

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 • Electrocardiography • 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.

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 (Akçakaya 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.

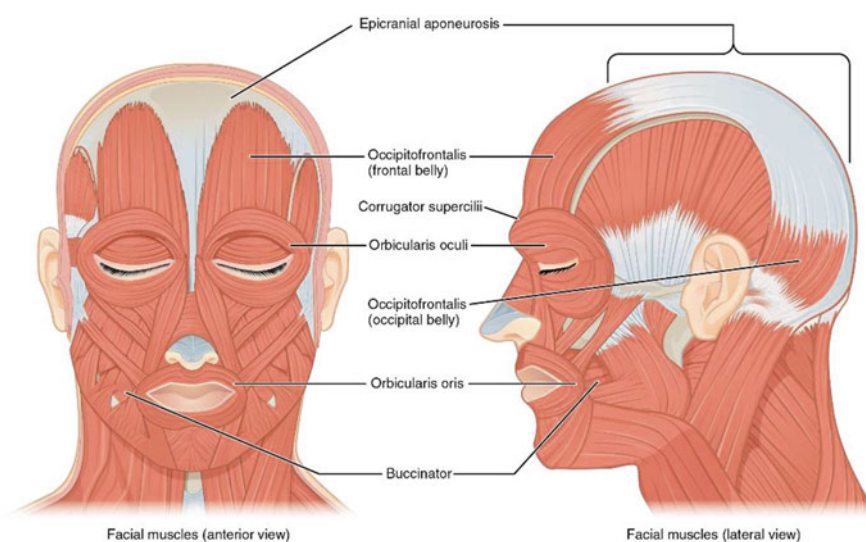


Fig. 2.2 Anterior and lateral view of the human facial muscles from (OpenStax_College 2013)

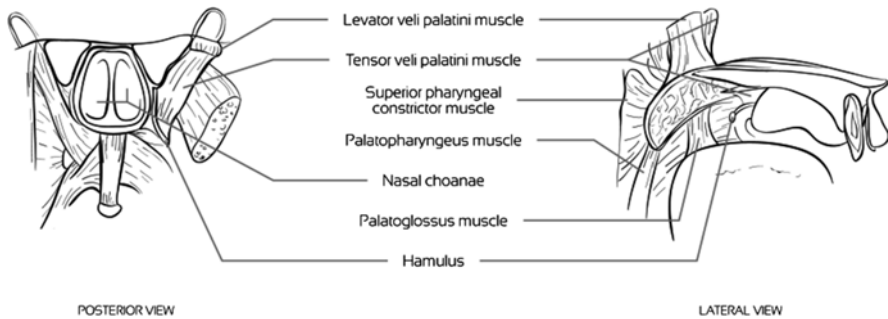


Fig. 2.3 Muscles of the soft palate from posterior (*left*) and the side (*right*) view

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 *orbicularis 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](https://doi.org/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](https://doi.org/10.1016/j.specom.2010.01.001)
- Brumberg JS, Wright EJ, Andreassen 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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 Pestiers 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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 of the 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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*. Universität 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*

An Introduction to Silent Speech Interfaces

Freitas, J.; Teixeira, A.; Dias, M.S.; Silva, S.

2017, IX, 102 p. 34 illus., 28 illus. in color., Softcover

ISBN: 978-3-319-40173-7