557

Mielke J (2011) An articulatory study of rhotic vowels in Canadian French. In: Proc. of the Canadian Acoustical Association

Miller AL, Finch KB (2011) Corrected high-frame rate anchored ultrasound with software alignment. J Speech Lang Hear Res 54:471–486

Nakajima Y (2005) Development and evaluation of soft silicone NAM microphone. In: Technical Report IEICE, SP2005-7

Nakajima Y, Kashioka H, Shikano K, Campbell N (2003a) Non-audible murmur recognition. Eurospeech 2601–2604

Nakajima Y, Kashioka H, Shikano K, Campbell N (2003b) 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

Noda K, Yamaguchi Y, Nakadai K, Okuno HG, Ogata T (2014) Lipreading using convolutional neural network. Proc Interspeech 2014

Otani M, Shimizu S, Hirahara T (2008) Vocal tract shapes of non-audible murmur production. Acoust Sci Technol 29:195–198

Perkell JS, Cohen MH, Svirsky MA, Matthies ML, Garabieta I, Jackson MTT (1992) Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. J Acoust Soc Am 92:3078–3096

Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audiovisual speech. Proc IEEE 91:1306–1326

Quatieri TF, Brady K, Messing D, Campbell JP, Campbell WM, Brandstein MS, Weinstein CJ, Tardelli JD, Gatewood PD (2006) Exploiting nonacoustic sensors for speech encoding. IEEE Trans. Audio. Speech. Lang. Processing 14. doi:10.1109/TSA.2005.855838

Raj B, Kalgaonkar K, Harrison C, Dietz P (2012) Ultrasonic Doppler sensing in HCI. IEEE Perv Comput 11:24–29. doi:10.1109/MPRV.2012.17

Scobbie JM, Wrench AA, 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.

Scott AD, Wylezinska M, Birch MJ, Miquel ME (2014) Speech MRI: morphology and function. Phys Medica 30:604–618. doi:10.1016/j.ejmp.2014.05.001

Shaikh AA, Kumar DK, Yau WC, Che Azemin MZ, 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.

Shin J, Lee J, Kim D (2011) Real-time lip reading system for isolated Korean word recognition. Pattern Recognit 44:559–571

Silva S, Teixeira A (2015) Unsupervised segmentation of the vocal tract from real-time MRI sequences. Comput Speech Lang 33:25–46. doi:10.1016/j.csl.2014.12.003

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

Stork DG, Hennecke ME (1996) Speechreading by humans and machines: models, systems, and applications. Springer, New York

Tao F, Busso C (2014) lipreading approach for isolated digits recognition under whisper and neutral speech. Proc Interspeech 2014

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.

Toth AR, 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 V-A, Bailly G, Lœvenbruck H, Toda T (2009) Multimodal HMM-based NAM-to-speech conversion. Interspeech 2009:656–659

Tran VA, 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 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

Turton D (2015) Determining categoricity in English /l/-darkening: A principal component analysis of ultrasound spline data. In: Proc. 18th ICPhS.

Vietti A, Spreafico L, Galatà V (2015) An ultrasound study of the phonetic allophony of Tyrolean/r. In: Proc. 18th ICPhS.

Wand M, Koutník J, Schmidhuber J (2016) Lipreading with long short-term memory. arXiv Prepr. arXiv1601.08188.

Wang J, Samal A, Green JR, 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, Green JR, Rudzicz F (2012b). Whole-word recognition from articulatory movements for silent speech interfaces. Proc Interspeech 2012

Wang J, Balasubramanian A, Mojica de la Vega L, Green JR, 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 JR (2014) Across-speaker articulatory normalization for speaker-independent silent speech recognition contribution of tongue lateral to consonant production. Proc Interspeech 2014:1179–1183

Whalen DH, McDonough J (2015) Taking the laboratory into the field. Annu Rev Linguist 1:395–415

Whalen DH, Iskarous K, Tiede MK, Ostry DJ, Lehnert-Lehouillier H, Vatikiotis-Bateson E, Hailey DS (2005) The haskins optically corrected ultrasound system (HOCUS). J Speech Lang Hear Res 48:543–553

Xu K, Yang Y, Stone M, Jaumard-Hakoun A, Leboullenger C, Dreyfus G, Roussel P, Denby B (2016) Robust contour tracking in ultrasound tongue image sequences. Clin. Linguist. Phon. 0, 1–15. doi:10.3109/02699206.2015.1110714

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.

Zhao G, Barnard M, Pietikainen M (2009) Lipreading with local spatiotemporal descriptors. IEEE Trans Multimedia 11:1254–1265

Zharkova N, Hewlett N (2009) Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /a/. J. Phon. 37:248–256. doi:http://dx.doi.org/10.1016/j.wocn.2008.10.005

Zharkova N, Hewlett N, Hardcastle WJ (2012) An ultrasound study of lingual coarticulation in/s V/ syllables produced by adults and typically developing children. J Int Phon Assoc 42:193–208

Zharkova N, Gibbon FE, Hardcastle WJ (2015) Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. Clin Linguist Phon 29:249–265

Zhu B (2008) Multimodal speech recognition with ultrasonic sensors. M.Sc. Thesis, Massachusetts Institute of Technology

Zhu B, Hazen TJ, Glass JR (2007) Multimodal speech recognition with ultrasonic sensors. Interspeech 2007:662–665

Chapter 4
Combining Modalities: Multimodal SSI

Abstract In previous chapters, we have seen how various silent speech interface (SSI) modalities gather information concerning the different stages of speech production, covering brain and muscular activity, articulation, acoustics, and visual speech features. In this chapter, the reader is introduced to the combination of different modalities, not only to drive silent speech interfaces, but also to further enhance the understanding regarding emerging and promising modalities, e.g., ultrasonic Doppler. This approach poses several challenges dealing with the acquisition, synchronization, processing and analysis of the multimodal data. These challenges lead the authors to propose a framework to support research on multimodal silent speech interfaces (SSIs) and to provide concrete examples of its practical application, considering several of the SSI modalities covered in previous chapters. For each example, we propose baseline methods for comparison with the collected data.

Keywords Multimodal SSI framework • Synchronization • Multimodal data • Data processing • Feature extraction • Feature selection • Ground truth • Surface electromyography • Ultrasound • Real-time magnetic resonance imaging

In the previous chapters, we have described how information generated during different stages of speech production, from brain activity to visual speech features, can be extracted, studied, and exploited for the design and development of silent speech interfaces (SSIs). Until now, our methodology has been to consider separately sensing technologies that address one of the discussed speech production stages. Since each of these provides a limited view of speech production, a possible approach is to address the fusion of multiple of those technologies when used simultaneously. This method yields a richer set of information and enables the minimization of the limitations exhibited by each individual modality, extensively mentioned in Chaps. 2 and 3.

In this chapter, we introduce the notion of multimodal SSI, which brings its own challenges, starting from how to synchronously acquire data from the considered sensing technologies, through to the processing and analysis of datasets with large dimensionalities in feature space, up to its fusion and practical use for silent speech interfaces.

This chapter presents a brief overview of the state-of-the-art regarding the simultaneous use of multiple modalities for silent speech interfaces, followed by a more detailed description of a multimodal framework developed by the authors, targeted to

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_4

support research on multimodal silent speech interfaces. The different characteristics of this framework are illustrated, without loss of generality, by an application of SSI to the European Portuguese language case. Our framework was designed to enable synchronized acquisition of data from the following modalities: SEMG, UDS, video, depth information, and audio. It also supports the innovative use of data collected from real-time MRI and ultrasound, used as "ground truth" modalities, i.e., modalities that provide baseline articulatory data to enable the analysis of less understood technologies such as UDS.

4.1 Silent Speech Interfaces Using Multiple Modalities

A literature survey on the topic of multimodal SSI finds a small number of studies addressing more than one modality (excluding audio) in SSI research. However, analyzing the trend of recent years, we see an increasing amount of multimodal initiatives.

Overall, results show that the use of more than one source of information improves the results of individual modalities. In the SSI field in particular, a first experiment was reported by Denby and Stone (2004), where two input modalities, in addition to speech audio, were used to develop an SSI. The authors employed ultrasound imaging of the tongue, lip profile video, and acoustic speech data with the goal of developing such 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). In 2006, a patent was granted for a helmet with video, audio, and ultrasonic data input, designed to increase transcription accuracy and/or to process silent speech (Lahr 2006). More recently, Florescu et al. (2010), using US and video, achieved a 65.3 % recognition rate only considering silent word articulation in an isolated word recognition scenario with a 50-word vocabulary using a classifier based on dynamic time warping (DTW). The reported approach also attributes substantially more importance to the tongue information, only considering a 30 % weight during classification for the lip information.

Using other modalities, Tran et al. (2009) reported a preliminary approach based on 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 authors (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) data (Galatas et al. 2012a), showing that the facial depth information can improve the system performance over audio-only and traditional audio–visual systems. On the importance of these visual features, 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" (i.e., visual expression of phonemes) 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.

The book authors have also contributed with several studies where multiple modalities were combined. These studies on multimodal data collections (Freitas et al. 2014b), feature selection (Freitas et al. 2014a), word recognition (Freitas et al. 2014a), nasal vowels (Freitas et al. 2015), and tongue detection (Freitas et al. 2014c) included other combinations of modalities, that differentiate from previous studies, not only in the mixture of SSI modalities, but also in its number, superior to prior literature experiments. The following sections and chapters will provide more details about these studies, describing how we have tackled the combination of different modalities.

4.2 Challenges: Which Modalities and How to Combine Them

When considering scenarios of natural HCI, input modalities such as SEMG and video, although distinct, hold important and complementary characteristics, like being noninvasive and capturing different stages of speech production, which are important characteristics to achieve the established objectives. It can nevertheless be argued whether SEMG and/or PMA are actually invasive when compared to other modalities. Hence, 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 implies some concerns and medical expertise.

Another challenge is how to improve performance of SSI systems. Despite the efforts and the available technology, the performance attained by SSI is still low when compared with ASR based solely on the acoustic signal. Thus, to achieve 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 noninvasive 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, if we use only SEMG,

it becomes unclear which tongue movements are actually being detected, and there is not enough information to derive which tongue movements are occurring during silent speech production (Freitas et al. 2014c).

To better understand the capabilities of each modality, we need reliable speech production data. Taking one of the main articulators in the speech production process, the tongue, as an example (Seikel et al. 2009), several technological alternatives that collected uttered data can be found in the literature: RT-MRI (Narayanan et al. 2011), US (Scobbie et al. 2008; Stone and Lundberg 1996; Hofe et al. 2013). Results from these authors have the potential to be used for further exploitation of the capabilities of other modalities, such as SEMG.

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 strengths of each one completing the other, and solving the weaknesses of every modality in the set, when considered individually.
- Synchronize data acquisition across modalities and ensure proper conditions for conducting correlational statistical studies.
- Include modalities that enable direct measures during different stages of speech production and during the motion of articulators.
- Collect and extract the relevant data from each modality, considering the large amounts of data involved.
- Find the best way to fuse or use the collected data.
- Analyze and/or classify high-dimensional cross-modality data.

In recent contributions (Freitas et al. 2014a, 2015), the authors have addressed SSI research considering a large set of modalities including the synchronous acquisition of ultrasound, video and depth information, SEMG, ultrasonic Doppler, and speech audio. The number of modalities involved and the heterogeneous nature of the resulting data have motivated the design and development of a dedicated framework to support the conceptual and technical aspects involved and to provide further research in multimodal SSI.

4.3 A Framework for Multimodal SSI

The proposed framework, supporting research in multimodal SSI, can be defined by a set of five stages, as depicted in Fig. 4.1:

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 SSI modalities

The final result can be applied to HCI and to address existing problems, such as nasality detection.

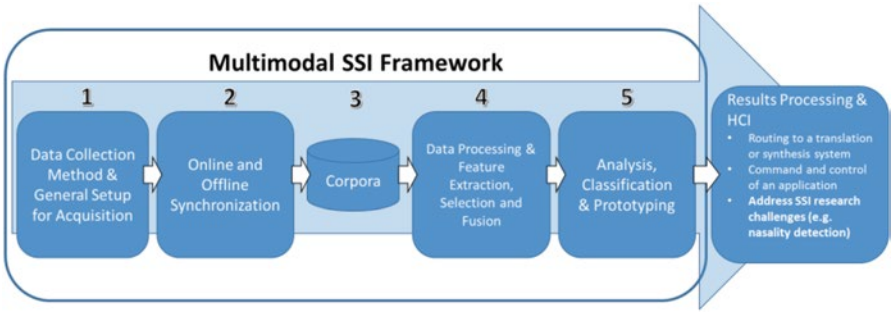

Fig. 4.1 Overview the proposed multimodal SSI framework with multiple stages and possible outcomes

The stages that constitute our framework are structured in a similar manner to conventional machine learning or ASR systems pipelines, and the difference between ours and these systems relies on its multimodal nature.

When first approaching the silent speech recognition problem (represented as stage 1 in Fig. 4.1), 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 the adoption of a single SSI modality is not sufficient to tackle the interaction barriers.

After defining our multimodal SSI method and the different components of our setup, we need to collect data samples. This is usually a cumbersome and time-consuming procedure, particularly for global-data techniques that try to generalize the data observations. To avoid future data collections, we need to ensure, not just the maximization of the amount of data extracted from different modalities, but also a synchronous acquisition of such data (represented as stage 2 in Fig. 4.1). Thus, our requirement is for an extendable data collection approach which includes, effortlessly, multiple modalities.

After acquiring the multimodal corpora, some operations related with the storage and processing of the respective metadata (e.g., annotation) may be needed. This is depicted as the intermediate stage 3 in Fig. 4.1.

At a posterior stage, after storing and processing the corpora, we need to transform the data into usable information, from an applied research perspective. This is depicted in Fig. 4.1 as stage 4 and includes, for each modality: processing the raw data; time alignment, enabling a synchronous relation between data coming from various modalities; extracting relevant information from the data; selecting the best data features, consequently reducing the dimensionality, which is beneficial for posterior classification stages; and, when applicable, fuse cross-modalities information.

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

In the following sections, we describe stages 1, 2, and 4 of the framework in more detail. Examples of stage 3—collected corpora—will be included in Chap. 5.

4.3.1 *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 sensing devices. For example, during RT-MRI acquisition, it is not possible, with the current technology, to acquire SEMG data; 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 that the data streams are correctly synchronized, for the reasons mentioned earlier. Regarding the data acquisition protocol, we propose two paths for synchronizing the data streams, here referred to as online and offline synchronization (described in detail in Sect. 4.3.2). In the first and most common case, we explore the possibility of using hardware or software markers to synchronize the data across streams. However, in some cases, such as when including data captured by RT-MRI, it becomes impossible to conduct a simultaneous multimodal acquisition due to technological restrictions. Therefore, the proposed framework takes advantage of an additional offline synchronization method, as depicted in Fig. 4.2.

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

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

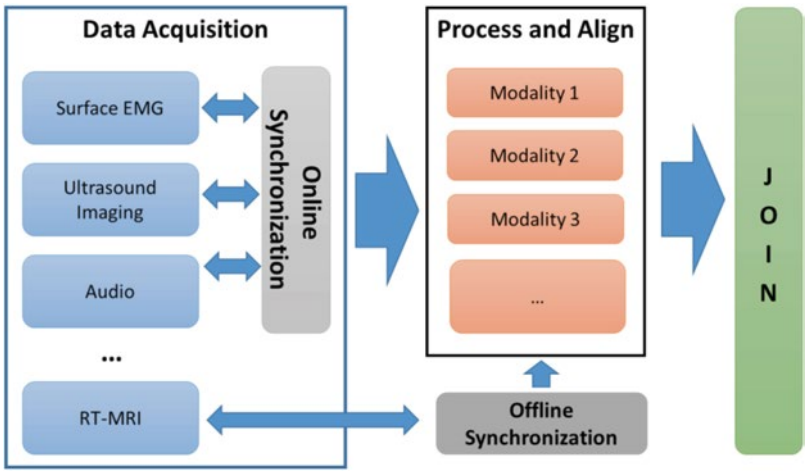

Fig. 4.2 Conceptual model of a multimodal data collection setup scenario and joint exploration of the selected modalities

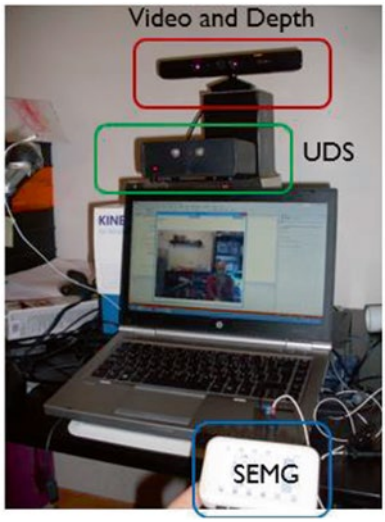

Fig. 4.3 Setup A: Acquisition devices and laptop with the multimodal data collection application running for setup A, based on video and depth information (provided by Kinect, which contains an infrared reflection sensor), ultrasonic Doppler sensing, and surface electromyography

4.3.1.1 Setup A: Base Data Collection

In line with the diagram depicted in Fig. 4.2, we can define a base multimodal data collection setup including four modalities: (1) facial information acquired from **Visual** and **Depth** sensors; (2) acquisition of muscle activity related with speech articulation, via **SEMG**; (3) capture of facial movements that occur during speech using **UDS** (Freitas et al. 2012b). Figure 4.3 depicts this first version of the data collection setup, referred to hereon as setup A.

The devices employed in setup A are the following:

- Microsoft Kinect for Windows that acquires RGB-D data.
- 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: two ultrasound transducers (400ST and 400SR working at 40 kHz), a crystal oscillator at 7.2 MHz, and frequency dividers to obtain 40 and 36 kHz, and all amplifiers and linear filters needed to process the echo signal (Freitas et al. 2012b).

We now detail some particular aspects of the configuration of setup A, which helps the reader replicating such data collection.

The Kinect sensor is placed at approximately 70.0 cm from the speaker. It is 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 640×480 pixel, and an 11-bit depth stream to code the Z dimension, with a resolution of 640×480 pixel, also at 30 frames per second. Kinect is configured to use the near depth range (i.e., range between 40.0 and 300.0 cm) and to track a seated skeleton.

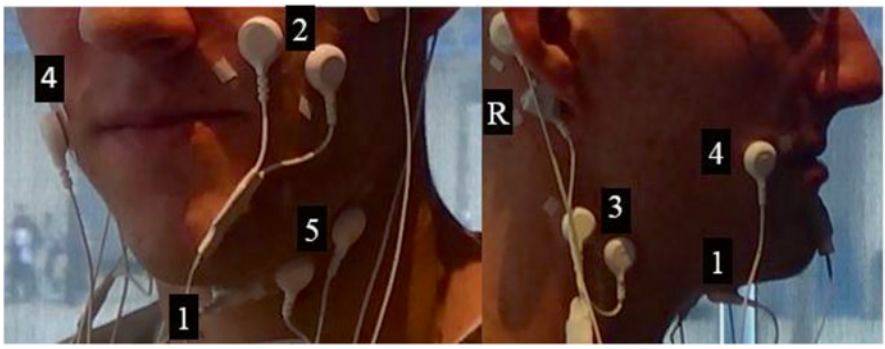

Fig. 4.4 Surface electromyography electrodes positioning and the respective channels (1–5) plus the reference electrode (R) in setup A

The SEMG acquisition system uses five pairs of EMG surface electrodes connected to a device that communicates with a computer via Bluetooth. As depicted in Fig. 4.4, each sensor is attached to the skin using a single use 2.5 cm diameter clear plastic self-adhesive surface, and also considering approximately 2.0 cm spacing between the electrodes center, for bipolar configurations.

For this setup, the five electrode pairs are 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 is placed below the ear between the mastoid process and the mandible. The SEMG channels 1 and 4 use 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 is based on previous work (e.g., Wand and Schultz 2011) and EMG electrode from channel 3 is placed according to findings regarding the detection of nasality in SSI (Freitas et al. 2014d).

4.3.1.2 Setup B: Data Collection Including "Ground Truth" Modalities

To further study the initial set of modalities and to obtain direct measures of articulators, such as the velum and the tongue, to serve as ground truth data, we can consider an extended setup (referred to as setup B), as depicted in Fig. 4.5, which features two new modalities added to the framework: **Speech Audio** and **Ultrasound Imaging**. The available technology allows for these two modalities to be acquired simultaneously with the modalities already considered for setup A. Ultrasound imaging is specifically added to the setup to gather information about the tongue movement.

Regarding setup B the following devices are added to setup A:

- Mindray DP6900 ultrasound system with a 65EC10EA transducer, an Expresscardl54 Video capture card, to capture the ultrasound video

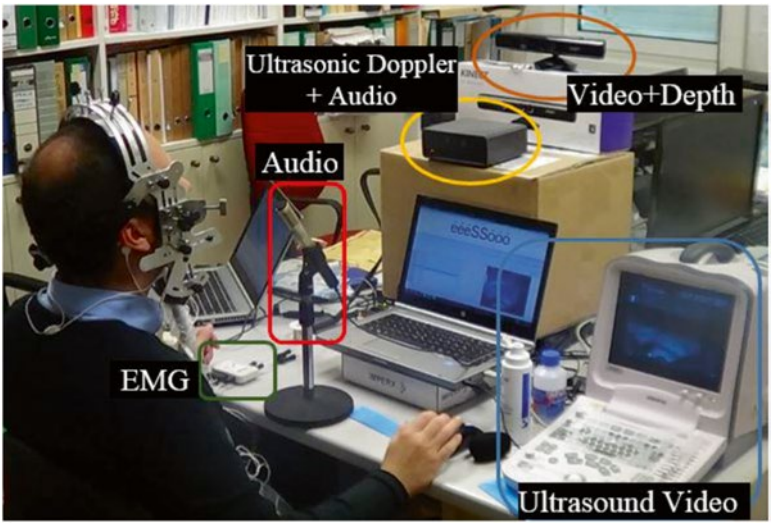

Fig. 4.5 Setup B: Augmented multimodal acquisition setup, based on video and depth information, ultrasonic Doppler sensing, surface electromyography, ultrasound imaging, and speech audio

- SyncBrightUp Audio–Video synchronization unit
- Directional microphone

To support all the devices listed in setup A and B, we use two sound cards, a TASCAM US-1641 (main sound board), a Roland UA-25 EX (secondary sound board), and two laptops. The hardware connection of all devices is depicted in the schema of Fig. 4.6.

For setup B, in order to capture tongue information, the same five pairs of SEMG electrodes are placed in the areas of the neck beneath the chin, somewhat limited by the ultrasound probe placed beneath the chin, as depicted in Fig. 4.7.

The UDS device should be placed at approximately 40.0 cm from the speaker and is connected to the main sound board, which in turn is connected to the laptop through a USB connection. The Doppler echo and the synchronization signals are sampled at 44.1 kHz and to facilitate signal processing, a frequency translation is 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 is used (Scobbie et al. 2008) to ensure that the relative position of the ultrasound probe towards the head is kept during the acquisition session, also securing the ultrasound probe below the participant's chin.

In terms of acquisition protocol, the inclusion of ultrasound imaging demands additional preparation. We start by placing the stabilization headset in the participant's

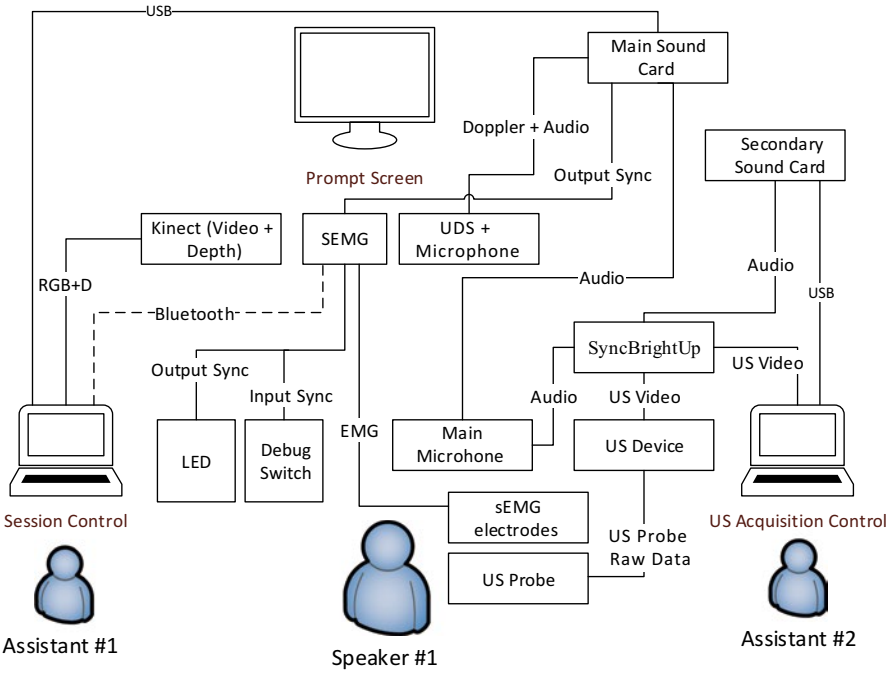

Fig. 4.6 Device schema for multimodal data acquisition of video and depth information, ultrasonic Doppler sensing, surface electromyography, ultrasound imaging and audio (two microphones)

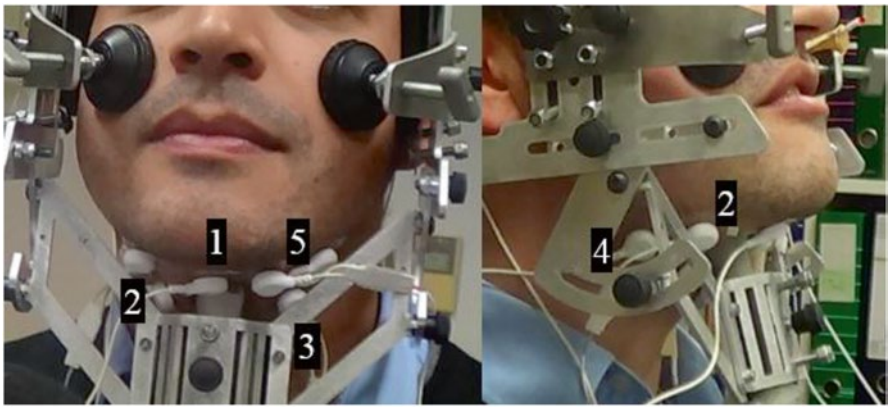

Fig. 4.7 Frontal (*left image*) and lateral (*right image*) view of the surface electromyography sensors (channels 1–5) in setup B

head and the ultrasound probe in place, below the chin, but left at some distance. This serves 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 should opt for not doing so beforehand. After SEMG sensor placement, the headset is properly secured. Finally, we should ask the participant to push the ultrasound probe against the chin as much as possible, keeping it within comfort levels, while the ultrasound is monitored to check for proper tongue imaging.

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

4.3.1.3 Scaling the Setup to Additional Modalities

With setup B, we already support capturing seven streams of data. However, to extend this 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 Fig. 4.8.

More restrictive solutions such as RT-MRI require using a separate setup and need to recur to an offline synchronization method, described in Sect. 4.3.2.1.

4.3.2 Synchronization

For a successful joint use of modalities, we need to either ensure the necessary conditions to record all signals with adequate synchronization, or have, as much as possible, time sync information at recording time to later synchronize modalities

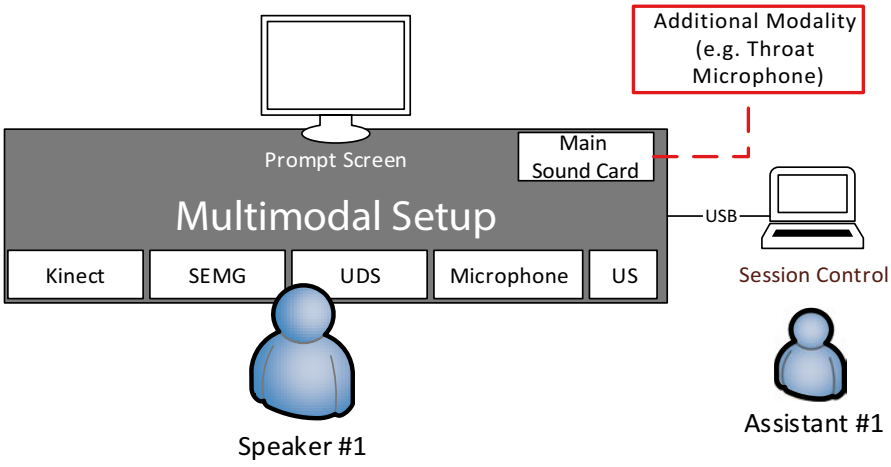

Fig. 4.8 Device scheme with an additional modality (*dashed red line*)

that cannot be recorded simultaneously. The challenge of synchronizing all signals resides in the fact that an effective synchronization event needs to be captured simultaneously by all input modalities.

In order to synchronize all input modalities from setup A (video, depth, SEMG, and UDS) via time alignment between all corresponding input streams, the framework instantiation uses 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 Fig. 4.6. Synchronization occurs when the output of a synch signal, programmed to be automatically emitted by the SEMG device at the beginning of each prompt, is used to drive a light-emitting diode (LED) and to provide an additional audio channel in an external sound card. The alignment between the video and depth streams is ensured by the Kinect SDK.

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

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

In setup B, the acquisition of ultrasound-related data is managed by Articulate Assistant Advanced (AAA) (Instruments A 2014), which is also responsible for recording audio and ultrasound video and triggering the SynchBrightUp unit. The SyncBrightUp unit, when triggered, introduces 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 is 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 use the audio signal provided by the SyncBrightUp unit, which contains the synchronization pulses, and we also record it using the main sound card along with the pulse emitted by the EMG device. The result was that we get the audio signal with the synchronization pulses recorded synchronously with the remaining data in both settings (SEMG and US). The audio signal is then used for synchronization between them.

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

4.3.2.1 Offline Synchronization

Offline synchronization comes into play when it is not possible to perform a simultaneous acquisition of modalities enabling a more straightforward synchronization. This happens, for example, with RT-MRI, which entails the use of a scanner inside which no metal is allowed.

As an example of offline synchronization, and to be able to profit from RT-MRI data, 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 audio 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 determine their correspondence. For that purpose, we use the DTW 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 us to take advantage of RT-MRI data, providing a greater insight on what is happening to the different articulators, and serving as ground truth data for the exploration of less known modalities. This offline synchronization enables studies such as Freitas et al. (2015).

4.3.3 Data Processing, Feature Extraction, and Feature Selection

After the multimodal data acquisition stage, we need to process the 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 examples of processing methods made available for the several modalities supported in the described setups (A and B), recently applied by the authors to SSI (Freitas et al. 2014a).

4.3.3.1 Processing Ultrasound Data for Ground Truth Definition

We have seen that the identification of tongue movement is quite important in SSI research. To this aim, ultrasound video data that observes such motion can be processed to identify the segments where tongue movement is present. The goal is then to examine the video sequence and annotate (tag) those segments where the tongue is in motion. Given the large amount of US data, it is important to ensure that the same criteria is used for movement annotation in all sequences, which is hard to accomplish when considering manual annotation. This limitation led us to consider an automatic annotation approach (Silva and Teixeira 2014). The inter-frame difference between each pair of video frames in the sequence is obtained by computing the difference between corresponding pixels. The pixel-wise differences are added

and the result is used as an indication of the amount of movement happening between the two frames. By computing these differences along the whole US sequence, we can conclude that when the tongue changes position, the inter-frame difference rises and this simple feature can be used, for example, to devise how SEMG sensors provide relevant tongue-related data. For more details about the proposed method, we forward the reader to Silva and Teixeira (2014).

4.3.3.2 Feature Extraction

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

Surface Electromyography: For feature extraction from SEMG, we used an approach that is based on temporal features, similar to the one described in (Freitas et al. 2012a, 2014a). For the particular case of the tongue gesture data (from setup B), the SEMG signals were first normalized, then a 50 Hz notch filter was applied to the signals, and they were also filtered using single spectrum analysis (SSA). The five features described in Freitas et al. (2012a) were extracted for each SEMG signal frame of 30 ms, and a frame shift of 10 ms was considered. A context width of 15 frames was used, generating a final feature vector of 155 dimensions per signal channel (5 features × 31 frames). Finally, we stacked all the channels in a single feature vector of 775 dimensions (5 channels × 155).

Ultrasonic Doppler Sensing: For UDS, we followed a similar approach to what is described in (Freitas et al. 2014a). Additionally, since in some cases no acoustic signal existed (e.g., first acquisition round for setup A, described in Sect. 4.3.3), we used the UDS signal to understand if facial movement occurred. For this purpose, we adopted a movement detection algorithm, inspired in the work of Kalgaonkar et al. (2007), that uses the energy of the UDS preprocessed 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 Fig. 4.9. Then, we applied a threshold to the resulting signal (depicted in the center plot of Fig. 4.9). 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 binary 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 Fig. 4.9).

Video and Depth: For RGB data, we extracted two types of features; the first used appearance-based methods applied to the RBG image, while in the second we

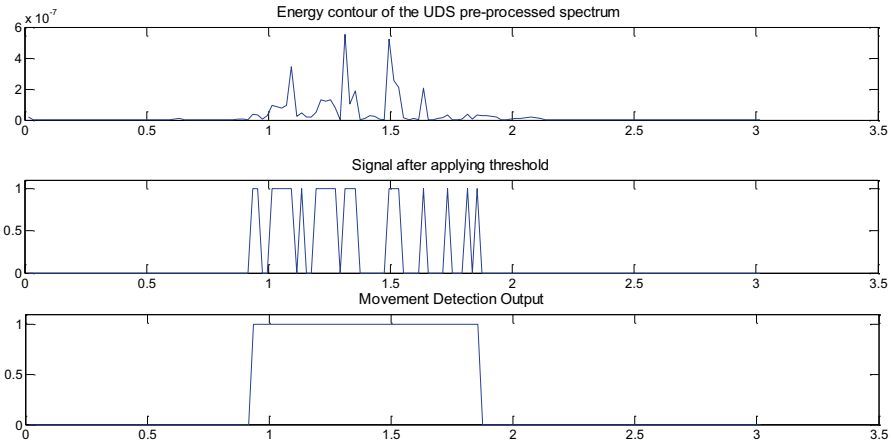

Fig. 4.9 Movement detection processing phases (from *top* to *bottom*) for the prompt "Voltar" ([voltar], to return): energy contour of the ultrasonic Doppler sensing preprocessed signal; signal after applying a threshold value and the signal output

extracted articulatory information from the lips (i.e., lip configuration). For depth, we have only used the extraction method based on appearance, applied to the "gray-scale" depth image.

For the first type of features, **appearance-based 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 64×64 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 64×64 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).

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.

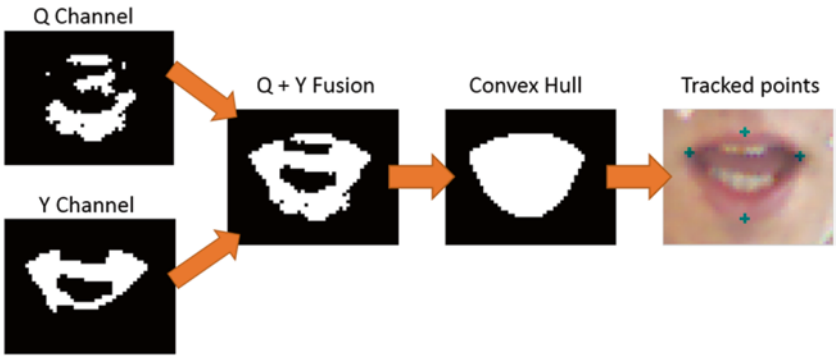

Fig. 4.10 Lips segmentation process based on the YIQ (luminance, in-phase, quadrature) color space of the region-of-interest. After noise removal, the binary representations of the Q and Y channels are fused and the convex hull of the resulting mask is determined, allowing the computation of relevant points in the lips

We developed a simple application that performs lip tracking (Abreu 2014), using four 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 may take some time (i.e., frames) to converge correctly, and are not able to keep up with some lip configurations during speech. Therefore, the tracked points although useful to derive an ROI are not appropriate to determine their coordinates.

The next step consists in the lips segmentation process (depicted in Fig. 4.10), which starts by converting the ROI to the YIQ (luminance, in-phase, quadrature) color space.

The Y channel, which 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 channel, 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 four points and to extract the characteristic width and height of the lips, computing the Euclidean distances between them (see Fig. 4.10).

As an example, Fig. 4.11 shows three plots of the lips' width and height variations for the words "Voltar" ([vOltar], to return), "Cato" ([katu], cactus), and "Ligar" ([ligar], turn on). 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 nonspeaking 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 Fig. 4.11.

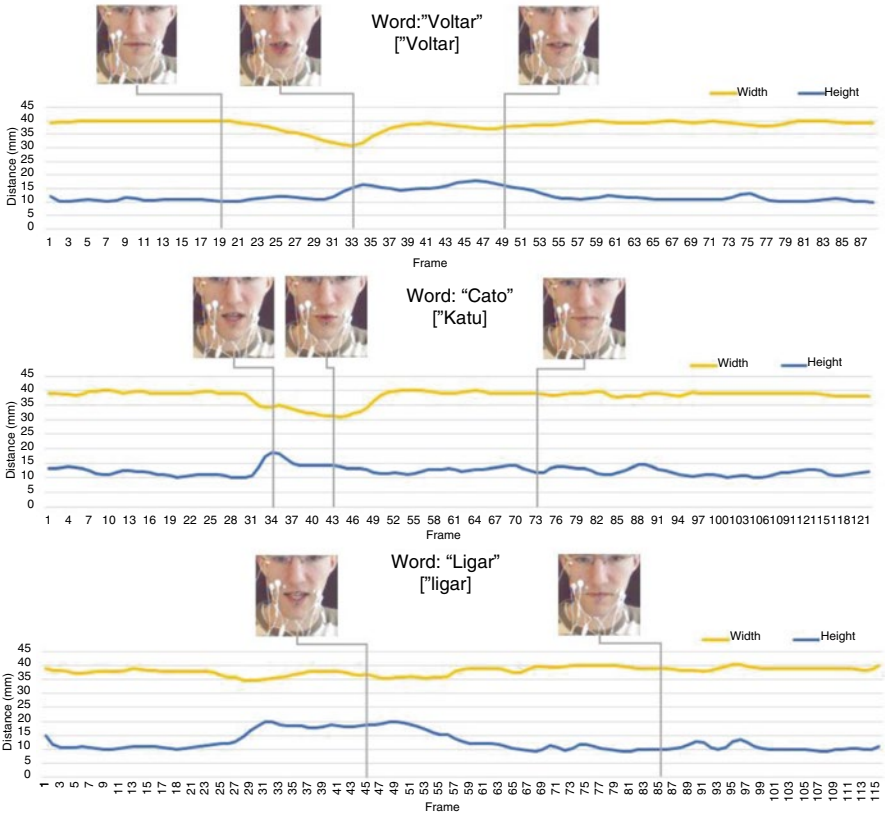

Fig. 4.11 Width and height characteristic values for the lips computed along the production of the words "Voltar," "Cato," and "Ligar." The presented image frames depict frames showing notable lip configurations

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."

4.3.3.3 Feature Selection

After the feature extraction phase, we obtain a large amount of data in our corpora. At this point, we need to apply a feature selection (FS) method in order to reduce the dimensionality of the input feature space (for the subsequent classification phase of our method). This generally allows us to achieve better learning performance with the data.

Regarding FS, several techniques are available in the literature and from these, we have selected two unsupervised and two supervised relevance measures

(Freitas et al. 2014a) based on their running-time and promising results with high-dimensional data (Ferreira and Figueiredo 2012). For the unsupervised case, we considered:

1. 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).
2. 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:

1. Shannon's mutual information (MI) (Cover and Thomas 2005), which measures the dependency between two random variables; and
2. Fisher's ratio (Fisher 1936), which measures the dispersion among classes.

To find 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 X_i and X_j, given by:

$$\mathrm{AC}_{ij} = \left|\cos\theta_{ij}\right| = \left| < X_i , X_j > \middle/ \left\| X_i \right\| \left\| X_j \right\| \right| \tag{4.1}$$

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).

4.4 Conclusions

Multimodal approaches improve SSI development, by enabling the access to data depicting multiple views, minimizing the problems found with each sensing technology and providing a wide panorama of speech production.

The effort of designing a framework to support the research in multimodal SSI proved to be relevant in addressing the challenges brought by our lack of understanding of how the different modalities can be used jointly. Furthermore, we consider it also as an essential contribution to the research agenda, since it enables faster deployment of new experiments and data collection setups, allowing synchronized acquisition and processing of several modalities, and reuse of previously acquired data.

The proposed multimodal SSI framework was instantiated using two different setups that share the important requirement of synchronous acquisition of all data streams. These two illustrative applications of the framework are an evidence of the versatility of our proposal to address the challenges that motivated our work. The setups presented as example instantiations of the framework 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 also be easily included, as long as synchronization is ensured.

The methods herein presented to process the acquired data consider the different particularities of each SSI modality and are an important component of the framework. These methods allow us to focus on key aspects of the data (e.g., a specific articulator) provide means to reduce the data dimensionality and create the conditions to add "meaning" to the data, with annotations of their relevant segments. With such approach, we are able, for instance, to extract velum movement curves from RT-MRI image sequences or to characterize the movement of the 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 serving as examples of what is at stake in each stage. As another example, and 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 articulators' movements. 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 noninvasive modalities (e.g., SEMG) and then perform classification.

The systematic approach to multimodal SSI, based on the proposed framework, also enables a more structured introduction into the technical aspects involved in the underlying multimodal data processing pipeline. Therefore, on the following chapter, the reader is introduced to the basics of actually developing an SSI including, for example, a hands-on tutorial to build a simple system based on data gathered using a video camera.

References

Abreu H (2014) Visual speech recognition for European Portuguese, M.Sc. thesis. University of Minho, Portugal

Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 23:681–685. doi:10.1109/34.927467

Cover TM, Thomas JA (2005) Elements of information theory. Wiley, New York. doi:10.1002/047174882X

Denby, B (2013. Down with Sound, the Story of Silent Speech. In: Workshop on Speech production in automatic speech recognition

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

Dubois C, Otzenberger H, Gounot D, Sock R, Metz-Lutz M-N (2012) Visemic processing in audio-visual discrimination of natural speech: a simultaneous fMRI–EEG study. Neuropsychologia 50:1316–1326

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 RA (1936) The use of multiple measurements in taxonomic problems. Ann Hum Genet 7:179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Florescu VM, 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

Freitas J, Teixeira A, Dias MS (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, Vaz F, Dias MS (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, pp 227–236. doi:10.1007/978-3-642-35292-8_24

Freitas J, Ferreira A, Figueiredo M, Teixeira A, Dias MS (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 MS (2014b) Multimodal corpora for silent speech interaction. In: 9th Language resources and evaluation conference, pp 1–5

Freitas J, Teixeira A, Silva S, Oliveira C, Dias MS (2014c) Assessing the applicability of surface EMG to tongue gesture detection. In: Proceedings of IberSPEECH 2014, lecture notes in artificial intelligence (LNAI). Springer, Berlin, pp 189–198

Freitas J, Teixeira A, Silva S, Oliveira C, Dias MS (2014d) 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 MS (2015) Detecting nasal vowels in speech interfaces based on surface electromyography. PLoS One 10, e0127040. doi:10.1371/journal.pone.0127040

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

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

Hofe R, Bai J, Cheah LA, Ell SR, Gilbert JM, Moore RK, Green PD (2013) Performance of the MVOCA silent speech interface across multiple speakers. In: Proc. of Interspeech, 2013, pp. 1140–1143

Holzrichter JF, Burnett GC, Ng LC, Lea WA (1998) Speech articulator measurements using low power EM-wave sensors. J Acoust Soc Am. doi:10.1121/1.421133

Instruments, A (2014) Articulate assistant advanced ultrasound module user manual, Revision 212. Articulate Instruments, Edinburgh

Kalgaonkar K, Hu RHR, Raj B (2007) Ultrasonic Doppler sensor for voice activity detection. IEEE Signal Proc Lett 14:754–757. doi:10.1109/LSP.2007.896450

Lahr RJ (2006) Head-worn, trimodal device to increase transcription accuracy in a voice recognition system and to process unvocalized speech. US 7082393 B2

Narayanan S, Bresch E, Ghoosh P, Goldstein L, Katsamanis A, Kim Y, Lammert AC, Proctor M, Ramanarayanan V, Zhu Y (2011) A multimodal real-time MRI articulatory corpus for speech research. In: Proc. Interspeech, 2011, pp. 837–840

Oppenheim AV, Schafer RW, Buck JR (1999) Discrete time signal processing. Prentice-Hall, Upper Saddle River

Plux Wireless Biosignals (n.d.) www.plux.info/. Accessed 30 October 2014

Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audiovisual speech. Proc IEEE 91:1306–1326

Scobbie JM, Wrench AA, 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

Silva S, Teixeira A (2014) Automatic annotation of an ultrasound corpus for studying tongue movement. In: Proc. ICIAR, LNCS 8814. Springer, Vilamoura, pp. 469–476

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

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

Tran V-A, Bailly G, Lœvenbruck H, Toda T (2009) Multimodal HMM-based NAM-to-speech conversion. Interspeech 2009:656–659

Tran VA, 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

Wand M, Schultz T (2011) Session-independent EMG-based Speech Recognition. In: International conference on bio-inspired systems and signal processing (BIOSIGNALS 2011), pp 295–300

Yau WC, Arjunan SP, Kumar DK (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

Chapter 5
Application Examples

Abstract In the previous chapters, we covered the main concepts and technologies behind silent speech interfaces (SSIs). While that material provides the reader with the necessary knowledge to understand the different concepts and proposed solutions, the technical aspects behind designing and developing interactive SSI systems are still a challenge given that they involve articulating different technologies and methods. In this chapter, we provide some examples to allow the reader to go from theory to practice. We start with a tutorial of a basic visual speech recognition system, using accessible hardware and resources. Then, we describe a more complex practical example that shows how to leverage the multimodal SSI concept introduced in Chap. 4. With this illustration, the reader has the opportunity to assess, hands-on, the capabilities of surface electromyography sensors. The last part of the chapter describes the creation of an SSI system that handles live data.

Keywords SSI applications • Tutorial • Visual speech recognition • Surface electromyography • Multimodal • Resources

In the first chapters of this text, we have provided the reader with the theoretical background and a review of the state-of-the-art in silent speech interfaces (SSIs). At this point, it becomes important to complement this knowledge with a more practical experience. Thus, in this chapter, we provide examples of applications and pointers to associated resources, which allow the reader to attain hands-on experience in SSI.

The first example, in the form of a tutorial, describes the steps required for building a small vocabulary SSI based on RGB images of the face. As long as the reader has access to MATLAB (The MathWorks 2013a) or Octave, he/she should be able to go through the tutorial and build his/her first SSI system.

The second example is more complex and requires access to SEMG sensors, as well as to an ultrasound imaging device, thus it will be harder, although still viable, for the reader to recreate this experience. Nevertheless, the objective is to

Electronic supplementary material: The online version of this chapter (doi:10.1007/978-3-319-40174-4_5) contains supplementary material, which is available to authorized users.

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_5

demonstrate how to use ground truth information obtained from specific modalities, so that the same recipe can be later applied to other scenarios.

The chapter ends with the brief presentation of a third practical example, an SSI system that handles live data.

5.1 Basic Tutorial: How to Build a Simple Video-Based SSI Recognizer

In this section, we cover the basic steps to create and test an interactive system based on a single SSI modality and capable of recognizing a small set of words.

To create a context for this example, the reader may consider an Ambient Assisted Living (AAL) scenario where we place our persona (targeted user) at home, working with a laptop, and where we can imagine such persona sitting back in a chair, watching a slideshow and invoking some of the control commands usually issued in that context. To keep the system and the tutorial simple, we will restrict it to support only five single-word commands ("Call", "Email", "Search", "Next", "Previous").

Generally speaking, to serve the human–computer interaction (HCI) requirements of this scenario, we could have designed a solution that considered speech recognition based on the audio signal, as the input HCI modality. However, in line with the topic of this book, we have considered a modality that more clearly serves the purpose of SSI design. Due to its computational simplicity and general availability, RGB video imaging of the face region is a good candidate and was chosen as the SSI modality for this basic tutorial case. This data stream can be collected from numerous sources, such as a simple external video camera, a laptop camera or, as in the case of this tutorial, a Microsoft Kinect device.

The main objective of this section is to provide a hands-on experience on how to build a prototype for a single modality SSI system without the need to have access to specific equipment. For that purpose, we provide both code[1] and small example databases that will be considered along this tutorial to explain the main requirements, processing steps, and expected outcomes.

We have adopted the MATLAB scripting language,[2] due to its simplicity and common use in the literature. MATLAB is 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. After reading this tutorial, the reader is encouraged to explore the provided scripts and examples to handle other datasets or use alternative processing methods.

[1] This code is provided as is, for illustrative purposes only, without any guarantee, and its use is the sole responsibility of the reader. Please refer to the source code for additional information regarding licensing.

[2] With simple adaptations the code can also be used in Octave.

5.1.1 Requirements and Setup

This tutorial was designed to provide the reader with a general idea of what is involved even if read without considering the companion scripts and databases. However, we still encourage the reader to consider both text and scripts, when addressing this chapter. To get the scripts running, these are the requirements that the reader needs to match:

1. An installation of MATLAB (Hanselman and Littlefield 2005; The MathWorks 2013b) with no particular demands regarding its version. The provided scripts have been tested in versions 2007, 2012, and 2013 without any issues.
2. The following MATLAB toolboxes: namely Image Processing Toolbox (The MathWorks n.d.), Voicebox (Brookes 1997) and MATLAB Toolbox for Dimensionality Reduction (v0.8.1b) (Van der Maaten et al. 2007). No action is needed regarding these last ones.
3. Have the support library LibSVM[3] installed and running on the reader's computer. For simplicity, this is included in the tutorial package, in folder libsvm-3.21, and its use should be straightforward to the user. However, any other version of the library should work. In case issues are found, the reader is forwarded to Chang and Lin (2011) and Hsu et al. (2003).
4. Download and extract the file TUTORIAL_SSI.zip from http://extras.springer.com. For the sake of simplicity, we recommend extracting it to a folder such as C:\SSI or C:\TUTORIAL, but any other folder should work too.
5. Change, in MATLAB, the working directory to the main directory created during the unzipping process of the tutorial code. For example, if extracted to C:\SSI the reader only needs to use "cd c:\ssi" in the MATLAB Command Window.

5.1.2 Overall Presentation of the Pipeline

The creation of an SSI recognizer directly follows the approach commonly used in the areas of pattern recognition and machine learning: first, features are extracted from a database of recorded examples, often referred to as the training set; then, with these features at hand, a model is trained; and, finally, the model is tested/evaluated on a different set of recordings, the test set, considering features extracted in a similar way to the ones used to train the model. In a real situation, a trained model that shows good enough results in the evaluation stage can be used to process features extracted from new data, possibly from a live stream (out of the scope of this tutorial). The complete pipeline is illustrated in Fig. 5.1.

[3] LibSVM package is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

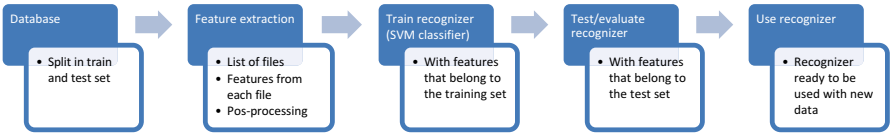

Fig. 5.1 Stages to create the recognizer

Due to the small amount of data and good results reported in many similar tasks, we have adopted SVM classifiers (Burges 1998), considering the LIBSVM implementation and its wrapper functions for MATLAB (Chang and Lin 2011). SVM classifiers are based on the concept of decision planes that separate between a set of objects from different classes. Since most classification tasks demand complex structures in order to make an optimal separation, the original objects are mapped using a set of mathematical functions, known as kernels. The mapped objects are then linearly separable and all that is needed is to find an optimal line that can separate the classes. In this tutorial, we assume the reader has minimum knowledge of how SVM classifiers work. Readers lacking information regarding SVM may find (Burges 1998; Hearst et al. 1998; Steinwart and Christmann 2008) useful.

The complete pipeline of this tutorial, encompassing the four first stages depicted in Fig. 5.1, can be easily run in MATLAB using:

```
>> tutorialSSI
```

The reader may run it just now and, if everything is in place and well configured, the tutorial code will evolve through its different steps, providing several intermediate outputs. However, to understand what all that means we advise the reader to first take a look at what is happening under the hood. In the following sections, we cover in more detail the four steps involved in creating an SSI recognizer, namely, database processing, feature extraction, model training, and model evaluation.

5.1.3 Step 1: Database Processing

To serve this processing step, we provide a small database of images obtained from recordings of several repetitions of the adopted word set. The words constitute the classes for the classifiers.

The database was recorded with a Microsoft Kinect using a simple data collection tool that manages informing the speaker about the word to be pronounced (usually designated as prompt) and advances to a new word after the pronunciation is finished while detecting the region of interest (mouth). The tool collects the time of start of each utterance which is used, after completion of the video recording, to segment the video into segments of several frames, from which images (PNG format) are extracted for each video frame. All the images that correspond to each prompt are saved in a separate folder considering the following rule of thumb:

```
<RootDir>\<Speaker>\<promptID>\*.png
```

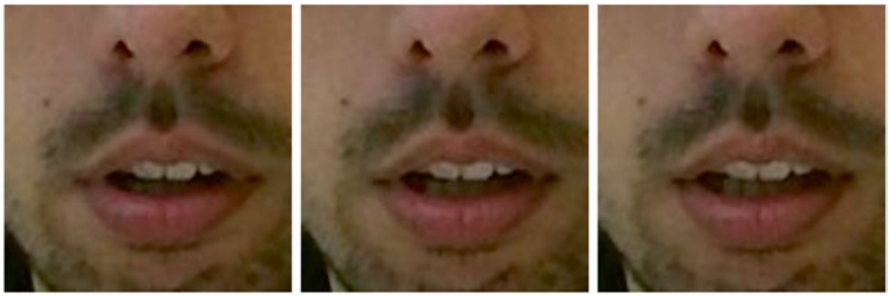

Fig. 5.2 Images extracted from the video sequence recorded for word "Next", that integrate de "TrainCorpus" of this tutorial. The images are very similar as they result from consecutive frames

For example, considering the database provided with this tutorial, in subfolder TrainCorpus, the images for the second repetition of word "Next", for speaker NV01 (the only available in this database) are stored in "`.\TrainCorpus\NV01\8-Next\`" and have sequential names such as 1-color.png, 2-color.png, etc. Examples of images from this database are depicted in Fig. 5.2.

Automatically obtaining a list of all the image files included in the database, along with their associated classes (in our case words), is essential to simplify all the processing. Given information on the root folder of the training set (or training corpus) to use, plus information on the speakers to process, and the set of words, the following function processes the database folders and returns a matrix with a row for each image including its associated prompt and class ids (in the first 2 columns of the matrix).

```
filesTrain = getFilesAndClasses(trainCorpusRoot, speakers, words);
```

This function derives the complete name of each data file and respective prompt and class ids from the names of folders and files that constitute the corpus.

5.1.4 Step 2: Feature Extraction

At this stage, data for training and testing has already been collected and is in place. It is now time to process it to extract features and store them in a matrix. We adopted the commonly used 2D discrete cosine transform (DCT) (Galatas et al. 2012; Potamianos et al. 2003) as the technique to extract features from each frame. The DCT compresses the pixel information by keeping the low spatial frequencies and selecting the first 64 coefficients contained in the upper left corner of the coefficients matrix. We have only considered the odd columns of the DCT, in order to take advantage of the facial symmetry and imposing horizontal symmetry to the image. The variation between speakers and recording conditions are smoothed by using feature mean normalization (FMN) (Potamianos et al. 2003).

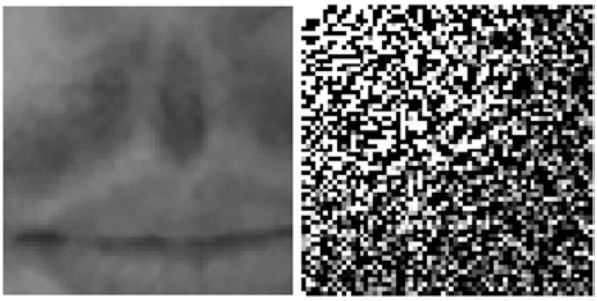

Fig. 5.3 An example of DCT feature extraction

In this tutorial, features of a dataset are extracted by calling:

```
features = extractFeatures(filesTrain, numberDCTCoeffs);
```

This last call uses as input the *filesTrain* matrix produced in Step 1 and the number of DCT coefficients to be computed.

During the process, the function provides some visual feedback of the processing, by showing, side by side, images from the database and of the extracted 2D DCT (see Fig. 5.3 for an example).

After feature extraction, we obtain a matrix with one row for each image (corresponding to a video frame). Each row contains information of the class (first column) and the 192 DCT features. After this stage of feature calculation, we can apply some post-processing in order to reduce dimensionality. In this tutorial, features are subjected to a two-stage LDA—linear discriminant analysis. In the first stage, we apply LDA to each frame reducing features per frame. In the second stage, LDA is applied to the stacking of adjacent frames, selecting the features with the highest eigenvalues. After applying the LDA, the first and second temporal derivatives are appended to the feature vector. Such post-processing is applied as follows:

```
featureMatrix = applyLDA(features, promptIDs, classIDs)
```

The application of LDA results in a reduction in the number of columns and provides the data transferred to the following step.

5.1.5 Step 3: Train the SSI Recognizer

With the features of the entire dataset available in a matrix (`featureMatrix`), and with the information on the corresponding classes stored in vector `trainClasses`, obtained directly from one of `filesTrain` columns, it is quite simple to train an SVM model that recognizes (classifies) pronounced prompts from a set of images whose extracted features are present in the mentioned dataset. This can be done by using the following MATLAB function call from LIBSVM:

```
svm_model = svmtrain(trainClasses, trainForClassif, '-t 0');
```

The t parameter used in the svmtrain call sets the kernel type to linear.

LIBSVM `svmtrain` implements the "one-against-one" approach for multi-class classification. For k classes, k (k-1)/2 classifiers are constructed and each one trains data from two classes.

At the end of this step, we obtain a model that is ready for testing.

5.1.6 Step 4: Test the SSI Recognizer

Before testing our recognizer, we need to have a test set of features. We considered a new set of data acquired in similar conditions and for the same words, but not considered for the training set. For the sake of simplicity, the test set reading and feature extraction steps, similarly to the process carried for the training set, uses a new function:

```
[testClasses testFeatures] = featuresFromTestSet(testCorpusRoot,
speakers, words)
```

The previous function call takes the root directory of the database to be used for test, information on speaker and words, and returns the feature matrix and a vector with class identifiers for the test set.

After this phase, the two outputs of this function are passed to *svmpredict* along with the SVM model obtained in the training step:

```
predicted_label= svmpredict(testClasses, testFeatures, svm_model)
```

As *svmpredict* outputs its predictions of the class identifier for each feature vector, the accuracy rate and confusion matrix are easy to obtain. The provided code depicts such rate and plots the confusion matrix as an image (see Fig. 5.4 for an example).

The results obtained for this example are very poor, mainly due to the reduced number of repetitions and the lack of a decent segmentation of the video. However, the main objective for this “toy” example tutorial is only to present the overall process and enable a first hands-on experience. Also, bear in mind that the obtained results are actually a fair depiction of the difficulty of the task. A more advanced tutorial, moving into a more realistic setting, is the subject of the next section.

5.1.7 Experimenting with Other Modalities

As previously mentioned, this tutorial was designed to serve the goal of being a first hands-on experiment, including the main steps required to build a basic SSI recognizer and, therefore, it is quite limited both in the number of modalities (only video) and considered data. Nevertheless, before closing this tutorial, we would like to provide the reader with some additional data sources that can be later explored in a scenario with multiple SSI modalities.

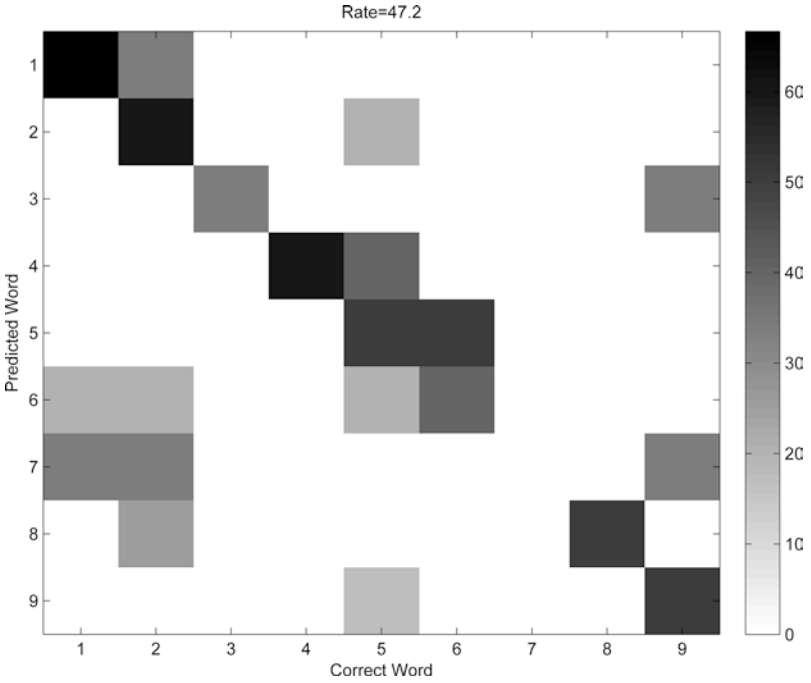

Fig. 5.4 Confusion matrix obtained for a simple SSI recognizer trained with SEMG using *tutorialSSIpart2*

Table 5.1 Column structure of the features file included with the tutorial

Column	Features of modality	Column	Features of modality
1	Word identifier (class)	6	EMG Channel 1
2	Surface EMG	7	EMG Channel 2
3	UDS (low freq.)	8	EMG Channel 3
4	Video	9	EMG Channel 4
5	Depth	10	EMG Channel 5

To keep it short and simple, we will skip the feature extraction steps (similar to the ones followed previously for the video frames) and provide the readers with precomputed features from a database comprising several modalities (collected with the framework presented earlier in this book). For each prompt, the different features are stored in different columns of a MATLAB matrix, as documented in Table 5.1. In this part of the tutorial, the words selected for recognition are the spoken numbers (“one”, “two”, etc. up to “nine”, followed by “zero”) and other words selected for AAL scenarios (Freitas et al. 2014b).

5.1.8 EMG-Based Recognizer

If we adapt the final steps (training and testing) of the script that implements the pipeline to process the features selected from a given features matrix, it is simple to create new recognizers and evaluate them.

As an example, to create a recognizer using the EMG features in column 2, simply run:

```
>> tutorialSSIpart2
```

This script loads the features from a .mat file, selects 80 % for training and 20 % for testing, trains a classifier using `svmtrain`, tests the classifier with the test set, and processes and visualizes the obtained results.

In the case shown in Fig. 5.4, the evaluation with the test set gives a rate of 50 %. The script also computes and graphically presents the confusion matrix.

Taking into consideration the other available features, simple changes to the script provided in the tutorial allow creating other SSI recognition systems. We encourage the reader to select other features and, even, combinations of features.

5.1.9 Other Experimentations

The provided features also allow experimentation with other classifiers by replacing the code relative to the function calls `svmtrain` and `svmpredict` with the adequate calls for other classifiers. In Freitas (2015), for example, DTW and kNN have also been used in addition to SVM. The code from these two MATLAB tutorials can also be used as a basis for your own modalities and datasets.

5.2 A More Elaborate Example: Assessing the Applicability of SEMG to Tongue Gesture Detection

After covering the basics of how to create a simple SSI system, we now present a more elaborate example using several modalities. Given the higher complexity of this example, the purpose is not to provide solely a hands-on approach, much like the adopted approach of the previous section, but to include more detailed information regarding the main technical choices and the interpretation of the outputs of the different stages. Nevertheless, the reader may notice that the pipeline for this example has similarities with the tutorial described in the previous section with an increase in complexity, both in considered data and methods.

The selected example (Freitas et al. 2014c) is an experimental SSI system developed by the authors to address a very specific challenge: tongue gesture detection with SEMG. The most promising approaches for SEMG-based speech interfaces

usually target tongue muscles (Schultz and Wand 2010). These studies consider data from 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 myoelectric signals are directly applied to the classification problem, in order to distinguish among different speech units (i.e., words or phonemes). Despite the interesting results obtained in small vocabulary tasks, it is yet unclear which tongue movements are actually being detected. Furthermore, there is a lack of information about different tongue movements during speech. This is due to the complex physiology of that area, the noise and the muscle cross talk observed on the face and neck. Finally, we do not know whether these movements can be correctly identified using SEMG.

To address these complex aspects, we have applied the multimodal framework described in Chap. 4. As described in Sect. 4.3.1.2 (of Chap. 4), we included US Imaging as the "ground truth" modality in our setup (details on how to connect the devices and positioning of the sensors can also be found in that section). Thus, based on synchronous acquisition of both SEMG and US of the tongue, we have addressed the applicability of EMG to tongue gesture detection. The US image sequences allowed us to gather data concerning tongue movement over time, providing the basis for the EMG analysis and for further development of more informed EMG classifiers that could be easily integrated in a multimodal SSI with an articulatory basis. This way, information about an articulator that is normally hidden, or is very difficult to extract, could be provided.

With the synchronized data, made available from the combined usage of the data collection setup and processing methods integrating our SSI framework, we present in this section an SSI detection experiment based on the probability of tongue movement. The adopted method consists in modeling the probabilistic distribution of classes' data. Then, based on these models, we derive the probability of tongue movement given the measured EMG signal. The considered classes are: "front to back movement," "back to front movement," and "non-movement." The first two represent the tongue movements found in each utterance and the third class denotes no tongue movement.

The following subsections describe the used corpora, the processing of information for analysis, and the respective results.

5.2.1 Corpora

Using setup B (described in Chap. 4, Sect. 4.3.1.2), we have synchronously acquired video, depth, UDS, and SEMG data along with ultrasound (US) imaging. In this example, for the sake of simplicity, we only cover the use of US and SEMG.

The corpus considered in this experiment was designed to cover several tongue transitions in the anterior–posterior axis (e.g., front-back and vice versa) as well as elevation and lowering of several tongue parts.

To define the corpus, we set 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).
3. Focus on a set of the most relevant tongue transitions to keep the length of the recording sessions below 30 min.

With these goals in mind, 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.

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 to two seconds of interval, yielding a total of six 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 text background turned green).

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

After acquiring the data and synchronizing all data streams, we determined and characterized the segments that contain tongue movement, based on the US data.

5.2.2 Data Processing

The raw data coming from the sensors was processed according to the methods described in Sect. 4.3.3 of Chap. 4. This section describes the US data processing (Silva and Teixeira 2014) and the SEMG feature extraction (Freitas et al. 2014a). Then, using a set of randomly selected utterances of each speaker and the automatic annotations from US, the probability mass functions for three speakers 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 the Gamma distribution were estimated using MATLAB with the Statistics Toolbox

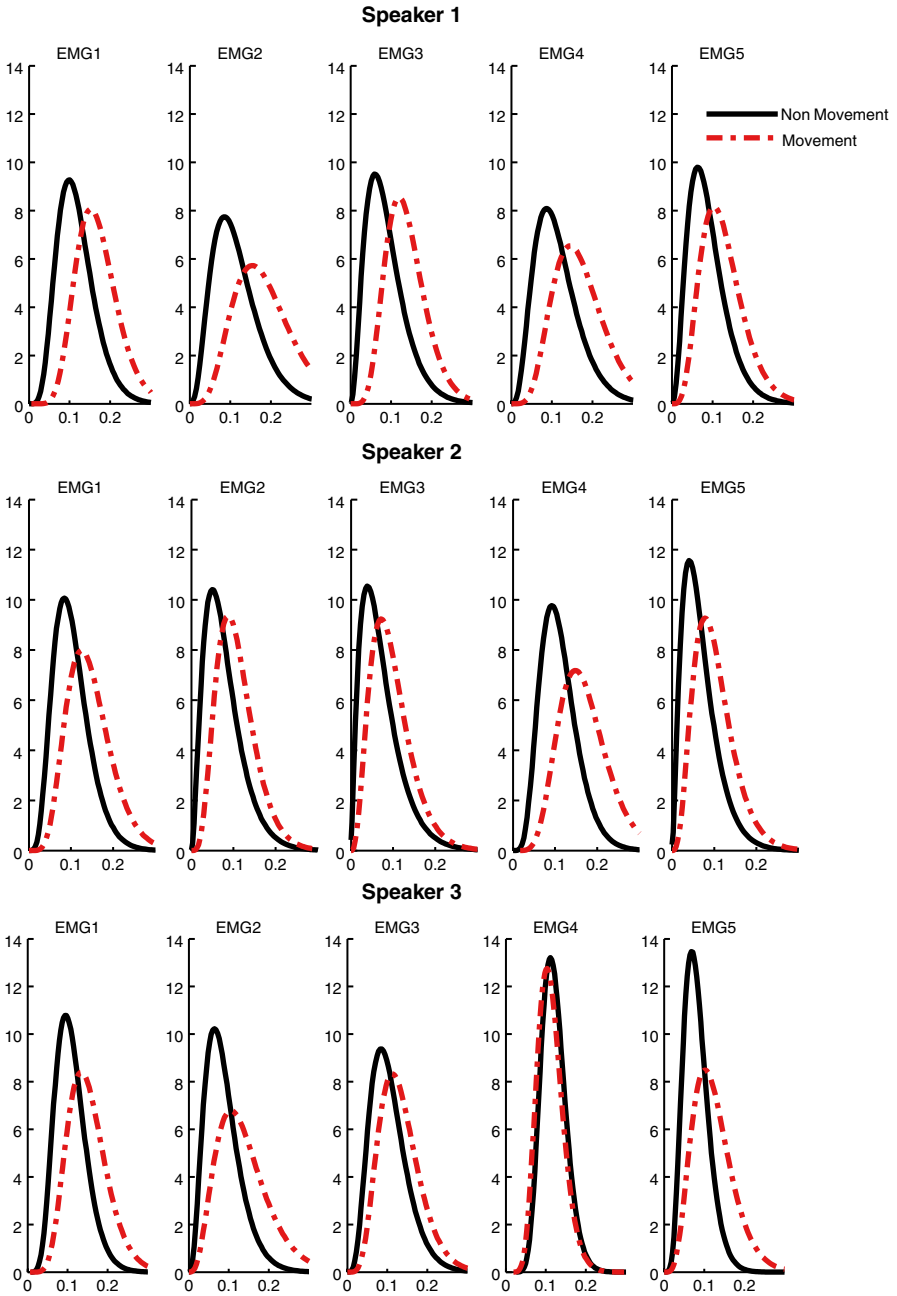

Fig. 5.5 Probability density functions for the five EMG channels of the three speakers, including curves for one of the movements (forward) and non-movement (Freitas et al. 2014c)

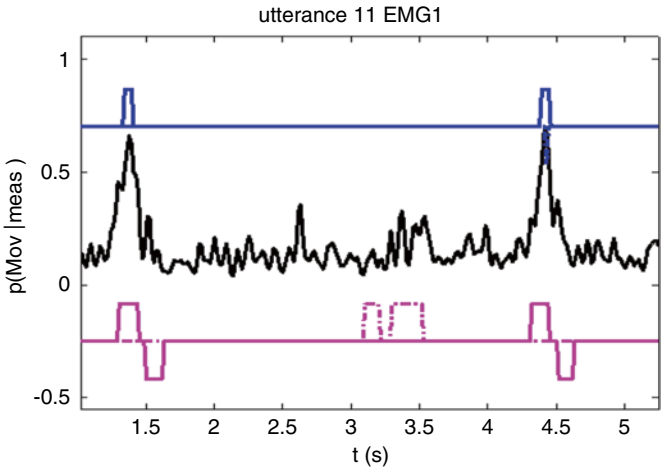

Fig. 5.6 Example of movement detection: top, the detected movements; middle, the probability of movement; and bottom, the US-based movement annotation, where forward movement is depicted by the segments represented above the middle line. The dashed line represents tongue movements not related with the utterances (Freitas et al. 2014c)

(The MathWorks 2013a, b). As depicted in Fig. 5.5, some differences in distributions between forward movement and non-movement were found for all speakers with some variations within EMG channels.

Based on the probability distribution functions, we estimated the probability of each tongue movement. Hence, considering the probability of the measured EMG (*measure*) given a movement (*movement*), stated as $p(\text{measure} \mid \text{movement})$, we can apply Bayes rules as follows:

$$p(\text{movement} \mid \text{measure}) = \frac{p(\text{measure} \mid \text{movement})\, p(\text{movement})}{p(\text{measure})} \tag{5.1}$$

Using the law of total probability to expand $p(\text{measure})$ yields:

$$p(\text{movement} \mid \text{measure}) = \frac{p(\text{measure} \mid \text{movement})\, p(\text{movement})}{(1 - p(\text{movement}))\, p(\text{measure} \mid \text{non movement}) + p(\text{movement})\, p(\text{measure} \mid \text{movement})} \tag{5.2}$$

The detection threshold was set to 0.5 and $p(\text{movement})$ to 0.3, which, based on an empirical analysis, presented a good balance between detections and false positives. Figure 5.6 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.

Table 5.2 Percentage of correct movement detections, by speaker and by SEMG sensor, and the corresponding averages

	SEMG channel					
	1	*2*	*3*	*4*	*5*	*Average*
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*	

To assess the applied technique, we compared the detection results with the US annotation in order to obtain information on correct detections (true positives), failures in detection (false negatives), and false detections (false positives). The outputs of each sample were analyzed to determine the number of correct detections and number of failures.

5.3 Results

The results as function of speaker and sensor are summarized in Table 5.2. The best results were attained by Speaker 1 in channel 3, with a detection of 80.0 % and an average of 67.1 % for the 5 channels. The other two speakers obtained the best detection result of 68.6 % and 66.8 % in SEMG channels 4 and 5, respectively.

A detailed analysis of the results reveals a strong variation of results across prompts. The best results, in terms of prompts, were achieved for [L, s, and t] in a / vCv/ context.

In terms of FP (false positives), we noticed that, although Speaker 1 presented the best results, it also showed a high rate of FP with 38.8 % of the utterances having many FPs. 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.

For the interested reader, more information can be found in (Freitas 2015; Freitas et al. 2014c).

5.4 An SSI System for a Real-World Scenario

In the previous examples, we used the multimodal SSI framework for analysis and classification of previously collected data. However, the deployment of SSI in real-world scenarios requires the ability to deal with data in real time. To that effect, this section follows an application-oriented approach for our third tutorial, proposing a modular solution for designing, developing, and testing a multimodal SSI system with live data.

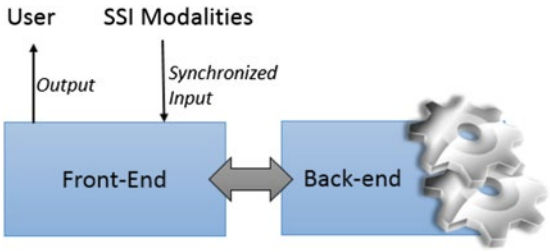

Fig. 5.7 High-level overview for a multimodal SSI system

5.4.1 Overall System Architecture

The envisaged system (depicted in Fig. 5.7) is composed by two main modules: a front end, which handles the synchronized input of SSI modalities and manages any relevant output modalities; and a back-end, where the processing of data coming from SSI modalities occurs.

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 includes data processing, extracting, and selecting the best features or reducing the feature vector dimensionality, as well as the stages for analyzing and classifying the data. The communication between both parts of the system should be performed in a loosely coupled manner, through an asynchronous service, agnostic of the underlying technology. This decoupled approach between the front end and the back-end allows, for example, the consideration of the methods already developed during previous research tasks using platforms that enable fast prototyping such as MATLAB (The MathWorks 2013a), which allows faster deployment of a prototype in real-world contexts.

The Unified Modeling Language (UML) sequence diagram in Fig. 5.8 describes a command and control type of application, which can be used as a testbed capable of evaluating different feature extraction, feature selection, and classification methods, from live data input. It provides a high-level example, where a prototype is used for word recognition, depicting its dynamics. The diagram can be interpreted in the following manner: After the system is initiated, the front end starts to acquire data; when a complete utterance is detected (using movement detection algorithms like the one described in Sect. 4.3.3.2 of chapter 4 or simply by clicking on a button), the data is sent to the back-end for processing and classification; the final step is to send the classification result back to the user.

5.4.2 A Possible Implementation

We can implement each part of the system in a different computer language to take advantage of the characteristics of different development environments that better match our purposes. For example, the front end may be implemented with a technology stack that enables access to consistent programming models and libraries suitable for rendering user interfaces.

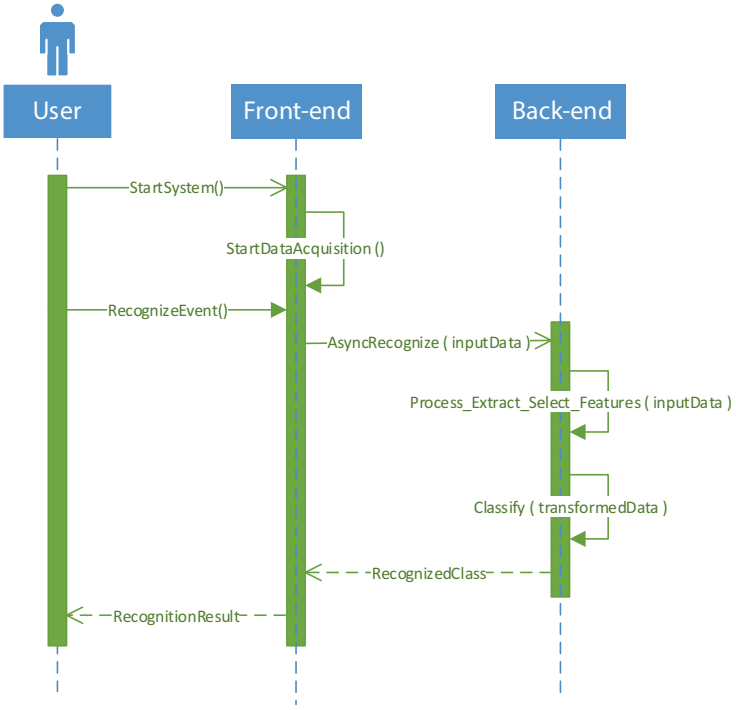

Fig. 5.8 UML sequence diagram with a high-level example of an HCI scenario: the recognition of an isolated word used as a command to control an application

To exemplify the implantation of our front end, we chose to use .NET technologies since it also allows using well-known third party Application Programming Interfaces (API), which provide easy access to low-level data from multiple devices, as depicted in Fig. 5.9. We divided the front end part of our prototype in the following modules:

- **Input Manager**—The input manager handles the input data from the input SSI modalities and respective buffering strategy.
- **Graphical User Interface**—The graphical user interface, for the operator to 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 back-end address.

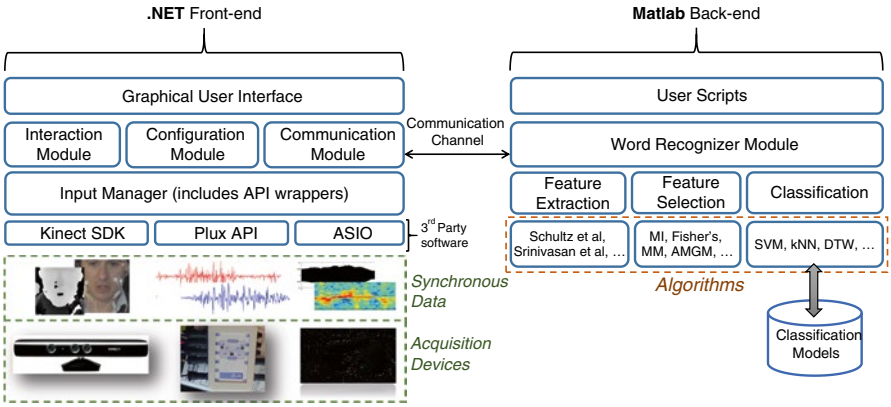

Fig. 5.9 Detailed architecture diagram of a multimodal SSI prototype for isolated word recognition. On the *left*, the front end part of the system (including devices and the corresponding data) is represented. On the *right*, the back-end part, including its modules and algorithms, is depicted

- **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.

Additionally, the front end also includes third party system software that allows low-level access to the hardware devices, such as the Kinect SDK,[4] Audio Stream Input/Output (ASIO) drivers, or the API to access the EMG recording device.

The back-end part of the system was developed using MATLAB (The MathWorks 2013a). As depicted in Fig. 5.9, both user scripts and the front end communication module (described above), access the interface provided by a module referred to as the word recognizer. This module receives the configuration information (i.e., which algorithms to use, vocabulary) 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.

This design of the back-end allows to reuse part of the code from previous examples. Additionally, 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. Note also that this architecture allows to easily

[4] Microsoft Kinect SDK can be accessed via the following URL: https://dev.windows.com/en-us/kinect.

replace, or complement, the back-end by other machine learning services from companies such as Microsoft (Azure Machine Learning[5]), Amazon (Amazon Web Services Machine Learning[6]), Google's TensorFlow[7], among others.

5.5 Conclusions

In this chapter, we have presented three illustrative examples of SSI applications, in the form of tutorials. The first, a very simple example, allows the reader to carry a hands-on experience by providing the steps, code and data to build a basic SSI recognition application based on video. Additionally, it enables the reader to explore other modalities whose data was previously collected.

With the second example, the reader gains insights on the importance of some of the extracted signals features, made available by the multimodal SSI framework described in Chap. 4. By allowing the synchronous acquisition of SEMG and US, this example provides direct data measurements of tongue movements (obtained after US annotation). This allows a more informed and novel exploration of the EMG data, namely concerning its suitability to detect tongue movements. The presented processing approach, although simple, shows how the acquired multimodal data can help deepen the knowledge regarding the different SEMG channels, and how it can inform the development of more complex processing methods, or the design of additional data collections.

By capturing synchronized modalities data, we gain the possibility of analyzing SEMG data with reliable ground truth information, enabling us to analyze how the speech production model is behaving. This example is more complex and costly to replicate than the first tutorial; however, the same idea can be extended to other modalities that are accessible to the reader. A similar example of this concept can be found in a study made by the authors where RT-MRI information (instead of US) was used to understand if SEMG would be able to detect velum movement (Freitas et al. 2015). These experiments, considering multiple input SSI modalities, create the conditions for in-depth studies with noninvasive modalities, and enable access to "ground truth" data for such studies. Such experiments, provide also a glimpse on how data, acquired through the framework presented in Chap. 4, can be used to serve the more complex goal of developing multimodal SSI using noninvasive modalities.

In the last part of this chapter, we follow a tutorial on how to build an application for a real scenario—handling live input from several SSI modalities. The presented solution allows to easily evaluate new parts of the processing pipeline with live data input.

The practical perspective of SSI adopted in this chapter aims at allowing the reader to gain a complete insight on the techniques, methods, libraries, and plat-

[5] More about Microsoft Azure Machine learning in https://azure.microsoft.com/services/machine-learning/

[6] More about Amazon Web Services Machine Learning in https://aws.amazon.com/machine-learning/

[7] More about Google's TensorFlow in https://www.tensorflow.org/

forms required to develop SSI applications. The following chapter completes our challenging journey through the expanding field of SSI, with some concluding remarks and perspectives for the future.

References

Brookes M (1997) Voicebox: speech processing toolbox for matlab. Software. March 2011. www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

Burges CJC (1998) A tutorial on support vector machines for pattern recognition. Data Min Knowl Discov 2:121–167

Chang C-C, Lin C-J (2011) LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol 2(27):1–27. doi:10.1145/1961189.1961199

Freitas J (2015) Articulation in multimodal silent speech interface for European Portuguese. Ph.D. thesis, University of Minho, University of Porto and University of Aveiro

Freitas J, Ferreira A, Figueiredo M, Teixeira A, Dias MS (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 MS (2014b) Multimodal corpora for silent speech interaction. In: 9th Language resources and evaluation conference, pp 1–5

Freitas J, Teixeira A, Silva S, Oliveira C, Dias MS (2014c) Assessing the applicability of surface EMG to tongue gesture detection. In: Proceedings of IberSPEECH 2014, lecture notes in artificial intelligence (LNAI), Springer, New York, pp 189–198

Freitas J, Teixeira A, Silva S, Oliveira C, Dias MS (2015) Detecting Nasal Vowels in Speech Interfaces Based on Surface Electromyography. PLoS One 10, e0127040. doi:10.1371/journal.pone.0127040

Galatas G, Potamianos G, Makedon F (2012) 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

Hanselman DC, Littlefield B (2005) Mastering Matlab 7. Pearson/Prentice Hall, Upper Saddle River

Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B (1998) Support vector machines. Intell Syst their Appl IEEE 13:18–28

Hsu C-W, Chang C-C, Lin C-J (2003) A practical guide to support vector classification. Tech. rep., Department of Computer Science, National Taiwan University

Potamianos G, Neti C, Gravier G, Garg A, Senior AW (2003) Recent advances in the automatic recognition of audiovisual speech. Proc IEEE 91:1306–1326

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

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

Steinwart I, Christmann A (2008) Support vector machines. Springer Science & Business Media, Berlin

The MathWorks (2013a) MATLAB

The MathWorks (2013b) MATLAB, statistics toolbox

The MathWorks (n.d.) Image processing toolbox—MATLAB

Van der Maaten LJP, Postma EO, van den Herik HJ (2007) Matlab toolbox for dimensionality reduction. MICC, Maastricht University, Maastricht

Chapter 6
Conclusions

Abstract In this book, we covered silent speech interfaces (SSIs), from the core principles and motivations that have fostered research on the topic, up to the enabling technologies that serve its development and the methods applied to analyze and process the data they provide. We have realized that research in silent speech interfaces is just at its infancy and evolving at a fast pace, boosted by novel technologies and views from a growing multidisciplinary community. In this chapter, the reader may find a reflection about upcoming steps and the future of SSI systems. After summarizing the topics addressed in this book, we discuss our views regarding the future of research in this area, a future brimming with challenges, but also with unquestionable potential for SSI to become a technological asset on our daily lives.

Keywords Silent speech interfaces • Trends • Future perspectives

In this book, we presented an introduction to silent speech interfaces (SSI) that supports human–computer communication even when no acoustic signal is available. Existing SSI modalities are aligned with the speech production process, as presented in Chap. 1, harvesting data from the brain (e.g., EEG), the muscles (e.g., SEMG), articulator movements (e.g., US), and subsequent effects (e.g., NAM).

After introducing a wide range of single modalities, analyzing their respective state-of-the-art and looking into the details of how they work (in Chaps. 2 and 3), we concluded on the advantages and disadvantages of each one. We then focused our attention on how SSI modalities could be combined, addressing the challenges of SSI modality fusion. Hence, we proposed a multimodal approach to SSI in the form of a framework (in Chap. 4). This multimodal framework provides ways to collect SSI data from several modalities, synchronize such data, create new SSI corpora, extract and select features, and analyze or classify the processing outcomes.

To complement the introduction in this area, we also provided a more practical part (in Chap. 5), giving readers the possibility of actually building their first (albeit simple) SSI recognizer. We provided different examples that go from a simple tutorial on how to build a word-recognizer based on video to the design of a system that handles live data.

In the following sections, we briefly highlight some pointers to recent advances and current trends in SSI-related technology, followed by our views on a set of routes that could be taken to enable further advances in SSI research.

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4_6

6.1 Current Trends

In recent years, we have witnessed to an increase in the amount of papers, projects and initiatives in the area of SSI. In this context, we have seen new devices specifically developed for SSI, such as the one based on permanent magnetic articulography (Cheah et al. 2015) that, even though still at an early prototype stage, exhibit a promising potential. Along that line, other types of techniques and devices have also been adapted with success to SSI, such as an array-based electromyography device (Wand et al. 2013), or visual speech using the Kinect depth sensor (Galatas et al. 2012). Yet, in some cases, access to these devices is cumbersome because they are either expensive or require additional knowledge to be built from scratch. Thus, the fact that SSI resources and datasets are now becoming openly available for the research community (Alghowinem et al. 2013; Freitas et al. 2014b; Telaar et al. 2014), reduces the cost of entry in this area and allows for more researchers to apply their algorithms without the need to collect data or acquire devices.

In terms of very recent publications, we identified some prominence in the following: application of techniques like neural networks to visual speech recognition (Wand et al. 2016), SEMG (Diener et al. 2015), and EMA (Bocquelet et al. 2015; Hahm and Wang 2015); control of wearables via tongue and jaw movements, by using magnets placed on the tongue or infrared proximity sensors place in the ear that measure the deformation of the ear canal (which is in turn caused by jaw movement) (Bedri et al. 2015a, b); mapping of the articulatory movements of the tongue and lips into audible speech using US and video (Hueber and Bailly 2016); determination of the best position for EMA sensors for continuous silent speech recognition (Wang et al. 2015); further exploitation of unspoken speech for communication (Li 2016; Yamaguchi et al. 2015); and the development of wearable SSI with the involvement of end-users (Cheah et al. 2015).

Areas relevant to SSI, such as machine learning are also evolving at a very fast pace, allowing easy access to new algorithms and models, via open frameworks (Abadi et al. 2016) or online services (Barga et al. 2015; Copeland et al. 2015).

6.2 Concluding Remarks

This book presented several ideas and examples which, although described in an introductory manner, reflect the authors' position in the SSI field. Throughout this book, we explored the concepts behind silent speech interfaces, its supporting technologies, and we had a glimpse on how these systems can be put together in practice. Now that we have seen what silent speech interfaces are, we fix our eyes into the future and share our personal views regarding the paths that can foster further advances in the field.

Multimodal silent speech interfaces—Beyond the desirable improvement of each modality (e.g., different feature extraction techniques), we strongly believe that the joint use of modalities may bring several benefits. However, these are also

associated with a few challenges, namely, finding the best way to fuse the SSI information coming from multiple sources without falling into problems such as high dimensionality, or maximizing the relevance of the extracted information.

In the multimodal context, the use of various SSI modalities may enable the simultaneous support of multiple stages of the speech production process, from which we can then extract information. Furthermore, easy addition of new and more accurate state-of-the-art sensors, such as time-of-flight depth cameras, may enable the extraction of 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.

The consideration of multiple SSI modalities may enable information disambiguation and complementarity, contributing to the design of better interfaces by improving their overall performance and fostering better adaptability to the user and his/her environmental diversity.

Understanding the full potential of current and new technologies—The adoption of a multimodal view for silent speech interfaces, supported on multimodal data collection settings, is not just valuable to explore the joint use of modalities when targeting the improvement of silent speech interaction. This approach provides also the basis for exploring the full potential of well-established and new technologies. For example, the use of technologies which typically are out of the scope of silent speech interfaces, due to their dependence on data that is not available in a given scenario (e.g., audio speech signal), to their cost or their invasive nature, may help understanding what kind of information other modalities are providing. Without this systematic assessment, we may fail to grasp the full capabilities of existing devices and remain limited in our analysis of a new technology.

Richer vocabularies and multilanguage support—Other roads we deem relevant to consider in SSI research include more extensive vocabularies, support of other languages with distinctive characteristics and continuous silent speech scenarios. In order to evolve in terms of vocabularies, we need more SSI data that observe an extension of silent speech units to be recognized. To minimize the impact in performance caused by the increased number of such units, we can take advantage of techniques that have presented good results in ASR, such as deep neural networks (Badino et al. 2016). We also find important to develop SSI systems for other languages. The main reason is that a speech interface in the users' native language leads to a more comfortable and natural interaction. However, one also needs to address and tackle the distinctive characteristics of each language (e.g., nasality in European Portuguese or French), which may have an impact in the performance of a given SSI system (Freitas et al. 2015). In this journey, articulatory information achieved by SSI systems can help to improve the baseline results or build real-time systems like the ones envisaged by Bocquelet et al. (2015).

Multidisciplinary research settings—Along this book we have seen how wide the range of disciplines involved in silent speech interfaces is (e.g., neuroscience, phonetics, signal processing, and machine learning, just to name a few). This naturally results from the different stages associated with speech production, from the brain to the resulting visual effects. Therefore, we should promote

bringing these different areas together to improve collaboration in the area of silent speech research. The challenge, as with any other multidisciplinary approach, is to enable the dialog among the different areas to develop better silent speech interfaces. It is highly desirable that the requirements set for silent speech interfaces move up the stream and motivate, for example, novel research on associated brain signals and auxiliary technology to collect them. And for this to happen, insights from experts in the different disciplines are vital.

Another aspect to have in mind is that the technologies employed in SSI systems, including the ones presented in this book, could be used for other purposes in the scope of communication in general (e.g., human–human communication, human–machine communication, speech therapy). For example, a speech therapist often needs to assess the speaking capabilities of an individual and, while relying mostly on the acoustical signal, the truth is that technological means to aid such tasks are scarce and frequently require a significant amount of manual input. Hence, we believe that a comprehensive investigation of speech capabilities in terms of articulation, like the ones we obtained using our multimodal framework, can be beneficial for this scenario. In this context, an extended or alternative version of the multimodal framework proposed, for example, addressing the adoption of fMRI and EEG as ground truth modalities, could improve and automate current speech therapy methods. This approach would promote a higher impact of SSI research by also contributing to adjoining fields of research.

Data availability and sharing—One important factor can motivate and increase the contributions from different fields of expertise to silent speech research, whether directly or through knowledge obtained for other application areas, as alluded above: the general availability of the datasets collected by various SSI research teams, and using different SSI technologies. For many contexts, researchers often face problems of accessing datasets to test novel methods, notably when using statistics and signal processing methodologies. Enabling access to SSI datasets, with adequate context, may promote the appearance of novel methods and insights for this area, even if silent speech applications are not the researcher's main goal.

Additionally, as already mentioned, the availability of SSI datasets can provide a bootstrap for researchers engaging in silent speech research, but it can also serve the important goals of replication and validation of experiments. It is not only a question of sharing data, but also the resources that allow its collection, including experimental protocols, data collection software, and equipment configuration settings. A wider dissemination of such databases brings an additional benefit: researchers can start having datasets that are actually comparable among them and that may be gathered to study particular aspects of the data. As highlighted in this book, one of the difficulties associated to some of the SSI modalities is the variability of the collected data among users (e.g., due to anatomical uncertainty). Being able to look into this diversity, while considering multiple datasets, will certainly bring additional light into this matter.

Usability evaluation—The main goal for silent speech interfaces is to assist users in their communication tasks with other people or with interactive systems.

While we strive to understand the basics of the technology and its application for silent speech, it is understandable that the usability of such interfaces (apart from considerations on comfort and perceived invasiveness), has not been, so far, a main concern of the community. We believe that this aspect should become central in future SSI research. Moving SSI into real scenarios will enable setting more ambitious goals and collect richer feedback on how the technology is performing, allowing the identification of shortcomings and, most importantly, obtaining a clearer idea on how users use and accept it. With such approach the focus now shifts from the technology to the user. Thus, the mindset should be to ask not what users can do for silent speech technology, but to ask what this technology can do for them.

Real and diverse application contexts—Having an SSI system working is just a part of the equation. Then, we need to understand how these systems are going to be used. Hence, another aspect (or challenge) to consider is how to take SSI systems beyond the research stage and integrate them in real-life applications. The widespread adoption of a multimodal interaction paradigm supported in standard architectures, such as the one proposed by the W3C (Dahl 2013; Silva et al. 2015), opens a path to an easier integration of silent speech modalities into the human–computer interaction ecosystem, side-by-side with other more commonly adopted modalities such as touch. One potential field of application of silent speech as an interaction technology is Ambient Assisted Living—AAL, that provides a diversity of challenging scenarios in which silent speech can be used by a wide range of users, including older adults. One example of recent research in this field is the project Marie Curie IAPP IRIS (Freitas et al. 2014a) that aims to improve communication in a domestic scenario for a family with different limitations regarding speech communication and diverse levels of proficiency with technology. Some approaches that may look promising in the lab may never be able to integrate in such scenarios, if we do not take into proper account the context of AAL environments.

To finalize, we expect that the concepts described in this book will not only increase the interest and work in the area of SSI, but also, directly or indirectly, lead to scientific advances and breakthroughs that help to tackle the research questions listed in Chap. 1 (e.g., speaker adaptation) and new challenges to come. More importantly, we hope that this book may lead to useful applications that can help and benefit our society, in particular speech-impaired users.

References

Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) TensorFlow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467

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

Badino L, Canevari C, Fadiga L, Metta G (2016) Integrating articulatory data in deep neural network-based acoustic modeling. Comput Speech Lang 36:173–195

Barga R, Fontama V, Tok WH (2015) Introducing Microsoft Azure Machine Learning. In: Predictive analytics with Microsoft Azure Machine Learning. Springer, New York, pp 21–43

Bedri A, Byrd D, Presti P, Sahni H, Gue Z, Starner T (2015a) Stick it in your ear: building an in-ear jaw movement sensor. In: Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing and proceedings of the 2015 ACM international symposium on wearable computers, ACM, pp 1333–1338

Bedri A, Sahni H, Thukral P, Starner T, Byrd D, Presti P, Reyes G, Ghovanloo M, Guo Z, (2015b) Toward silent-speech control of consumer wearables. Computer (Long Beach Calif) 54–62

Bocquelet F, Hueber T, Girin L, Savariaux C, Yvert B (2015) Real-time control of a DNN-based articulatory synthesizer for silent speech conversion: a pilot study. In: Sixteenth annual conference of the international speech communication association

Cheah LA, Gilbert JM, Gonzalez JA, Bai J, Ell SR, Fagan MJ, Moore RK, Green PD, Rychenko SI (2015) Integrating user-centred design in the development of a silent speech interface based on permanent magnetic articulography. In: Biomedical engineering systems and technologies. Springer, Berlin, pp 324–337

Copeland M, Soh J, Puca A, Manning M, Gollob D (2015) Microsoft Azure Machine Learning. In: Microsoft Azure. Springer, New York, pp 355–380

Dahl DA (2013) The W3C multimodal architecture and interfaces standard. J Multimodal User Interfaces. doi:10.1007/s12193-013-0120-5

Diener L, Janke M, Schultz T (2015) Direct conversion from facial myoelectric signals to speech using Deep Neural Networks. In: Neural Networks (IJCNN), 2015 Int. Jt. Conf. doi:10.1109/IJCNN.2015.7280404

Freitas J, Candeias S, Dias MS, Lleida E, Ortega A, Teixeira A, Silva S, Acarturk C, Orvalho V (2014a) The IRIS Project: a liaison between industry and academia towards natural multimodal communication. In: Iberspeech 2014

Freitas J, Teixeira A, Dias MS (2014b) Multimodal Corpora for Silent Speech Interaction. In: 9th Language resources and evaluation conference, pp 1–5

Freitas J, Teixeira A, Silva S, Oliveira C, Dias MS (2015) Detecting Nasal Vowels in Speech Interfaces Based on Surface Electromyography. PLoS One 10, e0127040. doi:10.1371/journal.pone.0127040

Galatas G, Potamianos G, Makedon F (2012) 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

Hahm S, Wang J (2015) Silent speech recognition from articulatory movements using deep neural network. In: Proc. of the International congress of phonetic sciences

Hueber T, Bailly G (2016) Statistical conversion of silent articulation into audible speech using full-covariance HMM. Comput Speech Lang 36:274–293

Li W (2016) Silent speech interface design methodology and case study. Chinese J Electron 25

Silva S, Almeida N, Pereira C, Martins AI, Rosa AF, e Silva MO, Teixeira A (2015) Design and development of multimodal applications: a vision on key issues and methods, Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). doi:10.1007/978-3-319-20678-3_11

Telaar D, Wand M, Gehrig D, Putze F, Amma C, Heger D, Vu NT, Erhardt M, Schlippe T, Janke M (2014) BioKIT-Real-time decoder for biosignal processing. In: The 15th Annual conference of the international speech communication association (Interspeech 2014)

Wand M, Schulte C, Janke M, Schultz T (2013) Array-based Electromyographic Silent Speech Interface. In: International Conference on bio-inspired systems and signal processing (BIOSIGNALS 2013)

Wand M, Koutník J, Schmidhuber J (2016) Lipreading with long short-term memory. arXiv Prepr. arXiv1601.08188

Wang J, Hahm S, Mau T (2015) Determining an optimal set of flesh points on tongue, lips, and jaw for continuous silent speech recognition. In: 6th Workshop on speech and language processing for assistive technologies (SLPAT), p 79

Yamaguchi H, Yamazaki T, Yamamoto K, Ueno S, Yamaguchi A, Ito T, Hirose S, Kamijo K, Takayanagi H, Yamanoi T (2015) Decoding silent speech in Japanese from single trial EEGS: preliminary results. J Comput Sci Syst Biol 8:285

Index

J. Freitas et al., *An Introduction to Silent Speech Interfaces*, SpringerBriefs in Speech technology, DOI 10.1007/978-3-319-40174-4