EMG-BASED SILENT SPEECH INTERFACES

Insights into the Challenge of Predicting Speech
from Articulatory Muscle Activity

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology
ahoLAB

Doctoral thesis

Inge Salomons

2024

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Supervisors

Prof. dr. Inmaculada Hernáez Rioja
Dr. Eva Navas Cordón

Cover illustration by David Fernández Rementería for Itelazpi (2023).

Abstract

This doctoral thesis is performed as part of the ReSSInt project, which aims to restore speech for Spanish alaryngeal speakers. An alaryngeal speaker has no larynx, which is an element of the speech production system containing the essential vocal cords. Using machine-learning-based technology, the main goal of the project is to develop a silent speech interface (SSI) based on non-acoustic biosignals. Biosignals are the product of biological processes during speech production, such as neural, muscular, or pulmonary activity. An SSI allows users to articulate without sound (silent speech) while a computer model interprets the biosignals related to the intended speech. The model is created using a large database of parallel speech signals and biosignals. To capture articulatory muscle activity, a technique called electromyography (EMG) is used, which measures the electrical pulses in activated muscles.

The topic of this thesis is EMG-based SSIs and focuses on the development of the first Spanish EMG-speech database, as well as research challenges associated with using this data for SSI development. The main contribution is the presentation of the ReSSInt-EMG database and its collection and validation procedure. Furthermore, using the data from this database, it aims to assess the effect of muscle activity variation between speakers, the impact of the absence of vocal cord and tongue information, and the differences in muscle activity between audible, silent, and alaryngeal speech. The results can be used to develop and improve an EMG-based SSI for Spanish alaryngeal speakers.

Acknowledgments

A few years ago I went to Bilbao to start my PhD journey at the Aholab Signal Processing Laboratory of the University of the Basque Country. It was a wholesome experience that taught me a lot in both my professional and personal life. This document is the product of that journey: my doctoral thesis. Many people were involved in it, and I would like to take a moment to express my gratitude.

First and foremost, I would like to acknowledge everyone who has been directly involved in this thesis. Without my supervisors and colleagues, I simply would not have been able to achieve it. Even though a PhD journey is a unique and individual project, it was the team effort that made it possible.

Dear **Eva** and **Inma**, I was lucky enough to have had not one, but two very involved supervisors. I appreciate you taking a chance on me, knowing that you would hire a linguist with the only signal processing background consisting of the courses you taught me during the master's program. Eva, we had already worked together, as you were the supervisor of my master's thesis. When I contacted you for more information about this PhD project, we had a video call and I was immediately excited. I took a few days to think whether I was willing to leave my home, family, and friends for three years and then decided to take the plunge. Your patient, understanding, and knowledgeable guidance during my master's thesis made me confident that you were the right person to guide me through my PhD as well. Inma, you were there every step of the way, but let me take as much responsibility as I could handle. You were determined in what you wanted me to achieve, which has helped me to take that extra step when necessary, as well as boosted my confidence. You have such a warm and open personality, and I knew I could always talk to you if there was anything on my mind. I thank you both for guiding me in becoming the researcher I am today.

Dear **Eder**, I have sometimes jokingly called you my work husband, but I see you more as my academic brother. You have the kindest of hearts and you inspire me with all your talents. I still do not know how you keep up doing a PhD, studying music, creating the most beautiful songs, and even remembering to water the office plant when no one else did. Working with you has been a great joy. I have enjoyed brainstorming, writing papers, presenting at conferences, having countless meetings, and collecting hours of data with you. I cannot imagine what my PhD journey would have looked like without you. Thank you for everything.

Dear **Sneha**; my colleague who became a friend. You were already a PhD student when I arrived, and welcomed me warmly, for which I'm still very grateful. We worked on different projects, so during work hours we both did our own thing, but on a personal level, we quickly connected. You invited me on my very first hike around Bilbao and from there our friendship grew. We have talked for hours and hours, and I have really missed those chats and our lunch breaks in the sun when you finished your PhD. But I'm happy to know that you now live happily with your husband in your home country. One day I'll come to visit!

Dear **Víctor, Iñigo, Itxasne, Xabier, Ibon, and Jon**; my Aholab colleagues. Thank you all so much for being so kind and welcoming, and for creating a great work atmosphere. You could have all talked Basque with each other but always made sure to speak Spanish or English when I was around, which I appreciate more than you may realize. A special thank you to Víctor, for coming up with the idea of the 3D mask and always being open for a brainstorming session; Iñigo for the interesting philosophical lunch chats; and Jon for inviting me to a real Athletic Bilbao experience.

Dear **Martijn**, thank you for adopting me for a while in your Speech Lab in Groningen. Even though I was only there for a few months, you included me as if I were one of you. (For this I am also very grateful to the entire Speech Lab crew.) You allowed me to approach my research from a more linguistic perspective, and I have learned a lot. No question was too much and you made me realize that there is a solution to every problem. One of my literal highlights was when you took me flying for one of your practice rounds in a two-seat plane; we landed on an island to eat some ice cream and circled my mom's house on the way back. It's something I'll never forget.

Secondly, but not less importantly, I would like to extend my gratitude to those who were always there to support me, to hear my cries when I was feeling alone or disheartened, and to share my excitement from one of the many adventures I got to experience; my friends and family.

Dear **Anne**, we met as two Dutch women living in Bilbao, born in the same year and region. We quickly bonded over those similarities, despite our opposite personalities. From all those people I've met outside of work, you are the one with whom I have the most special memories. I thank you for missing my loved ones back home more bearable.

Dear **Ilse, Monique, Laura, and Esther**, my friends of more than twelve years. This journey made me realize how special it is to have friends who have known you during different stages of your life. Thank you for being there for me, even from a distance, and for welcoming me back with open arms.

Dear family, and especially my parents **Theo** and **Marjan**, and brothers **Mark** and **Tim**; you have known me either my or your whole life, and watched me both struggle and grow. Your support (even if it was covered with jokes) has been the motivation for my life's journey. Thank you for your unconditional love.

Dear **Aron**, the one who has become very important to me in a very short time. You have believed in me from the moment we met, and from then on were there for me whenever I needed it. You are my rock. I am very grateful that you accept me as I am, and I am excited to embark on the future together as husband and wife soon.

Thank you, gracias, eskerrik asko, bedankt!

Inge Salomons
September 2024

Contents

Part I

Introduction and Background

1
Introduction

To speak is to be able to express one's feelings, emotions, desires, commands, and frustrations. It is a complex system involving our brain, lungs, vocal tract, and muscles, and yet most of us take it for granted. Try to imagine not having the act of speech available to you, and having to rely on alternative communication methods instead. What would you want this alternative method to comply with? Let me answer that for you. First, you would like others to understand you without much effort. Secondly, you would like to be able to convey the emotions that are attached to the message. Thirdly, you would like the conversation to go smoothly, so a fast processing time is essential. And lastly, you would not want to spend years learning how to use this method. And as a bonus, how nice would it be if you could communicate with the sound of your (old) voice?

This thesis focuses on a technological approach that has the potential to comply with all these requirements.

In the rest of this chapter, you will read more about the research goal and questions of this thesis, and how it is structured.

1.1 Research goal and questions

This thesis is performed as part of the ReSSInt project [1, 2], which aims to restore speech for Spanish people who have been deprived of the ability to speak. More specifically, the target group consists of alaryngeal speakers, who are people whose larynx has been removed. The larynx is a part of the speech production system that contains the vocal cords, which are essential for typical speech production. The main goal of the project is to develop a silent speech interface (SSI) based on non-acoustic biosignals using machine-learning-based technology. Biosignals are the product of biological processes during speech production and can be acquired from the brain, the tongue, or the muscles. Silent speech refers to the act of articulating as if a person were speaking, but without producing any sound. Therefore, biosignals can correspond to either silent (non-acoustic) or audible (acoustic) speech. An SSI is developed using a large database of parallel (silent or audible) speech signals and biosignals. It allows users to communicate without making any sound because a computer model interprets the biosignals related to the intended speech and outputs the predicted speech. This thesis focuses on predicting speech from articulatory muscle activity, and the challenges that are associated with it. To acquire biosignals from the muscles used to articulate, a method called electromyography (EMG) is used. This method measures the electricity in muscles using electrodes attached to the skin and the resulting signals represent the level of muscle activity.

This thesis aims to fill the research gap specifically related to EMG-based SSIs for Spanish alaryngeal speakers, by answering the following research questions:

1. What are the most important advances made in EMG-based SSI research, and in which areas is more research needed?
2. Which superficial muscles of the face and neck are involved in speech production?
3. What is the optimal acquisition setup and procedure for the development of the database?

4. What is the effect of variation in EMG signals between different speakers and sessions?

5. What is the effect of lack of information from the vocal cords and tongue, two important elements of speech production?

6. How does articulatory muscle activity of audible versus silent speech compare?

7. How does articulatory muscle activity of a typical versus an alaryngeal speaker compare?

1.2 Thesis guide

This thesis consists of four parts, each containing at least one chapter.

Part I: Introduction and Background.

After this introductory chapter, a background chapter follows, which provides detailed background information on the most important topics in this thesis, and an overview of state-of-the-art research in this field.

Part II: Data Collection.

This part focuses on the data collection and contains three chapters. Chapter 3 describes a pilot study to find the optimal way to acquire EMG signals from the articulatory muscles. Chapter 4 describes the data collection procedure and the resulting database. Chapter 5 validates the acquisition setup of the database.

Part III: Research Challenges.

This part focuses on understanding the depth of the challenges that arise with this research topic and also contains three chapters. Chapter 6 shows the effect of variability between speakers, and sessions of the same speaker. Chapter 7 shows the effect of lack of information from the vocal cords and tongue. Chapter 8 shows the effect of differences in muscle use between speakers and speech modes.

Part IV: Conclusion.
This part consists of one chapter, which provides a general discussion and conclusion.

Furthermore, after the bibliography, a list of abbreviations, a summary in English, Spanish, and Dutch, and a list of contributions can be found.

2 Background

This chapter aims to provide a wide theoretical background and literature overview so that the reader can read this thesis without knowledge of the topic of research presented here. First, we explain the typical process of human speech production, the changes in this process after the larynx with the vocal cords has been surgically removed, and the consequences of this procedure. Then, we provide a technical description of SSIs, and the method of EMG. Lastly, we present an overview of the history and recent advances in the research area of predicting speech from muscle movements. When applicable, a chapter has a more detailed topic-specific literature overview as well.

2.1 Speech production

The speech production system (Figure 2.1) consists of the sub-glottal part (the lungs and trachea), the vocal tract (the pharynx, the larynx with the vocal cords, and the oral cavity), and the nasal tract (the soft palate and the nasal cavity). In typical circumstances, speech is produced by pushing air from the lungs, through the vocal cords, to the mouth and nasal cavity. The vocal cords are responsible for the voicing of sounds. The vibration of the vocal cords results in a periodic interruption of the airflow, creating voiced sounds such as vowels. In unvoiced sounds, there is no vibration of the vocal cords, meaning that the airflow is free. The different sounds are created by moving the lips, tongue, and jaw (the articulators) uniquely, and releasing the air accordingly. To move the articulators, the speaker needs to activate the muscles in the face and neck.

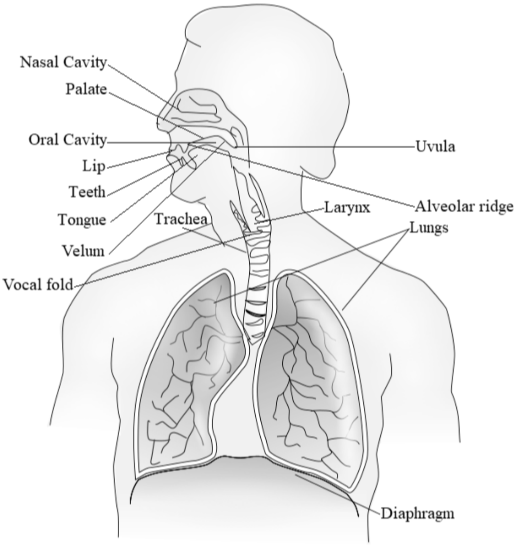


Figure 2.1: The speech production system. Copy right: Theresa Knott, CC BY-SA 2.5 https://creativecommons.org/licenses/by-sa/2.5, via Wikimedia Commons

The facial muscles that are related to speech are those that control the jaw, lips, and tongue. We call them speech or articulatory muscles in this thesis, but these muscles are also used for facial expressions, laughing, and eating. The most important muscle for speech production is the tongue. Figure 2.2 shows an image of the muscles of the face and Figure 2.3 shows two images of the muscles of the neck.

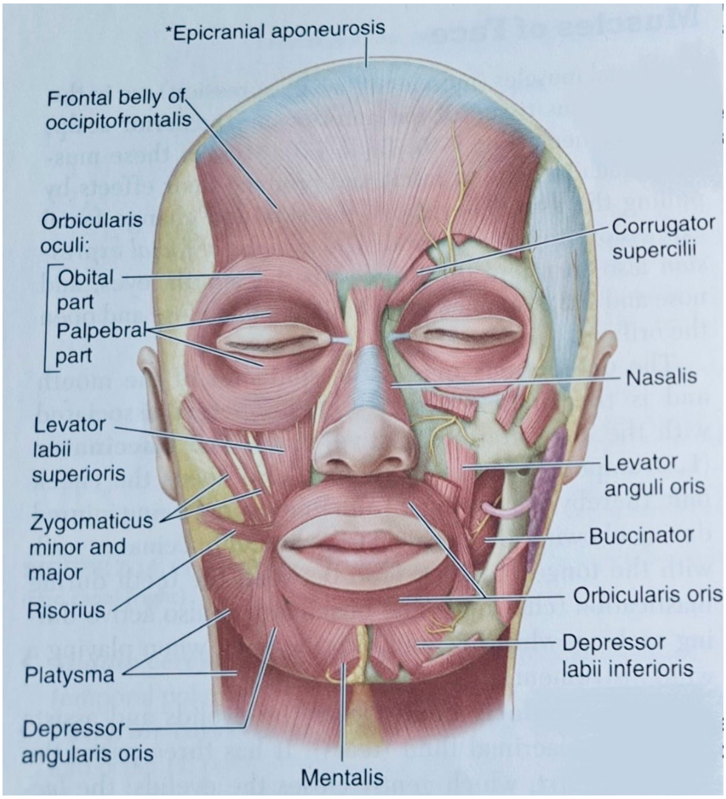
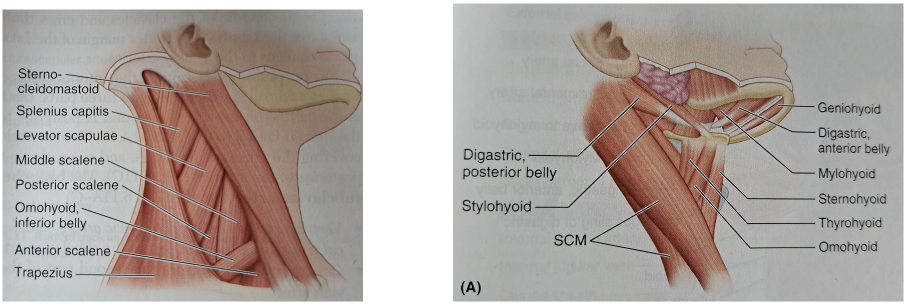


Figure 2.2: Muscles of the face [3].



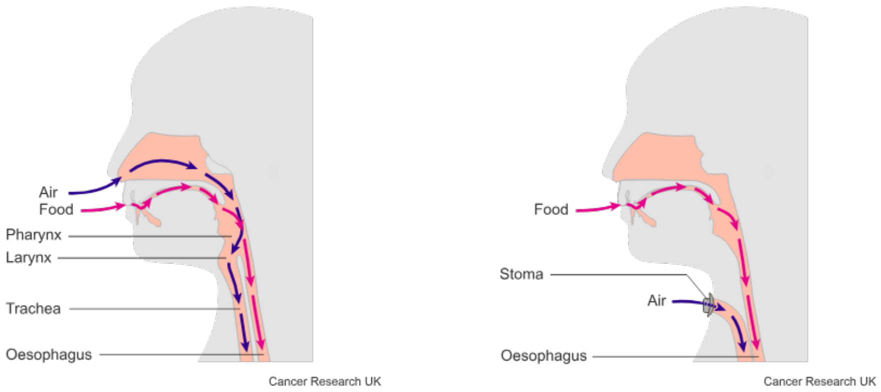
(a) (b)

Figure 2.3: Muscles of the neck [3].

2.2 Alaryngeal speech

Typical speech as described in the previous section is also referred to as laryngeal speech, since the larynx and especially the vocal cords, play an important role. If a person has undergone a larynx amputation surgery (laryngectomy), usually because of laryngeal cancer, the production of laryngeal speech is no longer possible. This is due to the absence of the vocal cords and the separation of the airway from the nasal cavity and the mouth. In order to breathe, these individuals receive a stoma in the throat, directly attached to the trachea. Figure 2.4 shows the difference in speech systems before and after a laryngectomy.

After surgery, these alaryngeal speakers lose the ability to produce speech naturally and rely on alternative communication methods. There are three common methods to produce alaryngeal speech [4]. The first is using a voice prosthesis or tracheo-esophageal puncture. An opening between the trachea and esophagus allows the speaker to push air from the lungs through this opening up into the mouth when covering the stoma with a finger, called tracheo-esophageal speech. Although this is the most common method, it occasionally results in speaking difficulty because the pharynx goes into spasm or there is swelling of the opening area. The second method is using an electrolarynx, which is a battery-operated machine that creates vibrations that normally the vocal cords



(a) Before

(b) After

Figure 2.4: Figures showing the airflow before and after a total laryngectomy. Copied from Cancer Research UK (https://www.cancerresearchuk.org/about-cancer/laryngeal-cancer/living-with/stoma/about, last accessed on 25/09/2024).

do, resulting in electro-laryngeal speech. Because it makes some noise, it is particularly used if a voice prosthesis is not (yet) an option. The last type of alaryngeal speech is esophageal speech. A speaker who produces speech this way pumps air from the mouth into the esophagus and the stomach, and when releasing this air, a vibrating tissue around the entrance of the esophagus simulates the vibration of the vocal cords. In general, each of these alaryngeal speaking methods has some limitations, of which the most prominent are that they are difficult to learn, or that the resulting voice can be difficult to understand by others [4–7].

2.3 Silent speech interfaces

Technological approaches to restore speech for alaryngeal speakers include personalized text-to-speech (TTS) systems [8], voice conversion [9, 10], bionic voices [11, 12], lean-AI approaches [13], and SSIs [14], among others [15].

An SSI uses non-acoustic biosignals to restore speech from nonverbal communication [16–18]. Biosignals are the product of chemical, electrical, physical, and biological processes during speech production, such as neural activity, articulator motor control, muscle activity, articulatory gestures, vibration of the vocal cords, and pulmonary activity. They are insensitive to environmental noise and independent of the acoustic speech signal.

There are two SSI approaches:

- Silent speech-to-text, where an automatic speech recognition (ASR) model decodes speech from features extracted from biosignals and outputs text, in combination with a TTS model that synthesizes speech from this text.
- Direct speech synthesis, where audible speech is generated directly from features extracted from biosignals, modeling the relationship between biosignals and the acoustic waveform.

Sensing techniques are used to retrieve the different types of biosignals related to speech production. SSIs can be based on biosignals retrieved from vocal tract imaging (i.e. [19]) to capture vocal tract movements, permanent magnet articulography (PMA) to capture speech articular movements (i.e. [20, 21]), EMG to capture the facial muscles' electrical activity (see Section 2.6), and electroencephalogram (EEG), to capture the neural activity in brain regions used for speech (i.e. [22–24]). Depending on the type of speech disorder, one method might be better suited than the other. We have selected EMG (Section 2.4), because the muscles and ability to articulate of alaryngeal speakers are still intact, and it is the least invasive of all methods.

The application area for the SSIs in the studies listed above is meant as a communication aid to provide a voice to people with speech disabilities. However, these interfaces can also be used in situations where private communications are required [25, 26] or in situations where audible speech would be masked by environmental noise [27].

2.4 Electromyography

EMG is a sensing technique used to measure and acquire muscle activity [28–31]. The literal meaning of EMG is "recording (*graphy*) of electricity (*electro*) of the muscle (*myo*)".

The muscles responsible for speech are skeletal muscles, meaning they are attached to bones via tendons in the tendon zone and are controlled voluntarily by the nervous system. The process required to move a muscle is as follows. First, a signal is generated by the motor cortex of the brain, which travels through a network of nerves to reach the muscle. At the end of this network are motor units consisting of motor nerve fibers, known as the innervation zone. Each muscle fiber within a motor unit has its own innervation. The power of muscle contraction depends on how many motor units are activated, with larger muscles typically requiring more motor units. The frequency of nervous impulses determines the extent of muscle contraction. The biological process of opening sodium and potassium channels in muscle cells, known as the motor unit action potential (MUAP), creates a myoelectric signal that can be detected and recorded by an EMG amplifier through electrodes.

Two types of electrodes can be used: those inserted in the muscle (invasive EMG; iEMG), and those attached to the skin (surface EMG; sEMG). Due to its non-invasive nature, we have selected sEMG, from now on simplified as EMG. In general, the EMG signal acquisition process consists of four steps: signal collection, signal amplification, signal filtering, and analog-to-digital conversion [31].

There are two ways to acquire EMG signals, namely in a monopolar or differential configuration. For monopolar acquisition, a reference electrode is required, which is placed in a location where no activity related to the muscle activity is expected, for example on the earlobe. Then the signal from the reference electrode is subtracted from the raw signal from the single monopolar electrode on the target muscle. Differential acquisition means that the difference between the signals acquired in two points is measured. This can be done using bipolar electrodes (made up of a pair of single electrodes) or an array of at least two electrodes. Two measuring points form one channel, whether that is between two bipolar electrodes, between the reference and a monopolar electrode, or between two electrodes in an array [31].

EMG recording recommendations by De Luca et al. [28] state that an electrode should be placed between an innervation zone and the tendon zone, or between two innervation zones, and along the longitudinal midline of the muscle. Furthermore, they suggest putting it in the middle of the muscle, and not on the outer edges. This is to avoid cross-talk as much as possible, which refers to the interference of muscle activity from surrounding or underlying muscles in the signal of the target muscle. These suggestions are in line with a study by Young et al. [32], in which they found that electrode shift increased the classification error, but that this error was lower for longitudinal channels compared to transverse channels. Secondly, the direction of electrode shift was a significant factor, with perpendicular shift resulting in higher error than parallel shift. Lastly, they found that the largest electrode (3x3 cm) performed worse in general, but was less sensitive to errors when shifted perpendicularly compared to smaller electrodes (2x2 and 1x1 cm).

The amplitude of an EMG signal can range from 0 to 10 millivolts (mV) [28]. EMG signals can be affected by many factors: the impedance of body skin, subcutaneous tissue layers, spread from the innervation zone, cross-talk from neighboring muscles, environmental noise, electrical power wires, and electrode size and position. The usable energy of an EMG signal is limited to the 0-500 Hertz (Hz) frequency range, with the dominant energy occurring between 50-150 Hz. Possible sources of noise that can be reflected in the signal are radiation from power sources around 50-60 Hz, and motion artifacts in the 0-20 Hz range [28].

For this reason, EMG signals are often pre-processed before using them. For example, low-pass and high-pass filters are applied to remove the information outside the ranges mentioned above. Depending on the equipment being used, this filtering can be applied by the hardware before the signal is digitized, or afterward using software [31].

2.5 EMG feature extraction

Since the resulting (filtered) EMG signal still can contain unwanted noise such as motion artifacts, features are often extracted from them to use as input data to the SSI instead of the raw signal. Several methods for EMG feature extraction have been explored in the context of silent speech studies. Maier-Hein et al. [33] first applied the Short-Term Fourier Transform (STFT) to the EMG signal, which is a mathematical technique used to analyze the frequency content of a signal over time, and is often applied to speech signals. From the STFT they then calculated the delta coefficients, which represent changes over time. These delta coefficients, together with the mean of the EMG signal in the time domain, formed the set of input features. Jou et al. [34] introduced a new feature extraction method, where they calculated the frame-based mean, power, and zero-crossing rate of the EMG signal in the time-domain (TD). They used contextual filters to model the context, such as a stacking filter to add frames according to k context width. Wand et al. [35] used this same method. Diener et al. [36] later developed these into the TD-15 features: the low-frequency signal power, the low-frequency signal mean, the high-frequency signal power, the high-frequency signal rectified mean, and the high-frequency signal zero-crossing rate. The TD0 (index) frame of all channels is combined and stacked into the past and the future for 15 frames each, to create the final TD-15 feature frames. Colby et al. [37] modeled Mel-frequency cepstral coefficients (MFCCs) and chose two co-activation features. Meltzner et al. [38, 39] concluded that the combination of MFCCs (and their corresponding delta features) and muscle co-activation levels (quantified amount of simultaneous firing activity between all possible pairs of EMG channels) yielded the best recognition performance. In later research, Meltzner et al. [40], only extracted MFCCs. Soon et al. [41] extracted features in the temporal domain, namely integrated EMG, mean absolute value (MAV), root mean square (RMS), variance, standard deviation (SD), and simple square integral, and in the time-frequency domain the MAV, RMS, variance, SD, and log RMS. Finally, Ma et al. [42] tried to reduce the feature set by only extracting the MAV and RMS.

2.6 EMG-based silent speech interfaces

As early as 1986, researchers found speech-related information in EMG signals retrieved from facial and neck muscles in one of the first studies on EMG-based speech recognition [43]. A pattern recognition algorithm based on the maximum likelihood algorithm was able to classify 17 words with an accuracy of 35% and two words with an accuracy of 97%, with four EMG channels. Later, they found that what the authors called the "average magnitude" (presumably referring to the mean amplitude of the rectified signal over time), provided the best information for recognition, with a 58% recognition accuracy of 10 words, using four channels [44].

This section provides an overview of the most important advances made in EMG-based speech-related research since then, in different areas and tasks.

2.6.1 Word recognition

Fifteen years after the initial EMG and speech study, Chan et al. [45] reached a classification error of 2.86% on a ten-word vocabulary (*zero* to *nine*) using a linear discriminant analysis (LDA) classifier, with five EMG channels. Later, they used a Hidden-Markov model (HMM) and achieved a 2.70% classification error [46]. On the same ten-digit vocabulary set, Maier-Hain et al. [33] reported a maximum accuracy of 98.8% using an HMM, with seven EMG channels. Other studies on word recognition showed 86.7% accuracy on 65 words [38], 87.07% on 65 words [47], and 92.64% on 110 words [48].

All of these experiments were performed session-dependently, however, Wand et al. [49] achieved a 21.93% word error rate (WER) on a vocabulary of 108 words with a session-independent recognition system.

2.6.2 Continuous speech recognition

Moving on to the recognition of continuous speech, Jou et al. [34] were the first to attempt this, to our knowledge. Their system resulted in a 32% WER with a 108-word vocabulary. Later, Meltzner et al. [39]

reported a 69.9% recognition rate on a continuous vocabulary of 200 words, Deng et al. [50] reached a 15.2% WER on 1200 utterances, and Wand et al. [51] showed 23.8% on 50 sentences.

2.6.3 Phoneme and syllable classification

In addition to recognizing isolated words and continuous speech, researchers have also zoomed in on smaller linguistic units, like phonemes and syllables. For instance, Zhou et al. [52] used an HMM to extract phonemic log-likelihoods, which were subsequently matched to their respective words using a word classifier, and reported an average word accuracy of 98.5%. Lopez-Larraz et al. [53] presented a system that recognizes 30 syllables from EMG signals with a mean accuracy of almost 70%. Wand et al. [54] used an LDA matrix of TD features as input and a Gaussian mixture model (GMM) to classify 45 English phones (+ silence class), which resulted in an accuracy of 19.24%. Furthermore, Schultz et al. [55] found that modeling co-articulation reduced the WER in a speaker-dependent 101-word recognition task from 47.15 to 31.49%. They modeled co-articulation by adding bundled phonetic features referring to place and manner of articulation such as *voiced fricative* or *rounded front vowel*.

2.6.4 EMG-to-speech conversion

Toth et al. [56] introduced a direct EMG-to-speech mapping approach based on a frame-based voice conversion model, which required the fundamental frequency (F0) values from the acoustic signals. Then, Nakamura et al. [57] used a support vector machine (SVM) that recognized whether a frame of an EMG signal was voiced or unvoiced with an accuracy of 84%. When attempting to estimate the F0 from EMG signals using a GMM-based voice conversion model, the result sounded unnatural. Zahner et al. [58] used a unit selection approach to convert EMG signals to audible speech, which yielded an average Mel-cepstral distortion (MCD) of 5.4. MCD is a metric to evaluate the quality of speech synthesis by measuring the difference in MFCCs between the original target signal and the converted synthesized signal. The lower the MCD value, the higher the similarity between the two signals, indicating a higher quality of the synthesized speech. Diener et al. [36] did a model comparison, and a subjective evaluation showed a preference for the result of deep neural network (DNN) feature mapping over a GMM. Janke et al. [14] used a DNN, which resulted in a mean MCD of 5.21 and a WER of 6.1-28.3%. However, they acknowledged that the overall intelligibility was low. Diener et al. [59] found that the conversion of continuous speech resulted in a higher MCD than isolated speech, even though the dataset of continuous speech was larger. More advances in this task were made by Gaddy et al. [60–62], who introduced a model that is trained on nearly 19 hours of open-vocabulary EMG signals acquired in both audible and silent mode from one typical male English speaker. Their approach consists of a transduction model to predict a set of audio features and a vocoder model to turn those features into waveforms. To be able to train on silent EMG signals, they applied a technique called dynamic time warping (DTW) to align these signals with the parallel audible signal. When testing the model on EMG signals from silent speech, they achieved a WER of 36.1% from automatic transcriptions and 32.3% by human evaluation. This is a large improvement compared to the 88.3% (automatic) and 95.1% (human) WER on their baseline model which was trained on EMG signals from audible speech only. Their work is considered the current state-of-the-art for EMG-to-speech conversion in English, mainly because of the acceptable intelligibility of the resulting signals when testing on silent speech.

2.6.5 Speaker-independent models

Most of the above-mentioned studies have acknowledged a large negative impact of inter-speaker variability on the recognition performance, and therefore have focused on improving speaker-dependent models only. However, a recent study by Zhang et al. [63] introduced a multi-speaker model using an improved conditional domain adversarial network (ICDAN). They reached an average accuracy of 86.32%.

2.6.6 Other languages

English is the most studied language in this field, but efforts towards other languages were made as well. For example, Mostafa et al. [64] achieved an overall accuracy of 82.3% on 11 Bangla vowels, using three channels. Soon et al. [41] compared four classifiers and found the best results (62.7% accuracy) with a deep neural network (DNN) using only one EMG channel, and performed this study for the Malay language. Ma et al. [42] reached a 72% classification accuracy using random forests, with four EMG channels, classifying 10 Chinese phonemes. Li et al. [65] proposed an optimized sequence-to-sequence approach to perform direct EMG-to-speech generation for Mandarin Chinese, and achieved a character error rate (CER) of 6.41% on average using human evaluation. They used the model by Gaddy and Klein for English [60] as a baseline, which resulted in a much higher average CER, namely 39.76%. Deng et al. [66] performed word classification experiments using a convolutional neural network (CNN) and achieved an 88.31% classification accuracy on 33 Chinese words. Li et al. [67] managed to reach 100% accuracy (and an average of 82.3%) when classifying ten Mandarin numeric words using a SVM. Nonetheless, Zhu et al. [68] found a higher accuracy when classifying the 10 digits in Chinese compared to English and argued that it is important to pay attention to language differences when deciding the optimal electrode setup for best practices of silent speech recognition. It suggests that the best-performing model might differ per language as well.

2.6.7 Speech-to-EMG and paralinguistics

Additionally, some other studies in this area are worth mentioning. First of all, as an effort towards the domain of acoustic-to-articulatory conversion, Botelho et al. [69] introduced a study of speech-to-EMG, where they tried to recover the EMG signal from the acoustic signal. They found this task was harder in a multi-speaker modality than in a multi-session modality. Related to the challenge of inter-speaker variability in EMG signals, is a study by Diener et al. [70], in which they were able to predict the speaker ID based on their EMG signals with an average recall of 73%, for 5 speakers. They were also able to predict the speaking mode of a speaker above chance level (audible: 50.6%, whispered: 50.3%, silent: 74.4%), which is also not surprising, as differences in signals between speaking modes were already established [35].

2.6.8 Alaryngeal speech

All of these studies were performed on healthy and typical individuals. This can be logical depending on the goal of the task, but when the goal is to restore speech for alaryngeal speakers, as is our case, it would be crucial to include this kind of speech as well. To our knowledge, Meltzner et al. [40] have been the only ones to compare the speech of these two types of speakers. They found a WER of 10.3% for the alaryngeal speakers. Furthermore, it appeared that the optimal sensor set was different from typical speakers: sensors located near the surgery site had less value in the case of alaryngeal speakers.

2.6.9 Summary

To summarize, many efforts have been made in the area of EMG-based speech interfaces. Depending on the language, we can say that currently, researchers have been able to directly predict speech from facial and neck muscles with acceptable intelligibility, but this is heavily speaker-dependent and relies on large amounts of data. Furthermore, there is a major lack of research into silent speech interfaces for Spanish alaryngeal speakers, which is the target group of our project. This means that more research is required, in all areas of this topic.

Part II

The Database

3
Finding the Optimal Electrode Setup

This chapter describes a series of pilot experiments designed to define the electrode setup for the new database described in Chapter 4. Motivated by the scarcity of information in related studies regarding this important decision-making process, we decided to carry out a set of experiments with multiple recording sessions and different setups. We included different electrode types (paired and concentric) and locations targeting different muscles in the face and neck involved in speech production. We then analyzed the results obtained in a phone classification task using frame-based phone accuracy. The final setup consists of eight channels with bipolar single-electrode pairs targeting eight specific muscles crucial for capturing speech-related information.

3.1 Introduction

As described in Chapter 1, a database of Spanish speech and EMG signals is required to develop an SSI for Spanish alaryngeal speakers. More specifically, an SSI is developed through machine learning, meaning that a computer model is trained and tested on large amounts of signals. The final interface takes the EMG signals acquired while articulating silently and translates these into a synthetic voice. For the best performance, it is essential to determine the optimal electrode setup with which the EMG signals are acquired. Decisions have to be made regarding the electrode type, number, and locations, while considering practicality and speaker comfort. The process and results of making these decisions are described in this chapter.

Over the years, several electrode setups have been used to acquire EMG from facial muscles, differing in configuration type (monopolar, bipolar, array electrodes, or a combination of them), shape (circular or rectangular), number of channels (ranging from one to more than 100), and locations of the electrodes. Table 3.1 provides an overview of previous studies and the electrode setups used. The first column contains a reference to the study, the second column shows the type and number of electrodes, and the third column lists the locations of those electrodes. The locations refer to anatomical regions, or muscles (see Figures 2.2 and 2.3 for images of the muscles in the face and neck). For a description of the different configuration types, see Section 2.4. When studies are grouped, the setups in these studies are the same. The table is divided by double lines into four sections, corresponding to four approaches used to select the electrode locations.

The first approach is targeting specific muscles. In a series of studies by different research groups [33, 34, 36, 45, 46, 54, 55, 71], a group of five muscles was targeted, namely the levator anguli oris (LAO), the zygomaticus major (ZYG), the platysma (PLT), the depressor anguli oris (DAO), and the anterior belly of the digastric (ABD), i.e. the superficial muscle most related to the tongue. In other studies, different muscles such as the buccinator (BUC), orbicularis oris (OBO), mentalis (MNT), levator labii superioris (LLS), mylohyoid (MLH), sternocleidomastoid

(SCM), or the risorius (RIS), i.e. the laughing muscle, were targeted [41, 42, 48, 64, 67, 72].

In the second approach, no specific muscles are targeted, but anatomical regions. A series of studies using this approach by a group of the same researchers include [37–40, 50, 73]. More recently, the same approach was used by Gaddy and Klein [60–62].

The third approach is a high-density electrode setup without targeting a particular muscle or anatomical region, either using electrode arrays [36, 71, 74], or all single electrodes [68, 75–78].

A fourth approach was proposed recently by [66] in which they used arrays to select eight electrodes to target specific muscles.

Table 3.1: Electrode number, type, and locations in previous studies. Each of the sections lists studies with a similar approach. Grouped studies used the same setup.

Reference	**Electrodes**	**Locations**
Chan et al. (2001, 2002) [45, 46]	5 pairs	LAO, ZYG, PLT, DAO, ABD
Maier-Hein et al. (2005) [33]	3 bi- and 4 monopolar pairs)	LAO, ZYG, PLT, DAO, ABD, Tongue
Jou et al. (2006) [34]	2 bi- and 4 monopolar pairs)	LAO, ZYG, PLT, ABD, Tongue
Schultz and Wand (2010) [55]; Wand and Schultz (2011) [54]	2 bi- and 3 monopolar pairs)	LAO, ZYG, PLT, ABD, Tongue
Diener et al. (2015) [36]; Diener (2021) [71]	2 bi- and 3 monopolar pairs)	LAO, ZYG, PLT, ABD, Tongue
Mostafa et al. (2016) [64]	3 electrodes	MAS, BUC, Depressor

Continues on next page

Table 3.1 – continued from previous page

Reference	Electrodes	Locations
Soon et al. (2017) [41]	1 pair	OBO
Ma et al. (2019) [42]	2 monopolar electrodes, 2 bipolar pairs	RIS, ABD, LIN, LAO
Wang et al. (2021) [48]	4 pairs	LAO, DAO, BUC, ABD
Wu et al. (2022) [72]	6 pairs	MNT, RIS, LLS, ABD, MLH, PLT
Li et al. (2023) [67]	6 tripolar	OBO, MAS, lower lip muscle, biabdominal anterior abdomen, inferior lateral muscle of the hyoid bone, SCM
Meltzner et al. (2008) [38]; Colby et al. (2009) [37]	11 bipolar bars	supralabial, labial, sublabial, submental neck, midline neck, lateral neck
Meltzner et al. (2011) [39]	8 single-differential bars	submental neck, ventromedial neck, supralabial face, infralabial face
Deng et al. (2014) [50]	4 sensors	above and below the oral commissure, submental surface, ventral neck surface
Meltzner et al. (2017) [40]	8 differential bars	submental, ventromedial, supralabial, infralabial
Meltzner et al. (2018) [73]	11 sensors	submental region, ventral neck, face
Gaddy and Klein (2020, 2021) [60, 61]; Gaddy (2022) [62]	8 monopolar electrodes	left cheek just above mouth, left corner of chin, below chin back 3 cm, throat 3 cm left from Adam's apple, mid-jaw right, right cheek just below mouth, right cheek 2 cm from nose, back of right cheek; 4 cm in front of ear

Continues on next page

Table 3.1 – continued from previous page

Reference	Electrodes	Locations
Wand et al. (2013) [74]	two 1x8 strips	cheek, chin
Wand et al. (2013) [74]; Diener et al. (2015) [36]; Diener (2021) [71]	4x8 grid, 1x8 strip	cheek, chin
Zhu et al. (2019, 2020, 2021) [68, 75, 76]; Wang et al. (2020, 2021) [77, 78]	120 high-density electrodes	cheeks, neck
Deng et al. (2023) [66]	8 electrodes within two 32-channel arrays	ZYG, RIS, DAO, SCM, ABD, PLT

From the four approaches, we believe that the one where specific muscles are targeted is the most accurate for the task of predicting speech from facial muscles, considering the physiology of muscles and the muscular anatomy of the face. First of all, while electrode arrays might be suitable for large muscles, we believe it is not the best approach when considering facial muscles. The arrays are rigid and therefore difficult to adjust to the movement of the muscles while speaking. Furthermore, knowing that activation potentials travel along the length of a muscle, longitudinal acquisition of the target muscle is required. However, most facial muscles are long, narrow, and close to each other, so using an array (or a high-density setup with single electrodes) increases the risk of cross-talk.

Figure 3.1 shows a diagram of the studies and how they are connected. Here you can see that, as is normal in academic research, often one study is a continuation of a previous study by a group of the same authors. However, we realized that the studies that follow the approach that we are interested in (see section one of Table 3.1) go back to one study [45] or do not provide a systematic approach. For this reason, we performed this pilot study to find the optimal electrode setup.

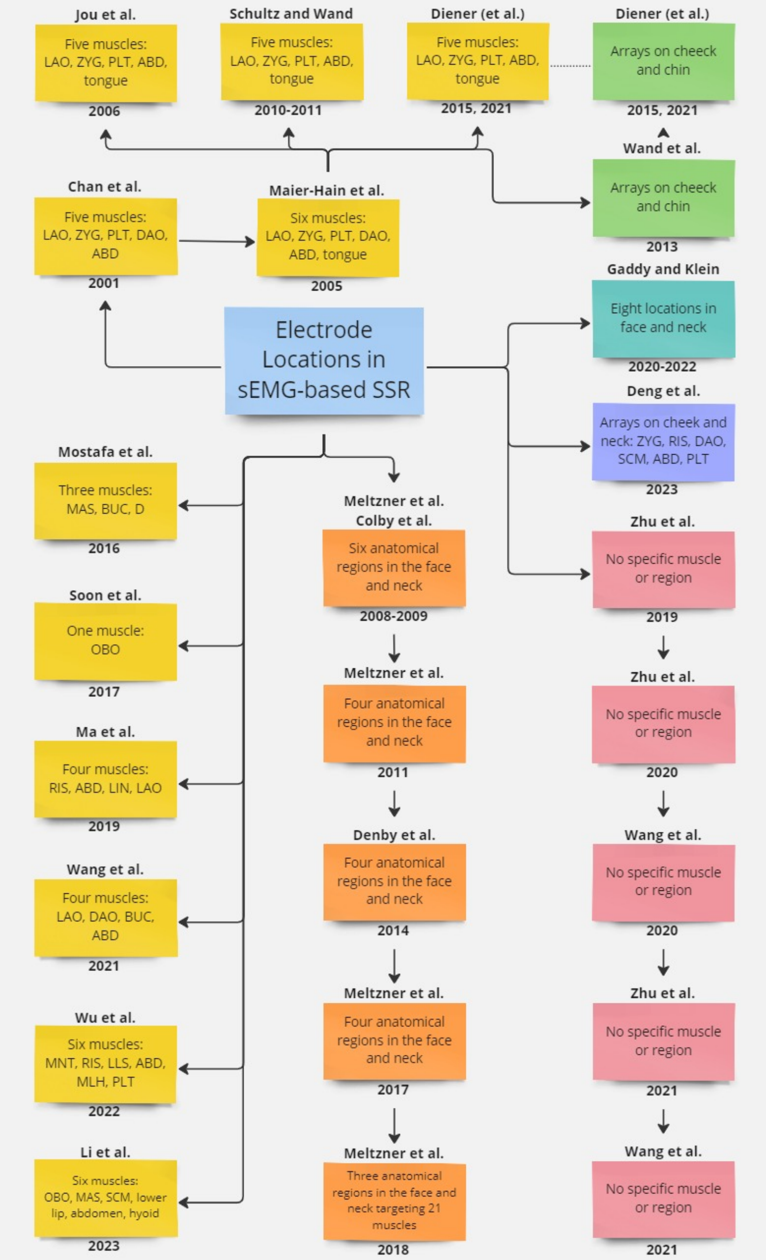

Jou et al.
Five muscles: LAO, ZYG, PLT, ABD, tongue
2006
Schultz and Wand
Five muscles: LAO, ZYG, PLT, ABD, tongue
2010-2011
Diener (et al.)
Five muscles: LAO, ZYG, PLT, ABD, tongue
2015, 2021
Diener (et al.)
Arrays on cheeck and chin
2015, 2021
Wand et al.
Arrays on cheeck and chin
2013
Chan et al.
Five muscles: LAO, ZYG, PLT, DAO, ABD
2001
Maier-Hain et al.
Six muscles: LAO, ZYG, PLT, DAO, ABD, tongue
2005
Gaddy and Klein
Eight locations in face and neck
2020-2022
Electrode Locations in sEMG-based SSR
Deng et al.
Arrays on cheek and neck: ZYG, RIS, DAO, SCM, ABD, PLT
2023
Mostafa et al.
Three muscles: MAS, BUC, D
2016
Meltzner et al. Colby et al.
Six anatomical regions in the face and neck
2008-2009
Zhu et al.
No specific muscle or region
2019
Soon et al.
One muscle: OBO
2017
Meltzner et al.
Four anatomical regions in the face and neck
2011
Zhu et al.
No specific muscle or region
2020
Ma et al.
Four muscles: RIS, ABD, LIN, LAO
2019
Denby et al.
Four anatomical regions in the face and neck
2014
Wang et al.
No specific muscle or region
2020
Wang et al.
Four muscles: LAO, DAO, BUC, ABD
2021
Meltzner et al.
Four anatomical regions in the face and neck
2017
Zhu et al.
No specific muscle or region
2021
Wu et al.
Six muscles: MNT, RIS, LLS, ABD, MLH, PLT
2022
Li et al.
Six muscles: OBO, MAS, SCM, lower lip, abdomen, hyoid
2023
Meltzner et al.
Three anatomical regions in the face and neck targeting 21 muscles
2018
Wang et al.
No specific muscle or region
2021

Figure 3.1: Diagram of studies and their electrode setups.

Regarding the distinction between monopolar and bipolar EMG acquisition configuration, the studies found [31, 79–81] are not very conclusive and not focused on small muscles such as muscles of the face. However, we were strongly advised by the equipment provider in favor of bipolar configuration, which we adopted. Bipolar acquisition ensures that there are two measuring points for the same muscle and, when placed correctly, activity from a non-target muscle can be canceled out. We compared two types of bipolar electrodes, namely concentric electrodes and pairs of single electrodes. The results of this experiment are included in this chapter. In addition, we tried cup electrodes that are usually used for the scalp, due to their smaller size, but they appeared to be too impractical to be used in the face. The main problem was that these electrodes would not stay in place, as a result of movements in the face and the weight of the cables.

To select the locations of the electrodes, we targeted 14 muscles in the face and neck of one participant and used the results of per-channel phone classification experiments to discard the least useful channels, resulting in an 8-channel setup targeting five muscles in the face and three in the neck: ABD, DAO, RIS, LLS, MAS, ZYG, depressor labii inferioris (DLI), and stylohyoid (SLH).

This chapter is organized as follows. In the next section, the materials and methodology of the pilot study are described in more detail, namely how the data is collected (Section 3.2.1) and processed (Section 3.2.2), and the experimental part (Section 3.2.3) of the classification tasks. Then, the results of the electrode type comparison (Section 3.3.1) and the channel selection (Section 3.3.2) are summarised. Finally, we provide a discussion and conclusion (Section 3.4).

3.2 Methods

This section describes the methodological part of this study. First, we explain how we collected the data, the materials we used, and the resulting pilot database (Section 3.2.1). Then, we describe how we processed the signals and extracted features from them (Section 3.2.2). Lastly, we explain the experiments we performed and the models we used, and how we evaluated the outcomes of these experiments (Section 3.2.3).

3.2.1 Data overview

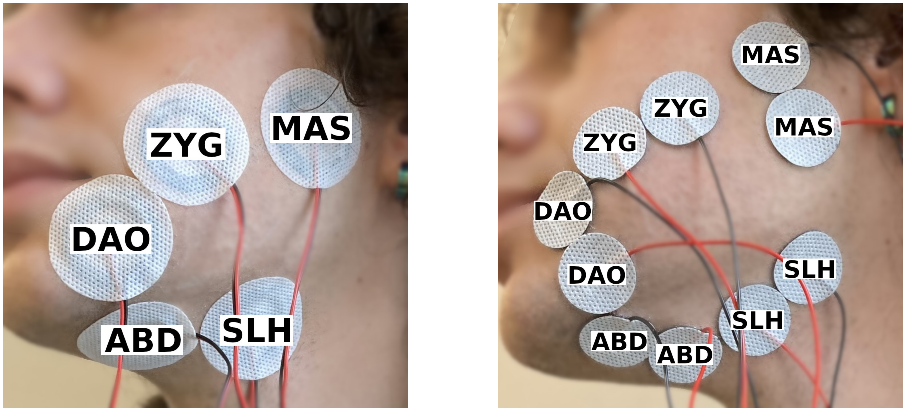
We collected data in three sessions on three different days by one male native Spanish speaker. During a session, the participant could take a short rest if required, but preferred not to. Each session was designed with a different goal in mind and served as data for different experiments (see Section 3.2.3). Table 3.2 provides an overview of the electrode setups per session, specifically the muscles that were targeted in each session.

Table 3.2: Electrode setups (targeted muscles) for each session in the pilot study plus the final setup for the official sessions of the database.

Muscle	**Pilot session**			**Official**
	1	**2**	**3**	**sessions**
anterior belly of the digastric (ABD)	X	X	X	X
depressor anguli oris (DAO)	X	X	X	X
depressor labii inferioris (DLI)		X	X	X
frontalis (FRT)			X	
levator anguli oris (LAO)		X		
levator labii superioris (LLS)			X	X
masseter (MAS)	X	X	X	X
orbicularis oris (OBO)		X		
platysma (PLT)		X		
posterior belly of the digastric (PBD)		X		
risorius (RIS)		X	X	X
sternocleidomastoid (SCM)		X		
stylohyoid (SLH)	X	X	X	X
sternothyroid (STR)		X		
superior belly of the omohyoid (SBO)		X	X	
zygomaticus major (ZYG)	X	X	X	X

We placed the electrodes in the middle of the muscle. In the case of the electrode pairs, the electrodes were placed next to each other in the direction of the muscle fiber. As a reference, we used images of the respective muscles[1] and an online 3D anatomy visualizer[2].

In terms of electrode type and size, we compared bipolar concentric electrodes (Figure 3.2a) to bipolar single-paired electrodes (Figure 3.2b), referred to as Session 1. On the one hand, the positions of the two electrodes in a concentric electrode were fixed, which could help reduce inter-session variability. On the other hand, a concentric electrode had a larger diameter (40 mm) than a single electrode (24 mm), which could result in more cross-talk. There was no inter-electrode distance (IED) between the two bipolar electrodes. The inner diameter of the concentric electrode was 10 mm and the outer diameter was 31 mm. The participant recorded 250 phonemically balanced short sentences taken from the Sharvard Corpus [82], once with each electrode setup, which consisted of five channels on the left side of the face.



(a) Concentric electrodes

(b) Single-paired electrodes

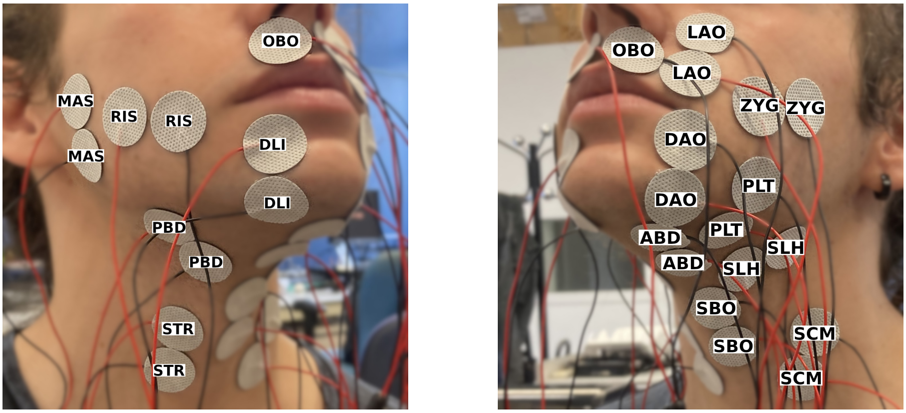
Figure 3.2: Electrode setup for Session 1, made up of five channels targeting the same set of five muscles but using two different types of electrodes.

To see which muscles were most significant, we did a session

[1] www.learnmuscles.com, last accessed on 25/09/2024

[2] www.zygotebody.com, last accessed on 25/09/2024

(Session 2) in which we placed 14 single-electrode pairs in an attempt to target 14 superficial muscles in the lower face, chin, and neck area (Figure 3.3). The initial plan was to make the setup symmetrical, but during the electrode placement, it turned out that the 14 channels had to be divided over both sides of the face due to lack of space, resulting in an asymmetrical setup. Each of the Spanish consonants was paired once with each of the five vowels in Spanish, and the participant recorded the resulting 105 consonant (C)-vowel (V) combinations three times in a row. Context was added to each combination, in the format *ata[C][V]ta*, to control for co-articulation.

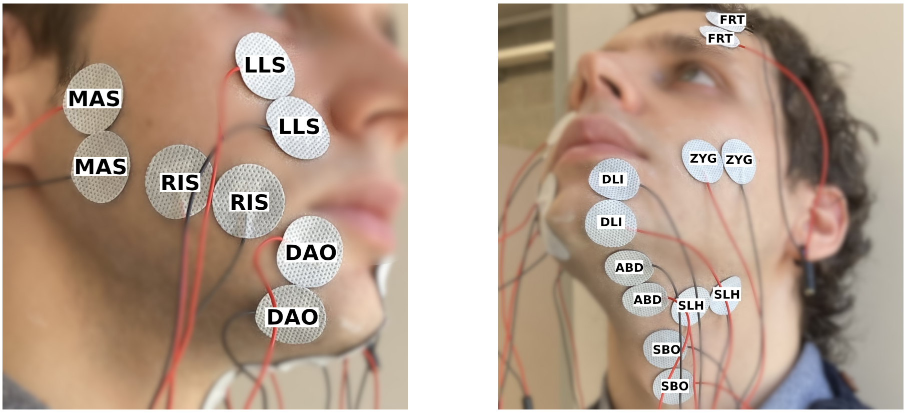


(a) Right side

(b) Left side

Figure 3.3: Electrode setup for Session 2, consisting of 14 channels targeting a different muscle each.

After analyzing and comparing the 14 channels of Session 2 (Section 3.3.2), we recorded another session (Session 3) to finalize the electrode setup. More specifically, we wanted to know if there were no large differences in performance between the electrodes we selected. The 250 sentences from the Sharvard Corpus were recorded two times. See Figure 3.4 for the electrode setup.



(a) Right side

(b) Left side

Figure 3.4: Electrode setup for Session 3, consisting of 10 channels targeting a different muscle each.

Section 4.1 describes in detail the acquisition setup in which the signals were acquired. In short, we collected the EMG data with a bio-electrical amplifier, and the speech data with a microphone and sound device, in a sound-proof room and using a silent computer. To be able to align the EMG and audio signals, a synchronization signal is shared between the amplifier and sound device.

The 26 phone classes present in the CV combinations are those presented in Table 4.2, except /L/, /j/, and /w/. The two Spanish semivowels /j/ and /w/ were left out of the CV combinations since they are neither consonants nor vowels. The lateral palatal /L/ is often replaced by central palatal /jj/ by many peninsular Spanish speakers (a linguistic phenomenon called 'yeismo'), which is why we did not consider it when creating the CV combination dictionary manually. The sentence set contains all the phones from Table 4.2, so 29 in total.

We split the data of each session into 80% for training and 20% for testing. For the CV words recorded in Session 2, we made sure that the balance of CV combinations was similar for the train and the test set. For the 250 sentences recorded in sessions 1 and 3, we assigned the last

20% of the sentences to the test set. See Table 3.4 for the amount of data in time for each subset of each session.

Table 3.4: Overview of the durations of the train and test data sets per session and electrode setup in the *mm:ss* format.

	Electrode setup	**Corpus**	**Train set**	**Test set**
Session 1	5 paired	250 sentences	09:12	02:16
	5 concentric	250 sentences	09:02	02:17
Session 2	14 paired	105 CV x3	01:00	00:15
Session 3	10 paired	250 sentences x2	17:16	04:23

3.2.2 Data processing

To perform the phone classification experiments, the raw EMG and audio signals needed to be processed and parameterized.

First, both audio and EMG signals were cut using the synchronization signal. Subsequently, each audio signal was automatically aligned with its phonetic labels using the Montreal forced aligner (MFA) [83].

Then, we parameterized the EMG signals by calculating a set of time-domain (TD) features. These features have been widely used in works related to EMG signals applied to speech recognition or generation [34, 60, 84]. In Section 2.4 a detailed description of EMG feature extraction in the relevant literature, including the use of TD features, is provided.

The initial step involves removing direct-current (DC) offsets from each EMG signal for the duration of each utterance, as defined by the synchronization signal. A DC offset is a baseline voltage that is not zero, which can be seen as a constant background noise. Removing it makes it easier to focus on the actual muscle activity. After this step, each signal is normalized by dividing it by its maximum absolute value. This ensures that all signal amplitudes are scaled consistently and easier to compare.

To obtain the TD features, we first separated the EMG signal ($x[n]$) into a low-frequency signal ($w[n]$) and a high-frequency signal ($p[n]$). The low-frequency signal was obtained by calculating a double average of $x[n]$ using a nine-point window. The calculation can be expressed as:

$$w[n] = \frac{1}{9}\sum_{k=-4}^{4} v[n+k], \quad \text{where } v[n] = \frac{1}{9}\sum_{k=-4}^{4} x[n+k] \tag{3.1}$$

The high-frequency signal, $p[n]$, was obtained by subtracting $w[n]$ from $x[n]$. This can be represented as:

$$p[n] = x[n] - w[n] \tag{3.2}$$

In addition, a rectified version of the high-frequency signal, $r[n]$, was calculated as follows:

$$r[n] = \begin{cases} p[n], & \text{if } p[n] \geq 0 \\ -p[n] & \text{if } p[n] < 0 \end{cases} \tag{3.3}$$

With the low-frequency signal ($w[n]$), high-frequency signal ($p[n]$), and rectified high-frequency signal ($r[n]$) obtained, we computed the set of five TD features for each frame, using a window with a duration of 25 ms and a frameshift of 5 ms. For $w[n]$ and $r[n]$, the frame-based power (P_w and P_r) and the frame-based time-domain mean ($\bar{w}$ and $\bar{r}$) were calculated, and for $p[n]$ the frame-based zero-crossing rate (z). These features are defined as:

$$TD0 = [\bar{w}, \bar{r}, P_w, P_r, z] \tag{3.4}$$

where:

$$\bar{w} = \frac{1}{N}\sum_{n=0}^{N-1} w[n], \quad \bar{r} = \frac{1}{N}\sum_{n=0}^{N-1} r[n] \tag{3.5}$$

$$P_w = \frac{1}{N}\sum_{n=0}^{N-1}|w[n]|^2, \quad P_r = \frac{1}{N}\sum_{n=0}^{N-1}|r[n]|^2 \tag{3.6}$$

$$z = \sum_{n=1}^{N-1} g(p[n]p[n-1]), \quad \text{where } g(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{if } x \geq 0 \end{cases} \tag{3.7}$$

where N denotes the number of samples in $x[n]$. To incorporate temporal context into the features, a stacking filter was used to concatenate the features of $2k + 1$ adjacent frames, where k represents the width of the stacking filter. We selected $k = 15$, resulting in a total of 31 frames being combined, with the analyzed frame in the center. The stacked TD0 vectors from all channels were then combined into a single array, which served as the input for the classifier. The length of the parameter vector assigned to each frame can be calculated as:

$$M \cdot 5 \cdot (2k + 1) \tag{3.8}$$

where M represents the number of channels.

To reduce the dimension of the parameter vector, we applied linear discriminant analysis (LDA) [85], as done in [54, 86]. The number of features is equal to the number of classes present in the data (phone labels) minus 1, which is the maximum allowed number of features in LDA reduction. In the case of the CV dataset, this resulted in 25 features, and in the case of the sentences, this resulted in 28 features.

3.2.3 Experiments

With the three experiments we performed, we had two goals. The first goal was to find out which type of bipolar electrodes would yield the highest accuracy. We used the data from Session 1 and did a phone classification task using the signals of all five channels, one time with the signals from the concentric electrodes and another time with the single-paired electrodes. The second goal was to select the optimal set of electrodes, regarding their number and locations. For this, we did two experiments with the data from sessions 2 and 3. To assess the amount of information provided by each muscle, a phone classification experiment was performed using the signals from one single channel each time. The muscles that achieved the highest accuracy were considered to contain the most useful information to do the task.

Since the size of the data set used in each experiment was limited to only one session, we did not want to base our conclusions on one classifier only and decided to compare three classifiers. The first was a Gaussian mixture model (GMM), which has been used in phone classification experiments before [54]. The second was a bagging classifier with decision trees (DT) as estimators, which we thought appropriate for the small data size. The third was a feed-forward neural network (NN), which we wanted to include since NNs are the most standard type of machine learning model used in recent years.

The maximum number of components in the GMMs was equal to the number of classes minus 1. Starting with 1 component, it continued adding components until the Bayesian Information Criterion (BIC) of the new model was higher than the last model´s BIC.

The number of decision trees for the DT models was 50 for Session 2, and 100 for sessions 1 and 3. The minimum number of samples in the leaf node was set to 5 for Session 2, and to 10 for sessions 1 and 3. These parameters were set following a parameter tuning experiment, in which we tried different combinations of parameter values and chose the one that resulted in the highest validation accuracy.

For the NN we used one hidden layer, with twice the number of features as input nodes, and the ReLU activation function. The output layer consisted of as many nodes as there were phone labels, and the softmax activation function. It was compiled using the cross-entropy loss function and the Adam optimizer. We used a batch size of 32 and a train size of 25 epochs for the experiments with data from Session 2, and a batch size of 64 and a train size of 50 epochs for the experiments with data from sessions 1 and 3. These parameter values were determined after training a classifier for 100 epochs with batch sizes 32, 64, and 128, and choosing the combination from the point the validation accuracy stopped increasing.

For each model, 5-fold cross-validation was implemented on the training set (which was 80% of the complete data set). We used the mean frame-based phone accuracy of the five validation sets as an evaluation measure to select the electrode locations and type. For Session 1 we applied a Wilcoxon Signed-Rank Test to check for statistical differences. We used the test accuracy to evaluate if the final setup was appropriate.

3.3 Results

This section summarizes the results of the experiments. First, the results of the comparison of electrode types (Section 3.3.1), and then the results that were used to select the channels (Section 3.3.2), are shown.

3.3.1 Electrode type

Figure 3.5 shows the mean validation accuracy per classifier and type of bipolar electrode, obtained with data from Session 1.

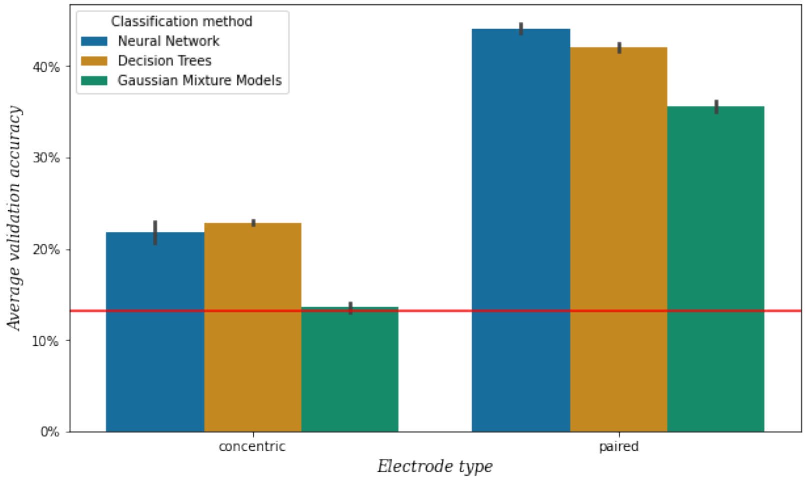


Figure 3.5: Mean frame-based validation accuracy after 5-fold cross-validation per electrode type and classification method, obtained from the data of Session 1. The vertical bars represent the confidence intervals, and the red line represents the baseline, which is the mean validation accuracy when always predicting the most frequent class [a].

It appears that, for all classifiers, the phone classification accuracy is significantly higher when using single-paired electrodes compared to concentric electrodes ($p < 0.001$). For this reason, we selected the paired electrodes for our optimal electrode setup.

3.3.2 Channel selection

Figure 3.6 shows the mean validation accuracy per classifier and channel, obtained with data from Session 2.

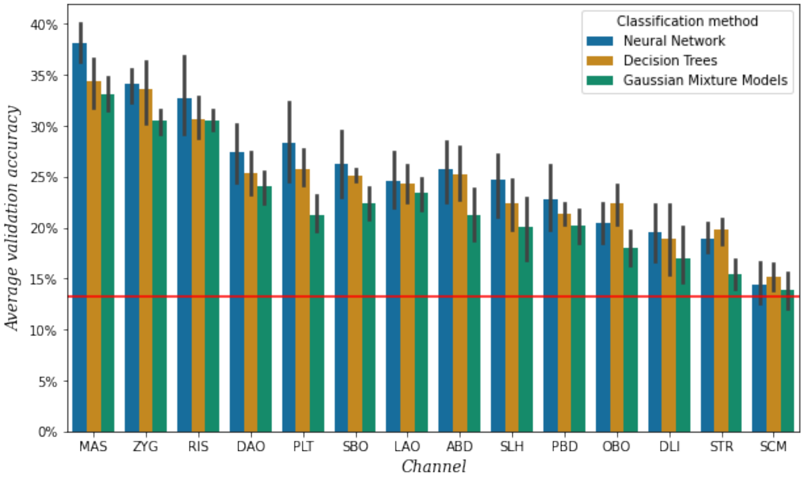


Figure 3.6: Mean validation accuracy after 5-fold cross-validation per channel and classification method, obtained from the data of Session 2. The vertical bars represent the confidence intervals, and the red line represents the baseline, which is the mean validation accuracy when always predicting the most frequent class [e].

It appears that for most channels, the highest accuracy is achieved by the NN and the lowest by the GMM. Interestingly, the order of the channel with the highest to the one with the lowest accuracy is different for each classifier. However, for all three classifiers (both separately and averaged), the six channels with the lowest accuracy are SLH, PBD, OBO, DLI, STR, and SCM.

It is important to mention that the electrodes of the three channels around the mouth, namely both electrodes of OBO, and the top electrodes of DAO and DLI, did not stick as well as the electrodes of the other channels. This was most likely due to the area under the electrodes being curved as a result of lip movement. We had to reattach these electrodes a few times during the session. The OBO channel had the most severe attachment problem as it was also affected by sweat and condensation of air coming from the nose.

We repeated the experiment with the data of the three rounds of the sessions separately, and it turned out that DLI belonged to the top 5 of highest accuracy in the first round, but decreased with each round. For this reason, we decided not to discard this channel yet.

After discarding OBO, STR, and SCM for their low performance, we took another look at the muscular anatomy of the remaining channels. The muscles SLH and PBD are located very close together and performed similarly as well, so we decided to only discard channel PBD and keep SLH, although in practice channel SLH most probably represents information from both muscles. In addition, we realized that the LAO is a very short muscle, but that the LLS is a closely located but longer muscle. So we decided to replace LAO with LLS because longer muscles are easier to target, and additionally to avoid the area directly above the lips. Furthermore, we saw that the PLT is a broad sheet of muscle instead of a muscle with a more specific location, making it difficult to know whether the information we are measuring belongs to this muscle. Therefore we decided to remove the channel corresponding to PLT.

Additionally, we added one new channel for the muscle of the forehead, the frontalis (FRT). The purpose of this channel was to be used as a reference, as it was not expected to provide any muscle information related to speech.

Finally, the set of 10 channels included in the next recording session (Session 3) was the following: MAS, ZYG, RIS, DAO, SBO, LLS (instead of LAO), ABD, SLH, DLI, and the new FRT (see Figure 3.4).

Figure 3.7 shows the test accuracy per channel and classifier after performing the classification experiment described in Section 3.2.3 on the data from Session 3. It can be seen that the channel with the lowest test accuracy is FRT with a performance similar to baseline. This result provides an extra assurance that the other channels indeed carry some information related to speech production. The highest test accuracy when using all the channels except FRT was achieved with an NN, at 48.42%.

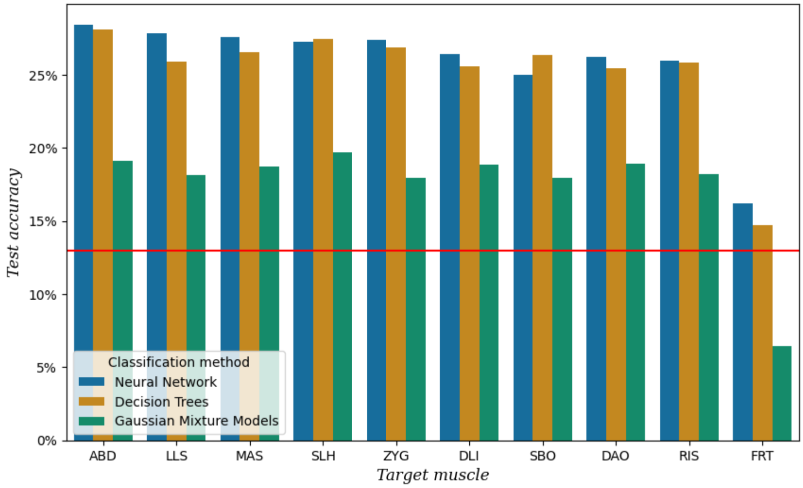


Figure 3.7: Test accuracy per channel and classification method for the data of Session 3. The red line represents the baseline, which is the test accuracy when always predicting the most frequent class [a].

For the final setup, we left out FRT for obvious reasons, but SBO as well. This channel is located in the area where the stoma is located in alaryngeal speakers. For studies with a different target group, this muscle might be a valuable addition, but for our study, we realized it was not practical.

The final setup, containing ABD, LLS, MAS, SLH, ZYG, DLI, DAO, and RIS, has been used to record the ReSSInt database (described in Chapter 4). For the experiments in the following chapters, data from this database is used.

3.4 Discussion and Conclusion

Following a common approach in previous studies where individual muscles in the face and neck are targeted [33, 34, 36, 41, 42, 45, 46, 48, 54, 55, 64, 67, 71, 72], we looked at the contribution of 14 individual muscles in a phone classification task. As a result of this study, we decided to include eight bipolar single-electrode pairs targeting one muscle each, of which five are located in the face and three in the neck, in an asymmetrical setup. Out of the eight muscles, six are present in the setups of at least one of the studies mentioned above as well, namely the ABD, DAO, RIS, LLS, MAS and ZYG. The LAO is more commonly used instead of LLS, but we chose LLS because it is a longer muscle. There are two more muscles that we included, namely the DLI and SLH. As far as we know these muscles have not been used in previous research, however, they have proven to be valuable in our experiments.

One important limitation of the study described in this chapter is that the experiments have been done with only one speaker. However, the channels that we discarded for reasons other than practicality had a noticeably lower performance than the channels we selected. We believe that the selected setup could be generalized to other speakers, as the muscles used for articulation are the same regardless of individual variance in the manner of articulation, but that some channels might be more useful than others depending on the speaker.

Additionally, due to the lack of space in the face of the speaker, we had to place the electrodes asymmetrically, and we assumed that this would not cause any difficulties, since the musculature of the face is in theory symmetrical. Multiple studies listed in Table 3.1 use an asymmetrical setup. However, we acknowledge that there is a possibility that the results could have turned out differently if we mirrored the setup, and that this has to be researched further.

Our setup consists of eight channels, which is more than the number used in related studies, which varies from one [41], three [64], four [42, 48], five [36, 45, 46, 54, 55, 71], six [34, 67, 72] to seven [33].

We selected single electrodes for our setup because we experienced that electrode arrays were not flexible enough, which made it harder to target the specific muscle we were interested in. However, we acknowledge that arrays from a different manufacturer could be less rigid and therefore more useful. Furthermore, a recent study [66] found a way to overcome the issue of not targeting specific muscles when using arrays, by selecting electrodes within the arrays that are located on the target muscles. An advantage of using arrays is that it is less prone to electrode shifts between sessions, reducing session variability, so it could be worth examining this option in future database developments.

Note that for the pilot study experiments described in this chapter, we looked at the impact of each channel individually. However, we assume that for the production of each (combination of) sound(s), not one, but at least a group of two muscles (channels) are responsible.

In this chapter, we presented a series of pilot experiments we conducted to find the optimal electrode setup for developing a database of EMG and (silent) speech data. The final setup consists of eight bipolar single-electrode pairs each targeting a muscle in the face or neck that has proven most valuable and practical in the experiments. Future recommendations when selecting the optimal electrode setup are to look into the potential effect of asymmetry and the contribution of a group of channels in different linguistic contexts. Furthermore, it could be interesting to test the assumption that everyone uses the same muscles for speech by comparing multiple speakers.

3.5 Contribution

This chapter contributes to this research field for several reasons. First, it compares multiple electrode types and setups, which can help other researchers in their search for the optimal setup. Second, its description of finding the setup is more elaborate than most similar studies, where often it was not described why a certain setup was chosen. Third, it is the first study performed with data from Spanish speakers.

4
The ReSSInt-EMG Database

This chapter describes all the steps involved in the design and development of the ReSSInt-EMG database. This database of parallel EMG and speech signals contains 22.5 hours of data from nine Spanish-speaking participants of different sexes and ages, both typical and alaryngeal. The signals correspond to either audible or silent speech mode. Different text corpora were recorded, to be used for different purposes. The complete and diverse database is a valuable contribution to the development of an EMG-based SSI (for Spanish) and forms the basis of the experiments described in this thesis.

4.1 Introduction

The development of a database requires a lot of preparation and a well-established methodology. First, it is important to determine the goal of the database. In our case, as explained in Chapter 1, the final goal is to train a model that can predict speech from the muscle movements of alaryngeal speakers. Our plan was to train the model on phonetically labeled EMG signals and test it using unlabeled EMG signals. To assign phonetic labels, it is essential to know which phonetic output corresponds to each EMG signal. Therefore, we aimed to synchronize the audio signal with the simultaneously recorded EMG signals from typical speakers who produced speech audibly. By doing so, we could transfer the phonetic labels from the audio signal to the corresponding EMG signals. This is not possible to do with empty audio signals from silent speech. However, we wanted to include silent speech data from both alaryngeal and typical speakers to use as test data. Furthermore, we planned to include data from multiple speakers so that we could develop a multi-speaker interface.

However, during the database recording stage (which lasted more than a year), novel research [62] revealed that it is possible to train a model on audible speech in combination with silent speech using the DTW technique. This resulted in better model performance than using audible speech only, but it required a lot of data from one speaker (see Section 2.6 for more details on this study). For this reason, we adjusted our initial plan and recorded additional sessions for one speaker, focusing more on silent speech than audible speech.

This chapter is organized as follows. In the methodology (Section 4.2), we start with the description of the different types of speech content (text corpora) we collected and why (Section 4.2.1). Then, we describe the acquisition setup (Section 4.2.2), namely which hardware and software we used, and how we ensured synchronization between the audio and EMG signals. We finish this section by explaining the recording protocol step-by-step (Section 4.2.3), which we followed to ensure the database's consistency as much as possible. For the results (Section 4.3), we present the meta-information of the database (Section 4.3.1), in terms of speaker information, session information, and duration, along with some data examples (Section 4.3.2).

4.2 Methods

This section provides a detailed overview of the data collection procedure to develop the ReSSInt-EMG database. It describes the corpora (4.2.1), acquisition setup (4.2.2), and the recording protocol (4.2.3).

4.2.1 Text corpora

We refer to a text corpus as a body of text in a certain format that is used to control which linguistic content the speaker intends to articulate (with or without sound) while recording. Each corpus can be used for different purposes. We describe the three types of corpora that we used in the recording of the ReSSInt-EMG database below. Table 4.1 shows which content belongs to which corpus ID.

Table 4.1: Type of content related to the different corpus IDs.

Corpus ID	Content
001	110 VCV combinations
002	100 isolated words
003	Sharvard sentences 1-100
004	Sharvard sentences 101-400
005	Sharvard sentences 401-700
006	Ahosyn sentences 1-150
007	Ahosyn sentences 151-300
008	Ahosyn sentences 301-400
009	Ahosyn sentences 401-500
010	Ahosyn sentences 501-505
011	Ahosyn sentences 506-570
012	Ahosyn sentences 571-635
013	Ahosyn sentences 636-700
014	Ahosyn sentences 701-765
015	Ahosyn sentences 766-830
016	Ahosyn sentences 831-895

Vowel-Consonant-Vowel combinations

Following the Speech Assessment Methods Phonetic Alphabet (SAMPA) [87] for Spanish[1], there are 22 consonants, 2 semi-vowels, and 5 vowels. The consonants are further divided into plosives, affricates, fricatives, nasals, and liquids. Table 4.2 lists each speech sound with the SAMPA symbol, as well as the corresponding International Phonetic Alphabet (IPA) symbol for future reference.

Leaving the semi-vowels out of consideration due to their acoustic complexity means that there are 110 consonant (C)-vowel (V) combinations. A specific CV corpus can be useful due to the equal distribution of phones and combinations. Possible applications are to research phone-specific topics such as phone confusion, manner and place of articulation with relation to specific muscles, and level of difficulty for each phone. For this reason, the first text corpus used for the database (referred to as corpus 001) contains 110 non-sense words with the following format: at[VCV]ta[2]. We included these contexts to make sure each VCV combination is affected by the same co-articulation effects. To reduce the number of combinations, only combinations where the same vowel occurred twice were considered. The idea behind this is that all possible phonetic transitions are included, from V to C and from C to V. See Appendix A for the complete list of text prompts for the speaker to articulate.

[1] https://www.phon.ucl.ac.uk/home/sampa/spanish.htm, last accessed on 25/09/2024

[2] Note that the CV corpus used in the pilot study (Section 3.2.1) had one less CV combination, namely with the consonant /L/. Despite the *yeismo* phenomenon, we decided to include it in the corpus for the official database nonetheless. We also improved the format from CV to VCV.

Table 4.2: Spanish speech sounds per category, in SAMPA and IPA format.

	SAMPA	IPA
Consonants		
Plosives	/p/	[p]
	/b/	[b]
	/t/	[t]
	/d/	[d]
	/k/	[k]
	/g/	[ɡ]
Affricates	/tS/	[tʃ]
	/jj/	[ɟ]
Fricatives	/f/	[f]
	/B/	[β]
	/T/	[θ]
	/D/	[ð]
	/s/	[s]
	/x/	[x]
	/G/	[ɣ]
Nasals	/m/	[m]
	/n/	[n]
	/J/	[ɲ]
Liquids	/l/	[l]
	/L/	[ʎ]
	/r/	[r]
	/rr/	[ɾ]
(Semi-)vowels		
Semi-vowels	/j/	[j]
	/w/	[w]
Vowels	/i/	[i]
	/e/	[e]
	/a/	[a]
	/o/	[o]
	/u/	[u]

100 most useful Spanish words

The second text corpus (referred to as corpus 002) is a list of 100 words in Spanish which we considered most useful in daily communication. The words were taken from a website[3] that is focused on improving daily communication for people with reduced communication means. From each category, a few general words were selected. Then we also checked the phonetic balance to see if every speech sound in the Spanish language was represented. Since some of the nouns included in the list have two genders, we made an equal division between masculine and feminine variations (for example *hermano* and *nieta*, and not *hermana* and *nieto*). See appendix B for the complete list of words.

There were two reasons why we included this corpus. The first is so that we had an isolated corpus of words to be used for experiments such as word classification. The second is that by representing these words relatively more in the data used for training, the model would have a better chance of performing well with these most important words.

Sentences

The last type of content that was included were sentences, which were taken from two existing corpora. The first is the Sharvard Corpus, which is a phonetically balanced corpus of 700 Spanish declarative sentences [82]. On average there are eight to nine (8.48) words per sentence. The second is the Ahosyn Corpus, which was developed to record text-to-speech (TTS) databases [88]. From this corpus, we extracted 895 declarative and interrogative sentences, with an average of almost 13 (12.88) words per sentence. The corpora that contain these sentences are referred to as 003 to 016.

Sentences most closely resemble continuous speech, which the model ultimately needs to handle for fluent use of the interface. For this reason, sentences are the focus of the three corpora described here and make up the largest part of each session.

[3] https://arasaac.org/pictograms/search, last accessed on 25/09/2024

4.2.2 Acquisition setup

This section describes in detail the devices and environment used to record the database, and the recording protocol. The protocol also includes our approach to mitigate inter-session variability, namely the use of reference points and personalized 3D masks.

Hardware

For the acquisition of EMG signals, we used a 96-channel Quattrocento bio-electrical amplifier, a 16-channel bipolar adapter with a jack connector, 8 single-channel bipolar adapters with a concentric connector, 8 bipolar electrodes (24 mm) with a concentric connector, a ground cable, and a wrist strap with a male clip connector. The amplifier was connected to the power source continuously, but during recording the device automatically switched to battery use only. The EMG signals were recorded with a sampling frequency of 2048 Hz.

To record the audio signals, a Neumann TLM103 diaphragm microphone connected to a sound interface was used. This was done with a sampling frequency of 16 kHz.

To ensure that the EMG and audio signals are synchronized, an extra signal was shared between the bio-electrical amplifier and the sound interface. This synchronization signal was generated by the amplifier at the beginning and the end of each prompt, and was registered into an extra EMG channel. The signal was also outputted by the amplifier as an analog signal and introduced into one of the channels of the sound interface. As a result, the stereo audio signals contain the speech signal in the left channel and the synchronization signal in the right channel.

We used a silent computer to reduce interference with the audio and EMG signals as much as possible.

Additionally, a camera captured a video of the facial movements, to provide supplementary data and allow multi-modal experiments, such as automatic lip reading. However, in the experiments described in this thesis, the video data were not considered.

Figure 4.1 shows an image of the complete acquisition setup.

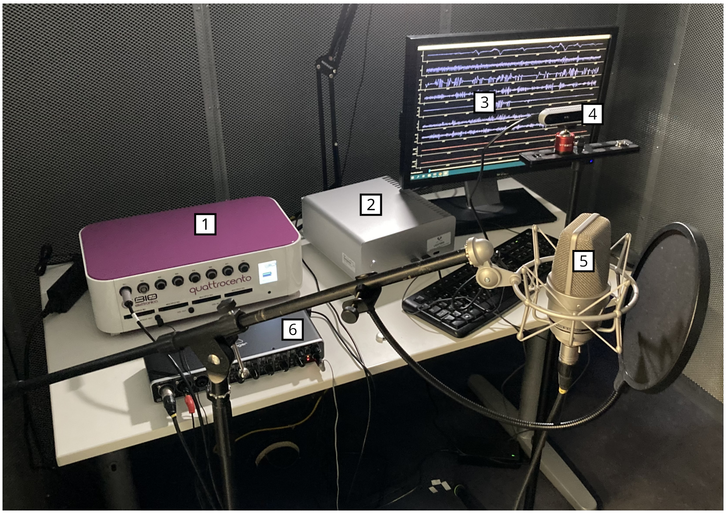


Figure 4.1: Acquisition setup: (1) bio-electrical amplifier; (2) silent computer; (3) computer screen; (4) camera; (5) microphone; and (6) audio interface.

Software

For the acquisition and synchronization of the audio and EMG signals, we used publicly available software[4], which also includes a user interface. The official software from the OT Bioelettronica EMG device company, OTBioLab+[5], was used to check the quality of the signals before each recording session.

Environment

Each session was recorded in a soundproof room. To reduce inter-session variability in audio and video as much as possible, the positions of the speaker, microphone, and video camera were kept constant for all sessions. We made sure there was a researcher present in the room at all times to check that the presented text prompts were pronounced

[4] https://github.com/cognitive-systems-lab/EMG-GUI, last accessed on 25/09/2024

[5] https://otbioelettronica.it/en/software/, last accessed on 25/09/2024

correctly. Furthermore, we constantly checked the quality of the EMG signals and replaced electrodes when they detached.

4.2.3 Recording protocol

The recording protocol was approved by the ethics committee of the University of the Basque Country.

The complete recording protocol of one session consists of the steps described below. The first session was different than the sessions following it because it was necessary to make a 3D scan to create the 3D mask. From the second session onwards, the 3D mask was used. So the first session followed steps 0, 1, 2, 4, 5, 6, and 7, and the steps for the second session and onwards were 3, 4, 5, 6, and 7.

Step 0: Signing consent form and giving instructions

Before the speaker came in for the first time, we sent them a document with instructions, so that they knew what to expect. The instructions were as follows:

- Try to articulate well
- Redo an utterance if you are not satisfied: not well articulated, not correctly pronounced, or moved during recording
- Take a break when necessary
- Don´t touch the cables if not necessary
- Don´t remove any cables
- Let us know when the reference band feels dry
- Let us know when you feel an electrode detaching
- Come shaved and without make-up or face cream

For ethical purposes, the speaker was required to sign a consent form on the first day.

Step 1: Marking the electrode locations

First, the locations of the electrodes were identified using facial landmarks and a measuring tape, using a procedure that was repeated for each speaker. For each muscle, these were the markers:

1. *Levator Labii Superioris*: From the top of the lip up, in the direction of the middle of the eye.

2. *Masseter*: The top is at the center of the left half of the distance between the middle of the ear and the nostril, in the direction of the neck.

3. *Risorius*: From the corner of the mouth, in the direction of the bottom of the ear.

4. *Depressor Labii Inferioris*: From the center of the distance between the corner of the mouth and the bottom of the lower lip, in the opposite direction of the nostril.

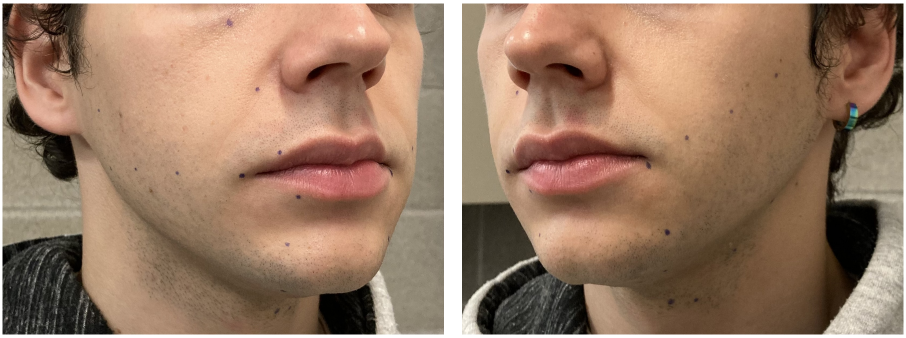
5. *Zygomaticus Major*: In the direction of the center of the distance between the corner of the eye and the bottom of the ear. The center of the distance between this point and the corner of the mouth is the actual center.

6. *Depressor Anguli Oris*: From the corner of the mouth in the opposite direction of the nostril.

7. *Anterior Belly of the Digastric*: The center is the center of the distance between the bottom of the chin and the thyroid cartilage or stoma, in the direction of the nostril.

8. *Stylohyoid*: The center is the center of the distance between the corner of the cheekbone and the thyroid cartilage or stoma.

The electrodes that were used are bipolar single electrodes, which means that two electrodes form one channel. So, the electrode setup consists of eight electrode pairs. The electrode locations were marked with three dots: in the middle between the two electrodes, and on the outer ends. This process was repeated for all eight electrode pairs, resulting in 24 reference points total (see Figure 4.2).



(a) Right side

(b) Left side

Figure 4.2: Reference points to be used to locate the electrodes.

The process of selecting the electrode type and locations is described in Chapter 3.

Step 2: Making a 3D scan

A personalized 3D mask (Figure 4.3) was used to ensure that the electrode locations remained constant throughout all sessions. After the reference points were marked in Step 1, a 3D-printing professional made a 3D scan of the face and printed a mask with holes corresponding to the reference points.

Step 3: Mark the electrode locations

The speaker was asked to hold the 3D mask tight and steady to their face. The researcher helped with adjustment if necessary, to ensure that the mask aligned with the face completely. Then, with a skin-friendly marker, the reference points were marked on the skin through the holes in the mask.

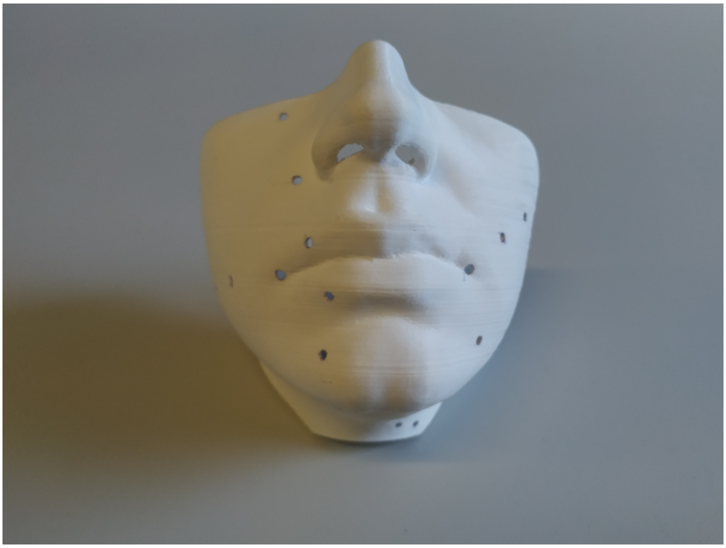


Figure 4.3: A personalized 3D mask. The holes were used as reference points to find the positions of the electrodes on the participant's face.

Step 4: Placing the electrodes

Using the reference points drawn in Step 3 (or Step 1 in case of the first recording session), the electrodes were placed on the face. We asked the speakers beforehand to shave, remove any make-up or cream, and clean the face. Also, before every electrode pair was placed, we cleaned the face one more time with alcohol. Then we applied conductive cream, which improved contact between the electrode and the skin. Figure 4.4 shows where the electrode pairs on each muscle were located for one of the participants.

Step 5: Connecting the cables

Every electrode pair was connected to a 16-channel bipolar adapter with a single-channel bipolar adapter. Strong armbands were used to hold the cables in place and reduce the pulling effect caused by their weights.

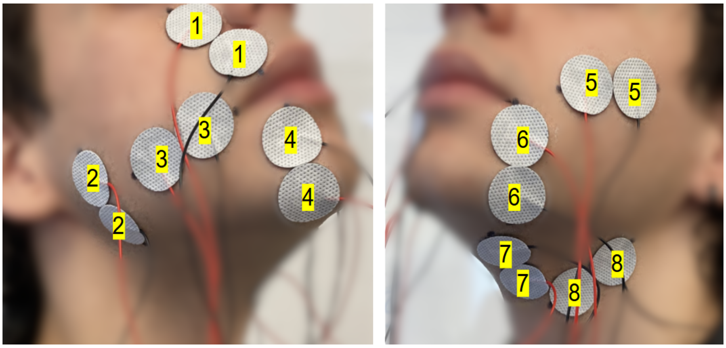


Figure 4.4: Electrode setup in the ReSSInt-EMG database. 1: Levator labii superioris, 2: Masseter, 3: Risorius, 4: Depressor labii inferioris, 5: Zygomaticus major, 6: Depressor anguli oris, 7: Anterior belly of the digastric, 8: Stylohyoid.

Then a wet reference strap was put on the left wrist, which is directly connected to the device with a ground cable. Once all the cables were connected, the speaker was asked if they could move their head freely and no cable was pulling. The speaker was instructed not to move their left arm. With their right arm, they had access to a mouse, which they needed to click to mark the start and end of each utterance. Then the OTBioLab+ software was used to check that all channels provided good signals.

Step 6: Recording

We made sure that all the hardware and software worked before the speaker came in for their session so that once the electrodes were placed and the cables were connected, the recording session could start immediately.

Step 7: Disconnecting the cables and removing the electrodes

When the recording session finished, the cables were disconnected and the electrodes were removed carefully. We provided the speaker with make-up remover to remove the marks on their face. We asked the speaker about their experience and wrote down everything that could affect the data. Examples are a decrease in motivation, tiredness, detached and replaced electrodes, and environmental influences such as extreme heat.

4.2.4 Quality control

As it is difficult to understand the characteristics of an EMG signal with the human eye, as opposed to speech signals, we had to rely on other methods to evaluate the signals' quality. The first step was to validate the acquisition setup by comparing the signals of the first few sessions with those of a reference database. This study is described in Chapter 5 and assured us that we could move forward with the selected setup.

From the beginning, we identified the problem of electrode detachment. First, our approach was to re-attach them every time they detached. It happened mostly to the electrodes around the mouth. Later, we used medical tape to secure them and replaced the entire electrode pair instead of re-attaching them.

After the first more extensive study with our data (see Chapter 6) we realized that for a few sessions, the EMG signal quality was not satisfactory. We decided to pause the recording process, find the source of the problem, and perform a more detailed quality check on the already recorded data. With help from an expert from the EMG equipment supplier, we found out that we were supplied with a bad batch of electrodes, which is why some sessions had low-quality signals and others did not. After receiving new electrodes, we continued recording and watched the EMG signals more closely while recording. The latter was especially useful in identifying a detachment problem before the electrode fell off since it showed either high noise or no signal at all.

Additionally, a quality check was performed by a colleague after each recording and consisted of evaluating the signal's minimum, maximum, mean, median, and root mean square (RMS). Each measure was calculated per utterance and channel and then averaged. Sessions that represented outliers in one of these measures did not pass the check and were recorded again.

4.3 Results

This section shows the output of the data collection process, by providing all the detailed meta-information, such as speaker information and signal duration, and examples of recorded data.

4.3.1 Meta-information

In total, nine Spanish-speaking people participated as speakers in the data collection process of the database, of which six typical speakers and three alaryngeal speakers. Initially, a fourth alaryngeal speaker came in to record as well, but their data was of too low quality to be able to include in the database. The number of sessions recorded by each speaker varies from 1 to 15 sessions. Table 4.3 shows the relevant information per speaker.

Table 4.3: ID, type, sex, age, and number of recorded sessions of each speaker in the database.

Speaker ID	Speaker type	Sex	Age	Number of recorded sessions
001	typical	male	29	15
002	typical	female	29	8
003	typical	male	51	4
004	typical	female	46	4
005	typical	male	45	8
006	typical	female	61	4
007	alaryngeal	female	61	2
008	alaryngeal	male	77	1
009	alaryngeal	male	64	2

The database contains a total of 22.5 hours of recordings, that are distributed among the different speech modes and corpus types as specified in Table 4.4. See Appendix C for a detailed overview per speaker.

Table 4.4: Duration of the database per speech mode (audible or silent) and corpus type (VCV, words or sentences) in the format hh:mm:ss.

	Audible	**Silent**	**Total**
VCV	1:22:01	0:39:04	2:01:05
Words	1:11:16	1:02:49	2:14:05
Sentences	13:33:22	4:44:21	18:17:43
Total	16:06:39	6:26:14	22:32:53

4.3.2 Data examples

The language in which the speech data was acquired is Castilian Spanish. Two speech modes were used: audible and silent. Audible speech refers to typical speech produced with sound, and silent speech refers to articulated-only speech where no sound is produced, also called mouthed speech. Simultaneously, EMG data was recorded, namely eight signals from eight locations in the face and neck. An example of an audio signal and an example signal of each EMG channel are shown in Figure 4.5, together with the synchronization signals. Note the difference in amplitude height for the EMG signals, since not every muscle has the same amount of muscle activation. These signals correspond to an utterance produced audibly. When an utterance is produced silently, the audio signal is still included in the database, but it is of course empty.

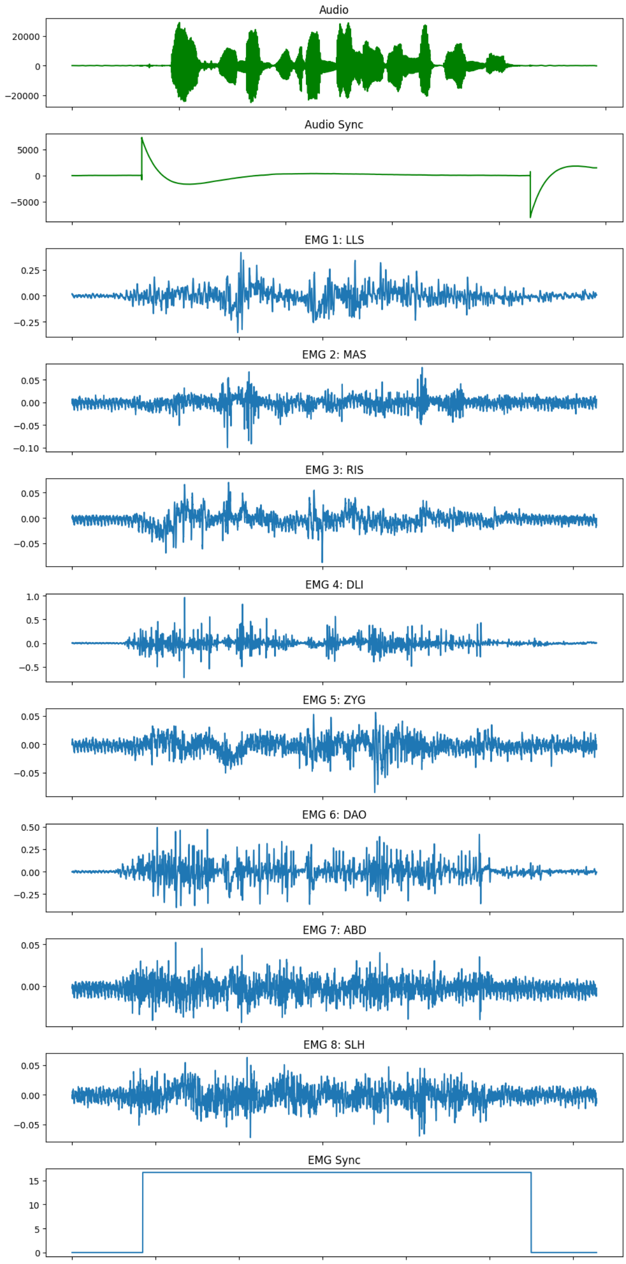


Figure 4.5: Audio, EMG, and synchronization signals corresponding to "Hay gemas de gran valor en la tienda." in audible mode from speaker 001.

4.4 Discussion and Conclusion

In this chapter, we have described the newly developed ReSSInt-EMG database and the design and execution of the data collection process. The final database contains 22.5 hours of EMG and speech signals recorded in several different contexts.

We removed or repeated sessions when we found that the quality did not adhere to our standards. This was often due to the detachment of electrodes during recording, but sometimes we could not find a direct reason.

While meticulous efforts were invested to ensure the overall quality and reliability of all EMG signals within the database, it is important to acknowledge the challenging nature of EMG signal acquisition. Despite our effort to prevent the inclusion of signals associated with detached electrodes, the database might contain signals that deviate from the intended standard.

For future research, (parts of) the content of this database can be used for several applications. In the following chapters of this thesis, it will become evident how we used the data for tasks such as phone classification and statistical analysis to study the effect of linguistic content and variations between speakers and recording conditions on a model's ability to predict speech from muscle activity. Colleagues have also been involved in tasks more directly related to SSIs, such as EMG-to-text, EMG-to-speech, and introducing lip movements from the video images to these tasks. The presence of different text corpora, speakers, and sessions, makes it possible for other researchers in the field to select the relevant data for their tasks.

Although we followed the instructions from the manufacturer of the EMG hardware, there are two points we did not consider, which could possibly improve the process of EMG acquisition in the future. First, we did not pay attention to the polarity of the EMG channel since we were told by the EMG manufacturer that this did not matter. If it does matter, the matter of polarity could be included in the model, but we suggest to already keep the consistency during recording. Second, even though we used the same method to locate the muscle for each speaker and were very careful doing so, new information was provided to us that in the medical field, apparently muscles are located using the technique of palpation. We highly recommend investigating this technique before starting EMG acquisition, since it could improve the accuracy of locating the right muscle and reduce the risk of cross-talk.

4.5 Contribution

The ReSSInt-EMG database is a valuable contribution to the field of EMG-based silent speech recognition and other research related to the complex interaction between speech and muscle activity. First of all, it includes more than 22 hours of data, and it is also the first to focus on the Spanish language. Furthermore, it contains data from both typical and alaryngeal speakers, and it allows for the development of multi-modal applications due to the inclusion of video images. Despite challenges like electrode detachment, efforts were made to ensure the database's quality. Looking ahead, the database promises a variety of applications in the field. Recommendations for future EMG data acquisition include attention to EMG channel polarity and the use of palpation techniques for muscle localization.

The database will be publicly available on the website of the European Land Registry Association (ELRA)[6]. An official report can be found on the website of the ReSSInt project[7].

[6] https://www.elra.eu/, last accessed on 25/09/2024

[7] https://aholab.ehu.eus/ressint/resultados/, last accessed on 25/09/2024

5
Validation of the Acquisition Setup

This chapter aims to validate the acquisition setup of the new ReSSInt-EMG database, by comparing the phone classification performance with that of a reference database. The reference database is the EMG-UKA Trial Corpus, which is a comparable SSI database of parallel EMG and audio signals, but for English. The results show an average classification accuracy of 40.85% for a small amount of ReSSInt-EMG data, compared to an accuracy of 28.32% for the same amount of reference data. Although a direct comparison cannot be made since the databases were recorded in different circumstances, the large difference in accuracy suggests that the data acquisition procedure of the new database is valid.

5.1 Introduction

As described in Chapter 4, we have created the first known SSI database for the Spanish language. We selected the electrode setup for acquiring the EMG signals of this database after performing a pilot study described in Chapter 3. The complete acquisition setup, including the bio-electrical amplifier, the computer and screen, the microphone, and the audio interface, was determined based on the relevant instruction manuals. The goal of the research described in this chapter was to validate the setup before continuing to record the entire database.

The structure of this chapter is as follows. First, we describe the methodology in Section 5.2. In short, we selected a public database with similar contents to use as a reference database. Although the signals were acquired in English, they are parallel audio and EMG signals from multiple speakers and sessions, acquired with a single-electrode setup. Then, we performed the same data processing procedure and phone classification experiments on both databases. The results of these experiments are listed in Section 5.3, which are discussed in Section 5.4. To summarize, although a direct comparison between the two databases could not be made, we found that the classifier was able to predict the correct phone label with much higher accuracy using the signals from our database than those of the reference database. This provided us with enough confidence to continue recording our database with the selected acquisition setup.

5.2 Methods

This methodology section includes three subsections, namely an overview of the selected data (Section 5.2.1), how we processed these data (Section 5.2.2), and a description of the phone classification experiments (Section 5.2.3).

5.2.1 Data overview

In this section, we describe the reference database and the part of the ReSSInt-EMG database we used for the experiments in this chapter.

The EMG-UKA Trial Corpus

The EMG-UKA Trial Corpus is a subset of the complete EMG-UKA Corpus[1] [89] and includes four speakers and 13 recorded sessions (see Table 5.1 for details). It contains three types of speech utterances: audible, whispered, and silent. For the experiments described in this chapter, we only used the EMG signals corresponding to the audible utterances. Each session includes 50 audible sentences, except for session 101 from speaker 002, which contains 520 sentences. The sessions with 50 sentences are divided into a training set of 40 sentences and a test set of 10 sentences. Session 101 from speaker 002 is divided into 500 sentences for training and 20 sentences for testing.

The muscles targeted in the EMG-UKA corpus, using six channels, are: the LAO (channels 2 and 3), ZYG (channels 2 and 3), PLT (channel 4), ABD (channel 1), DAO (channel 5), and the tongue (channels 1 and 6). It must be noted that channel 5 was not used in most of the phone recognition experiments performed with this database [51, 90–92]. However, it has been used in [93] for word recognition, speaker identification [70], and speech-to-EMG conversion [69, 94]. We decided to use all the channels, including channel 5.

We found that the phonetic transcriptions included in the EMG-UKA Trial Corpus were partially incorrect, and therefore we created new phonetic transcriptions, using the Librispeech lexicon [2] and additional transcriptions for words that were initially not included in the lexicon. Additionally, we did a new phonetic alignment using the MFA [83]. In total, a set of 40 different phone labels was identified. Initial and final silences were removed, but short pauses inside the sentences were included in the phone classification experiments.

[1] https://catalog.elra.info/en-us/repository/browse/ELRA-S0390/, last accessed on 25/09/2024.

[2] https://www.openslr.org/11/, last accessed on 25/09/2024.

Table 5.1: Session information of EMG-UKA and ReSSInt-EMG databases.

Database	Speaker	Session	Sex	Duration
EMG-UKA	002	001	M	3:29
		003		3:24
		101		26:04
	004	001	F	3:23
	006	001	M	3:45
	008	001	M	3:19
		002		3:06
		003		3:02
		004		2:50
		005		2:41
		006		2:40
		007		2:38
		008		2:42
ReSSInt-EMG	001	001	M	16:51
		002		17:31
		003		17:00
		004		19:24
	002	001	F	25:25
		002		26:50
		003		25:55
		004		27:06
	003	001	M	24:38
	004	001	F	26:05

The ReSSInt-EMG database

The selected data from the ReSSInt-EMG database consists of the first few sessions recorded by four speakers, with a total of ten sessions. The lower part of Table 5.1 shows the details of these sessions. The duration corresponds to the sum of the audio signals per session, after synchronization with the EMG signals.

During each recording session, three different kinds of items were recorded, namely non-sense words with VCV structures, isolated words, and sentences. The sentences correspond to either one of the two existing corpora Sharvard or Ahosyn. The number of Ahosyn sentences in each session is smaller than the number of Sharvard sentences because they are generally longer. See Section 4.2.1 for more details of the text corpora.

Table 5.2: Corpus information of ReSSInt-EMG sessions. All the signals in these sessions were acquired in audible speech mode.

Session	Corpus
all	110 VCV combinations 100 isolated words Sharvard sentences 1-100
001	Sharvard sentences 101-400
002	Sharvard sentences 401-700
003	Ahosyn sentences 1-150
004	Ahosyn sentences 151-300

Table 5.2 shows that the first 100 sentences of the Sharvard corpus were recorded in each session. We assigned the final 20 of these sentences to the test set to have a similar test set size as in the long session of the EMG-UKA database (002-101). The training set consists of the remaining sentences recorded in that session, either 230 or 380, depending on the session number.

One session from one speaker (002-002) was left out of the experiments after we found that the synchronization signal was not recorded correctly during part of the session. Without a good synchronization signal it is impossible to align the EMG and audio signals. This session was later repeated and the final database now includes this good-quality session.

For the ReSSInt-EMG database, a partially different electrode setup was used, compared to the EMG-UKA Trial Corpus. First of all, there are eight channels, each targeting a different muscle. The following three targeted muscles are similar: the ABD, ZYG, and DAO. Two others are close to, or underneath, the muscle targeted in the EMG-UKA setup: the LLS and DLI. The remaining three are additional: the RIS, MAS, and SLH.

Each speech utterance was aligned with the corresponding phonetic transcription using the MFA. The phonetic dictionary was created using the Aholab transcriber, which uses the SAMPA phone set consisting of 29 different phones (see Table 4.2 for an overview). As with EMG-UKA, initial and final silences have been removed but short pauses inside the sentences have been taken into account in the classification experiments.

For the EMG-UKA corpus, both monopolar and bipolar channels were used, whereas the ReSSInt-EMG database only contains signals from bipolar channels. Furthermore, the type of electrodes is different, in terms of form, size, and manufacturer.

5.2.2 Data processing

The extraction of the TD features from the EMG signals has been done as in the pilot study described in Chapter 3. The five frame-based features are the following: the time-domain mean of the nine-point double-averaged signal, the power of the nine-point double-averaged signal, the time-domain mean of the rectified high-frequency signal, the power of the rectified high-frequency signal, and the zero-crossing rate. A window of 25 ms duration and 5 ms frame shift was used to extract the EMG features. A total of 5 x N features were calculated for each frame, where N is the number of channels (N=6 for EMG-UKA and N=8 for ReSSInt-EMG). To add context information to the frames, the features of surrounding frames and the features of the current frame were stacked together with a stacking filter. The width of the stacking filter indicated the number of adjacent frames before and after the actual frame. Since 5 TD features were calculated for each of the N EMG channels, the length of the parameter vector assigned to each frame can be calculated as $N \cdot 5 \cdot (2k+1)$, where $k = 15$ was the width of the stacking filter. A more detailed description of these features can be found in Section 3.2.2.

To reduce the dimension of the parameter vector, we applied LDA [85], as in [54] and [84]. For EMG-UKA, a reduction from 930 to 32 features was performed, while for ReSSInt-EMG the features were reduced from 1240 to 28. The difference in the number of phone classes for each database partially explains the difference in the number of features, since the maximum allowed number of features in LDA reduction is the number of classes minus 1. We adopted the number 32 from previous work with the EMG-UKA database [54].

Furthermore, Mel-frequency cepstral coefficients (MFCCs) were extracted from the audio signals using a Hamming window, calculating 13 coefficients for each frame. To obtain the coefficients, we used a 30-filter filter bank.

5.2.3 Experiments

The experiments consisted of the classification of phone labels from EMG-TD features or audio-MFCCs. Phone classification experiments have been previously performed with the EMG-UKA database [54, 84], so we decided to use these experiments for this study as well.

We used a bagging classifier [95] with 100 DTs as estimators and a required minimum of 50 samples in the leaf node. We compared other classifiers, such as an NN, a GMM, and bagging classifiers with different estimators, however, the bagging classifier with DTs yielded the highest validation accuracy. The classification was performed in speaker- and session-dependent modality, which means that the training and test were derived from the same speaker and session. Cross-validation was performed using the K-fold method, dividing the utterances in the training subset into five groups ($K = 5$). Five classifiers were trained, using four different folds each time and testing them with the unseen fold. Then the obtained results were averaged. Finally, a new classifier was trained using all the training data, which was tested with the test subset to obtain the test accuracy.

5.3 Results

Tables 5.3 and 5.4 show the obtained results of the phone classification experiments performed on the EMG-UKA Trial Corpus and the ReSSInt-EMG database, respectively. Since the classification was performed speaker- and session-dependently, the results are shown for each session separately. The tables show the following three types of frame-based accuracy values:

- The test accuracy based on the MFCCs (acoustic signals). These values are provided as a practical reference for the general classifier performance.
- The validation accuracy based on the EMG signals. Since the K-fold validation method was used, the mean and SD of the accuracy results obtained with the 5 folds are shown.
- The test accuracy based on the EMG signals.

Table 5.3: Phone accuracy obtained with EMG-UKA database.

Speaker	**Session**	**MFCC acc. (%)**	**EMG Valid. acc. (%)**	**EMG Test acc. (%)**
002	101	52.87	26.82±0.35	28.32
002	001	44.04	23.26±0.97	19.99
	003	45.07	28.11±1.20	25.69
004	001	40.81	20.63±0.75	16.14
006	001	45.44	22.11±2.03	22.26
008	001	44.13	29.75±0.50	25.94
	002	43.47	29.50±0.88	24.08
	003	41.25	27.08±0.92	22.78
	004	43.97	28.55±1.29	23.63
	005	44.77	29.37±1.29	25.81
	006	42.99	29.43±1.67	25.79
	007	43.05	28.33±1.29	23.17
	008	40.05	27.21±0.48	24.77
mean±sd		**43.25±1.65**	**26.94±3.02**	**23.33±2.76**

The results for the long EMG-UKA session 101 from speaker 002 have been underlined in Table 5.3. Due to the different duration of this session compared to the other sessions, the averages shown in the last line of this table do not include that long session.

Table 5.4: Phone accuracy obtained with ReSSInt-EMG database.

Speaker	Session	MFCC acc. (%)	EMG Valid. acc. (%)	EMG Test acc. (%)
001	001	69.98	45.92±1.24	44.01
	002	71.93	44.17±0.57	45.25
	003	71.50	42.07±1.18	43.98
	004	67.64	40.06±1.10	36.42
002	001	70.57	40.52±0.58	40.91
	003	68.17	35.86±1.04	37.65
	004	73.02	36.44±0.98	40.76
003	001	71.01	42.84±0.39	41.80
004	001	69.55	38.78±0.68	36.84
mean±sd		**70.37±1.64**	**40.74±3.18**	**40.85±3.09**

5.4 Discussion and Conclusion

To create the ReSSInt-EMG database, we had to determine the acquisition setup, namely the type, number, and locations of the electrodes, the acquisition equipment, and the contents and duration of the recording sessions (see Chapters 3 and 4). The classification experiments in this chapter aimed to validate the acquisition procedure, so we performed the experiments under similar experimental conditions using both the new database and a reference database. More specifically, we aligned the characteristics of the ReSSInt-EMG sessions with those of the longest EMG-UKA session in terms of duration and employed the same classifier with identical features and parameters.

Although the experimental design was different in terms of phonetic labeling and the type of classifier used, the results we obtained with the EMG-UKA Trial Corpus were in the range of those presented in the literature [54, 84]. Interestingly, the validation accuracy obtained for the short EMG-UKA sessions was on average similar to that of the long session, despite the differences in the amounts of training material. However, as expected, the highest test accuracy was obtained with the largest session (101) from speaker 002.

The results of the ReSSInt-EMG sessions show some variability among speakers as well as among different sessions from the same speaker, despite the effort of using a personalized 3D mask to avoid session variability within speakers. However, this variability was also observed in the results for the acoustic data, so it cannot be directly attributed to a variation in the electrode locations.

A side-by-side comparison between the results of the experiments performed on the two databases is not feasible, due to the differences in the experimental setup. The languages differ, resulting in a difference in the number and type of phone classes (40 classes for EMG-UKA vs. 30 for ReSSInt-EMG). Furthermore, the recording conditions and materials used were not the same. This means that the following comparisons aim to contextualize the performance of the ReSSInt-EMG database in relation to the EMG-UKA database.

First of all, the difference of almost 30 percentage points in the acoustic classification accuracy (43.25% for EMG-UKA and 70.37% for ReSSInt-EMG) assured us that the acoustic signals were of sufficient quality. We believe that the 70.37% accuracy could potentially improve by adding more training data and possibly adjusting the classification method and parameter settings, but that was not the aim of this study.

Secondly, the results of the classification with EMG signals show that each of the sessions in the ReSSInt-EMG database scored higher (on average 40.85%) than the best-performing session of the EMG-UKA database, which was the largest session and achieved an accuracy of 28.32%. The idea was that if the new database obtained at least the same results as the reference database, then the new data would be considered valid. The results confirm this, leading us to decide to continue recording the rest of the sessions of the ReSSInt-EMG database with the established acquisition setup.

5.5 Contribution

The chapter presents the classification results of data from two EMG-audio databases. It offers the first results on the ReSSInt-EMG database and provides a baseline for future research.

Part of this chapter has been published as:

Del Blanco, Eder; **Salomons, Inge**; Navas, Eva; Hernáez, Inma. “Phone Classification Using Electromyographic Signals”. In: *IberSPEECH 2022*. ISCA, Nov. 2022, pp. 31–35. DOI: 10.21437/IberSPEECH.2022-7.

Part III

Research Challenges

6 The Impact of Speaker and Session Variability

This chapter evaluates the impact of variability between speakers and sessions on the development of an EMG-based SSI. The methodology consists of a series of phone classification experiments, where phonetic labels of unlabeled EMG signals were predicted with a classification model trained on phonetically labeled EMG signals. We evaluate and compare the performance of each model using classification accuracy. Results show that the models can predict the phonetic label best when they are trained and tested on data from the same session. The accuracy drops drastically when the model is tested on data from a different session, yet it improves when more data are added to the training data. Similarly, when the same model is tested on data from a different speaker, the accuracy decreases. This suggests that using larger amounts of data could help to reduce the impact of inter-session variability, but more research is required to understand if this approach would suffice to account for inter-speaker variability as well.

6.1 Introduction

To develop an SSI that can predict speech from facial muscle movements, a large database of (parallel) EMG and speech data is required. In Chapter 4 we already described how we created such a database for Spanish, called the ReSSInt-EMG database. Due to speakers getting tired and electrodes detaching over time, there is only a limited amount of data available per session. Generally, speakers were able to record for one hour, resulting in approximately 30 minutes of data per session.

This calls for the effort to mitigate variations between signals from different sessions, to be able to train an SSI model with as much data as possible. Several factors increase the speaker and session dependency on the model performance and decrease its generalization capability [16–18]. Some of these cannot be controlled, namely individual speaker characteristics such as variations in anatomy, articulatory patterns, and speaking style. Even the same speaker may exhibit variability in articulation across different sessions due to factors such as fatigue, mood, motivation, and health. However, other factors related to session variability can be controlled to some extent. For example, we tried to keep the EMG electrode placement across sessions as constant as possible by using a personalized 3D mask for each speaker (see Section 4.3). Furthermore, we followed a recording protocol (see Section 4.2.3) to use the same standard for each speaker and session. Environmental factors are more difficult to control, such as background noise or uncomfortable levels of temperature or humidity. For this reason, we occasionally postponed sessions or interrupted them to record later in better conditions.

Nevertheless, despite these efforts, and because of the influence of uncontrollable factors, speaker and session variability is inevitable. In this chapter, we study the impact of this variability, specifically when using data from our database. We performed a set of phone classification experiments using data from different speakers and sessions. The task of classifying phones is similar to those performed in Chapters 3 and 5, but in this chapter, we improved the classification and feature reduction methods.

The structure of this chapter is as follows. Section 6.2 describes the data from the ReSSInt-EMG database we used as well as the experimental design. The results of the experiments are described in Section 6.3, which are then interpreted and discussed in Section 6.4.

6.2 Methods

This section describes this study's methodology, namely the data used for the experiments, how these data were processed, and the experiments themselves.

6.2.1 Data overview

Table 6.1 shows the details of the sessions of the ReSSInt-EMG database that we used for this study, namely, 16 sessions in total from six different speakers. Note that this is only a part of the complete database, as recordings were still ongoing at the time the experiments in this study were performed. Furthermore, all of these sessions were acquired in audible mode. See Section 4.2.2 for more information about the data acquisition, and an image of the electrode setup in Figure 4.4.

For the experiments described in this chapter, we only used the signals corresponding to the Sharvard and Ahosyn sentences and not the VCV combinations or isolated words (see Table 5.2).

Each session was split into 80% training and 20% test data. During the recording process, the utterances were presented in a unique and random order. To ensure consistency, we assigned the final 20% of each set of sentences as test data before the experiment. This approach ensured that the time of recording within each session was unrelated to the train–test split, and the utterances designated as the test set remained constant for each speaker.

6.2.2 Data processing

A phonetic dictionary was created using the Aholab transcriber, which uses the Spanish SAMPA phone set, comprising 29 phones. Then, each audio signal was segmented into phone labels using the MFA.

Table 6.1: Speaker and session information for the part of the ReSSInt-EMG database used in this study. The duration is expressed in mm:ss format and is limited to the part of each session that includes audible sentences.

Speaker	Sex	Age	Session	Duration	Train	Test
001	M	29	001	16:51	13:28	03:23
			002	17:32	14:04	03:28
			003	17:00	13:48	03:12
			004	19:22	15:14	04:08
002	F	29	001	25:25	20:20	05:05
			002	30:34	24:27	06:07
			003	22:36	18:17	04:19
			004	27:06	21:18	05:48
003	M	51	001	24:38	19:50	04:48
			002	21:43	17:27	04:16
004	F	46	001	26:04	20:46	05:18
			002	24:09	19:17	04:52
005	M	45	001	23:39	18:56	04:43
			002	22:31	18:00	04:31
006	F	61	001	32:57	26:21	06:36
			002	29:01	23:21	05:40

Initial and final silences were removed, while short pauses between words were kept, resulting in an extra class. This results in a total of 30 classes.

From the EMG signals, we extracted five TD features, after removing the direct-current offsets and normalizing them. The TD features were calculated as proposed in [34], and are explained in more detail in Chapter 3.2.2.

We used a window with a duration of 25 ms and a frame shift of 5 ms to extract the EMG features. Since five TD features were calculated for each of the eight EMG channels, the length of the parameter vector assigned to each frame resulted in 1240 features for a width of the stacking filter of 15 and 8 channels.

To reduce the dimension of the parameter vector, we applied LDA. To select the optimum dimension, we analyzed the effect of the number of features on the frame-based phone classification accuracy. Figure 6.1 shows the average validation accuracy per number of LDA features for the first session of each speaker. Based on this graph, we selected 21 LDA features because the average accuracy reaches a plateau at that value. Selecting a higher number of features would result in a more complex model and a longer training time. The classifier used to search for the optimal LDA value was an NN with a batch size of 128 and 20 epochs.

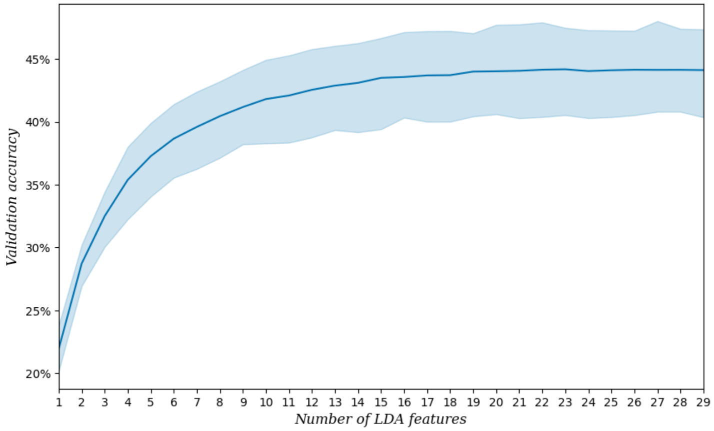


Figure 6.1: Validation accuracy per number of LDA features averaged over session 001 of all speakers. The classification method is an NN with a batch size of 128 and 20 epochs. The solid line represents the average accuracy, and the area above and below the line shows the standard deviation range.

6.2.3 Experiments

This section describes the experimental part of the study, namely the classifier used and its configuration and how we considered speaker and session dependency for the experiments.

Classification Method

The classifier used for the experimental part of this study is a feed-forward NN with one hidden layer using a batch size of 256 and 100 training epochs. We chose these parameters based on a hyper-parameter search by tracking the validation accuracy during 250 training epochs for three batch sizes: 64, 128, and 256. We repeated this for session 001 of all six speakers and averaged the results (see Figure 6.2). We selected 100 training epochs because at that point the performance reaches a plateau. Additionally, we selected the largest batch size (256) because there is no difference between the three batch sizes, and a larger batch size means lower training time. The network has an input dimension equal to the number of features (21 nodes) and a dense layer with twice as many nodes as features (42 nodes in total) using a rectified linear units activation function [96]. The output layer has the same number of nodes as the number of classes (30, which includes the Spanish 29 phones and a silence class) and uses a *softmax* activation function [97]. Furthermore, a categorical cross-entropy loss function and the *Adam* optimizer [98] with a learning rate of 0.001 were applied.

We used 5-fold cross-validation to obtain the validation accuracy.

Speaker and Session Dependency

This study involves three separate rounds of experiments, each varying in terms of speaker and session dependency. In the first round, the data were both speaker- and session-dependent, which means that the training and test data were taken from the same session. In the second round, the data were speaker-dependent but session-independent. This means that the training data came from a different session or different sessions than the test data, but that all sessions were recorded by the

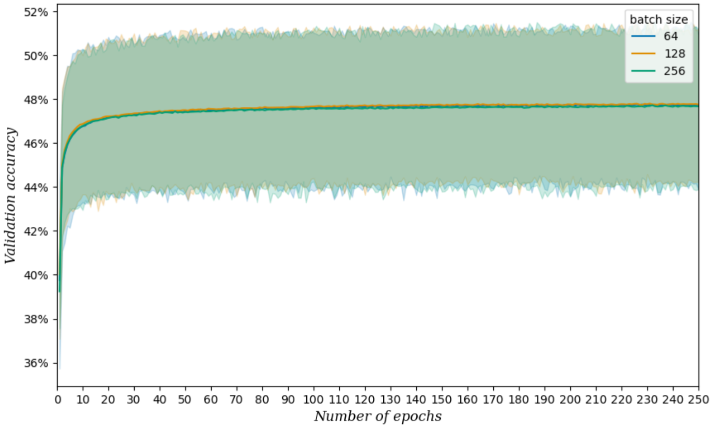


Figure 6.2: Validation accuracy per number of epochs and three batch sizes averaged over session 001 of all speakers. The solid line represents the average accuracy, and the area above and below the line shows the standard deviation range.

same speaker. This method allows for the evaluation of the effect of increasing the amount of data from the same speaker on the performance of the model as well as the impact of inter-session variability on the accuracy. In the third round, we used speaker-independent data by training the model using data from multiple sessions of one speaker and testing it using data from another speaker.

The test session contains a session-specific corpus that was not included in the sessions used to train the model, making the experiment speaker-independent but also text-independent. The goal was to assess the potential of creating a model that can be applied to new speakers without the need for adaptation by training it only on data from the actual database.

6.3 Results

In this section, we show the results of the experiments, first from those in the session-dependent mode and then from the ones we performed in the session-independent mode, which are both speaker-dependent. Lastly, we also show the results from the speaker-independent experiments.

6.3.1 Speaker-dependent, session-dependent

Table 6.2 shows the results of the session-dependent experiments, for which the model was trained and tested with data from the same speaker and session. Some speakers show higher classification accuracy (speakers 001 and 005) than other speakers, but speaker 006 has the lowest results, in particular for session 002. After reviewing the data from sessions with relatively lower results, we realized that some channels presented low-quality signals, related to the performance and detachment of the electrodes. Specifically, in sessions 003-002, 004-002, 005-001, and 006-002, we observed this quality issue. For all these sessions, the results are between 5.82 and 10.55% lower than the other session of the same speaker, except for 004-002. However, it does affect the session-independent experiments, as will become clear in Section 6.3.2.

The identification of this issue resulted in more strict quality control and the re-recording of these sessions to include in the final database, as described in Section 4.2.4.

Table 6.2: Speaker-dependent, session-dependent classification results.

Speaker	**Session**	**Validation accuracy**	**Test accuracy**
001	001	50.48±1.01	46.42
	002	49.12±0.86	47.15
	003	45.80±0.66	45.53
	004	50.41±1.05	50.54
002	001	43.71±0.48	42.61
	002	42.80±0.96	42.52
	003	38.76±1.35	38.05

Continues on next page

Table 6.2 – continued from previous page

Speaker	Session	Validation accuracy	Test accuracy
	004	39.39±0.77	39.64
003	001	46.73±1.12	45.27
	002	42.41±1.07	39.45
004	001	43.22±1.50	38.44
	002	41.29±1.37	39.62
005	001	43.61±1.56	41.19
	002	51.45±0.54	50.40
006	001	35.92±1.17	35.27
	002	28.39±1.31	24.72
mean±sd		**43.34±5.80**	**41.68±6.14**

6.3.2 Speaker-dependent, session-independent

To evaluate session-independent classification, we first used the models from the previous section (session-dependent experiments) and tested them using the test set from another session. Then, we trained new models using a variable number of training sessions from the same speaker. The results in Table 6.3 show that the test accuracy decreases in a session-independent configuration. This decrease in test accuracy is not the same for every speaker. However, when additional sessions are included in the set of training data, the test accuracy increases. Nevertheless, it is always lower than the test accuracy obtained with session-dependent classification.

On the other hand, contrary to the test accuracy, the validation accuracy decreases as more training sessions are added. This is an indication of less over-fitting, as it shows better generalization capability.

The effect of some low-quality signals in a few sessions (mentioned in Section 6.3.1) is challenging to assess in these experiments because both training and test sessions include some of these defective signals, in the case of speakers 003 to 006. For instance, both experiments for speaker 004 showed poor results because session 004-002 was used either for training or testing.

Table 6.3: Speaker-dependent, session-independent classification results.

Speaker	Train session(s)	Test session	Validation accuracy	Test accuracy
001	002	001	49.12±0.86	23.40
	002,003		45.08±0.89	27.89
	002,003,004		42.50±0.74	30.41
	001	002	50.48±1.01	19.57
	001,003		46.85±1.07	22.11
	001,003,004		43.81±0.70	24.54
	001	003	50.48±1.01	14.19
	001,002		48.16±1.14	18.09
	001,002,004		45.00±0.34	18.25
	001	004	50.48±1.01	15.86
	001,002		48.16±1.14	22.38
	001,002,003		44.49±0.37	24.93
002	002	001	42.80±0.96	10.00
	002,003		39.69±0.53	18.32
	002,003,004		37.90±0.61	21.93
	001	002	43.71±0.48	20.90
	001,003		41.23±1.09	23.81
	001,003,004		37.79±0.80	24.19
	001	003	43.71±0.48	17.79
	001,002		42.46±1.02	18.03
	001,002,004		39.63±0.53	16.73

Continues on next page

Table 6.3 – *continued from previous page*

Speaker	**Train session(s)**	**Test session**	**Validation accuracy**	**Test accuracy**
	001	004	43.71±0.48	19.01
	001,002		42.46±1.02	20.84
	001,002,003		39.42±0.51	22.92
003	002	001	42.41±1.07	20.66
	001	002	46.73±1.12	15.05
004	002	001	41.29±1.37	10.95
	001	002	43.22±1.50	8.63
005	002	001	51.45±0.54	11.83
	001	002	43.61±1.56	23.61
006	002	001	28.39±1.31	16.02
	001	002	35.92±1.17	8.30

6.3.3 Speaker-independent

To evaluate speaker-independent classification, we employed the models trained with three sessions from the session-independent experiments and tested them with the remaining session from each of the other speakers. The results, presented in Table 6.4, show that the classification accuracy varies greatly, despite all models being trained on similar amounts of data. This table only shows the test accuracy, since the validation accuracy of these models is already shown in Table 6.3.

When one of the low-quality sessions is used as a test session, the test accuracy is also low, so that effect is clear.

Table 6.4: Speaker-independent classification results.

Train speaker	Train sessions	Test session	Test speaker	Test accuracy
001	002,003,004	001	002	19.47
			003	14.47
			004	12.08
			005	9.33
			006	8.41
	001,003,004	002	002	18.28
			003	15.10
			004	6.91
			005	19.90
			006	8.51
	001,002,004	003	002	8.36
	001,002,003	004	002	10.47
002	002,003,004	001	001	14.07
			003	15.71
			004	14.78
			005	8.11
			006	10.53
	001,003,004	002	001	15.95
			003	18.09
			004	10.26
			005	16.90
			006	7.21
	001,002,004	003	001	16.79
	001,002,003	004	001	20.43

6.4 Discussion and Conclusion

This chapter presents the results of the frame-based phone classification experiments performed to assess the impact of speaker- and session-variability in the data of the ReSSInt-EMG database.

The session-dependent classification results show varying outcomes not only across speakers but also across multiple sessions from the same speaker. Furthermore, the session-independent results indicate a large decrease in test accuracy when the model was tested with data from sessions that were not included in the training data, with the effect of this difference depending on the speaker.

Unfortunately, this means that the impact of inter-session variability in the EMG signals from our database when predicting speech sounds from them is quite substantial. We have already established the risk factors of EMG acquisition at different moments and our approaches to address them in Section 6.1, however, we will now elaborate on these issues in greater depth.

First, despite the use of a 3D mask, small variations in electrode placement can occur between sessions. Unless a robot can be used to place the electrodes, differences of a few millimeters are inevitable. Second, the physical or mental state of the speaker may lead to slight differences in articulation between sessions, as each is recorded on a different day. For instance, a person may articulate with less effort when feeling exhausted, resulting in reduced muscle activation. Third, environmental conditions such as temperature and humidity can affect the speaker's state and the contact between the electrodes and the skin. High temperatures may cause increased sweating and decreased motivation. These factors can result in each session being recorded under unique circumstances, which impacts the recorded EMG signals. Consequently, a model that can identify patterns in the EMG signals of one session may struggle to recognize those same patterns in signals from a different session.

Interestingly, when more sessions were added to the training data, the test accuracy increased. Given a corresponding decrease in validation accuracy, we believe that the improvement is due to enhanced diversity and representation of the data, allowing the model to better generalize beyond the training data. These results suggest that developing an EMG-based SSI with sufficient performance for real-world applications requires a large and diverse database. While using a larger set of training data may potentially slow down the experiments and require additional resources, we firmly believe that it is crucial to use as much training data as possible, provided that sufficient processing capabilities are available and the addition of new data leads to improved model performance.

The speaker-independent classification results demonstrated a substantial decrease in test accuracy when trained with data from other speakers, even when the amount of training data was comparable to the speaker-dependent, session-independent models. This suggests that the differences between the EMG signals of different speakers are substantial, making it challenging for the model to generalize to a different speaker. These differences can be attributed to various factors, such as differences in speakers' physiognomy, articulation manner, or speaking pace. Further experiments are needed to investigate whether training the model with a more extensive database from a single speaker or with data from multiple speakers can enhance speaker-independent performance.

6.5 Contribution

This chapter provides an analysis of the impact of signal variation between speakers and sessions on the phone classification performance with data from the ReSSInt-EMG database.

Part of this chapter has been published as:

Salomons, Inge; Del Blanco, Eder; Navas, Eva; Hernáez, Inma; De Zuazo, Xabier. "Frame-Based Phone Classification Using EMG Signals". In: *Applied Sciences* 13.13 (June 2023), p. 7746.
DOI: 10.3390/app13137746.

7
Phone Confusion Analysis

This chapter describes a phone confusion analysis of a set of phone classification experiments based on EMG signals. Understanding the relationship between speech and the muscles used for speaking is essential to learn the possibilities and limitations of an EMG-based SSI. When considering only information from the muscles of the face and neck, important information from the tongue and vocal cords is missing. This is reflected in the results, which show confusion between pairs of phones that only differ in the tongue's position, or the voicing feature.

7.1 Introduction

In this chapter, we continue with a series of phone classification experiments, but this time with the aim to analyze phone confusion after using EMG signals to classify phones and to gain insight into the relationship between muscle movement and speech. Previous phone confusion analysis for English showed that detecting voicing as well as the manner of articulation when using EMG signals is challenging [54]. This chapter focuses on the Spanish language, but we expect a similar trend since surface EMG electrodes are located in the face and neck, which makes capturing the inner movement of the tongue difficult, and the tongue is an important part of speech production in either language. However, Spanish uses a different phone set, so we believe that for the development of an EMG-based SSI for Spanish, a language-specific phone confusion analysis is necessary because it could provide new insights.

The outline of this chapter is as follows. In Section 7.2.1, we describe the dataset used in this chapter, and in Section 7.2.2 we explain the steps involved in the data processing procedure. Section 7.2.3 describes the experimental part of this study. In Section 7.3, the results of the experiments are presented. These results are analyzed and discussed in Section 7.4, together with a summary of the findings.

7.2 Methods

The methodology consists of a classification task aimed at predicting the correct phone label of frames of EMG signals. We used data from the ReSSInt-EMG database and trained a one-layer feed-forward NN using cross-validation with features extracted from those signals.

7.2.1 Data overview

The part of the database (Chapter 4) that we used in this chapter consists of 28 sessions recorded by six speakers (3 males and 3 females) aged 29 to 61, with a total of 11.5 hours of audible speech data. The number of recorded sessions differs per speaker (see Table 7.1 for an overview of the sessions and the total audio duration per speaker). In each session, a consistent base set of utterances was recorded, consisting of three distinct sets: 110 VCV combinations, 100 isolated words, and 100 sentences. Additionally, each session included another set of sentences, which was unique for each session, but remained consistent across speakers. During data processing, each set of words or sentences was split into an 80% training set and a 20% test set. This division process was applied uniformly to each session to ensure consistency. It is important to note that the test set was derived from the 100 sentences within the base set. As a result, the utterances in the test set remained the same across all sessions.

Table 7.1: Overview of the database: speakers, sessions, and total audio duration (hh:mm:ss) per speaker. All the signals in these sessions were acquired in audible speech mode.

Speaker	**Sessions**	**Total Duration**
001	001-005, 007, 008	2:17:28
002	001-005, 007, 008	3:11:54
003	001, 002	1:00:02
004	001, 002, 005	1:09:26
005	001-003, 005	1:32:39
006	001-005	2:28:28
all		11:39:57

As described in detail in Chapter 4, to each audio signal belongs one synchronization signal and eight EMG signals, each targeting one of eight muscles in the face and neck.

7.2.2 Data processing

The first step in the data processing was cropping the EMG and the audio signals using the synchronization signal. Then, the audio signal was segmented using the MFA, to obtain the labels that were used to perform the classification. The phonetic transcriptions were obtained with a transcriber created by the Aholab Signal Processing Laboratory using the SAMPA phone set. In this chapter, we refer to the phones using the International Phonetic Alphabet (IPA) for more clarity. The silences at the start and the end of every utterance were discarded, but the short pauses between words were kept (represented as ’sil’). The distribution of the phone labels is shown in Figure 7.1.

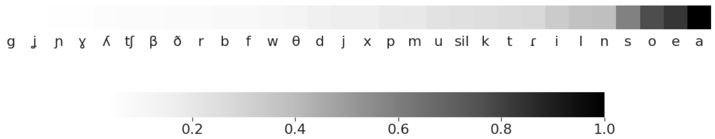


Figure 7.1: Frequency of labels in the database, normalized with respect to the most frequent label, [a]. The darker the shade, the higher the relative number of frames labeled with that label.

Instead of using the raw EMG and audio signals, we extracted features from them to use as input data. We calculated a set of TD features as described in Section 3.2.2, with a rectangular window of 25 ms duration and a 5 ms step size.

The addition of temporal context information is essential as the signals related to the movement of muscles are not necessarily simultaneous to the generated speech. This means that relevant information might not be in the central frame but in the surrounding frames. To incorporate temporal context information into each frame, a stacking filter was applied, which allows to combine the features of the current frame with those of adjacent frames. We selected a stacking filter width of 15 frames, which means that the actual frame is stacked with the 15 preceding frames and the 15 subsequent frames, so that information from a total of 31 frames (i.e. 135 ms) was used. This resulted in a high-dimensional feature vector for each frame, with a length of $n \cdot 5 \cdot (2k+1)$, where $n = 8$ is the number of EMG channels and $k = 15$ is the width of the stacking filter, resulting in a total of 1240 features per feature vector.

To perform phone classification using acoustic features for the purpose of identifying the top performance of the classifier, we computed MFCCs using a Hamming window and a filterbank of 30 filters, calculating 13 features for each window. The window length and the frame shift were identical to those used for TD feature extraction. As with the EMG features, we applied a stacking filter with a width of 15 to each frame, resulting in a feature vector of 403 audio features.

To reduce the dimension of the feature vector, we applied LDA, as done in previous chapters. The maximum number of features allowed for LDA is the number of classes minus 1, which in this case was 29 since there were 30 phone classes. In order to find the optimal number of LDA features, we selected session 002 of each speaker and performed a simple classification task on the EMG data following the model architecture described in Section 7.2.3, using a batch size of 64 samples and 10 epochs and iterating over 1 to 29 features. See Figure 7.2 for the validation accuracy of each feature averaged over all speakers. Based on this graph, we selected 17 features for all experiments described further in this chapter.

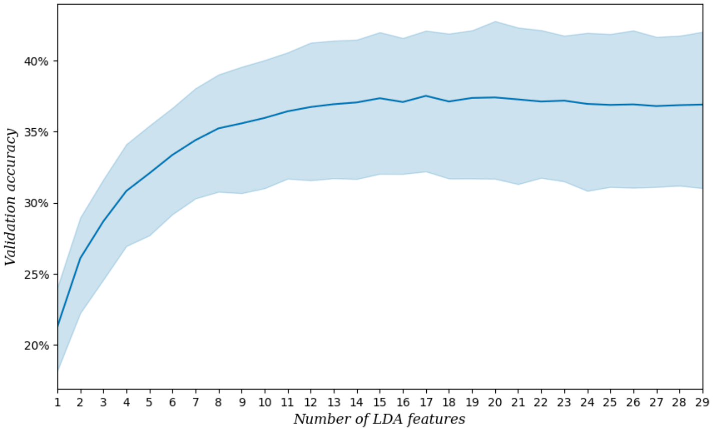


Figure 7.2: Validation accuracy (in %) per number of LDA features averaged over session 002 of each of the six speakers. The solid line represents mean accuracy, and the area above and below the line shows the standard deviation range.

7.2.3 Experiments

This section describes the classifiers we used for the experiments, and how we applied cross-validation.

Baseline

To function as a baseline, a dummy classifier was used to achieve chance-level accuracy. This dummy classifier chooses the most common class (so the class with the most EMG frames). Due to the unbalanced label distribution as shown in Figure 7.1, using a baseline that represents random selection (in this case 3.33%) would not be fair.

Feed-forward neural network

The NN used to perform the phone classification consisted of one feed-forward hidden layer with 34 nodes (double the number of inputs), and an output layer with 30 nodes (the number of phone classes). The activation function for the hidden layer was ReLU, while the output layer had a softmax activation function. As a metric to measure the multi-class classification accuracy of the network, the categorical cross-entropy loss function was used. The network was trained using an Adam optimizer and a learning rate of 10^{-3}. For a similar task in previous work (Chapter 5), we have also compared other classification models, such as a bagging classifier and a GMM. We learned that a bagging classifier outperforms an NN when using small datasets, but that a neural network is more effective when working with larger datasets.

We performed a hyper-parameter search in order to find the optimal batch size and number of epochs. Figure 7.3 shows the validation loss for 40 epochs and three different batch sizes: 32, 64, and 128. It can be seen that the batch size had no significant effect, so we selected 128, which had the shortest training time. Since the learning curve flattens after about 20 epochs, we selected this number for the final configuration.

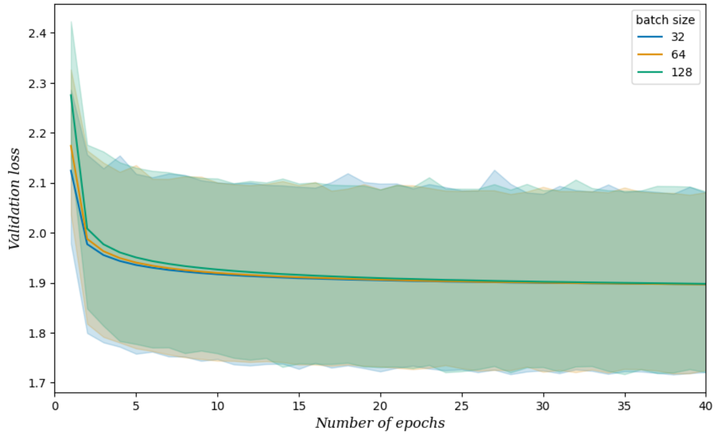


Figure 7.3: Validation loss per number of epochs averaged over session 002 of each of the six speakers, using a feed-forward neural network after LDA reduction with 17 features. Results are shown for three different batch sizes. The solid line represents the mean validation loss, and the area above and below the line shows the standard deviation range.

Cross-validation

As described in Section 7.2.1, we split the data into 80% train and 20% test sets. The test accuracy mentioned further in this chapter represents results based on the test set. However, in the training phase, we used a cross-validation procedure using 5 folds, and the average accuracy of these folds is referred to as the validation accuracy.

7.3 Results

The mean validation and test accuracy for all experiments are shown in Table 7.2. All the accuracy values mentioned in this chapter are frame-based. For the results per speaker, see Figure 7.4. The average time per model run was 13 minutes per session (without considering the dummy classification experiments). It can be seen that there is some variation between speakers, especially between speaker 006 and the other speakers. When speaker 006 is not taken into account, the mean test accuracy based on EMG features increases to 38.12%

Table 7.2: Validation (including standard deviation) and test accuracy for all experiments, averaged over all speakers and sessions.

Input features	Validation accuracy	Test accuracy
None (baseline)	13.86±1.17%	13.10%
EMG-TD	37.52±5.34%	35.98%
Audio-MFCCs	67.03±6.75%	69.68%

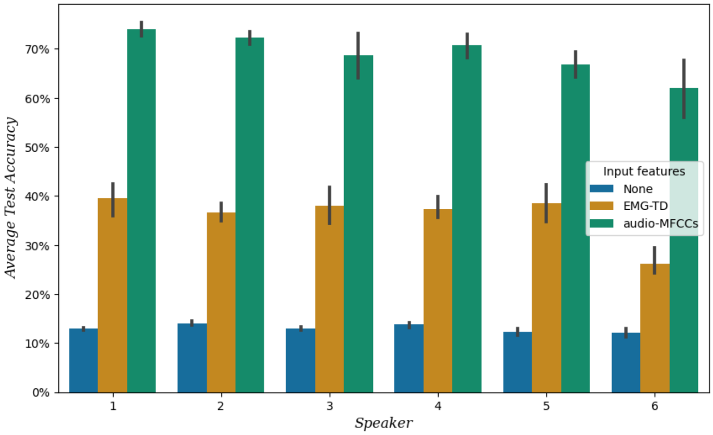


Figure 7.4: Test accuracy averaged over sessions per speaker for different types of input features: none (baseline; most common class), EMG-TD, and audio-MFCCs. The solid lines show the confidence intervals.

Figure 7.5 shows the confusion matrix for all phone classes for the classification on the test set of EMG features. The matrix is normalized by the true labels, to account for the imbalance of phone classes. The matrix is organized by a shared phonetic feature, namely the manner of articulation, resulting in the following phone groups: vowels ([a], [e], [i], [o], [u]), semivowels ([j], [w]) and consonants. The consonants are further divided into plosives ([b], [p], [t], [d], [k], [g]), fricatives ([β], [f], [θ], [ð], [s], [x], [ɣ]), affricates ([tʃ], [ɟ]), nasals ([m], [n], [ɲ]) and liquids ([l], [ʎ], [ɾ], [r]). The label 'sil' refers to the short pauses between words. These silences were predicted correctly in 28.94% of the cases. Table 7.3 shows the phone confusion pairs within each group, of cases where the confusion was higher than the accuracy of the true label.

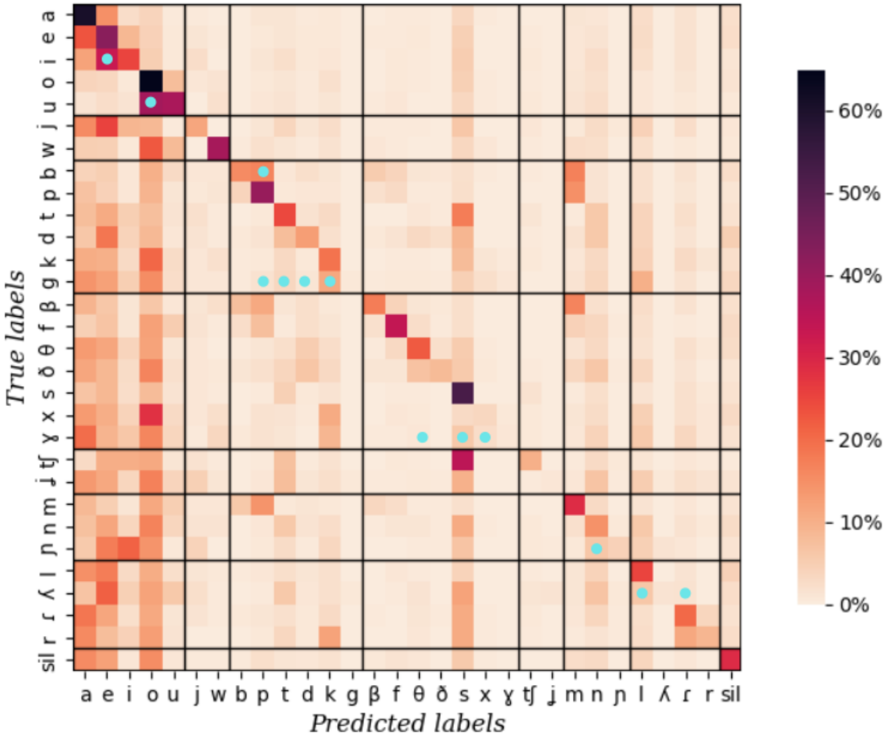


Figure 7.5: Confusion matrix of the results of the test sets averaged over all speakers and sessions for the EMG features. The blue dots show the phone confusions as listed in Table 7.3.

Table 7.3: Table of within-phone group confusions, showing instances where the confusion between phones was higher than the accuracy of the true phone.

True	Accuracy	Predicted	Confusion	Group
i	25.18%	e	32.86%	vowels
u	37.57%	o	38.68%	
ɲ	4.50%	n	6.13%	nasals
ʎ	1.41%	l	6.32%	liquids
		ɾ	1.57%	
r	9.02%	ɾ	11.31%	
b	15.67%	p	16.49%	plosives
ɡ	0.57%	k	11.39%	
		t	3.43%	
		p	1.72%	
		d	1.13%	
ɣ	0.95%	s	5.93%	fricatives
		x	2.35%	
		θ	1.00%	

7.4 Discussion and Conclusion

In the previous section, we presented the results of our phone classification experiments. To start with, they showed that the mean chance level is 13.10%, and the mean accuracy based on EMG features is 35.98% (see Table 7.2), which implies that phones can to some extent be differentiated using information from the muscles. We also presented the results of the same experiment but this time using features from the audio signals, which led to a mean accuracy of 69.68%. This result validates the model architecture but is not used for analysis since it does not contribute to the goal of this study.

We found a Pearson´s correlation between the phone accuracy and label counts of 0.79. We can observe in Figure 7.5 that the phones [a], [e], [o] and [s] are predicted more often than other phones, to be recognized as vertical lines in the confusion matrix in Figure 7.5. As can be seen in Figure 7.1, these phone classes are the ones with the highest counts, so the false positive predictions in this case are most likely an effect of the correlation. Using the same reasoning, we can observe that those labels whose presence is rare, like [ɣ], [ʝ], [x] or [ʎ], are almost never predicted, what can be recognized as white columns.

When two phones that show confusion are members of the same phone group, this confusion can be explained in terms of their phonetic features. For example, as can be seen in Table 7.3, the vowel [i] is more often incorrectly predicted as [e] than correctly as [i]. Similarly, the vowel [u] is more often predicted incorrectly as [o] than correctly as [u]. These vowel pairs are indeed very close in their manner of articulation and the difference in muscle movements is subtle enough to explain this confusion. From the nasals group it is not surprising that the [n] and the Spanish [ɲ] show some level of confusion since the biggest difference between those two phones is the movement of the tongue, which is hard to capture with EMG. The same is true for the two different r´s in Spanish, the [ɾ] and the more tongue-rolling [r], and the two different l´s, within the group of liquids. When looking at the plosives, two unvoiced-voiced pairs ([p]-[b] and [k]-[g]) show confusion among them, which is as expected, since they only differ in the voicing feature, and the EMG electrodes are unlikely to pick up on that. They also share the place of articulation, so there is very little phonetic difference between them. Similar confusion between voiced-unvoiced phone pairs was also reported in previous work [54].

The analysis of the classification accuracy of phones based on EMG signals shows that it is possible to derive certain information from them, yet the results also revealed some level of phone confusion. More specifically, confusion between two phones is more likely to occur when they share the manner of articulation, or only differ in voicing. We are confident that part of this issue can be addressed in developing an SSI, for example by applying language models.

7.5 Contribution

This chapter contributes to the research community with an analysis of phone confusion in EMG muscle activity between Spanish phones similar in manner of articulation, which is valuable information to know when developing a SSI for Spanish speakers. The level of confusion seems so significant that it must be addressed during development, for example, by including a language model.

Part of this chapter has been published as:

Salomons, Inge; Del Blanco, Eder; Navas, Eva; Hernáez, Inma. "Spanish Phone Confusion Analysis for EMG-Based Silent Speech Interfaces". In: *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1179–1183. DOI: 10.21437/Interspeech.2023-1881.

8
Statistical Analysis of EMG Patterns

This chapter analyses the muscle activity of eight muscles in the face and neck during speech production, by statistically modeling the activity levels of EMG signals acquired in different contexts. More specifically, using a generalized additive model (GAM), the root mean square (RMS) patterns of EMG signals acquired by different speakers (typical or alaryngeal), during multiple sessions and in two different speech modes (audible vs. silent), are compared. The results show that EMG signals of silent speech have significantly higher RMS levels than EMG signals of audible speech, suggesting that the speaker compensates for the lack of auditory feedback by articulating more. However, a subsequent qualitative comparison with the patterns associated with alaryngeal speech (showing lower RMS levels) suggests that the audible speech of laryngeal speakers may be more suitable for developing an SSI for alaryngeal speakers. Further analysis into the different muscles that the EMG signals were acquired from, and a comparison of phonetic outputs, indicate that a GAM analysis can be useful in understanding the relationship between muscle use and speech production.

8.1 Introduction

Most SSIs are trained on parallel EMG and audio signals and then tested with EMG signals in silent mode from laryngeal speakers [71, 99]. However, differences between EMG signals from audible and silent speech have been reported, and are attributed to the lack of auditory feedback in the case of silent speech, making this approach rather challenging [35, 99–102]. To overcome this discrepancy, in the current state-of-the-art, SSIs have been trained on silent speech instead, achieving a WER of 28.8% for direct EMG-to-speech in English [62] and 95.5% accuracy when classifying 10 Chinese digits [67], using data from a single speaker. For English alaryngeal speakers, a study has shown that EMG-based alaryngeal speech recognition is feasible, with a 10.3% average word error rate [40]. However, although models using voice conversion exist to restore speech for Spanish alaryngeal speakers [10, 103], there is a lack of studies focusing on EMG-based SSIs for this group of speakers.

The purpose of this chapter is to gain more insight into muscle activity in different contexts, to ultimately get closer to developing an SSI for alaryngeal speakers. We used the EMG signals from the new database we developed, described in Chapter 4. We calculated the RMS of each signal to represent the activity pattern of the muscle [31] and used a statistical method that can identify non-linear patterns (GAM; [104]) to study the effect of different variables on the RMS patterns of the EMG signals. Our aim was to identify how individual muscles used for speech production compare to each other, and if it matters how the speech was produced (audibly or silently), by whom (laryngeal or alaryngeal speaker), and how small differences in the (intended) phonetic output (Spanish *noche* compared to *leche*) affect the EMG signals.

The next section provides an overview of the data used for the experiments in this chapter, how it was processed, and an explanation of GAMs (Section 8.2). Then, we list the results of the experiments in Section 8.3, and discuss them in Section 8.4.

8.2 Methods

In this section, we provide details on the dataset we used in this chapter (8.2.1), how we processed the data (8.2.2), and the model we used to perform the experiments (8.2.3).

8.2.1 Data overview

The data used in this chapter is part of the ReSSInt database, namely a list of 100 common Spanish words (corpus 002; see Section 4.2.1) that was recorded by nine speakers multiple times, either in audible or silent speech mode. Out of the nine speakers, six were laryngeal speakers and three were speakers who had undergone a laryngectomy (alaryngeal speakers). In total, there are 48 repetitions of each word, namely 27 repetitions in audible mode, 17 repetitions in silent mode by laryngeal speakers, and 4 repetitions in silent mode by alaryngeal speakers. Every repetition refers to one session, in which all 100 words were recorded once. For each word there is one EMG signal for each of the eight muscles: anterior belly of the digastric (ABD), depressor anguli oris (DAO), depressor labii inferioris (DLI), levator labii superioris (LLS), masseter (MAS), risorius (RIS), stylohyoid (SLH), and zygomaticus major (ZYG). In summary, there are 38,400 signals in the selected data, namely eight EMG channels times 100 words times 48 repetitions.

8.2.2 Data processing

The first step of data processing is filtering out all signals that showed values below -30 or above 30 millivolts (mV). When a signal has values below or above these values, it means it is saturated and is often a result of insufficient contact of the electrode with the skin. We grouped the dataset in combinations of speakers, sessions, and channels beforehand, so that if a channel showed unreasonably high mV values for one or more words, all the words in that session were removed. After filtering, there were 28,574 signals left, meaning that 25.6% of the total available signals were filtered out. Per group, these percentages are 23.3% for audible speech, 27.2% for silent laryngeal speech, and 34.4% for silent alaryngeal speech. Table 8.1 shows the number of remaining signals per channel and speech mode, as well as the total. As can be seen, the DLI channel suffered the most loss by filtering and even resulted in zero signals in the case of alaryngeal speakers. This channel is located under the mouth, and the upper electrode was often affected by the arching of the lower lip, resulting in the detachment of the electrode. When one or more electrodes are not attached completely, this is reflected in the EMG signal with high amplitudes.

Table 8.1: Number of EMG signals per muscle and in total after filtering, per group. The numbers before filtering are calculated by multiplying the number of repetitions (sessions) by the number of words by the number of muscles (channels). The uneven number in the audible data set results from four missing words in the original data set.

	Total	**Audible**	**Silent larynx**	**Silent no larynx**
Before	**38,368**	**21,568**	**13,600**	**3,200**
SLH	4,696	2,596	1,700	400
MAS	4,596	2,496	1,700	400
ABD	4,397	2,497	1,600	300
ZYG	4,196	2,396	1,400	400
RIS	3,996	2,396	1,300	300
LLS	3,298	1,898	1,200	200
DAO	2,698	1,798	800	100
DLI	697	497	200	0
After	**28,574**	**16,574**	**9,900**	**2,100**

In the second step, the RMS of the mV values for each signal was calculated with a rectangular window size of 25 ms and a window shift of 5 ms.

During the third and final step, each signal was time-normalized between 0 and 1, to account for varying speech rates. The average signal duration was 1.57 seconds (1.56 seconds audible; 1.61 seconds silent; 1.42 seconds alaryngeal) with a standard deviation of 0.42 seconds.

8.2.3 Experiments

For the experiments, we used GAM [104], which is a non-linear regression method that can be used to model non-linear patterns. A GAM estimates the relationship between the response variable and each of the predictor variables using non-linear smooth functions. In this study, the responsible variable is always the time-normalized RMS function of the EMG signal, and the predictor variables depend on the task and can be muscle, word, or speech mode. The use of non-linear smooth functions allows for more flexibility than a traditional linear regression model. Consequently, GAMs are able to capture more complex patterns while simultaneously measuring the significance of potential effects or differences. To prevent overfitting the training data, a penalty factor is included, considering each smooth function's effective degrees of freedom (EDF). Higher EDF means a more complex or elaborate smooth function, which can be penalized to improve the model's generalization. In contrast to a linear regression model, visualization is required to interpret the patterns.

For this chapter, we used the *bam* function from the *mgcv* library [104] in *R* [105] to fit the models and the *itsadug* library [106] to interpret the results. We followed the approach described in [107] for model fitting.

8.3 Results

In this section, we report the effects of the different variables that each EMG signal differs in, namely word (8.3.1), speech mode (8.3.2), and muscle (8.3.3).

8.3.1 Effect of word

In the experiments following this section, we have included all of the 100 words. However, before proceeding to do that, it is essential to ensure the EMG data collected adequately is able to distinguish between words as well. For this validation step, we selected a minimal pair, namely the words *noche* [notʃe] and *leche* [letʃe], which differ in pronunciation in the first part and are the same in the last part. In this experiment, we fitted a model with the data from only these two words, assessing whether a significant difference was captured at the beginning of the word pronunciations (i.e. where these differed in their pronunciation). Our results show a difference in the activation pattern in general (averaged across all muscles; see Figure 8.1), but also specifically for muscles LLS (Figure 8.2a), DAO (Figure 8.2b), ABD (Figure 8.2c), RIS (Figure 8.2d) and ZYG (Figure 8.2e).

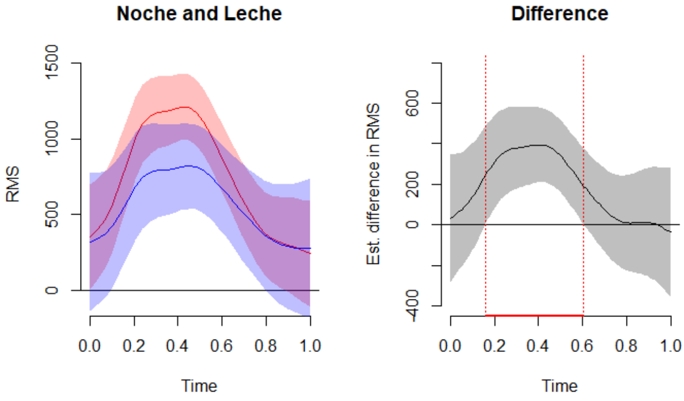


Figure 8.1: RMS patterns for noche (blue) and leche (red) on the left, and the difference plot on the right. The time window of the significant difference is 0.16 - 0.60.

In sum, these results show that GAMs are adequately able to capture differences in the EMG patterns, and therefore can be suitably used to compare different groups in the following sections.

8.3.2 Effect of speech mode

To understand the effect of speech mode (audible or silent, for the laryngeal speakers), we fitted a model with the RMS values as the response variable dependent on potential non-linear patterns over time. We explicitly assessed the difference between the time-based patterns for the two speech modes via a binary difference curve [107].

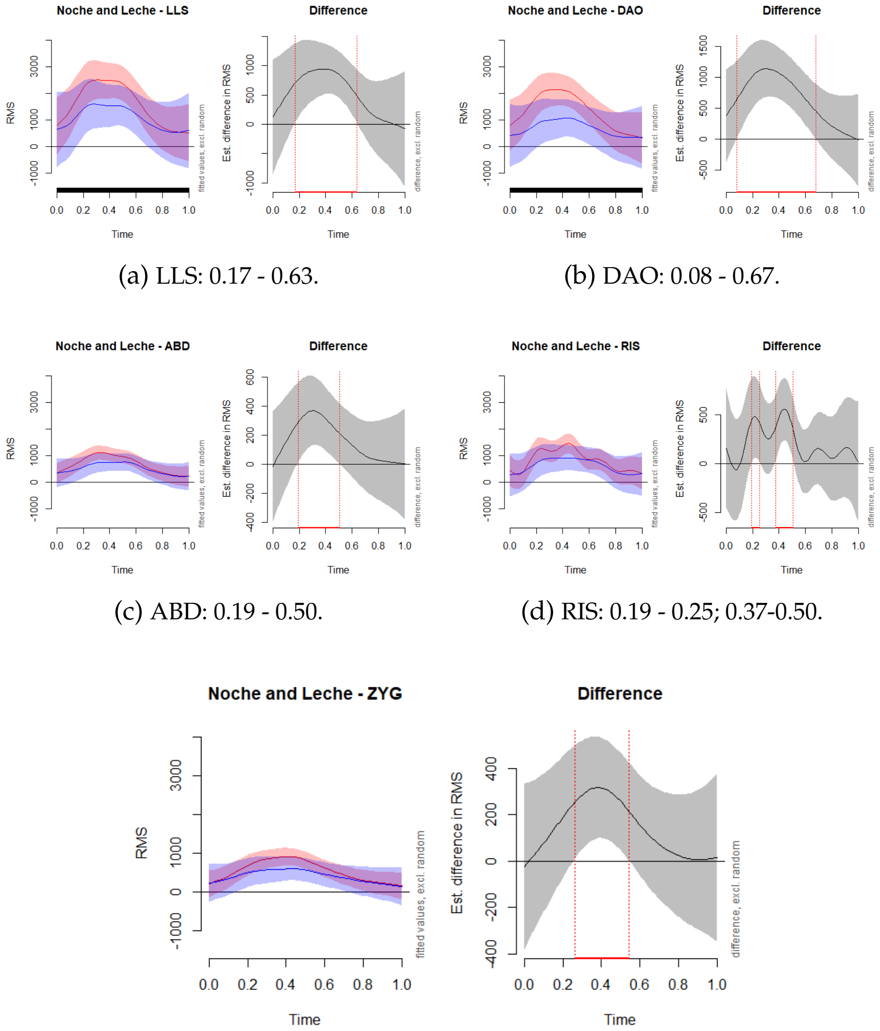
The results of this model show that the binary curve reflecting the difference between audible speech and silent speech was significant. Figure 8.3 shows the two RMS patterns and the difference plot showing where the patterns significantly differ. Clearly, the pattern for silent speech shows higher RMS values than for audible speech, with the difference being largest (and significant) during the first third part of the word pronunciation.

As explained in Section 8.2.1, each EMG signal belongs to one of three groups: the speakers with larynx (laryngeal speakers) produced speech either audibly or silently, and the laryngectomized (or alaryngeal) speakers spoke only in silent mode.

We fitted a second model, where we assessed the difference between the alaryngeal speakers and the two speech modes of the laryngeal speakers, again via binary difference curves. Figure 8.4 shows the smoothed RMS patterns over time for each group. Although these patterns seem to appear different, the results of the fitted model show no significant difference for both comparisons (audible vs. alaryngeal: $p = 0.195$; silent vs. alaryngeal: $p = 0.260$). However, the lack of a significant difference is caused by the low number of speakers in the alaryngeal group (only three speakers). Figure 8.4 shows that the pattern for the alaryngeal group is much closer to the audible speech than the silent speech of the laryngeal group.

8.3.3 Effect of muscle

We were also interested in the RMS patterns for each muscle in the two different speaking modes. For this experiment, we exclusively used the data from the speakers with a larynx and fitted a model where for each of the eight muscles, the speaking modes were contrasted. The results show that, except for RIS and SLH, the patterns differ significantly along part of the trajectory between the audible and silent production in the EMG signals of the muscle. In all cases, the RMS was higher for the silent production. Figure 8.5 and Figure 8.6 visualize the RMS patterns of each muscle in both modes and the difference between them.



(a) LLS: 0.17 - 0.63.

(b) DAO: 0.08 - 0.67.

(c) ABD: 0.19 - 0.50.

(d) RIS: 0.19 - 0.25; 0.37-0.50.

(e) ZYG: 0.26 - 0.54.

Figure 8.2: RMS patterns for noche (blue) and leche (red) on the left, and the difference plot on the right, for the muscles LLS, DAO, ABD, RIS, and ZYG. The numbers in the description show the time window(s) of the significant difference.

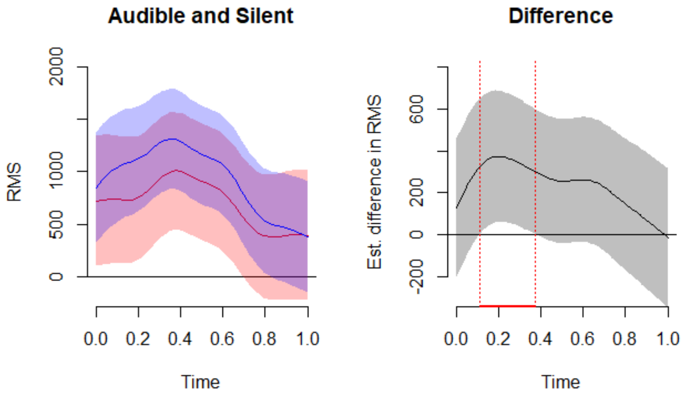


Figure 8.3: RMS patterns for audible speech (red) and silent speech (blue) on the left, and the difference plot on the right. The time window of the significant difference is 0.11 - 0.37.

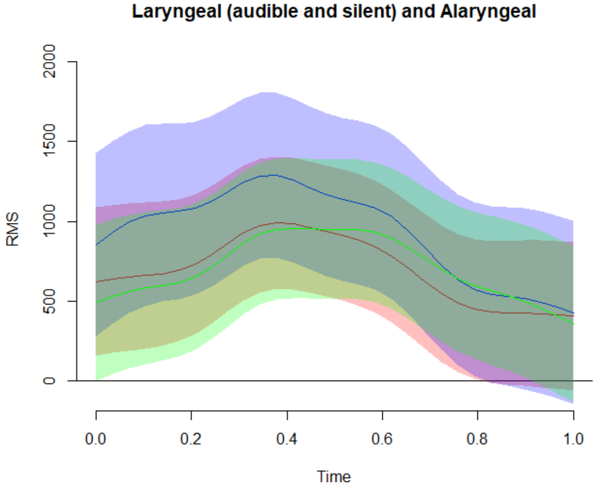


Figure 8.4: RMS patterns over time for each of the three groups. There is no significant difference between alaryngeal speech (green) and the two types of laryngeal speech: audible (red) and silent (blue).

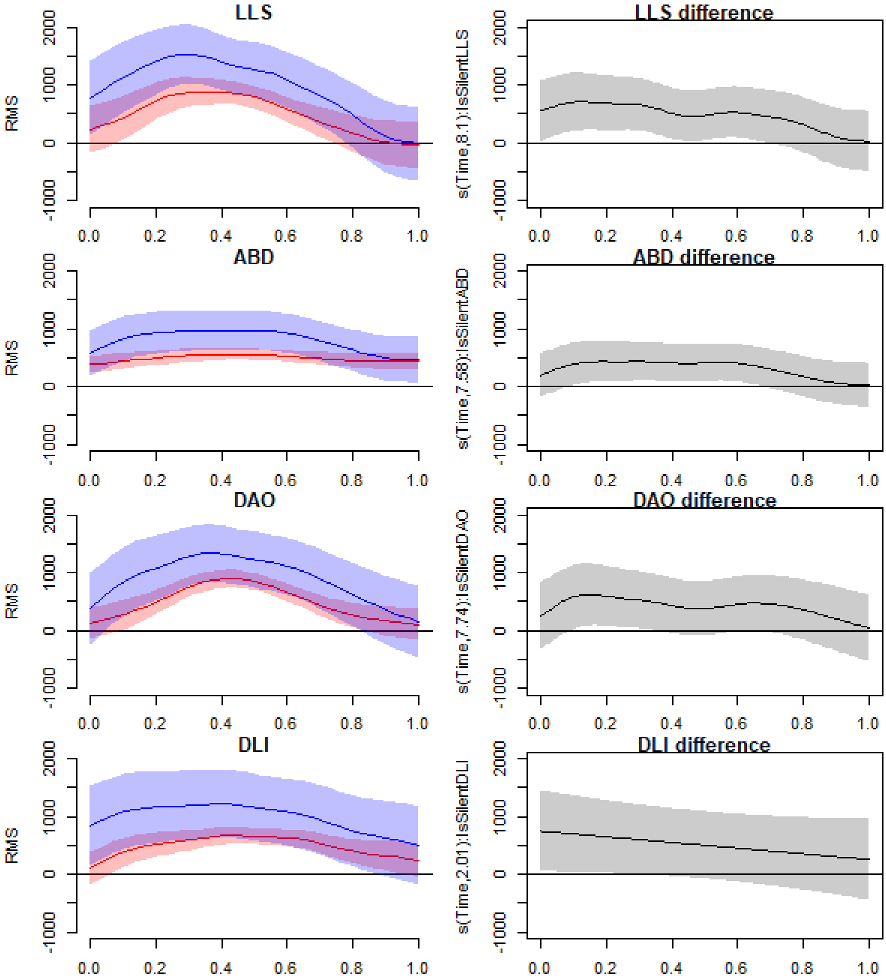


Figure 8.5: Individual muscle differences between audible (red) and silent (blue) mode for speakers with larynx, for the muscles LLS, ABD, DAO and DLI.

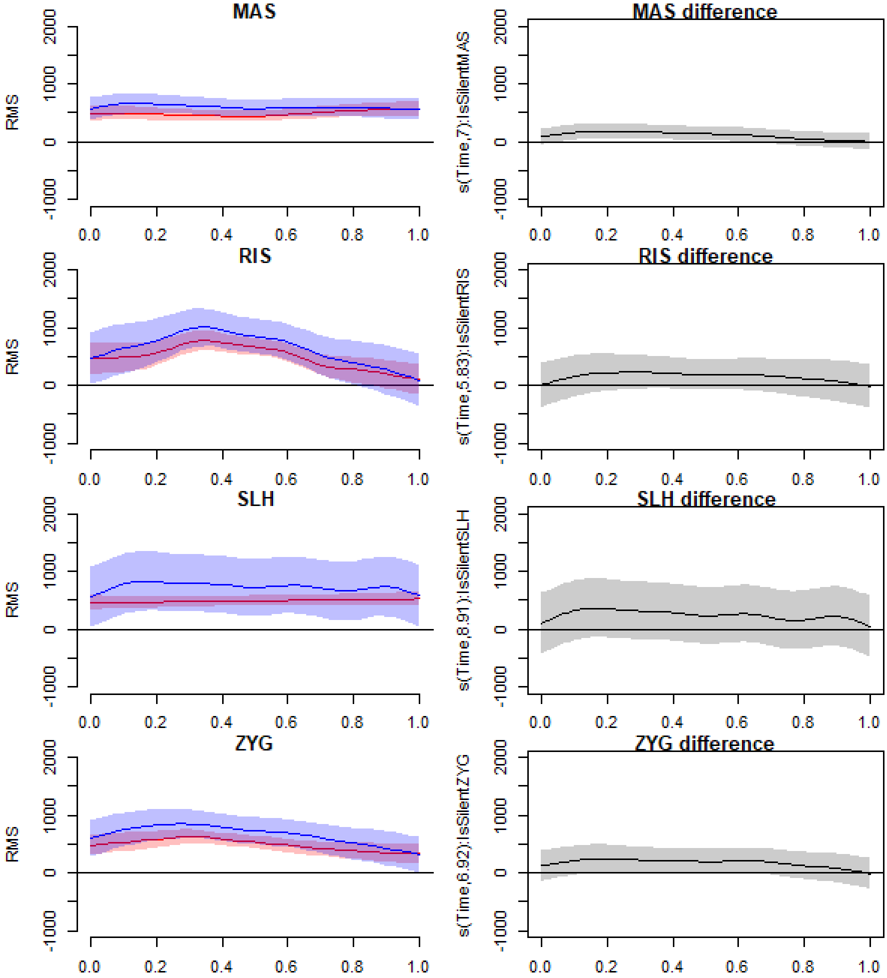


Figure 8.6: Individual muscle differences between audible (red) and silent (blue) mode for speakers with larynx, for the muscles MAS, RIS, SLH and ZYG.

8.4 Discussion and Conclusion

This chapter addresses several questions regarding the properties of EMG signals during audible and silent speech production among Spanish speakers. Through the analysis of EMG data collected from both laryngeal and alaryngeal speakers, the chapter aims to identify differences in muscle activation patterns in different contexts.

Our analysis reveals a significant difference in RMS patterns between audible and silent speech modes of laryngeal speakers, with higher RMS values for silent speech. This finding is in line with previous studies, which have also reported differences between these two speech modes [35, 99–102]. We believe that due to the absence of auditory feedback, the speaker tries to compensate by articulating more when speaking silently. When extending this analysis to the individual muscles, it appears that this difference is present in all muscles except RIS and SLH.

Further analysis reveals no significant difference between audible and silent speech from laryngeal speakers in comparison to alaryngeal speech. However, given the limited data from alaryngeal speakers, an analysis with more data is required before drawing any representative conclusions. Interestingly, the pattern of alaryngeal speech seems to be more similar to that of audible speech from laryngeal speakers as opposed to silent speech. If this trend proves to be significant in a larger dataset, it could mean that using audible speech might be a better approach to training an SSI for alaryngeal speakers than using silent speech from laryngeal speakers.

We were rather strict in our filtering process, as we decided to remove all signals of a channel in one session if one or more signals of that channel showed abnormally low or high mV values. We made this decision to make sure that all signals in the data used are of good quality, but we acknowledge that we might have filtered out good signals as well. Nevertheless, we were able to identify clear significantly different patterns, even while being relatively conservative in our data selection process.

For future research, it could be interesting to study the RMS patterns of signals of different phoneme groups. The first experiment in this chapter, which compared a minimal word pair to evaluate the suitability of using GAMs, already showed that this kind of analysis can be used to highlight which muscles are activated based on differences in the phonetic output. Further analysis of the activity of one or more muscles in specific contexts might be useful to identify potential relationships between muscle use and the production of specific sounds. Similarly, it would be worth it to evaluate if the potential relationships are similar across speakers, or whether each speaker has a unique way of using their muscles to produce certain phone groups.

8.5 Contribution

This chapter uses GAM to analyze EMG signals during speech. It finds higher EMG activity in silent speech compared to audible speech, suggesting compensatory muscle use. Furthermore, audible speech patterns in laryngeal speakers are potentially better suited for SSIs for alaryngeal speakers. It also shows a promising method to study more closely individual muscle use.

Part of this chapter has been accepted for publication:

IberSPEECH 2024. "Analyzing Speech Muscle Activity Using a Generalized Additive Model". **Salomons, Inge**; Hernáez, Inma; Navas, Eva; Wieling, Martijn. *Accepted for publication.*

Part IV

Conclusion

9
General Discussion and Conclusion

In this thesis, we have described a new database that we developed, and experiments we performed with this data. The database was the first important step in the development of an EMG-based SSI for alaryngeal Spanish speakers, and the experiments aimed at understanding the challenges that arise in the development of such a technology.

In the field of EMG-based SSIs, major advances have been made since the first study almost four decades ago. It started with the successful identification of speech-related information in EMG signals [43]. Currently, speech can be generated from EMG signals with a WER of 28.8% using a single-speaker model trained on nearly 20 hours of data from a typical male English speaker [62]. More research is needed to determine whether it is possible to lower that error rate using a multi-speaker model trained on less data from a variety of speakers (including those who experience speech difficulties) speaking different languages.

To develop an SSI for our target group, alaryngeal Spanish speakers, there was not yet a database available. For this reason, we first conducted a pilot experiment to find the optimal EMG electrode setup (Chapter 3). The study analyzed the contribution of specific muscles in the face and neck to the performance of a phone classifier. Building on previous research, we initially targeted 14 muscles and subsequently identified eight key muscles based on their contributions. The final setup consists of eight bipolar single-electrode pairs, each targeting one of the muscles ABD, DAO, RIS, LLS, MAS, ZYG, DLI and SLH. The DLI and SLH are new additions compared to other studies in this field, whereas the other six have been frequently used as well.

Although we believe we made the right choice with the information available to us at the time of this pilot study, we encourage other researchers in this field to explore other options as well, such as electrode arrays, the effect of speaker differences, and facial asymmetry.

The development of the ReSSInt-EMG database, involving a recording procedure and regular data quality assessments, resulted in a collection of 22.5 hours of EMG and speech signals (Chapter 4). These signals were recorded across various contexts, in terms of speaker type, speech mode, and linguistic content. The database is a major contribution to the field because it allows for the extension of SSI development for non-English speaking people with speech difficulties.

Despite the precautions that were taken to ensure data quality, occasional deviations from the intended quality standard may have occurred due to the complex and sensitive nature of EMG data acquisition.

Validation experiments conducted on the ReSSInt-EMG database aimed to establish the efficacy of the acquisition setup (Chapter 5). By comparing phone classification results with those from a public English speech and EMG database, we validated the robustness of our data acquisition procedure. Despite differences in experimental setup and languages, our results demonstrated comparable performance, affirming the validity of our new database.

Further analysis addressed an important concern in EMG-based SSI research, namely the effect of variability across speakers and sessions (Chapter 6). This phenomenon of inter-speaker and inter-session variability is inevitable and is attributed to factors such as electrode placement, speaker physiology, or environmental conditions. The main challenge it poses is the development of a multi-speaker interface, for which it needs to be able to generalize. However, even though we found that variability had a negative effect on the classification performance using our data, it appeared that increasing the amount of training data downsized this effect.

Moreover, we performed a phone confusion analysis to assess the potential of using EMG signals to differentiate between speech sounds that are phonetically similar (Chapter 7). For example, in the case of [p] and [b], phonetically speaking, the only difference is the voicing feature. Without this information, will the interface be able to pick up on what might be minuscule differences in muscle use when producing these sounds? The same can be asked for sounds that differ only in tongue position, of which there is no data, for example, the [r] and [ɾ]. For English, a study already confirmed this assumption and revealed confusion occurring mostly between voiced-unvoiced pairs [54]. Our analysis shows that also for Spanish, phone confusion can be a challenge for EMG-to-Speech research, but we are confident that this can be overcome.

Lastly, muscle activation patterns of EMG signals acquired in different conditions were compared, using a statistical method that can model non-linear patterns (Chapter 8). It revealed a higher muscle activity in silent speech compared to audible speech. We attributed this to the lack of acoustic feedback, and that the silent speaker is making up for this by articulating more. Notably, alaryngeal (silent) speech exhibited similarities to audible speech from laryngeal speakers instead of silent speech. However, we had very little data available for alaryngeal speech, so more analyses are needed to confirm this. In general, we believe that each dataset is unique and that this kind of analysis is crucial to understanding the data's characteristics.

Overall, our findings underscore the complexity of recognizing speech from EMG signals and highlight the importance of comprehensive data collection, thorough validation and quality assessments, and in-depth analyses to advance the field of SSIs. Further research addressing inter-speaker variability and phonetic nuances will be instrumental in realizing the full potential of EMG-based SSIs for Spanish alaryngeal speakers.

Appendices

Appendix A List of VCV words

	a	**e**	**i**	**o**	**u**
p	atapata	atepeta	atipita	atopota	atuputa
m**b**	atambata	atembeta	atimbita	atombota	atumbuta
t	atatata	ateteta	atitita	atotota	atututa
n**d**	atandata	atendeta	atindita	atondota	atunduta
k	atakata	ateketa	atikita	atokota	atukuta
n**g**	atangata	atengueta	atinguita	atongota	atunguta
S	atachata	atecheta	atichita	atochota	atuchuta
jj	atallata	atelleta	atillita	atollota	atulluta
f	atafata	atefeta	atifita	atofota	atufuta
B	atabata	atebeta	atibita	atobota	atubuta
T	atazata	ateceta	aticita	atozota	atuzuta
D	atadata	atedeta	atidita	atodota	atuduta
s	atasata	ateseta	atisita	atosota	atusuta
x	atajata	atejeta	atijita	atojota	atujuta
G	atagata	ategueta	atiguita	atogota	atuguta
m	atamata	atemeta	atimita	atomota	atumuta
n	atanata	ateneta	atinita	atonota	atunuta
J	atañata	ateñeta	atiñita	atoñota	atuñuta
l	atalata	ateleta	atilita	atolota	atuluta
L	atayata	ateyeta	atiyita	atoyota	atuyuta
r	atarata	atereta	atirita	atorota	aturuta
rr	atarrata	aterreta	atirrita	atorrota	aturruta

Table 1: List of VCV words. The top row shows the vowels (V) and the left column the consonants (C) in SAMPA notation. The resulting combinations in reading format are the words that were presented to the speaker. The plosives /b/, /d/, and /g/ are induced by an extra nasal sound in front of the plosive sound.

Appendix B List of 100 most useful Spanish words

día	mayo	coche	otoño	entrar	autobús	adelante
dos	niño	enero	padre	hombre	derecha	bicicleta
fin	ocho	error	perro	idioma	domingo	calendario
mes	seis	fruta	primo	inicio	escuela	diciembre
pan	tres	julio	salir	jueves	familia	invierno
sol	vida	junio	siete	lluvia	febrero	izquierda
tío	vino	leche	tarde	mañana	hermano	mediodía
uno	abril	lugar	abuelo	martes	octubre	miércoles
agua	árbol	lunes	agosto	planta	pescado	noviembre
café	avión	madre	camión	pueblo	próximo	primavera
casa	barco	marzo	ciudad	sábado	sobrina	septiembre
diez	calle	mujer	comida	semana	trabajo	temporada
flor	calor	nieta	cuatro	tiempo	verdura	transporte
frío	carne	noche	detrás	verano	viernes	
hija	cinco	nueve				

Table 2: List of 100 most useful Spanish words sorted alphabetically and by word length.

Appendix C Detailed database information

Table 3 shows detailed information about the ReSSInt database. It lists which sessions were recorded per speaker, and the duration of signals for each corpus included in that session. It also shows totals per session, per speaker, and for all speakers and sessions (in the end). Information about the kind of content that each corpus contains can be found in Table 4.1. The explanations of the abbreviations are as follows:

- NP:A = non-parallel audible; duration of audible signals in a session where each signal of this corpus was only recorded in audible mode.
- NP:S = non-parallel silent; duration of silent signals in a session

where each signal of this corpus was only recorded in silent mode.

- P:A = parallel audible; duration of audible signals in a session where each signal of this corpus was recorded in both modes.
- P:S = parallel silent; duration of silent signals in a session where each signal of this corpus was recorded in both modes.

Table 3: Detailed database information. Duration format is (hh:)mm:ss.

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
001	001	001	2:36	–:–	–:–	–:–	19:24
		002	–:–	–:–	–:–	–:–	
		003	4:27	–:–	–:–	–:–	
		004	12:21	–:–	–:–	–:–	
	002	001	2:24	–:–	–:–	–:–	21:50
		002	1:57	–:–	–:–	–:–	
		003	4:31	–:–	–:–	–:–	
		005	12:58	–:–	–:–	–:–	
	003	001	2:15	–:–	–:–	–:–	21:02
		002	1:50	–:–	–:–	–:–	
		003	4:42	–:–	–:–	–:–	
		006	12:15	–:–	–:–	–:–	
	004	001	2:56	–:–	–:–	–:–	24:43
		002	2:27	–:–	–:–	–:–	
		003	5:21	–:–	–:–	–:–	
		007	13:59	–:–	–:–	–:–	
	005	001	–:–	–:–	2:54	2:57	21:01
		002	–:–	–:–	2:10	2:11	
		003	–:–	–:–	5:19	5:30	
	006	001	–:–	3:09	–:–	–:–	28:56
		002	–:–	2:12	–:–	–:–	
		003	–:–	5:37	–:–	–:–	
		004	–:–	17:58	–:–	–:–	

Continues on next page

Table 3 – *continued from previous page*

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
	007	002	–:–	–:–	2:03	2:02	25:01
		003	–:–	–:–	5:08	5:45	
		008	10:03	–:–	–:–	–:–	
	008	002	–:–	–:–	2:27	2:16	25:43
		003	–:–	–:–	5:20	6:03	
		009	9:37	–:–	–:–	–:–	
	010	010	–:–	–:–	0:31	1:12	22:53
		011	–:–	–:–	6:31	14:39	
	011	010	–:–	–:–	0:30	1:07	22:52
		012	–:–	–:–	6:33	14:42	
	012	010	–:–	–:–	0:27	0:58	20:20
		013	–:–	–:–	6:02	12:53	
	013	010	–:–	–:–	0:29	1:00	20:31
		014	–:–	–:–	6:01	13:01	
	014	010	–:–	–:–	0:26	0:58	18:59
		015	–:–	–:–	5:36	11:59	
	015	010	–:–	–:–	0:26	0:56	18:25
		016	–:–	–:–	5:25	11:38	
		Total	1:46:39	28:56	1:04:18	1:51:47	5:11:40
002	001	001	2:19	–:–	–:–	–:–	29:09
		002	1:29	–:–	–:–	–:–	
		003	5:50	–:–	–:–	–:–	
		004	19:31	–:–	–:–	–:–	
	002	001	–:–	–:–	–:–	–:–	33:25
		002	2:55	–:–	–:–	–:–	
		003	7:22	–:–	–:–	–:–	
		005	23:08	–:–	–:–	–:–	
	003	001	3:22	–:–	–:–	–:–	28:28
		002	2:34	–:–	–:–	–:–	
		003	6:39	–:–	–:–	–:–	
		006	15:53	–:–	–:–	–:–	

Continues on next page

Table 3 – *continued from previous page*

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
	004	001	4:39	–:–	–:–	–:–	37:00
		002	3:41	–:–	–:–	–:–	
		003	8:13	–:–	–:–	–:–	
		007	20:27	–:–	–:–	–:–	
	005	001	–:–	–:–	4:29	4:32	31:52
		002	–:–	–:–	3:00	3:07	
		003	–:–	–:–	7:28	9:16	
	006	001	–:–	4:22	–:–	–:–	42:50
		002	–:–	3:11	–:–	–:–	
		003	–:–	7:57	–:–	–:–	
		004	–:–	27:20	–:–	–:–	
	007	002	–:–	–:–	3:53	3:36	40:04
		003	–:–	–:–	8:54	8:54	
		008	14:47	–:–	–:–	–:–	
	008	002	–:–	–:–	3:42	3:04	36:51
		003	–:–	–:–	8:33	8:10	
		009	13:22	–:–	–:–	–:–	
		Total	2:36:11	42:50	39:59	40:39	4:39:39
003	001	001	4:28	–:–	–:–	–:–	32:05
		002	3:03	–:–	–:–	–:–	
		003	5:55	–:–	–:–	–:–	
		004	18:39	–:–	–:–	–:–	
	002	001	2:55	–:–	–:–	–:–	25:01
		002	2:51	–:–	–:–	–:–	
		003	4:36	–:–	–:–	–:–	
		005	14:39	–:–	–:–	–:–	
	005	001	–:–	–:–	3:01	3:12	22:06
		002	–:–	–:–	2:18	2:50	
		003	–:–	–:–	5:08	5:37	
	006	001	–:–	2:52	–:–	–:–	27:45
		002	–:–	2:27	–:–	–:–	

Continues on next page

Table 3 – *continued from previous page*

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
		003	–:–	5:18	–:–	–:–	
		004	–:–	17:08	–:–	–:–	
		Total	57:06	27:45	10:27	11:39	1:46:57
004	001	001	–:–	–:–	–:–	–:–	28:23
		002	2:23	–:–	–:–	–:–	
		003	6:16	–:–	–:–	–:–	
		004	19:44	–:–	–:–	–:–	
	002	001	2:44	–:–	–:–	–:–	29:28
		002	2:39	–:–	–:–	–:–	
		003	5:43	–:–	–:–	–:–	
		005	18:22	–:–	–:–	–:–	
	005	001	–:–	–:–	2:42	2:32	21:17
		002	–:–	–:–	2:22	2:13	
		003	–:–	–:–	5:28	6:00	
	006	001	–:–	3:08	–:–	–:–	31:55
		002	–:–	2:18	–:–	–:–	
		003	–:–	5:58	–:–	–:–	
		004	–:–	20:31	–:–	–:–	
		Total	57:51	31:55	10:32	10:45	1:51:03
005	001	001	3:01	–:–	–:–	–:–	29:04
		002	2:33	–:–	–:–	–:–	
		003	5:51	–:–	–:–	–:–	
		004	17:39	–:–	–:–	–:–	
	002	001	3:31	–:–	–:–	–:–	28:13
		002	2:14	–:–	–:–	–:–	
		003	5:38	–:–	–:–	–:–	
		005	16:50	–:–	–:–	–:–	
	003	001	3:41	–:–	–:–	–:–	23:27
		002	2:17	–:–	–:–	–:–	
		003	5:14	–:–	–:–	–:–	
		006	12:15	–:–	–:–	–:–	

Continues on next page

Table 3 – *continued from previous page*

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
	004	001	2:57	–:–	–:–	–:–	25:20
		002	2:46	–:–	–:–	–:–	
		003	6:10	–:–	–:–	–:–	
		007	13:27	–:–	–:–	–:–	
	005	001	–:–	–:–	2:52	3:13	20:12
		002	–:–	–:–	–:–	2:32	
		003	–:–	–:–	5:26	6:09	
	006	001	–:–	2:51	–:–	–:–	31:49
		002	–:–	2:20	–:–	–:–	
		003	–:–	6:25	–:–	–:–	
		004	–:–	20:13	–:–	–:–	
	007	002	–:–	–:–	3:13	3:27	28:43
		003	–:–	–:–	6:05	7:06	
		008	8:52	–:–	–:–	–:–	
	008	002	–:–	–:–	2:50	2:52	27:02
		003	–:–	–:–	5:43	6:34	
		009	9:03	–:–	–:–	–:–	
		Total	2:03:59	31:49	26:09	31:53	3:33:50
006	001	001	2:30	–:–	–:–	–:–	31:15
		002	2:03	–:–	–:–	–:–	
		003	6:19	–:–	–:–	–:–	
		004	20:23	–:–	–:–	–:–	
	002	001	2:45	–:–	–:–	–:–	33:54
		002	2:13	–:–	–:–	–:–	
		003	7:14	–:–	–:–	–:–	
		005	21:42	–:–	–:–	–:–	
	005	001	–:–	–:–	3:04	3:36	28:06
		002	–:–	–:–	2:13	2:28	
		003	–:–	–:–	8:29	8:16	
	006	001	–:–	3:42	–:–	–:–	34:32
		002	–:–	3:08	–:–	–:–	
		003	–:–	7:08	–:–	–:–	

Continues on next page

Table 3 – *continued from previous page*

Speaker	Session	Corpus	NP:A	NP:S	P:A	P:S	Total
		004	–:–	20:34	–:–	–:–	
		Total	1:05:09	34:32	13:46	14:20	2:07:47
007	006	001	–:–	–:–	–:–	–:–	46:44
		002	–:–	4:14	–:–	–:–	
		003	–:–	12:07	–:–	–:–	
		004	–:–	30:23	–:–	–:–	
	009	001	–:–	2:48	–:–	–:–	48:47
		002	–:–	2:36	–:–	–:–	
		003	–:–	10:11	–:–	–:–	
		005	–:–	33:12	–:–	–:–	
		Total	–:–	1:35:31	–:–	–:–	1:35:31
008	006	001	–:–	2:44	–:–	–:–	32:47
		002	–:–	2:08	–:–	–:–	
		003	–:–	6:29	–:–	–:–	
		004	–:–	21:26	–:–	–:–	
		Total	–:–	32:47	–:–	–:–	32:47
009	006	001	–:–	3:40	–:–	–:–	36:48
		002	–:–	2:47	–:–	–:–	
		003	–:–	7:20	–:–	–:–	
		004	–:–	23:01	–:–	–:–	
	009	001	–:–	3:42	–:–	–:–	36:51
		002	–:–	2:00	–:–	–:–	
		003	–:–	7:26	–:–	–:–	
		005	–:–	23:43	–:–	–:–	
		Total	–:–	1:13:39	–:–	–:–	1:13:39
		Total	9:26:55	6:39:44	2:45:11	3:41:03	22:32:53

Bibliography

[1] Hernaez, Inma; González-López, Jose Andrés; Navas, Eva; Pérez Córdoba, Jose Luis; Saratxaga, Ibon; Olivares, Gonzalo; Sánchez de la Fuente, Jon; Galdón, Alberto; García Romillo, Víctor; González-Atienza, Míriam; Schultz, Tanja; Green, Phil; Wand, Michael; Marxer, Ricard; Diener, Lorenz. "Voice Restoration with Silent Speech Interfaces (ReSSInt)". In: *IberSPEECH 2021*. ISCA, Mar. 2021, pp. 130–134. DOI: 10.21437/IberSPEECH.2021-28.

[2] Hernaez, Inma; Gonzalez Lopez, Jose Andres; Navas, Eva; Pérez Córdoba, Jose Luis; Saratxaga, Ibon; Olivares, Gonzalo; Sanchez de la Fuente, Jon; Galdón, Alberto; Garcia, Victor; Castillo, Jesús del; Salomons, Inge; Blanco Sierra, Eder del. "ReSSInt Project: Voice Restoration Using Silent Speech Interfaces". In: *IberSPEECH 2022*. ISCA, Nov. 2022, pp. 226–230. DOI: 10.21437/IberSPEECH.2022-46.

[3] Moore, Keith L.; Agur, Anne M. R.; Dalley Arthur F., II. *Essential Clinical Anatomy*. Fifth. Philadelphia: Wolters Kluwer Health, 2015.

[4] Tang, Christopher G.; Sinclair, Catherine F. "Voice Restoration After Total Laryngectomy." In: *Otolaryngol Clin North Am* 48.4 (Aug. 2015), pp. 687–702. DOI: 10.1016/j.otc.2015.04.013.

[5] Weinberg, Bernd. "Acoustical Properties of Esophageal and Tracheoesophageal Speech". In: *Laryngectomee rehabilitation* (1986), pp. 113–127.

[6] Most, Tova; Tobin, Yishai; Mimran, Ravit Cohen. "Acoustic and Perceptual Characteristics of Esophageal and Tracheoesophageal Speech Production". In: *Journal of communication disorders* 33.2 (2000), pp. 165–181.

[7] Bell, R. Bryan; Andersen, Peter; Fernandes, Rui. *Oral, Head and Neck Oncology and Reconstructive Surgery*. 1st ed. Elsevier, 2016.

[8] Repova, Barbora; Zabrodsky, Michal; Plzak, Jan; Kalfert, David; Matousek, Jindrich; Betka, Jan. “Text-to-Speech Synthesis as an Alternative Communication Means after Total Laryngectomy”. In: *Biomedical Papers* 165.2 (June 2021), pp. 192–197. DOI: 10.5507/bp.2020.016.

[9] Nakamura, Keigo; Toda, Tomoki; Saruwatari, Hiroshi; Shikano, Kiyohiro. “Speaking-Aid Systems Using GMM-based Voice Conversion for Electrolaryngeal Speech”. In: *Speech Communication* 54.1 (Jan. 2012), pp. 134–146. DOI: 10.1016/j.specom.2011.07.007.

[10] Serrano, Luis; Raman, Sneha; Tavarez, David; Navas, Eva; Hernaez, Inma. “Parallel vs. Non-Parallel Voice Conversion for Esophageal Speech”. In: *Interspeech 2019*. ISCA, Sept. 2019, pp. 4549–4553. DOI: 10.21437/Interspeech.2019-2194.

[11] Fuchs, Anna Katharina; Hagmüller, Martin; Kubin, Gernot. “The New Bionic Electro-Larynx Speech System”. In: *IEEE Journal of Selected Topics in Signal Processing* 10.5 (Aug. 2016), pp. 952–961. DOI: 10.1109/JSTSP.2016.2535970.

[12] Ahmadi, Farzaneh; Kobayashi, Kazuhiro; Toda, Tomoki. “Development of a Real-time Bionic Voice Generation System Based on Statistical Excitation Prediction”. In: *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. Pittsburgh PA USA: ACM, Oct. 2019, pp. 655–657. DOI: 10.1145/3308561.3354591.

[13] Zieliński, Konrad; Szamburski, Ryszard; Machnacz, Ewa. *Post-Laryngectomy Interaction Restoration System*. Mar. 2019, p. 164. DOI: 10.1145/3308557.3308731.

[14] Janke, Matthias; Diener, Lorenz. “EMG-to-Speech: Direct Generation of Speech from Facial Electromyographic Signals”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017), pp. 2375–2385. DOI: 10.1109/TASLP.2017.2738568.

[15] Zieliński, Konrad; Rączaszek-Leonardi, Joanna. "A Complex Human-Machine Coordination Problem: Essential Constraints on Interaction Control in Bionic Communication Systems". In: *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–8. DOI: 10.1145/3491101.3519672.

[16] Denby, B.; Schultz, T.; Honda, K.; Hueber, T.; Gilbert, J. M.; Brumberg, J. S. "Silent Speech Interfaces". In: *Speech Communication* 52.4 (2010), pp. 270–287. DOI: 10.1016/j.specom.2009.08.002.

[17] Freitas, João; Teixeira, António; Dias, Miguel Sales; Silva, Samuel. *An Introduction to Silent Speech Interfaces*. Springer Cham, 2016.

[18] Gonzalez-Lopez, Jose A.; Gomez-Alanis, Alejandro; Martin Donas, Juan M.; Perez-Cordoba, Jose L.; Gomez, Angel M. "Silent Speech Interfaces for Speech Restoration: A Review". In: *IEEE Access* 8 (2020), pp. 177995–178021. DOI: 10.1109/ACCESS.2020.3026579.

[19] Chung, Joon Son; Senior, Andrew; Vinyals, Oriol; Zisserman, Andrew. "Lip Reading Sentences in the Wild". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (July 2017), pp. 3444–3453. DOI: 10.1109/CVPR.2017.367.

[20] Gonzalez, Jose A.; Cheah, Lam A.; Gomez, Angel M.; Green, Phil D.; Gilbert, James M.; Ell, Stephen R.; Moore, Roger K.; Holdsworth, Ed. "Direct Speech Reconstruction From Articulatory Sensor Data by Machine Learning". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (Dec. 2017), pp. 2362–2374. DOI: 10.1109/TASLP.2017.2757263.

[21] Gonzalez, Jose A.; Green, Phil D. "A Real-Time Silent Speech System for Voice Restoration after Total Laryngectomy". In: *Revista de Logopedia, Foniatría y Audiología* 38.4 (Oct. 2018), pp. 148–154. DOI: 10.1016/j.rlfa.2018.07.004.

[22] Anumanchipalli, Gopala K.; Chartier, Josh; Chang, Edward F. "Speech Synthesis from Neural Decoding of Spoken Sentences". In: *Nature* 568.7753 (Apr. 2019), pp. 493–498. DOI: 10.1038/s41586-019-1119-1.

[23] Herff, Christian; Diener, Lorenz; Angrick, Miguel; Mugler, Emily; Tate, Matthew C.; Goldrick, Matthew A.; Krusienski, Dean J.; Slutzky, Marc W.; Schultz, Tanja. "Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices". In: *Front. Neurosci.* 13 (Nov. 2019), p. 1267. DOI: 10.3389/fnins.2019.01267.

[24] Willett, Francis R.; Kunz, Erin M.; Fan, Chaofei; Avansino, Donald T.; Wilson, Guy H.; Choi, Eun Young; Kamdar, Foram; Glasser, Matthew F.; Hochberg, Leigh R.; Druckmann, Shaul; Shenoy, Krishna V.; Henderson, Jaimie M. "A High-Performance Speech Neuroprosthesis". In: *Nature* 620.7976 (Aug. 2023), pp. 1031–1036. DOI: 10.1038/s41586-023-06377-x.

[25] Sun, Ke; Yu, Chun; Shi, Weinan; Liu, Lan; Shi, Yuanchun. "LipInteract: Improving Mobile Device Interaction with Silent Speech Commands". In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. 2018, pp. 581–593.

[26] Zhang, Ruidong; Chen, Mingyang; Steeper, Benjamin; Li, Yaxuan; Yan, Zihan; Chen, Yizhuo; Tao, Songyun; Chen, Tuochao; Lim, Hyunchul; Zhang, Cheng. "SpeeChin: A Smart Necklace for Silent Speech Recognition". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.4 (2021), pp. 1–23.

[27] Krishna, Gautam; Tran, Co; Yu, Jianguo; Tewfik, Ahmed H. "Speech Recognition with No Speech or with Noisy Speech". In: *ICASSP*. IEEE, 2019, pp. 1090–1094.

[28] De Luca, Carlo J. *Surface Electromyography: Detection and Recording*. Tech. rep. DelSys Incorporated, 2002, p. 10.

[29] Ghapanchizadeh, Hossein. "Review of Surface Electrode Placement for Recording Electromyography Signals". In: *Biomed Research* Special Section: Complex World of Neuroscience.Special Issue: S1-S7 (2017), p. 7.

[30] Rodriguez-Tapia, Bernabe; Soto, Israel; Martinez, Daniela M.; Arballo, Norma Candolfi. "Myoelectric Interfaces and Related Applications: Current State of EMG Signal Processing–A Systematic Review". In: *IEEE Access* 8 (2020), pp. 7792–7805. DOI: 10.1109/ACCESS.2019.2963881.

[31] Vojtech, Jennifer M.; Stepp, Cara E. "Electromyography". In: *Manual of Clinical Phonetics*. Ed. by Martin J. Ball. 1st ed. Routledge, Feb. 2021, pp. 248–263. DOI: 10.4324/9780429320903-20.

[32] Young, A. J.; Hargrove, L. J.; Kuiken, T. A. "The Effects of Electrode Size and Orientation on the Sensitivity of Myoelectric Pattern Recognition Systems to Electrode Shift". In: *IEEE Trans. Biomed. Eng.* 58.9 (Sept. 2011), pp. 2537–2544. DOI: 10.1109/TBME.2011.2159216.

[33] Maier-Hein, L.; Metze, F.; Schultz, T.; Waibel, A. "Session Independent Non-Audible Speech Recognition Using Surface Electromyography". In: *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* San Juan, Puerto Rico: IEEE, 2005, pp. 331–336. DOI: 10.1109/ASRU.2005.1566521.

[34] Jou, Szu-Chen; Schultz, Tanja; Walliczek, Matthias; Kraft, Florian; Waibel, Alex. "Towards Continuous Speech Recognition Using Surface Electromyography". In: (2006), p. 4. DOI: 10.21437/Interspeech.2006-212.

[35] Wand, Michael; Jou, Szu-Chen Stan; Toth, Arthur R; Schultz, Tanja. "Impact of Different Speaking Modes on EMG-Based Speech Recognition". In: *InterSpeech*. Brighton UK: ISCA, 2009, p. 5.

[36] Diener, Lorenz; Janke, Matthias; Schultz, Tanja. "Direct Conversion from Facial Myoelectric Signals to Speech Using Deep Neural Networks". In: *2015 International Joint Conference on Neural Networks (IJCNN)*. Killarney, Ireland: IEEE, July 2015, pp. 1–7. DOI: 10.1109/IJCNN.2015.7280404.

[37] Colby, Glen; Heaton, James T.; Gilmore, L. Donald; Sroka, Jason; Yunbin Deng; Cabrera, Joao; Roy, Serge; De Luca, Carlo J.; Meltzner, Geoffrey S. "Sensor Subset Selection for Surface Electromyograpy Based Speech Recognition". In: *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Taipei, Taiwan: IEEE, Apr. 2009, pp. 473–476. DOI: 10.1109/ICASSP.2009.4959623.

[38] Meltzner, Geoffrey S; Sroka, Jason; Heaton, James T; Gilmore, L Donald; Colby, Glen; Roy, Serge; Chen, Nancy; Luca, Carlo J De. "Speech Recognition for Vocalized and Subvocal Modes of Production Using Surface EMG Signals from the Neck and Face". In: *InterSpeech*. Brisbane Australia: ISCA, 2008, p. 4. DOI: 10.21437/Interspeech.2008-661.

[39] Meltzner, G. S.; Colby, G.; Yunbin Deng; Heaton, J. T. "Signal Acquisition and Processing Techniques for sEMG Based Silent Speech Recognition". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Boston, MA: IEEE, Aug. 2011, pp. 4848–4851. DOI: 10.1109/IEMBS.2011.6091201.

[40] Meltzner, Geoffrey S.; Heaton, James T.; Deng, Yunbin; De Luca, Gianluca; Roy, Serge H.; Kline, Joshua C. "Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy". In: *IEEE/ACM Trans. Audio Speech Lang. Process.* 25.12 (Dec. 2017), pp. 2386–2398. DOI: 10.1109/TASLP.2017.2740000.

[41] Soon, Mok Win; Anuar, Muhammad Ikmal Hanafi; Abidin, Mohamad Hafizat Zainal; Azaman, Ahmad Syukri; Noor, Norliza Mohd. "Speech Recognition Using Facial sEMG". In: *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. Kuching: IEEE, Sept. 2017, pp. 1–5. DOI: 10.1109/ICSIPA.2017.8120569.

[42] Ma, Siyuan; Jin, Dantong; Zhang, Ming; Zhang, Bixuan; Wang, You; Li, Guang; Yang, Meng. "Silent Speech Recognition Based

on Surface Electromyography". In: *2019 Chinese Automation Congress (CAC)*. Hangzhou, China: IEEE, Nov. 2019, pp. 4497–4501. DOI: 10.1109/CAC48633.2019.8996289.

[43] Morse, Michael S.; O'Brien, Edward M. "Research Summary of a Scheme to Ascertain the Availability of Speech Information in the Myoelectric Signals of Neck and Head Muscles Using Surface Electrodes". In: *Computers in Biology and Medicine* 16.6 (Jan. 1986), pp. 399–410. DOI: 10.1016/0010-4825(86)90064-8.

[44] Morse, M.S.; Day, S.H.; Trull, B.; Morse, H. "Use of Myoelectric Signals to Recognize Speech". In: *Images of the Twenty-First Century. Proceedings of the Annual International Engineering in Medicine and Biology Society*. Seattle, WA, USA: IEEE, 1989, pp. 1793–1794. DOI: 10.1109/IEMBS.1989.96459.

[45] Chan, A. D. C.; Englehart, K.; Hudgins, B.; Lovely, D. F. "Myo-Electric Signals to Augment Speech Recognition". In: *Medical & Biological Engineering & Computing* 39.4 (July 2001), pp. 500–504. DOI: 10.1007/BF02345373.

[46] Chan, A.D.C.; Englehart, K.; Hudgins, B.; Lovely, D.F. "Hidden Markov Model Classification of Myoelectric Signals in Speech". In: *IEEE Eng. Med. Biol. Mag.* 21.5 (Sept. 2002), pp. 143–146. DOI: 10.1109/MEMB.2002.1044184.

[47] Lee, Ki-Seung. "EMG-based Speech Recognition Using Hidden Markov Models with Global Control Variables". In: *IEEE Transactions on biomedical engineering* 55.3 (2008), pp. 930–940.

[48] Wang, Youhua; Tang, Tianyi; Xu, Yin; Bai, Yunzhao; Yin, Lang; Li, Guang; Zhang, Hongmiao; Liu, Huicong; Huang, YongAn. "All-Weather, Natural Silent Speech Recognition via Machine-Learning-Assisted Tattoo-like Electronics". In: *npj Flex Electron* 5.1 (Dec. 2021), p. 20. DOI: 10.1038/s41528-021-00119-7.

[49] Wand, Michael; Schultz, Tanja. "Session-Independent EMG-based Speech Recognition". In: *Biosignals*. 2011, pp. 295–300.

[50] Deng, Yunbin; Heaton, James T; Meltzner, Geoffrey S. “Towards a Practical Silent Speech Recognition System”. In: (2014), p. 6. DOI: 10.21437/Interspeech.2014-296.

[51] Wand, Michael; Schmidhuber, Jürgen. “Deep Neural Network Frontend for Continuous EMG-Based Speech Recognition”. In: *Interspeech*. 2016, pp. 3032–3036.

[52] Zhou, Quan; Jiang, Ning; Hudgins, Bernard. “Improved Phoneme-Based Myoelectric Speech Recognition”. In: *IEEE Transactions on Biomedical Engineering* 56.8 (2009), pp. 2016–2023.

[53] Lopez-Larraz, Eduardo; Mozos, Oscar M.; Antelis, Javier M.; Minguez, Javier. “Syllable-Based Speech Recognition Using EMG”. In: *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*. 2010, pp. 4699–4702. DOI: 10.1109/IEMBS.2010.5626426.

[54] Wand, Michael; Schultz, Tanja. “Analysis of Phone Confusion in EMG-based Speech Recognition”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Prague, Czech Republic: IEEE, May 2011, pp. 757–760. DOI: 10.1109/ICASSP.2011.5946514.

[55] Schultz, Tanja; Wand, Michael. “Modeling Coarticulation in EMG-based Continuous Speech Recognition”. In: *Speech Communication* 52.4 (Apr. 2010), pp. 341–353. DOI: 10.1016/j.specom.2009.12.002.

[56] Toth, Arthur R; Wand, Michael; Schultz, Tanja. “Synthesizing Speech from Electromyography Using Voice Transformation Techniques”. In: *Tenth Annual Conference of the International Speech Communication Association*. 2009.

[57] Nakamura, Keigo; Janke, Matthias; Wand, Michael; Schultz, Tanja. “Estimation of Fundamental Frequency from Surface Electromyographic Data: EMG-to-F0”. In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. May 2011, pp. 573–576. DOI: 10.1109/ICASSP.2011.5946468.

[58] Zahner, Marlene; Janke, Matthias; Wand, Michael; Schultz, Tanja. "Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach". In: *Interspeech 2014*. ISCA, Sept. 2014, pp. 1184–1188. DOI: 10.21437/Interspeech.2014-300.

[59] Diener, Lorenz; Bredehoeft, Sebastian; Schultz, Tanja. "A Comparison of EMG-to-Speech Conversion for Isolated and Continuous Speech". In: *Speech Communication; 13th ITG-Symposium*. 2018, pp. 1–5.

[60] Gaddy, David; Klein, Dan. "Digital Voicing of Silent Speech". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Oct. 2020, pp. 5521–5530. DOI: 10.48550/arXiv.2010.02960.

[61] Gaddy, David; Klein, Dan. "An Improved Model for Voicing Silent Speech". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Ed. by Chengqing Zong; Fei Xia; Wenjie Li; Roberto Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 175–181. DOI: 10.18653/v1/2021.acl-short.23.

[62] Gaddy, David. "Voicing Silent Speech". PhD thesis. Berkeley: University of California, Nov. 2022.

[63] Zhang, Yakun; Cai, Huihui; Wu, Jinghan; Xie, Liang; Xu, Minpeng; Ming, Dong; Yan, Ye; Yin, Erwei. "EMG-Based Cross-Subject Silent Speech Recognition Using Conditional Domain Adversarial Network". In: *IEEE Trans. Cogn. Dev. Syst.* 15.4 (Dec. 2023), pp. 2282–2290. DOI: 10.1109/TCDS.2023.3316701.

[64] Mostafa, S. S.; Awal, M. A.; Ahmad, M.; Rashid, M. A. "Voiceless Bangla Vowel Recognition Using sEMG Signal". In: *SpringerPlus* 5.1 (Dec. 2016), p. 1522. DOI: 10.1186/s40064-016-3170-9.

[65] Li, Huiyan; Lin, Haohong; Wang, You; Wang, Hengyang; Zhang, Ming; Gao, Han; Ai, Qing; Luo, Zhiyuan; Li, Guang. "Sequence-to-Sequence Voice Reconstruction for Silent Speech in a Tonal Language". In: *Brain Sciences* 12.7 (June 2022), p. 818. DOI: 10.3390/brainsci12070818.

[66] Deng, Zhihang; Zhang, Xu; Chen, Xi; Chen, Xiang; Chen, Xun; Yin, Erwei. "Silent Speech Recognition Based on Surface Electromyography Using a Few Electrode Sites Under the Guidance From High-Density Electrode Arrays". In: *IEEE Trans. Instrum. Meas.* 72 (2023), pp. 1–11. DOI: 10.1109/TIM.2023.3244849.

[67] Li, Wei; Yuan, Jianping; Zhang, Lu; Cui, Jie; Wang, Xiaodong; Li, Hua. "sEMG-based Technology for Silent Voice Recognition". In: *Computers in Biology and Medicine* 152 (Jan. 2023), p. 106336. DOI: 10.1016/j.compbiomed.2022.106336.

[68] Zhu, Mingxing; Zhang, Haoshi; Wang, Xiaochen; Wang, Xin; Yang, Zijian; Wang, Cheng; Samuel, Oluwarotimi Williams; Chen, Shixiong; Li, Guanglin. "Towards Optimizing Electrode Configurations for Silent Speech Recognition Based on High-Density Surface Electromyography". In: *J. Neural Eng.* 18.1 (Feb. 2021), p. 016005. DOI: 10.1088/1741-2552/abca14.

[69] Botelho, Catarina; Diener, Lorenz; Küster, Dennis; Scheck, Kevin; Amiriparian, Shahin; Schuller, Björn W.; Schultz, Tanja; Abad, Alberto; Trancoso, Isabel. "Toward Silent Paralinguistics: Speech-to-EMG — Retrieving Articulatory Muscle Activity from Speech". In: *Interspeech*. 2020, pp. 354–358. DOI: 10.21437/Interspeech.2020-2926.

[70] Diener, Lorenz; Amiriparian, Shahin; Botelho, Catarina; Scheck, Kevin; Küster, Dennis; Trancoso, Isabel; Schuller, Björn W.; Schultz, Tanja. "Towards Silent Paralinguistics: Deriving Speaking Mode and Speaker ID from Electromyographic Signals". In: *Interspeech*. 2020, pp. 3117–3121. DOI: 10.21437/Interspeech.2020-2848.

[71] Diener, Lorenz. "The Impact of Audible Feedback on EMG-to-Speech Conversion". PhD thesis. University of Bremen, 2021.

[72] Wu, Jinghan; Zhang, Yakun; Xie, Liang; Yan, Ye; Zhang, Xu; Liu, Shuang; An, Xingwei; Yin, Erwei; Ming, Dong. "A Novel Silent Speech Recognition Approach Based on Parallel Inception Convolutional Neural Network and Mel Frequency Spectral Coefficient". In: *Frontiers in Neurorobotics* 16 (2022).

[73] Meltzner, Geoffrey S.; Heaton, James T.; Deng, Yunbin; De Luca, Gianluca; Roy, Serge H.; Kline, Joshua C. "Development of sEMG Sensors and Algorithms for Silent Speech Recognition". In: *Journal of Neural Engineering* 15.046031 (2018). DOI: 10.1088/1741-2552/aac965.

[74] Wand, Michael; Schulte, Christopher; Janke, Matthias; Schultz, Tanja. "Array-Based Electromyographic Silent Speech Interface". In: *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing*. Barcelona, Spain: SciTePress - Science and and Technology Publications, 2013, pp. 89–96. DOI: 10.5220/0004252400890096.

[75] Zhu, Mingxing; Huang, Zhen; Wang, Xiaochen; Zhuang, Jiashuo; Zhang, Haoshi; Wang, Xin; Yang, Zijian; Lu, Lin; Shang, Peng; Zhao, Guoru; Chen, Shixiong; Li, Guanglin. "Contraction Patterns of Facial and Neck Muscles in Speaking Tasks Using High-Density Electromyography". In: *2019 13th International Conference on Sensing Technology (ICST)*. Sydney, Australia: IEEE, Dec. 2019, pp. 1–5. DOI: 10.1109/ICST46873.2019.9047731.

[76] Zhu, Mingxing; Wang, Xiaochen; Wang, Xin; Wang, Cheng; Yang, Zijian; Williams Samuel, Oluwarotimi; Chen, Shixiong; Li, Guanglin. "The Effects of Electrode Locations on Silent Speech Recognition Using High-Density sEMG". In: *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. Roma, Italy: IEEE, June 2020, pp. 345–348. DOI: 10.1109/MetroInd4.0IoT48571.2020.9138289.

[77] Wang, Xiaochen; Zhu, Mingxing; Cui, Han; Yang, Zijian; Wang, Xin; Zhang, Haoshi; Wang, Cheng; Deng, Hanjie; Chen, Shixiong; Li, Guanglin. "The Effects of Channel Number on Classification

Performance for sEMG-based Speech Recognition". In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Montreal, QC, Canada: IEEE, July 2020, pp. 3102–3105. DOI: 10.1109/EMBC44109.2020.9176260.

[78] Wang, Xiaochen; Zhu, Mingxing; Samuel, Oluwarotimi Williams; Yang, Zijian; Lu, Lin; Cai, Xingxing; Wang, Xin; Chen, Shixiong; Li, Guanglin. "A Pilot Study on the Performance of Time-Domain Features in Speech Recognition Based on High-Density sEMG". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. Mexico: IEEE, Nov. 2021, pp. 19–22. DOI: 10.1109/EMBC46164.2021.9630541.

[79] Reucher, H.; Rau, G.; Silny, J. "Spatial Filtering of Noninvasive Multielectrode EMG: Part I–Introduction to Measuring Technique and Applications". In: *IEEE Trans Biomed Eng* 34.2 (Feb. 1987), pp. 98–105. DOI: 10.1109/tbme.1987.326034.

[80] De Luca, C. J.; Merletti, R. "Surface Myoelectric Signal Cross-Talk among Muscles of the Leg". In: *Electroencephalogr Clin Neurophysiol* 69.6 (June 1988), pp. 568–575. DOI: 10.1016/0013-4694(88)90169-1.

[81] Mohr, Maurice; Schön, Tanja; von Tscharner, Vinzenz; Nigg, Benno M. "Intermuscular Coherence Between Surface EMG Signals Is Higher for Monopolar Compared to Bipolar Electrode Configurations". In: *Frontiers in Physiology* 9 (2018).

[82] Aubanel, Vincent; Lecumberri, Maria Luisa García; Cooke, Martin. "The Sharvard Corpus: A Phonemically-Balanced Spanish Sentence Resource for Audiology". In: *International Journal of Audiology* 53.9 (Sept. 2014), pp. 633–638. DOI: 10.3109/14992027.2014.907507.

[83] McAuliffe, Michael; Socolof, Michaela; Mihuc, Sarah; Wagner, Michael; Sonderegger, Morgan. "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi." In: *Interspeech*. Vol. 2017. 2017, pp. 498–502.

[84] Wand, Michael. *Advancing Electromyographic Continuous Speech Recognition: Signal Preprocessing and Modeling*. KIT Scientific Publishing, 2015. DOI: 10.5445/KSP/1000040667.

[85] Fisher, R. A. "The Use of Multiple Measurements In Taxonomic Problems". In: *Annals of Eugenics* 7.2 (Sept. 1936), pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

[86] Wand, Michael; Schultz, Tanja. "Pattern Learning with Deep Neural Networks in EMG-based Speech Recognition". In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Aug. 2014, pp. 4200–4203. DOI: 10.1109/EMBC.2014.6944550.

[87] Wells, J.C. "SAMPA Computer Readable Phonetic Alphabet". In: *Handbook of Standards and Resources for Spoken Language Systems*. Ed. by D. Gibbon; R. Moore; R. Winski. Berlin and New York: Mouton de Gruyter, 1997, Part IV, section B.

[88] Sainz, Iñaki; Erro, Daniel; Navas, Eva; Hernáez, Inma; Sanchez, Jon; Saratxaga, Ibon; Odriozola, Igor. "Versatile Speech Databases for High Quality Synthesis for Basque". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. May 2012.

[89] Wand, Michael; Janke, Matthias; Schultz, Tanja. "The EMG-UKA Corpus for Electromyographic Speech Processing". In: *Interspeech 2014*. ISCA, Sept. 2014, pp. 1593–1597. DOI: 10.21437/Interspeech.2014-379.

[90] Wand, Michael; Schultz, Tanja. "Towards Real-Life Application of EMG-based Speech Recognition by Using Unsupervised Adaptation". In: *Interspeech*. 2014, pp. 1189–1193. DOI: 10.21437/Interspeech.2014-301.

[91] Wand, Michael; Schultz, Tanja; Schmidhuber, Jürgen. "Domain-Adversarial Training for Session Independent EMG-based Speech Recognition". In: *Interspeech*. 2018, pp. 3167–3171. DOI: 10.21437/Interspeech.2018-2318.

[92] Proroković, Krsto; Wand, Michael; Schultz, Tanja; Schmidhuber, Jürgen. "Adaptation of an EMG-Based Speech Recognizer via Meta-Learning". In: *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. 2019, pp. 1–5. DOI: 10.1109/GlobalSIP45357.2019.8969231.

[93] Abdullah, Asif; Chemmangat, Krishnan. "A Computationally Efficient sEMG Based Silent Speech Interface Using Channel Reduction and Decision Tree Based Classification". In: *Procedia Computer Science* 171 (2020), pp. 120–129. DOI: 10.1016/j.procs.2020.04.013.

[94] Sharma, Manthan; Gaddam, Navaneetha; Umesh, Tejas; Murthy, Aditya; Ghosh, Prasanta Kumar. "A Comparative Study of Different EMG Features for Acoustics-to-EMG Mapping". In: *Interspeech*. 2021, pp. 616–620. DOI: 10.21437/Interspeech.2021-2240.

[95] Breiman, Leo. "Bagging Predictors". In: *Machine learning* 24.2 (1996), pp. 123–140.

[96] Nair, Vinod; Hinton, Geoffrey E. "Rectified Linear Units Improve Restricted Boltzmann Machines". In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 807–814.

[97] Bridle, John. "Training Stochastic Model Recognition Algorithms as Networks Can Lead to Maximum Mutual Information Estimation of Parameters". In: *Advances in neural information processing systems* 2 (1989).

[98] Kingma, Diederik P; Ba, Jimmy. "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[99] Wand, Michael; Janke, Matthias; Schultz, Tanja. "Tackling Speaking Mode Varieties in EMG-Based Speech Recognition". In: *IEEE Trans. Biomed. Eng.* 61.10 (Oct. 2014), pp. 2515–2526. DOI: 10.1109/TBME.2014.2319000.

[100] Janke, Matthias; Wand, Michael; Schultz, Tanja. "Impact of Lack of Acoustic Feedback in EMG-based Silent Speech Recognition". In: *Interspeech 2010*. ISCA, Sept. 2010, pp. 2686–2689. DOI: 10.21437/Interspeech.2010-712.

[101] Wand, Michael; Janke, Matthias; Schultz, Tanja. "Investigations on Speaking Mode Discrepancies in EMG-based Speech Recognition". In: *Interspeech 2011*. ISCA, Aug. 2011, pp. 601–604. DOI: 10.21437/Interspeech.2011-241.

[102] Janke, Matthias; Wand, Michael; Nakamura, Keigo; Schultz, Tanja. "Further Investigations on EMG-to-speech Conversion". In: *ICASSP*. 2012, pp. 365–368. DOI: 10.1109/ICASSP.2012.6287892.

[103] Raman, Sneha; Sarasola, Xabier; Navas, Eva; Hernaez, Inma. "Enrichment of Oesophageal Speech: Voice Conversion with Duration–Matched Synthetic Speech as Target". In: *Applied Sciences* 11.13 (Jan. 2021), p. 5940. DOI: 10.3390/app11135940.

[104] Wood, Simon N. *Generalized Additive Models: An Introduction with R, Second Edition*. 2nd ed. Boca Raton: Chapman and Hall/CRC, May 2017. DOI: 10.1201/9781315370279.

[105] Team, R Core. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022.

[106] van Rij, Jacolien; Wieling, Martijn; Baayen, R. Harald; van Rijn, Hedderik. *itsadug: Interpreting Time Series and Autocorrelated Data Using GAMMs*. 2022.

[107] Wieling, Martijn. "Analyzing Dynamic Phonetic Data Using Generalized Additive Mixed Modeling: A Tutorial Focusing on Articulatory Differences between L1 and L2 Speakers of English". In: *Journal of Phonetics* 70 (Sept. 2018), pp. 86–116. DOI: 10.1016/j.wocn.2018.03.002.

List of abbreviations

ABD anterior belly of the digastric. 24, 29, 30, 42, 67, 69, 107, 108, 110, 114, 122

ASR automatic speech recognition. 12

BIC Bayesian Information Criterion. 37

BUC buccinator. 24

C consonant. 32, 48

CER character error rate. 19

CNN convolutional neural network. 19

DAO depressor anguli oris. 24, 29, 30, 40, 42, 67, 69, 107, 108, 110, 114, 122

DC direct-current. 34

DLI depressor labii inferioris. 29, 30, 39, 40, 42, 69, 107, 108, 114, 122

DNN deep neural network. 18

DT decision trees. 37, 71

DTW dynamic time warping. 18, 46

EDF effective degrees of freedom. 109

EEG electroencephalogram. 12, 153

ELRA European Land Registry Association. 63

EMG electromyography. iii, , 4, 7, 12–15, 17, 19, 20, 24, 29, 33–35, 43, 45, 46, 51–53, 58, 60–63, 65, 68, 70, 72, 74, 80, 82, 83, 91, 93–97, 100, 102, 103, 105–112, 114, 116, 117, 121–123, 153–157, 168, 171

FRT frontalis. 30, 40

GAM generalized additive model. 105, 106, 109, 110, 117, 157

GMM Gaussian mixture model. 17, 18, 37, 39, 71, 98

HMM Hidden-Markov model. 16

Hz Hertz. 14, 51

IED inter-electrode distance. 31

IPA International Phonetic Alphabet. 48, 49, 96

LAO levator anguli oris. 24, 30, 42, 67

LDA linear discriminant analysis. 16, 36, 70, 83, 97

LLS levator labii superioris. 24, 29, 30, 42, 69, 107, 108, 110, 114, 122

MAS masseter. 29, 30, 42, 69, 107, 108, 115, 122

MAV mean absolute value. 15

MCD Mel-cepstral distortion. 17, 18

MFA Montreal forced aligner. 34, 67, 70, 81, 95

MFCCs Mel-frequency cepstral coefficients. 15, 17, 71, 72, 96

MLH mylohyoid. 24

MNT mentalis. 24

MUAP motor unit action potential. 13

mV millivolts. 14, 107, 108, 116

NN neural network. 37, 39, 41, 71, 83, 84, 94, 98

OBO orbicularis oris. 24, 30, 39, 40

PBD posterior belly of the digastric. 30, 39

PLT platysma. 24, 30, 67

PMA permanent magnet articulography. 12, 153

RIS risorius. 25, 29, 30, 42, 69, 107, 108, 110, 111, 115, 116, 122

RMS root mean square. 15, 58, 105, 106, 108–113, 116

SAMPA Speech Assessment Methods Phonetic Alphabet. 48, 49, 70, 81, 95, 125

SBO superior belly of the omohyoid. 30

SCM sternocleidomastoid. 24, 30, 39

SD standard deviation. 15, 72

SLH stylohyoid. 29, 30, 39, 42, 70, 107, 108, 111, 115, 116, 122

SSI silent speech interface. iii, 4, 7, 11, 12, 15, 24, 45, 62, 65, 66, 79, 80, 92–94, 103–106, 116, 117, 121–123, 153–157, 159, 168, 169

STFT Short-Term Fourier Transform. 15

STR sternothyroid. 30, 39

SVM support vector machine. 17, 19

TD time-domain. 15, 17, 34–36, 70, 82, 83, 96, 97, 154

TTS text-to-speech. 11, 12, 50

V vowel. 32, 48

WER word error rate. 16–18, 20, 121

ZYG zygomaticus major. 24, 29, 30, 42, 67, 69, 107, 108, 110, 115, 122

Summary

A silent speech interface (SSI) has the potential to restore the ability to speak for those who have lost their voice. This machine-learning-based interface translates biosignals from the speech production system other than speech itself into actual speech signals. These biosignals can be acquired from the brain, the tongue, or the muscles, with techniques called electroencephalogram (EEG), permanent magnet articulography (PMA), and electromyography (EMG) (respectively). Silent speech refers to the act of articulating as if a person were speaking, but without producing any sound. So, biosignals can correspond to either silent speech or audible speech. This thesis focuses on the research questions and challenges related to developing an SSI based on activity from the articulatory muscles (EMG signals). For example, first, it is essential to find the most optimal method of acquiring the signals, such as the selection of the EMG electrode locations. Then, it would be most helpful to know the impact of speaker variability, or the absence of information from the tongue and vocal cords, on the model's prediction performance. By answering these questions, this thesis aims to fill the research gap specifically related to EMG-based SSIs developed for Spanish alaryngeal speakers. An alaryngeal speaker has no larynx, which is a part of the speech production system that contains the vocal cords, an essential element of typical speech production. To help them communicate again is the ultimate goal of the research described in this thesis.

Part I of this thesis provides an elaborate background of all the topics related to EMG-based SSIs for alaryngeal speakers.

Chapter 1 contains a brief thesis introduction, providing a description of the research goal and questions of this thesis, and a thesis guide.

Chapter 2 explains the basis of speech production, alaryngeal speech, SSIs, and EMG. This chapter also contains a literature overview

of studies related to EMG-based SSIs. As shown in the chapter, there is a lack of research focused on the Spanish language and the target group of alaryngeal speakers. However, this is not the case for the English language and typical speakers. For example, during my PhD trajectory, a model was made available by other researchers with which fairly intelligible speech was converted from EMG signals of audible and silent speech produced by an English typical speaker.

Part II of this thesis presents a new database of Spanish speech and EMG signals, and the collection and validation method.

Chapter 3 describes a pilot study we performed to find the optimal acquisition setup. It aimed to find the best type, number, and locations of EMG electrodes. For the experiments described in this Chapter (and Chapters 5- 7), we did a phone (speech sound) classification task. First, we calculated a set of five frame-based features from the raw EMG signals known as time-domain (TD) features, which have been extensively used in related studies. Then, we automatically transcribed the speech signals, so that for each frame we also had a phone label. With the EMG features and the phone label as input information of a varying number of words or sentences (depending on the specific experiment), we trained the classifier. In the testing phase, we only provided the EMG features as input, so that the trained classifier had to predict the corresponding phone labels based on the relations between certain features and phone labels it learned during training. We measured its performance using the accuracy measure, which represents how often the predicted label was accurate (in percentage). In this chapter about the pilot study, we compared the accuracy values for two types of electrodes, but also for different locations all over the face and neck. Based on these comparisons, we selected a setup consisting of eight bipolar single-electrode pairs, where each electrode pair targets a specific muscle in the lower facial or upper neck area.

Chapter 4 presents the database, which contains signals belonging to words or sentences, spoken audible or silently. We collected over 22 hours of data in total from six typical speakers and three alaryngeal speakers. The amount of sessions differs per speaker, but all of them recorded a base set of sessions. This chapter also describes the complete acquisition procedure, for example, the 3D mask we used for each speaker to reduce session variability, and provides some data examples.

Chapter 5 describes a study we performed to ensure that our acquisition procedure was valid, and compares classification performance with signals from our database and that of an English reference database. When predicting the phones corresponding to parts of EMG signals using a simple classifier, the test accuracy of one long session from the reference database was 28.32%, considering 40 English phone classes. Repeating this experiment with nine sessions from our Spanish database with approximately the same duration, the average test accuracy we reached was 40.85%. Even though the Spanish language has 30 phone classes, which explains part of the higher accuracy, we were confident that we could continue acquiring data with the selected acquisition setup.

Part III of this thesis focuses on assessing the potential and limitations of developing an SSI with our newly acquired data.

Chapter 6 addresses a well-known limitation of SSI development, namely the impact of variability between different speakers or different data acquisition sessions from the same speaker. Due to this variability, there can be a strong dependency of the model on the data it was trained on. In this chapter, we describe a study that consisted of experiments in which the train and test data of the phone classifier were from the same speaker or session, or not. We started with data from the same session (which automatically means from the same speaker) and repeated this experiment for 16 sessions from six different speakers. This resulted in a maximum test accuracy of 50.54%. The data in this session consists of audible speech signals and their corresponding EMG signals. Then, we trained the classifier with data from other sessions of the same speaker, allowing us to increase the amount of training data. It appeared that when the classifier was trained on another session, the test accuracy drastically lowered to a maximum accuracy of 23.40%. However, there was a slight improvement when training data was added. With the addition of two more training sessions, the maximum test accuracy was 30.41%. The effect of session variability is concerning, but it is promising to know that increasing the amount of data can be a solution. When repeating the last experiment, but using training data from one speaker and testing data from another, the maximum test accuracy was reduced even further to 20.43%. It should be noted that the accuracy can vary depending on the specific train-test data combination, implying that some speakers and sessions are better matches than others. Either way, we found that speaker variability also has an impact on our data's potential, and should be considered in the final SSI development.

Chapter 7 aims to identify the effect of similarities between speech sounds on the ability to predict speech from the EMG signals of those sounds. In Spanish, there are seven groups of speech sounds, and each group has a particular manner of articulation. Within the plosives, minimal pairs exist where the only difference is realized by adding vibration with the vocal cords (called voicing), or not. These are [b-p], [d-t], and [ɡ-k], where the first in each pair is voiced and the second is unvoiced. In other cases, the only difference can be the position of the tongue, as in [r-ɾ]. Also within the groups of vowels, some are minimally different from each other, like [i-e]. Since EMG signals do not contain information related to the use of the tongue or the vocal cords, some confusion between sounds with minimal articulation difference is to be expected. Therefore, we performed a phone confusion analysis, which is the study described in this chapter. Again, we did phone classification experiments, but this time we looked with more detail at the phone predictions. The results showed that for some minimal pairs, the confusion was indeed higher than the accuracy. For example, the [u] was more often predicted inaccurately as [o] (38.68%) than accurately as [u] (37.57%). However, it is important to add that some of the confusion can be explained by a high correlation we found between the phone accuracy and label count, which resulted in the most common phones being predicted relatively more often as well. The main takeaway of this study it could be useful to not look at the general SSI performance but at the contribution of individual phonetic features, and add a linguistic component such as a language model.

Chapter 8 presents a statistical analysis of EMG signals to get a better understanding of muscle activity in different contexts. Using a non-linear regression method called generalized additive model (GAM), we compared the activity patterns of typical and alaryngeal speakers, audible and silent speech, and different words. First, we concluded that the method can be suitable for future research into the relationship between muscle use and specific sound production, based on the finding that a difference in muscle activity showed up at the part where two analyzed words differed phonetically as well. However, the most important finding was that when people speak silently, they show more activity in their muscles, which can be attributed to trying to compensate for the lack of auditive feedback. This means that an SSI trained on EMG signals from audible speech only might not be the best approach, and EMG signals from silent speech should be included in the training process.

Part IV is the final part of this thesis and consists of Chapter 9, which is the general discussion and conclusion.

To conclude, in the context of EMG-based SSI research, this thesis focuses on the design and development of a Spanish EMG-speech database and its collection and validation procedure, as well as analyses of the effect of speaker variability, minimal articulation differences of speech sounds, and speech mode. The findings can be used to develop and improve an SSI for Spanish alaryngeal speakers.

Resumen

Una interfaz de habla silenciosa (a la que nos referiremos como SSI, de su nombre en inglés, silent speech interface (SSI)) tiene el potencial de devolver la capacidad de hablar a quienes han perdido la voz. Estas interfaces, utilizan técnicas de aprendizaje automático para convertir bioseñales generadas por el apartado fonador en señales de voz. Estas bioseñales pueden generarse en diferentes puntos en el proceso de producción del habla, de forma que pueden obtenerse del cerebro (por electro-encefalografía u otras técnicas), o de los diferentes músculos que intervienen en la producción del habla como la lengua, los músculos de la cara y el cuello etc. El término 'Habla silenciosa' hace referencia al acto de articular como si se estuviera hablando, pero sin producir sonido. Así, las bioseñales pueden obtenerse tanto con habla audible como con habla silenciosa. Se han estudiado diversas técnicas para capturar las señales. De entre ellas, esta tesis está centrada en los retos asociados al desarrollo de interfaces silenciosas utilizando señales electromiográficas (EMG) capturadas desde los músculos articulatorios. El estudio de estas interfaces requiere enfrentarse a diferentes retos o preguntas de investigación. Por ejemplo, en primer lugar, será necesario encontrar el método óptimo para la adquisición de las señales: las selección de los electrodos y sus posiciones. Además, es necesario analizar las diferencias entre los diferentes hablantes, los diferentes modos de habla, o el efecto de la ausencia de información de las cuerdas vocales o del movimiento de la lengua. En esta tesis se analizan todos estos retos asociados a las interfaces silenciosas basadas en EMG y de forma específica para el español y personas sin laringe. Ayudar a estas personas en su proceso de comunciación es el objetivo último del trabajo que se describe en esta tesis.

La Parte I en esta tesis describe las bases y fundamentos de diferentes aspectos relativos a las interfaces silenciosas basadas en señales EMG.

El Capítulo 1 contiene una breve introducción a la tesis, descri-

biendo los objetivos de la investigación, así como una guía del documento.

El Capítulo 2 describe los fundamentos de la la producción del habla, el habla alaríngea, las interfaces de habla silenciosa y las señales EMG. Este capítulo también contiene un resumen bibliográfico de los estudios relativos a las interfaces de habla silenciosa basadas en EMG. Tal y como se describe en el capítulo, las investigaciones desarrolladas para la lengua española y en especial con el grupo de hablantes alaríngeos son escasas. Sin embargo, esto no es así para el inglés y para hablantes típico. Por ejemplo, durante mi trayectoria doctoral, otros investigadores desarrollaron un modelo con el que se obtenía habla inteligible a partir de señales EMG de un hablante típico de inglés.

La Parte II de esta tesis presenta una nueva base de datos de voz y señales EMG, describiendo la metodología seguida para la recopilación y validación de los datos.

En el Capítulo 3 se describe el estudio piloto llevado a cabo para encontrar la configuración de adquisición óptima. El objetivo era encontrar el mejor tipo de electrodos, así como su número y ubicación en la cara y cuello del hablante. En los experimentos descritos en este capítulo (y en los capítulos 5- 7) realizamos una tarea de clasificación de fonemas. En primer lugar, con las señales EMG procedentes de los electrodos se calculan un conjunto de características temporales, conocidas como *TD-features*, que han sido utilizadas ampliamente en la bibliografía. A continuación, se segmentan en fonemas automáticamente las señales de voz, de modo que a cada trama se le asocia una etiqueta de fonema. Así, con las características EMG y la etiqueta del fonema como información de entrada, obtenidos para un número variable de palabras o frases (dependiendo del experimento), se entrena el clasificador. En la fase de prueba, se proporcionan las características EMG a la entrada, de modo que el clasificador ya entrenado predice las etiquetas de los fonemas correspondientes. Para medir el rendimiento del sistema se utiliza el porcentaje de aciertos para cada fonema. En este capítulo dedicado al estudio piloto, se comparan los porcentajes de acierto para dos tipos de electrodos y para distintas ubicaciones de los electrodos en la cara y el cuello. Basándonos en estas comparaciones, se seleccionó una configuración consistente en ocho pares de electrodos individuales bipolares, en los que cada par de electrodos se ubica en un músculo específico de la parte inferior de la cara o de la parte superior del cuello. La base de datos descrita más adelante fue adquirida con esta configuración final.

El Capítulo 4 presenta la base de datos obtenida. Esta base de datos contiene señales EMG y señales de voz correspondientes a palabras o frases pronunciadas de forma audible (audio y EMG) o de forma silenciosa (únicamente EMG). Se recogieron aproximadamente 22 horas de datos de seis hablantes típicos y de tres alaríngeos. El número de sesiones difiere por hablante, pero todos ellos grabaron un conjunto común de frases y palabras. En este capítulo también se describe el procedimiento completo de adquisición, incluyendo por ejemplo el uso de la máscara 3D que utilizamos para cada hablante con el fin de reducir la variabilidad de las sesiones, y se ofrecen algunos ejemplos de datos.

El Capítulo 5 describe el estudio realizado para validar el procedimiento de adquisición. Para ello compara el comportamiento de un mismo clasificador cuando se entrena y prueba con señales de nuestra base de datos y con las de una base de datos (en inglés) que se toma como referencia. Al predecir los fonemas correspondientes a señales EMG utilizando un clasificador simple, el porcentaje de aciertos obtenido para una sesión larga de la base de datos de referencia fue del 28,32%, considerando 40 clases o fonemas del inglés. Repitiendo el experimento con aproximadamente la misma cantidad de datos de nuestra base de datos en español, con 30 clases o fonemas, el porcentaje de aciertos promedio alcanzado fue del 40,85%. Aunque el menor número de fonemas del español explica un parte de la mejora en el porcentaje de acierto, los resultados ofrecen una validación positiva de la configuración de adquisición.

La Parte III de esta tesis se centra en evaluar el potencial y las limitaciones del desarrollo de una interfaz de habla silenciosa con la base de datos recién adquirida.

En el Capítulo 6 se aborda una limitación bien conocida en la elaboración de una interfaz de habla silenciosa, a saber, el impacto de las variabilidad entre diferentes hablantes o diferentes sesiones de adquisición de datos del mismo hablante. Debido a esta variabilidad, puede producirse en el comportamiento del modelo una fuerte dependencia del conjunto de datos con el que ha sido entrenado. En este capítulo, describimos un conjunto de experimentos de clasificación de fonemas que analizan precisamente esta dependencia del modelo con el hablante y con la sesión. Empezamos con datos de una misma sesión (lo que automáticamente significa del mismo hablante) y repetimos el experimento para 16 sesiones de seis hablantes distintos. La sesión denominada 001-004 (sesión 004 del hablante 001) obtuvo el mejor porcentaje de acierto (50,54%). Los datos de esta sesión consisten en señales de voz audibles y sus correspondientes señales EMG. A continuación, entrenamos el clasificador con datos de otras sesiones del mismo locutor (001), aumentando así la cantidad de datos de entrenamiento. Cuando el clasificador se entrenó con otra sesión (001-001), la precisión de la prueba disminuyó drásticamente hasta el 23.40%. Sin embargo, hubo una ligera mejora al ir aumentando los datos de entrenamiento y al añadir dos sesiones más los resultados subieron al 30.41%. Es decir, que el efecto de la variabilidad de las sesiones puede compensarse aumentando la cantidad de datos de entrenamiento. Al repetir el último experimento, pero sustituyendo los datos de entrenamiento por los de otro hablante, el porcentaje de aciertos se redujo aún más hasta el 20,43%. Sin embargo, hay que señalar que este porcetaje puede variar mucho en función de la combinación específica de datos de entrenamiento y prueba, lo que implica que algunos locutores y sesiones se comportan mejor que otros. En cualquier caso, hemos observado que la variabilidad de los hablantes también influye en el potencial de nuestros datos, por lo que debería tenerse en cuenta en el desarrollo final.

El Capítulo 7 tiene como objetivo identificar el efecto que tiene la similitud entre ciertos sonidos sobre la capacidad de predicción de dichos sonidos a partir de las señales EMG. En español hay siete grupos de fonemas según el modo de articulación. Dentro del grupo de las oclusivas encontramos pares mínimos en los que la única diferencia consiste en la vibración (o no vibración) de las cuerdas vocales (lo que se denomina sonoridad). Se trata de [b-p], [d-t] y [g-k], en donde el primero sonido de cada par es sonoro y la segundo es sordo. En otros casos, la única diferencia puede ser la posición de la lengua, como en [r-ɾ]. También dentro del grupo de las vocales encontramos pares constrastivos, como [i-e]. Dado que las señales EMG no contienen información directamente relacionada con el empleo de la lengua o de las cuerdas vocales, es de esperar cierta confusión entre estas parejas de sonidos con diferencias mínimas de articulación. Por ello se ha realizado un análisis de confusión entre fonemas, que es el estudio descrito en este capítulo. Una vez más, hicimos experimentos de clasificación de sonidos, esta vez analizando con más detalle las predicciones. Los resultados mostraron que para algunos pares mínimos la confusión era efectivamente superior al acierto. Por ejemplo, [u] se predijo con mayor frecuencia de forma inexacta como [o] (38,68%) que de forma correcta como [u] (37,57%). Sin embargo, es importante añadir que parte de la confusión puede explicarse por la alta correlación que encontramos entre el porcentaje de acierto del fonema y el recuento de etiquetas, lo cual provoca que los fonemas más comunes también se predijeran con relativa mayor frecuencia. La principal conclusión de este estudio es que podría ser útil tener en cuenta la contribución de las características fonéticas individuales y añadir un componente lingüístico, como un modelo de lenguaje.

El Capítulo 8 presenta un análisis estadístico de las señales EMG para comprender mejor la actividad muscular en diferentes contextos. Utilizando un método de regresión no lineal denominado Modelo aditivo generalizado (GAM) se comparan los patrones de actividad de hablantes típicos y hablantes alaríngeos tanto para habla audible como para habla silenciosa, y para diferentes palabras. En primer lugar, concluimos que el método puede ser adecuado para futuras investigaciones sobre la relación entre el uso de los músculos y la producción de sonidos específicos, basándonos en el hallazgo de que ciertas diferencias en la actividad muscular aparecía también en la parte en la que dos palabras analizadas diferían fonéticamente. Sin embargo, el hallazgo más importante fue que cuando las personas hablan en silencio muestran más actividad en sus músculos, lo que puede atribuirse a que intentan compensar la falta de retroalimentación auditiva. Esto significa que un entrenamiento basado únicamente en señales de habla audible podría no ser el mejor enfoque, y que las señales de habla silenciosa deberían incluirse en el proceso de entrenamiento.

La Parte IV es la parte final de esta tesis y consta de el Capítulo 9, con una discusión general y las conclusiones de la tesis.

Para concluir, en el contexto de la investigación en interfaces de habla silenciosas basadas en EMG, esta tesis se centra en el diseño y desarrollo de una base de datos en español y en el procedimiento de recopilación y validación, así como en el análisis del efecto de la variabilidad del hablante, las diferencias mínimas de articulación de los sonidos del habla y el modo de habla. Los resultados de la tesis pueden ser útiles para desarrollar y mejorar una SSI para hablantes alaríngeos en español.

Samenvatting

Dit proefschrift vat het werk samen van een promotieonderzoek dat is uitgevoerd als onderdeel van het ReSSInt project. Dit project heeft als doel om, met behulp van technologie, Spanjaarden zonder stembanden weer een stem te geven. Van deze mensen zijn de stembanden operatief verwijderd, waardoor ze niet meer kunnen spreken zoals voorheen. Hierdoor zijn ze toegewezen op alternatieve methoden om te kunnen communiceren.

Een stille-spraakinterface (SSI) is een alternatieve communicatiemethode gebaseerd op een technologie waarbij spraak kan worden herkend zonder dat er daadwerkelijk geluid wordt geproduceerd. Terwijl de gebruiker articuleert zonder geluid (oftewel stil spreekt), zet de SSI spraak-gerelateerde informatie uit de tong, gezichtsspieren, lippen of hersenen, om naar spraak. Om bijvoorbeeld informatie uit de gezichtsspieren te halen wordt gebruik gemaakt van een methode genaamd elektromyografie (EMG). Dit is een techniek waarbij, doormiddel van elektroden op de huid, de elektrische activiteit van spieren wordt gemeten wanneer ze samentrekken.

Het onderwerp van dit proefschrift is het onderzoeken van de mogelijkheden om een SSI te ontwikkelen gebaseerd op EMG, oftewel een systeem dat spieractiviteit rondom de mond vertaalt naar spraak. In voorgaande jaren is er al veel onderzoek gedaan naar dit onderwerp, maar dit was voornamelijk gericht op de Engelse taal. Dit proefschrift presenteert de eerste onderzoeken naar EMG-SSI's voor Spaanse mensen zonder stembanden.

Een SSI wordt ontwikkeld met behulp van machinaal leren. Dit houdt in dat een computermodel patronen leert herkennen in data (gegevens) en aan de hand van de geleerde patronen voorspellingen kan doen. In dit geval bestaat de data uit spraaksignalen (audio) en spiersignalen (EMG). Omdat iemand zonder stembanden geen geluid kan produceren, wordt data gebruikt van zowel mensen met als mensen zonder stembanden. In de volgende stappen wordt uitgelegd hoe het computermodel precies wordt ontwikkelt:

1. Data verzamelen: Spraaksignalen en spiersignalen worden opgenomen van mensen met stembanden. Van mensen zonder stembanden worden alleen spiersignalen opgenomen, terwijl ze stil spreken.

2. Trainen van het model: Het computermodel wordt gevoed met spiersignalen en corresponderende spraaksignalen, en leert de relatie tussen deze signalen. Dit gebeurt door middel van algoritmen die patronen kunnen herkennen.

3. Testen van het model: Na het trainen wordt het model getest met spiersignalen die niet in het trainingsproces gebruikt zijn. Het model voorspelt en genereert (gesproken) tekst op basis van deze spiersignalen met behulp van de patronen die het heeft geleerd. De accuraatheid van deze voorspellingen toont aan hoe goed het model onbekende spiersignalen kan omzetten naar spraak. Dit kunnen dus ook de spiersignalen van mensen zonder stembanden zijn.

Om een SSI te ontwikkelen dat echt gebruikt kan worden, zullen eerst de volgende vragen beantwoord moeten worden, welke de basis vormen van dit proefschrift.

- Welke spieren van het gezicht en de nek zijn betrokken bij spraakproductie?

- Wat is het effect van variatie in EMG-signalen tussen verschillende sprekers, en verschillende opname-sessies van dezelfde spreker?

- Wat is het effect van het ontbreken van informatie van de stembanden en tong, twee belangrijke elementen van spraakproductie?

- Hoe verhoudt de articulatorische spieractiviteit van hoorbare spraak zich tot die van stille spraak, en die van mensen met en zonder stembanden in het geval van stille spraak?

Verschillende onderzoeken zijn uitgevoerd om deze vragen te kunnen beantwoorden. Het proefschrift bestaat daarom ook uit meerdere hoofdstukken. Hieronder volgt een beschrijving van ieder hoofdstuk.

Hoofdstuk 2 biedt achtergrondinformatie over alle relevante onderwerpen die in dit proefschrift aan bod komen, zoals hoe het spraakproductiesysteem werkt, wat er verandert na verwijdering van de stembanden, hoe een SSI precies werkt, een beschrijving van de EMG techniek, en welke resultaten andere studies in dit onderzoeksgebied hebben laten zien.

Hoofdstuk 3 beschrijft een kort vooronderzoek waarin we hebben onderzocht waar we de EMG elektroden het beste konden plaatsen om zoveel mogelijk relevantie informatie te verkrijgen. We hebben gekozen voor acht spieren in de onderkant van het gezicht en een deel van de hals. Met de signalen van deze acht spieren (van de veertien in totaal) kon een simpel computermodel het beste voorspellen welke spraak erbij hoorde. Het model kwam niet in de buurt van een goedwerkend SSI, maar voldeed wel om verschillende spieren te vergelijken.

Hoofdstuk 4 beschrijft de complete dataverzameling. We hebben ruim 22 uur aan signalen verzameld van zes sprekers met en drie sprekers zonder stembanden. De sprekers met stembanden hebben we zowel met geluid als zonder geluid woorden en zinnen laten articuleren, en de sprekers zonder stembanden (uiteraard) alleen zonder geluid. Tegelijkertijd meetten we de spieractiviteit van die acht spieren.

Hoofdstuk 5 had als doel het verifiëren van de dataverzamelmethode. Hiervoor vergeleken we onze signalen met die van een al bestaande verzameling voor het Engels en het bleek dat we uit onze signalen veel meer informatie konden halen, waardoor we door zijn gegaan met verzamelen.

Hoofdstuk 6 beschrijft een onderzoek waarin we een model hebben laten trainen en testen met data van dezelfde of andere spekers of opnamesessies van dezelfde spreker. Hiermee wilden we analyseren in hoeverre variatie in signalen invloed heeft op de prestatie van het computermodel. Het bleek dat de accuraatheid van de voorspellingen minder werden zodra een model was getraind met signalen van een sessie en getest met signalen van een andere sessie. Hetzelfde effect was zichtbaar wanneer er signalen van verschillende sprkers werden gebruikt. Echter, het bleek wel dat als er meer data werd toegevoegd waarmee het model kon trainen, de voorspellingen tijdens de testfase stap voor stap beter werden.

Hoofdstuk 7 beschrijft hoe we onderzochten of het computermodel de juiste van twee klanken weet te voorspellen als deze qua articulatie veel op elkaar lijken. Een voorbeeld hiervan zijn de [g] en [k], die enkel verschillen in het trillen van de stembanden ([g]) of niet ([k]). Het blijkt dat in dergeljke gevallen het model inderdaad meer moeite heeft. Een manier om dit probleem aan te pakken is om bijvoorbeeld extra contextinformatie aan te bieden, zodat het model op basis van kansberekening een meer ondersteunde voorspelling kan doen. Als het model tijdens het trainen geleerd heeft dat een [g] vaker voorafgegaan wordt door een klinker, en een [k] door een medeklinker, wordt de voorspelling makkelijker.

Hoofdstuk 8 beschrijft een analyse van de EMG signalen met behulp van een statistische methode. We wilden namelijk graag weten hoe we konden onderzoeken welke spieren bijdragen aan de productie van specifieke (combinaties van) klanken. Dit onderzoek heb ik uitgevoerd bij het Speech Lab in Groningen, waar ze gespecialiseerd zijn in deze methode. Samengevat hebben we gekeken naar de gemiddelde activatiepatronen van de signalen in verschillende contexten. Zo hebben we bijvoorbeeld gevonden dat, zoals verwacht, het patroon anders is in het begin van twee woorden die alleen in de eerste helft van elkaar verschillen (*noche* en *leche*). Ook vonden we dat er over het algemeen meer activatie is van de spieren als een woord zonder geluid wordt uitgesproken dan met. Wij hadden hiervoor als verklaring bedacht dat als er geen geluid is, mensen onbewust beter gaan articuleren om dat te proberen te compenseren. Over het algemeen concludeerden we dat deze methode geschikt kan zijn om de spierbewegingen van specifieke klanken preciezer in kaart te brengen.

Samengevat richt dit proefschrift zich, in de context van EMG-SSIs, op het ontwerp en de ontwikkeling van een Spaanse EMG-spraakdatabase en de verzamel- en validatieprocedure ervan, evenals op analyses van het effect van sprekervariabiliteit, minimale articulatieverschillen van spraakklanken, en spraakmodus. De bevindingen kunnen worden gebruikt om een SSI voor Spaanse alaryngeale sprekers te ontwikkelen en verbeteren.

Contributions

Listed below are the main contributions of this doctoral thesis to research in EMG-based SSIs.

- A database of EMG and (silent) speech signals from typical and alaryngeal Spanish speakers, called ReSSInt-EMG.

- A literature review of EMG electrode setups used in previous studies about EMG-based SSIs.

- A pilot study to identify the most useful muscles to extract speech-related information from.

- An effect analysis of two known challenges in this research area: variation in EMG signals between speakers and sessions, and the absence of information from two important speech production elements, namely the vocal cords and the tongue.

- A statistical analysis of EMG activity patterns in different contexts, which can be used as a method to map muscle activity and linguistic units with more detail.

The following works have been peer-reviewed and published:

- **Salomons, Inge**; Del Blanco, Eder; Navas, Eva; Hernáez, Inma; De Zuazo, Xabier. "Frame-Based Phone Classification Using EMG Signals". In: *Applied Sciences* 13.13 (June 2023), p. 7746. DOI: 10.3390/app13137746.

- **Salomons, Inge**; Del Blanco, Eder; Navas, Eva; Hernáez, Inma. "Spanish Phone Confusion Analysis for EMG-Based Silent Speech Interfaces". In: *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1179–1183. DOI: 10.21437/Interspeech.2023-1881.

- Del Blanco, Eder; **Salomons, Inge**; Navas, Eva; Hernáez, Inma. "Phone Classification Using Electromyographic Signals". In: *IberSPEECH 2022*. ISCA, Nov. 2022, pp. 31–35.
  DOI: 10.21437/IberSPEECH.2022-7.

The following works have been peer-reviewed and accepted for publication:

- **Salomons, Inge**; Hernáez, Inma; Navas, Eva; Wieling, Martijn. "Analyzing Speech Muscle Activity Using a Generalized Additive Model". Submitted to IberSPEECH 2024.

- Del Blanco, Eder; **Salomons, Inge**; García, Víctor; Navas, Eva; Hernáez, Inma. "Comparative Analysis of Mono-speaker and Multi-speaker Models for EMG-to-Speech Conversion". Submitted to IberSPEECH 2024.

The following work has been peer-reviewed and accepted for presentation:

- **Salomons, Inge**; Hernáez, Inma; Navas, Eva; Wieling, Martijn. "Mapping Speech and Facial Muscles: Using Generalized Additive Modeling to Understand Speech Production through Electromyography". In: *International Seminar on Speech Production (ISSP)*. 2024.