A NOVEL WIRELESS TONGUE TRACKING SYSTEM FOR SPEECH APPLICATIONS

A Dissertation
Presented to
The Academic Faculty

By

Nordine Sebkhi

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical & Computer Engineering

Georgia Institute of Technology

December 2019

A NOVEL WIRELESS TONGUE TRACKING SYSTEM FOR SPEECH APPLICATIONS

Approved by:

Dr. Omer T. Inan, Advisor
School of Electrical & Computer Engineering
Georgia Institute of Technology

Dr. David V. Anderson
School of Electrical & Computer Engineering
Georgia Institute of Technology

Dr. Nina M. Santus
Department of Communication Sciences & Special Education
University of Georgia

Dr. Benjamin D.B. Klein
School of Electrical & Computer Engineering
Georgia Institute of Technology

Dr. Pamela T. Bhatti
School of Electrical & Computer Engineering
Georgia Institute of Technology

Date Approved: September 31, 2019

ACKNOWLEDGEMENTS

First and foremost, I would like to dedicate this work to my parents. Although they have never been to school, they understood the importance of a good education for the success of their children. Rewarding their effort and sacrifices with a Ph.D degree from one of the best universities in the world has been a motivation during the toughest moments in this journey.

Secondly, I would like to acknowledge my colleagues: Dr. Nazmus Sahadat that has been my closest friend and collaborator, Dr. Sarah Ostadabbas that has been a wonderful mentor, and all the members of the Inan Research Lab for their warm welcome. Furthermore, I would like to thank Dr. Omer Inan not only for being my academic advisor but also for his invaluable mentoring and support. Also, I would like to express my gratitude to the following professors for accepting to be part of my thesis committee: Dr. David Anderson, Dr. Nina Santus, Dr. Benjamin Klein, and Dr. Pamela Bhatti.

It would be very remiss of me not to thank Dr. Jun Wang for giving me the chance to continue my research post-graduation. Dr. Jun Wang is not only believing in me but also on the potential of the system developed in this doctoral thesis.

Also, I would like to mention the Opportunity Research Scholars (ORS) program in which I was a PhD mentor to ECE undergraduate students. Particularly, I would like to thank Julie Ridings for her support, as well as Arpan Bhavsar and Shayan Siahpoushan for their great work on this project as undergraduate scholars.

Finally, I would like to conclude this acknowledgment by mentioning my business mentors that instilled in me their passion for entrepreneurship: Dr. Bob Gemmell and Richard DiMonda from the TI:GER program (Scheller College of Business, Georgia Tech) for giving me a solid foundation in entrepreneurship, and Dr. Paul Lopez and Tim Oltzer for their coaching that allowed me to win the inaugural Global Pitchfest competition organized by TiE Silicon Valley.

TABLE OF CONTENTS

LIST OF FIGURES

SUMMARY

The technological innovation of MagTrack is its ability to track the position of a magnetic tracer wirelessly, and with an accuracy of less than 2 mm, by implementing a novel permanent magnet localization method. Combining its wireless and high-accuracy tracking capability with the millimetric dimensions of the tracer, MagTrack is very well-suited for tongue tracking. Compared to commercially available tongue tracking systems, MagTrack has many of their advantages but without most of their shortcomings. For instance, MagTrack can be embedded in a headset to be portable, is affordable by using off-the-shelf components and custom-designed parts that are 3D-printed in our lab, and the tongue can be tracked wirelessly, thus enabling it to move without any hindrance.

Having access to tongue motion is important for many applications. Chief among them is the treatment of speech sound disorders in which the errant placement of the tongue is responsible for the decreased speech intelligibility. A preliminary study was conducted to assess the feasibility of MagTrack to be used in gamified visual feedback that would help patients better practice correct placement of their tongue by providing objective and quantitative measures of speech performance based on the error between their tongue placement and carefully-selected visual targets.

Another promising application for MagTrack is as the input of a silent speech interface that can recognize speech solely from tongue motion to generate synthesized voice for people that cannot produce sounds. Pilot human studies were conducted to obtain a preliminary assessment of speech recognition accuracy and the quality of synthesized speech from tongue motion recorded on hundreds of utterances by MagTrack.

Finally, since MagTrack is an improved version of the Tongue Drive System, it has the potential to be a practical assistive technology to help people with quadriplegia to regain autonomy for some tasks of their daily life such as driving their powered wheelchair, or even control their computer and mobile devices without the help of their caregivers.

CHAPTER 1
INTRODUCTION

1.1 Motivation

The ability to communicate with others is crucial for an individual's quality of life because it is an important component in everyday activities such as working, learning, and socializing. Chief among all forms of communication is speech which is an integral part of daily life. The mechanisms that are responsible for producing speech are not only complicated but are also still being studied today because the underlying processes are not fully understood. At a high level, speech sounds are originated from the air exhaled from the lungs which is first modulated by the vocal folds and subsequently by the articulators. Specifically, the motion of the articulators is responsible for generating the many phonemes that are part of a specific language. Thus, the articulators are essential components for our ability to be understood by others. Among all the articulators (tongue, lips and jaw), the tongue is the most important since it plays a key role in the production of most phonemes [1]. In addition to being important for speech, the tongue has also many unique characteristics such as not being prone to fatigue even after being used for a long time, and being able to be controlled with a high degree of accuracy. These unique characteristics enable the tongue to be used in many interesting applications. MagTrack, a portmanteau of ”Magnetic” and ”Tracking”, is a tongue tracking system that is being developed in our lab and is the object of this thesis. Figure 1.1 provides an overview of the system and a subset of the applications for which MagTrack can be used. The next chapters will describe each component of the system in more details, and will show how MagTrack has been used in some applications. Specifically, the focus of this work is on speech-related applications for which tracking the motion of the tongue is of utmost importance. First, the main driver for the

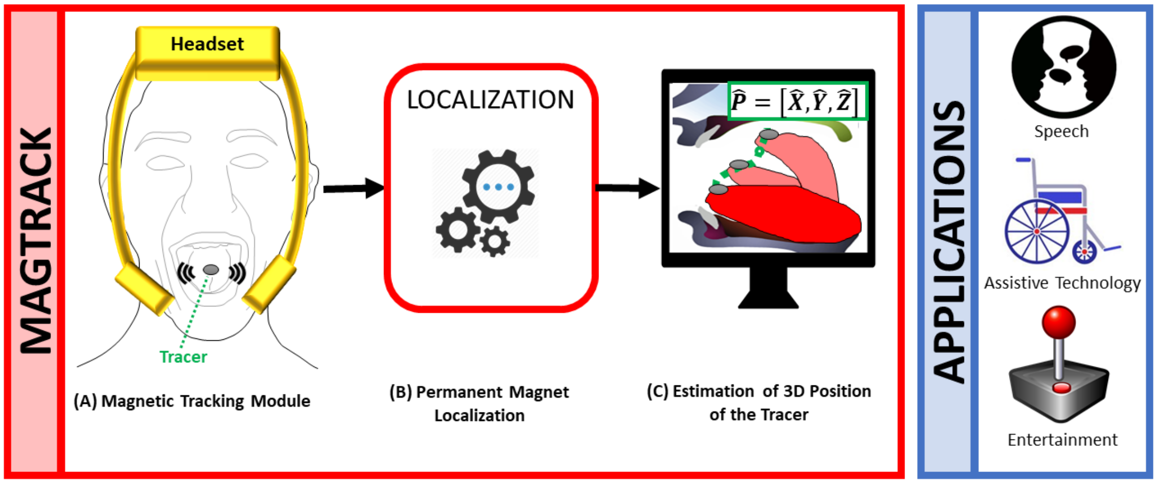

MAGTRACK
Headset
Tracer
(A) Magnetic Tracking Module
LOCALIZATION
(B) Permanent Magnet Localization
$\hat{P} = [\hat{X}, \hat{Y}, \hat{Z}]$
(C) Estimation of 3D Position of the Tracer
APPLICATIONS
Speech
Assistive Technology
Entertainment

Figure 1.1: Overview of MagTrack composed of (A) a magnetic tracking module embedded in a headset that measures the magnetic field generated by a tracer attached on the tongue. The magnetic readings are fed into (B) a permanent magnet localization algorithm that generates (C) an estimation of the position of the tracer in real-time. The captured tongue motion can then be used in many applications related to speech, assistive technology, and entertainment.

development of MagTrack has been to help people with speech sound disorders (SSD) to improve their intelligibility. During language acquisition in childhood, proper articulatory motion is learned and retained throughout one's life. However, as a result of developmental deficiencies or brain damage [2, 3], SSD impede the proper execution of articulatory motion which can render speech incomprehensible. In an attempt to regain a satisfactory level of intelligibility, people suffering from SSD undergo speech therapy under the care of a speech-language pathologist (SLP). During therapy sessions, access to articulatory motion is crucial not only for the SLP that must identify the root causes of errant speech to derive a treatment plan but also for the patient to understand and correct their own faulty motion through practice. As we already mentioned, the tongue is the most important articulator, consequently, having access to its motion is a critical aspect of the treatment of SSD. However, tongue motion is also the most challenging to observe since, unlike the lip and jaw that can be seen, it is hidden from sight in the oral cavity. This technical challenge was the motivation for the development of MagTrack and will be the topic of discussion in chapter 4.

The treatment of SSD is not the only speech-related application that MagTrack can be used for. Silent speech interface (SSI) is a new type of assistive technology that recognizes speech from articulatory motion and generates a synthesized voice for people that are unable to produce speech sounds, mainly due to the removal of the larynx as a consequence of a laryngeal cancer. Currently, the only remediation available to these individuals are tracheoesophageal and esophageal speech, which are difficult to master, and electrolarynx which generates a robotic-sounding voice that is difficult to understand and can result in social stigma. As an alternative, SSI is designed to generate a synthesized speech in quasi real-time with a more natural-sounding voice that is usually close to the individual's original voice. MagTrack is being used as the tracking system that provides the articulatory motion from the tongue, and our preliminary results will be discussed in chapter 5.

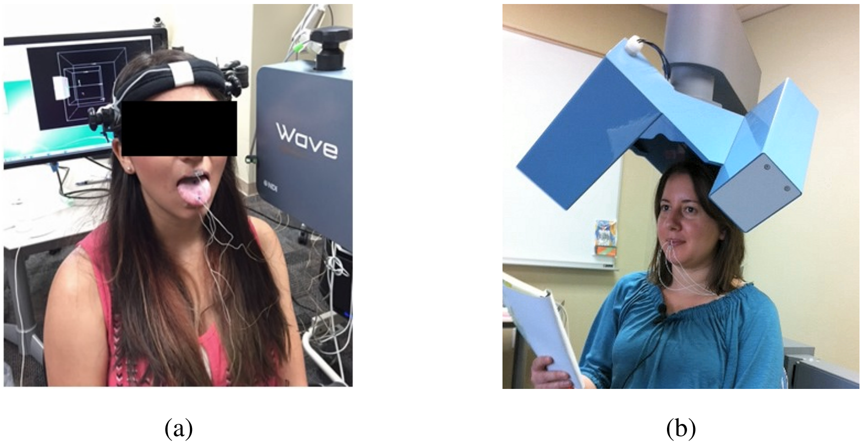
As an example of a non-speech application, chapter 5 will also describe the use of MagTrack to help people with quadriplegia (i.e. complete paralysis of the four limbs) to leverage the tongue as a joystick to be more autonomous for some tasks of daily living, which include driving their powered wheelchairs [4] and emulating a mouse to control a computer [5]. In entertainment, some applications that will not be discussed in this thesis but could benefit from MagTrack include playing video games with the tongue as a controller, being an added input controller for augmented and/or virtual reality devices, improving the quality of speech rendering in movie animation, and even facilitating 3D design by implementing a permanent magnet localization concept to a new type of wireless stylus. These are just scratching the surface of the possibilities that tracking the tongue with MagTrack can offer.

1.2 State-of-the-art

The next chapter will provide more detailed description on how MagTrack works. But, to better understand the innovation that MagTrack provides, this section provides an overview of the tongue tracking systems that are currently available on the market. Our discussion will be restricted to the systems that are mostly used in this field: the electromagnetic articulograph (EMA), the electropalatograph (EPG), and the ultrasound tongue imaging (UTI).

1.2.1 Electromagnetic Articulograph

EMA is the gold standard for tongue tracking because it is capable of estimating the 3D position and 2D orientation of many tracers simultaneously. The tracers are glued on the articulators (Figure 1.2a) and are made of wired receiver coils in which a current is induced by the magnetic field generated by an array of external transmitter coils. In [6], a mathematical model relates the induced current in the receiver coil to its relative position (3D) and orientation (2D), commonly referred as its 5D state, to a transmitter coil. As a result, EMA is capable of tracking up to 24 tracers simultaneously because the coils can measure the induced current independently from each other. Indeed, each receiver coil is connected to a controller and its signal is processed by first separating each magnetic source (i.e. transmitter coil) in the frequency domain since each source has a different carrier frequency, then estimating the relative 5D state of the receiver coil to each source using a voltage-to-distance function [7], and finally using triangulation to derive the tracer's 5D state in the global frame of reference [6]. This estimation of the 5D state of a tracer is executed at a sampling rate between 100 Hz and 400 Hz depending on the EMA manufacturer and model. Figure 1.2 shows the two EMA models that are mostly used by researchers: AG501 by Carstens (Figure 1.2b) and Wave by NDI (Figure 1.2a). The tracking accuracy of these EMA systems have been reported in [8, 9, 10] and these studies show that sub-millimeter average errors can be achieved. However, the methods of assessment used in these studies are not comprehensive and mostly relies on manual operations that reduce reproducibility between data collection sessions. Although EMA is a popular motion tracking system in speech research because it can track any articulator and many tracers simultaneously, it is not used outside of few research labs due to major shortcomings. Chief among them, the receiver coils are wired which has been shown to impede natural speech [11], and EMA is cost-prohibitive (>$40,000), not portable, and its setup is complex and time-consuming [12].


[redacted]

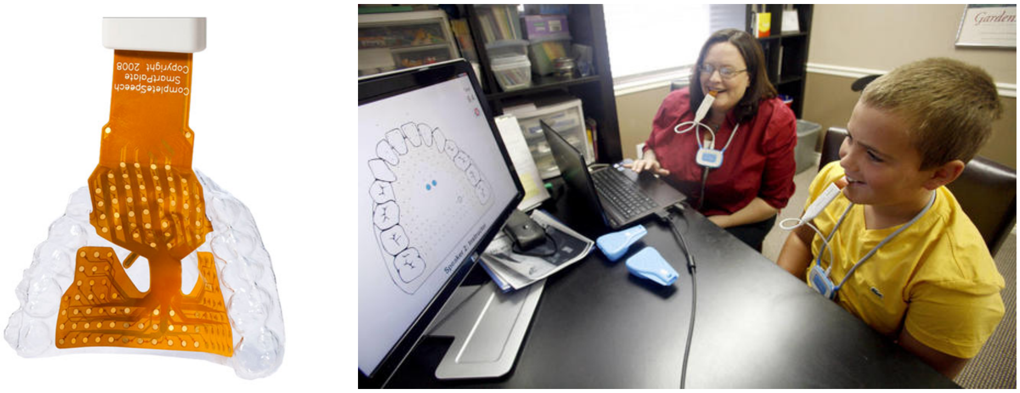
(a)

(b)

Figure 1.2: Electromagnetic articulograph: (a) tracers glued on the articulators for the Wave system from NDI [13], (b) the AG501 from Carstens [14].

1.2.2 Electropalatograph

EPG is able to detect the points of contact between the tongue and the palate. A custom-designed mouthpiece (Figure 1.3a) is inserted in the user's mouth and placed over the palate. Electrodes are embedded in the mouthpiece and are activated when the tongue is in contact by closing an electric circuit that conducts a small current through the user's body and to the EPG controller [15]. Because each electrode can be activated independently from each other, EPG can provide a coarse visualization of the shape of the tongue when in contact with the palate, which can be mapped and compared to a reference pattern as a visual biofeedback [16]. However, the mouthpiece has been shown to hinder natural speech and is costly to manufacture because it must be custom-designed to fit the user's palatal dimensions. Also, EPG cannot track any tongue motion that does not involve a contact between the tongue and the palate which excludes virtually all vowels and some consonants. Despite these shortcomings, EPG is being used in SLPs' private practices (Figure 1.3b) and it was shown that EPG-based speech therapy can help patients improve the production of some phonemes such as /s,t,d/ [17]. Though, its adoption by the SLP community remains greatly limited.



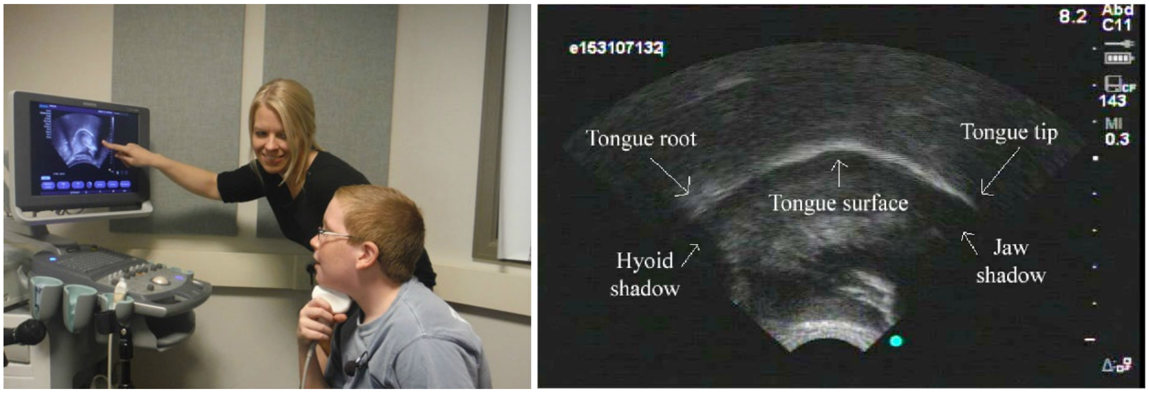
(a)

(b)

Figure 1.3: Electropalatograph: (a) custom-designed mouthpiece with embedded electrodes [18], and (b) EPG being used as part of a speech therapy [19].

1.2.3 Ultrasound Tongue Imaging

UTI generates grey images of the oral cavity from the reflection pattern of ultrasound waves. An ultrasound probe (Figure 1.4a) is typically placed under the chin and directed towards the mouth. The processed ultrasound patterns are then displayed in a user interface as 2D grey images (Figure 1.4b) in which the surface of the tongue is usually visible as a light-colored line. The main advantage of UTI is the fact that its tongue tracking method is wireless and noninvasive, especially when compared to EMA and EPG. It is also portable and affordable. However, a major drawback is that the ultrasound probe should not move in order to produce stable and reliable images [20] which is impossible to achieve without affixing the probe on a helmet. But, the probe must be placed as close to the chin as possible to enable the delineation of the tongue surface to be visible on the images which, consequently, restricts natural jaw motion during speech. Furthermore, the tongue tip is usually not visible because it is obstructed by the jaw and hyoid bone [21] while being the most important part of the tongue in speech production [22]. Because of these shortcomings, UTI is not widely used in speech research and, as far as we know, not found in private practices.



(a) (b)

Figure 1.4: Ultrasound Tongue Imaging: (a) a subject with an ultrasound probe placed under the chin [23], and (b) a grey image generated from the ultrasound patterns showing the inside of an oral cavity with the tongue surface visible as a light-colored line [24].

1.3 Need for MagTrack

The aforementioned systems are well suited for some specific tongue tracking applications, but they have not been widely adopted by researchers and end-users. For instance, the EMA and EPG are relying on wired tracking methods that hinder natural speaking behavior since the wires impede the normal motion of the articulators. Additionally, even though their intra-oral parts are coated with non-toxic materials, users might be uncomfortable knowing that electronic components are made of toxic materials that could potentially leak after a prolonged use. Of course, the intra-oral parts can be renewed before the coating would wear off, but this will add to their already high cost.

As mentioned earlier, the EMA is outside of the budget of most research labs and virtually all private practices and end-users. It is also not portable in the case of the AG-series or difficult to move for the Wave system. To be noted, EMA is an attractive tracking system for forefront research in speech due to its high accuracy and its ability to track multiple probes simultaneously, but it is not suitable for applications outside of a research lab.

Regarding EPG, the tongue tracking is limited to a 2D plane (i.e. palate) and is performed at a lower resolution than the other systems because the visual feedback is only composed of a finite number of discrete points. This substantially limits the speech sounds that EPG can recognize since the vast majority of vowels are spoken without any tongue contact to the palate.

For UTI, the major limitation is the fact that the probe must be in contact with the skin, usually under the chin, in order to produce images with sufficient resolution to identify the tongue's delineation. Furthermore, since the tracking is performed in 2D only, it is not intuitive to understand the components in the images, and it seems challenging to integrate this system into a user friendly device that can be used by the end-users. However, the fact that the tracking is performed entirely outside the mouth is undoubtedly a unique feature that could be helpful in other applications.

MagTrack was designed to combine many advantages of the aforementioned systems while addressing the limitations. The main objective is to implement a wireless tracking approach that can be embedded in wearable form factor. To do so, MagTrack relies on a new type of tracking technology that has been developed in recent years: Permanent Magnet Localization (PML). More details about how PML works will be provided in the next chapter, but as a summary, PML estimates in real-time the 5D state (3D position and 2D orientation) of a permanent magnet (more commonly referred as ”*tracer*”). The tracer is typically shaped as a cylinder with millimetric dimensions, and it generates a constant magnetic field that is recorded by an array of magnetometers. The magnetic measurements are then fed into a localization algorithm that outputs an estimation of the tracer's state (Figure 1.1). By placing a tracer on the tongue, PML is well suited for tongue tracking applications because it can track any motion of the tongue like EMA but without interfering with natural tongue motion like UTI since neither wires nor electronics are located inside the mouth. The tracer is safe to be used inside the mouth and on the tongue because it is coated with materials that is non-toxic, and in the event of being swallowed, it is exhausted naturally via our digestive tract because of its millimetric-sized and light-weight (<1g). For speech-related applications, PML has the potential to track millimetric movements of the tongue that can produce different phonemes [25]. Furthermore, PML is affordable: magnets and magnetometers are inexpensive since they are produced in high volume for mass markets. For instance, magnetometers can be found in all mobile devices and magnets in electric motors, appliances, furniture, apparels, among others.

In conclusion, MagTrack has the potential to be the best suited solution to be a practical tongue tracking system that would enable a wider segment of the end-users to benefit from its applications.

CHAPTER 2

SYSTEM DESCRIPTION

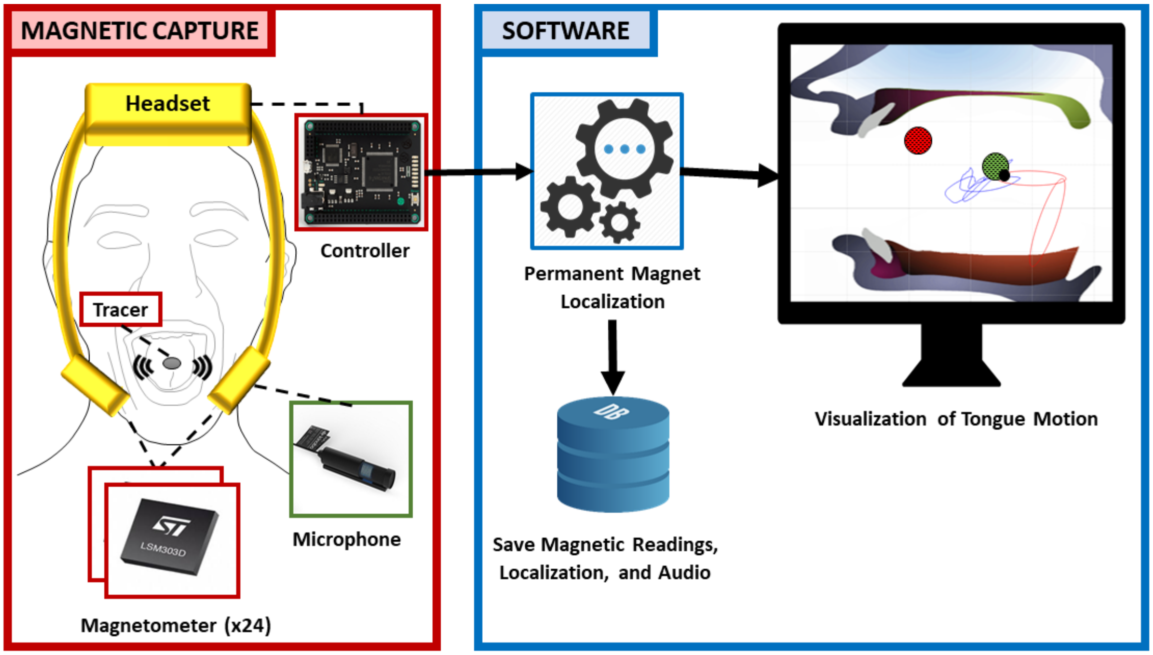
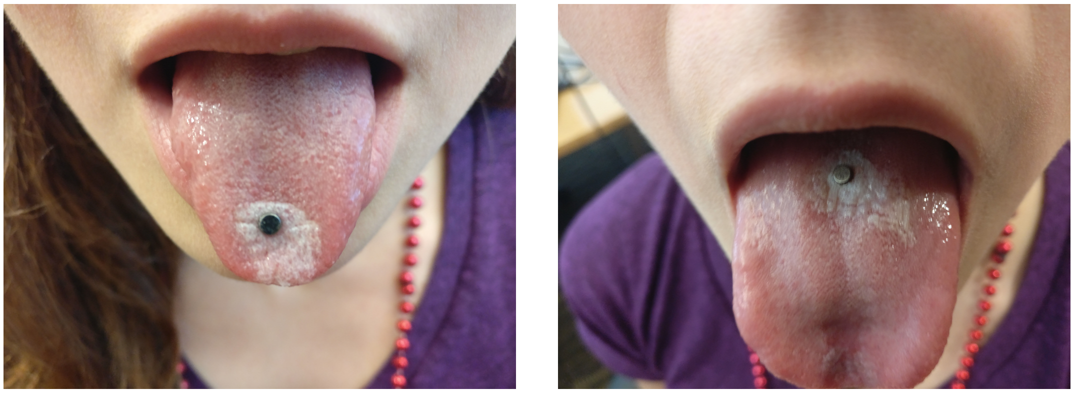

MAGNETIC CAPTURE
Headset
Controller
Tracer
Microphone
Magnetometer (x24)
SOFTWARE
Permanent Magnet Localization
Save Magnetic Readings, Localization, and Audio
Visualization of Tongue Motion

Figure 2.1: Overview of MagTrack's system components. A magnetic capture module transmits raw magnetic recordings from 24 magnetometers to a localization algorithm that estimates The 5D state of the tracer. The motion of the tracer is displayed on a user interface and saved into a database. Additionally, the voice of the user is recorded from a microphone.

MagTrack is a complete solution for a tongue tracking system. Figure 2.1 provides an overview of the different components of the system. MagTrack is capable of tracking tongue motion wirelessly by estimating the 5D state (3D position and 2D orientation) of a tracer glued on the tongue. Its magnetic capture module records the variation of the magnetic field generated by the moving tracer thanks to an array of 24 magnetometers. A controller creates data packets from the magnetic measurements that are fed into a permanent magnet localization algorithm whose output is the estimation of the tracer's state. These states are displayed in real-time on a custom-designed user interface and saved for later review and processing. Additionally, MagTrack records the user's voice from a microphone. In the current version of MagTrack, its body is a stationary headset that is anchored on an external support to remain fixed in position, and is secured to the user's head by headgear to reduce any undesired head motion artifacts.

The first section of this chapter will focus on the permanent magnet localization method to understand how the magnetic field recordings are translated into the tracer's state. The second and third sections will describe in more details the hardware and software components of the system, respectively. Finally, the last section will provide some pointers into future improvements of the system.

2.1 Permanent Magnet Localization



(a)

(b)

Figure 2.2: Tracer glued in different locations: (a) blade and (b) dorsum.

Tongue tracking is achieved through a permanent magnet localization algorithm that estimates the 3D position and 2D rotation of the magnetic tracer from its induced magnetic field. Figure 2.2 shows the tracer used in MagTrack (D21B-N52, K&J Magnetics, Pipersville, PA, USA) as it is attached on different positions on the tongue. This tracer is made of a N52-grade Neodymium which is the strongest magnetic material that can be found in the market, and it is disc-shaped (3 mm diameter and 1.5 mm thickness). The glue that is used to attach the tracer on the tongue is an FDA-approved dental adhesive (PeriAcryl, GluStitch, British Columbia, Canada). The tracer is considered to be a magnetic dipole that generates a constant and static magnetic field. Magneto-resistive magnetometers can measure its magnetic flux density $\overrightarrow{B}$ with high resolution (up to 61μgauss). In the literature and for sake of simplicity, the term ”*magnetic flux density*” is often reduced to ”*magnetic field*”. Since the tracer is attached on the tongue, the tracer moves with the tongue which results in the magnetic measurements from the magnetometers to vary according to their relative distance to the tracer. These raw magnetic measurements are then fed into our PML that estimates the 5D state of the tracer.

Historically, PML has been implemented as a multi-parametric nonlinear optimization problem that attempts to find the solution of the inverse of the point-source dipole equation (PSDE) [26]. In summary, the PSDE outputs the magnetic field that a magnetometer should read provided a 5D tracer's state as an input. However, the variable of interest is the tracer's state which cannot be directly derived by inverting the PSDE because of its high nonlinearity. Therefore, many studies have focused on implementing derivative-free nonlinear optimization methods to estimate the optimal 5D state that matches the magnetometers' readings [27, 28, 29, 30, 31]. Initially, MagTrack's PML was based on this method MagTrack with its implementation and tracking accuracy published in [32, 26]. More details about this method is provided below in section 2.1.1.

This nonlinear optimization-based PML has many shortcomings that limited its performance when used with real datasets. More recently, we have implemented an alternative method of localization by leveraging the capability of machine learning to find an optimal nonlinear model that best fits the magnetic measurements from the magnetometers to the state of the magnet. Among all the advantages of machine learning, one of the most attractive is that once a model is trained, the localization is completed much faster than in the previous method because the algorithm is non-iterative. Indeed, the optimization-based method is iterative since it has to try many guesses before finding the the optimal solution to the PSDE. A feedforward neural network was chosen as the localization model for this method which is described in more details in section 2.1.2.

2.1.1 Nonlinear Optimization

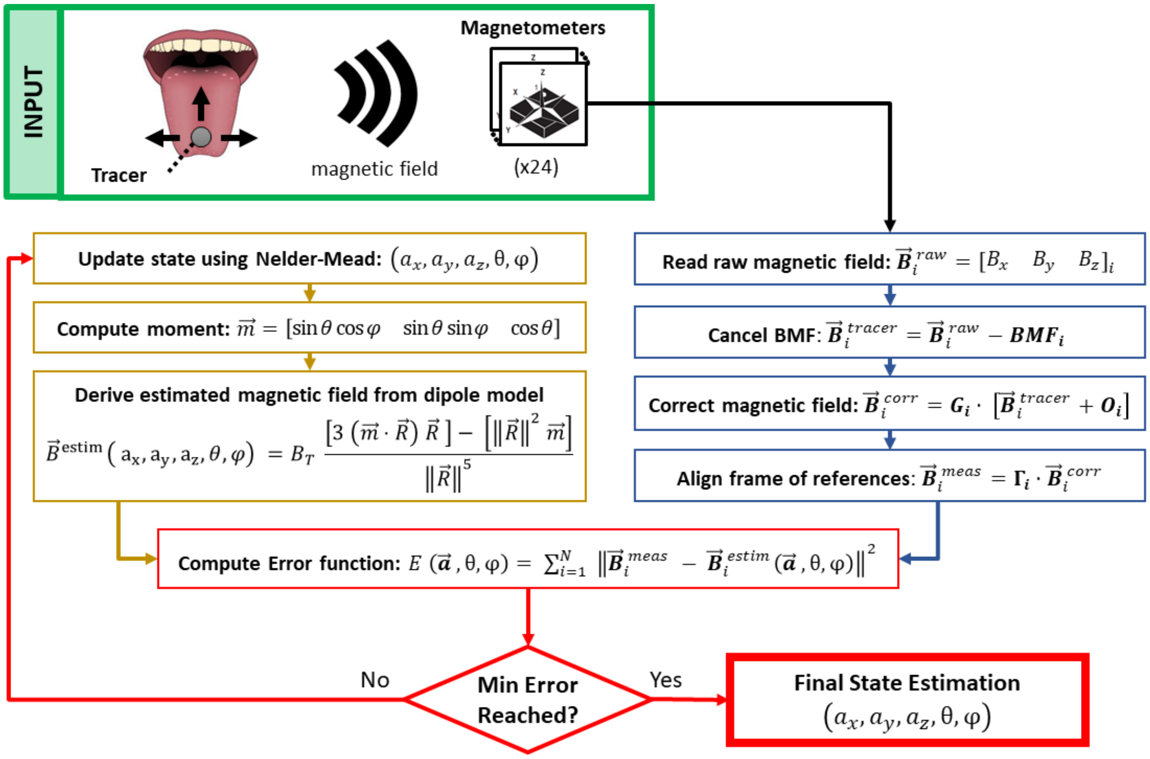

INPUT
Tracer
magnetic field
Magnetometers
(x24)
Update state using Nelder-Mead: $(a_x, a_y, a_z, \theta, \varphi)$
Compute moment: $\vec{m} = [\sin\theta\cos\varphi \quad \sin\theta\sin\varphi \quad \cos\theta]$
Derive estimated magnetic field from dipole model
$\vec{B}^{\text{estim}}(a_x, a_y, a_z, \theta, \varphi) = B_T \frac{[3(\vec{m}\cdot\vec{R})\vec{R}] - [\|\vec{R}\|^2 \vec{m}]}{\|\vec{R}\|^5}$
Read raw magnetic field: $\vec{\boldsymbol{B}}_i^{raw} = [B_x \quad B_y \quad B_z]_i$
Cancel BMF: $\vec{\boldsymbol{B}}_i^{tracer} = \vec{\boldsymbol{B}}_i^{raw} - \boldsymbol{BMF_i}$
Correct magnetic field: $\vec{\boldsymbol{B}}_i^{corr} = \boldsymbol{G_i} \cdot [\vec{\boldsymbol{B}}_i^{tracer} + \boldsymbol{O_i}]$
Align frame of references: $\vec{\boldsymbol{B}}_i^{meas} = \boldsymbol{\Gamma_i} \cdot \vec{\boldsymbol{B}}_i^{corr}$
Compute Error function: $E(\vec{\boldsymbol{a}}, \theta, \varphi) = \sum_{i=1}^{N} \|\vec{\boldsymbol{B}}_i^{meas} - \vec{\boldsymbol{B}}_i^{estim}(\vec{\boldsymbol{a}}, \theta, \varphi)\|^2$
No
Min Error Reached?
Yes
Final State Estimation
$(a_x, a_y, a_z, \theta, \varphi)$

Figure 2.3: Overview of our localization algorithm based on a nonlinear optimization method.

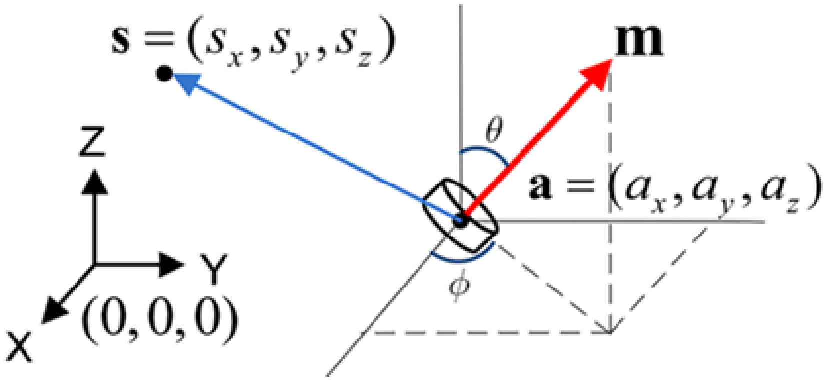

$\mathbf{s} = (s_x, s_y, s_z)$
$\mathbf{m}$
θ
$\mathbf{a} = (a_x, a_y, a_z)$
ϕ
Z
Y
X
(0,0,0)

Figure 2.4: Key parameters used by the PSDE for a magnetometer located at ***s***, and a tracer positioned at ***a*** with a dipole moment rotated by θ and ϕ.

An overview of this localization method is provided in Figure 2.3. The point-source dipole equation is a nonlinear model that predicts the theoretical magnetic field $\overrightarrow{\boldsymbol{B}}^{theo}$ that should be read at a magnetometer (located at $\overrightarrow{s}$) for a tracer placed at a 3D position $\overrightarrow{\boldsymbol{a}}$ and with a dipole moment $\overrightarrow{M}$ whose orientation is described by a zenith (θ) and an azimuth (ϕ) [27]. These key parameters are illustrated in Figure 2.4 and the formal equation of the PSDE shown below,

$$\overrightarrow{\boldsymbol{B}}^{theo}\left(\overrightarrow{\boldsymbol{r}},\theta,\phi\right) = \frac{\mu_0}{4\pi}\frac{\left[3\left(\overrightarrow{M}\cdot\overrightarrow{\boldsymbol{r}}\right)\overrightarrow{\boldsymbol{r}}\right]-\left[\|r\|^2\overrightarrow{M}\right]}{\|r\|^5} \tag{2.1}$$

where $\overrightarrow{\boldsymbol{r}} = \left(\overrightarrow{\boldsymbol{s}} - \overrightarrow{\boldsymbol{a}}\right)$ is the distance vector between the magnetometer and the tracer, and μ_0 is the magnetic permeability of free space. The following equation provides an expression of the dipole moment in terms of its strength and direction,

$$\overrightarrow{M} = \frac{B_r\, d^2\, l\, \pi}{4\,\mu_0}\,\overrightarrow{m} \tag{2.2}$$

$$\overrightarrow{m} = [\ \sin(\theta)\cos(\phi),\ \ \sin(\theta)\sin(\phi),\ \ \cos(\theta)\ ] \tag{2.3}$$

where B_r, d and l are the following parameters of the D21B-N52 permanent magnet: residual flux density (14,800 gauss), diameter (3 mm) and length (i.e. thickness: 1.5 mm). The unit vector $\overrightarrow{m}$ is the normalized dipole moment that provides the information about the tracer's orientation relative to the magnetometer. Equation 2.1 can be simplified by using the expression of $\overrightarrow{M}$,

$$\overrightarrow{\boldsymbol{B}}^{theo}\left(\overrightarrow{\boldsymbol{r}},\theta,\phi\right) = B_T\,\frac{\left[3\left(\overrightarrow{m}\cdot\overrightarrow{\boldsymbol{r}}\right)\overrightarrow{\boldsymbol{r}}\right]-\left[\|r\|^2\overrightarrow{m}\right]}{\|r\|^5} \tag{2.4}$$

$$B_T = \frac{B_r\, d^2\, l}{16}$$

where B_T is a constant that depends only on the tracer's characteristics. To be noted, the PSDE is only an approximation of its underlying physical model and performs optimally only when the highest dimension of the tracer is negligible compared to its distance to the magnetometer (hence the term "*point-source*" in PSDE). In [33], a ratio of 0.4 between the largest dimension of the tracer and its distance to a magnetometer is reported to be acceptable. In MagTrack, the ratio is 0.15 with the largest dimension of the tracer being its diameter (3 mm) and its smallest distance to a magnetometer in a typical setting for speech data collection is 2 cm. Lower ratio cannot be easily reached because the signal-to-noise ratio (SNR) will drop drastically which will result in decreased tracking accuracy. The drop in SNR is due to the fact that the tracer is made of the strongest magnetic material commercially available (Neodymium, Grade N52), thus, no smaller magnets can be found with similar magnetic strength. Increasing distance between the tracer and the magnetometers is also difficult to implement because the tracer's magnetic field becomes weak enough that it cannot be extracted from the sensor's noise.

Equation 2.4 provides an estimation of the magnetic field measured at a magnetometer. However, our objective is to estimate the position $\overrightarrow{a}$ and orientation (θ,ϕ) of the tracer from the magnetic measurement $\overrightarrow{\boldsymbol{B}}^{meas}$ of a magnetometer. Because the dipole equation is nonlinear, there is no analytical solution. Some numerical solutions have been developed to find a closed-loop formula [34] or a linear estimation [27]. Although these methods are less computationally intensive, they typically result in lower accuracies than nonlinear optimization algorithm [29]. For speech applications, very small displacement of the tracer must be tracked, consequently, increased accuracy is needed. A traditional approach for this optimization problem is the nonlinear least squares method where the objective is to minimize the sum of squared residuals [35] in an over-determined system (i.e. more equations than unknowns). Because the range of operation of a magnetometer is limited to a few centimeters, the magnetic measurements are captured by an array of 24 three-axial magnetometers to ensure a better coverage of the oral cavity in which the tongue moves.

More magnetometers cannot be added at this stage because all available general-purpose input/output (GPIO) pins on the controller are used in the current design of MagTrack. The error function E is the following sum of squared residuals,

$$E\left(\overrightarrow{\boldsymbol{a}}, \theta, \phi\right) = \sum_{i=1}^{N} \left\| \overrightarrow{\boldsymbol{B}}_i^{meas} - \overrightarrow{\boldsymbol{B}}_i^{estim} \left(\overrightarrow{\boldsymbol{a}}, \theta, \phi\right) \right\|^2 \tag{2.5}$$

where N is the number of magnetometers (N=24) and i the index of a magnetometer. The residual is the difference between the measured magnetic field $\overrightarrow{\boldsymbol{B}}_i^{meas}$ at a magnetometer and the estimated field $\overrightarrow{\boldsymbol{B}}_i^{estim}$ if the tracer was placed at a state $(a_x, a_y, a_z, \theta, \phi)$. $\overrightarrow{\boldsymbol{B}}_i^{estim}$ is used instead of $\overrightarrow{\boldsymbol{B}}^{theo}$ because the PSDE in equation 2.4 is established in a magnetometer's frame of reference (FoR), while the tracer's state $(a_x, a_y, a_z, \theta, \phi)$ is relative to a global FoR attached to MagTrack's body. Therefore, the state parameters in the global FoR must first be projected to each magnetometer's FoR before being the input to equation 2.4,

$$\overrightarrow{\boldsymbol{B}}_i^{estim} \left(\overrightarrow{\boldsymbol{a}}, \theta, \phi\right) = \boldsymbol{R_i}\, \overrightarrow{\boldsymbol{B}}_i^{theo} \left(\overrightarrow{\boldsymbol{s}}_i - \overrightarrow{\boldsymbol{a}}, \theta, \phi\right) \tag{2.6}$$

where $\overrightarrow{\boldsymbol{s}}_i$ is the position of the i^{th} magnetometer in the global FoR, $\overrightarrow{\boldsymbol{s}}_i - \overrightarrow{\boldsymbol{a}}$ the distance vector in the i^{th} magnetometer's FoR , and $\boldsymbol{R_i}$ the rotation matrix of the i^{th} magnetometer in the global FoR.

To provide the best tracking accuracy, the magnetometers must be calibrated to account for process variations during manufacturing and undesired soft-iron magnetic effects. A linear model is commonly used where the raw magnetic measurements $\overrightarrow{\boldsymbol{B}}_i^{raw}$ of a magnetometer are transformed as follows,

$$\overrightarrow{\boldsymbol{B}}_i^{meas} = G_i \left(\overrightarrow{\boldsymbol{B}}_i^{raw} + O_i - BMF_i\right) \tag{2.7}$$

where G_i and O_i are the gain (3x3 matrix) and offset (1x3) of the i^{th} magnetometer, respectively. BMF_i is the background magnetic field that is a disturbance that must be canceled to isolate the magnetic field generated by the tracer. As long as the device is positioned far enough from dynamic magnetic sources (e.g. power outlet, magnetic surfaces), the BMF can be considered as a constant vector. It is calculated by averaging 100 raw magnetic measurements (1 sec) collected for each magnetometer without any tracer in their vicinity.

As a summary, the objective of this localization method is to find the optimal set of state parameters that minimizes the error function E (equation 2.5). There are many different search strategies to find the optimal state, and these are the responsibility of an optimizer. In previous studies [35, 36], a benchmark of nonlinear optimizers was carried out to estimate their tracking accuracy and execution time. The benchmarked optimizers included Particle Swarming Optimization (PSO), DIRECT, Powell, and Nelder-Mead. It was concluded that Nelder-Mead provides the best trade-off between tracking accuracy and execution time in our application [36].

2.1.2 Neural Network

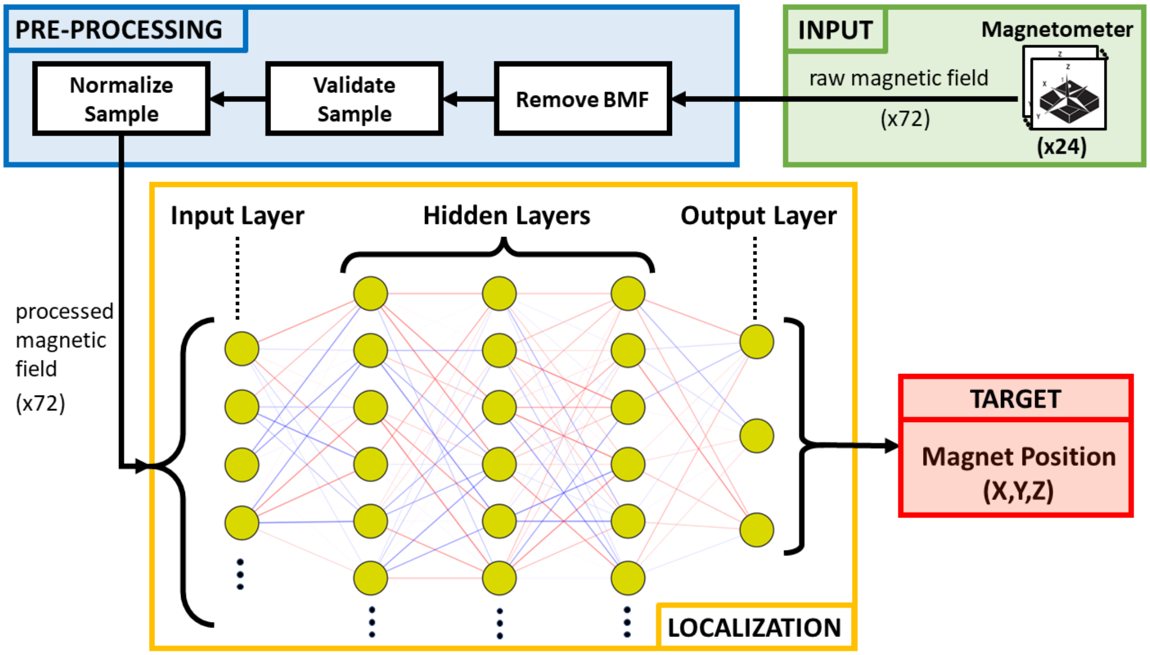


Figure 2.5: Overview of the localization algorithm based on a feedforward neural network.

In the optimization-based localization method, the nonlinearity of the PSDE restricts our choice for an optimal optimization algorithm. Furthermore, this method is sensitive to the initial estimation since each new estimation is initialized by the previous estimation of the tracer's state to be more accurate and reduce execution time. Also, we observed in our pilot studies that many iterations are needed for the optimizer to converge back to a proper estimation when the magnet was inadvertently placed in an area where its magnetic field is either weak (e.g. far from all magnetometers) or saturating (e.g. too close to a magnetometer). During this time, the PML generates incorrect state estimations even though the magnet is placed back to a working area. This behavior is not compatible with real-time tongue tracking for which even a small number of incorrectly localized states would affect its outcome.

With the recent advances in machine learning, and particularly the fact that the computation power needed to train such models is now more accessible than ever, there is a new opportunity to do away with the theoretical and approximation of the PSDE and generate new models that are more representative of the actual underlying physical phenomenon. There has been a recent body of research that implemented artificial neural networks (ANN) as the PML. In [37], the 1D position of a magnet moving along one dimension (total distance = 12 mm) was tracked by an array of 9 magnetometers using an ANN. The model was predefined to have 2 hidden layers, and the researchers studied the influence on the tracking error of feature reduction using principal component analysis and the number of neurons per layer. In [38], an ANN was also used for the localization of a tracer with a diameter of 9.5 mm and a thickness of 3.2 mm. This tracer was tracked by 8 magnetometers with an average 3D positional error of 1 mm.

Feedforward neural networks (FNN) are a type of ANN that are well suited for localization. Their architecture is simple but quite flexible to fit nonlinear models. Other machine learning algorithms such as support vector machine (SVM) and k-nearest neighbors (kNN) were tested but with no success because the tracking error was high enough to render the localization unusable. As shown in Figure 2.5, our model was composed of 72 magnetic field values (24 x 3-axial magnetometers) as predictors and three target variables (x,y,z) corresponding to the tracer's position. As an added bonus of using machine learning to generate a model, the estimation of the tracer's orientation is not needed to predict its position, as opposed to optimization-based PML where the angles must be estimated to calculate the estimated magnetic field (equation 2.6). Therefore, we chose to exclude the prediction of the angles θ and ϕ because their values are not being used in any of our tongue tracking applications.

The raw magnetic field measured by the magnetometers are pre-processed before being fed to the FNN. Similarly to the optimization-based localization, the background magnetic field is removed to isolate the magnetic field generated solely by the tracer. The next step is to validate the magnetic sample by comparing the magnetic measurements to a saturation and a noise threshold. If the tracer is within a proximity range from a magnetometer ($<$1 cm), its output saturates because the magnetic field generated by the tracer is much higher than the sensor's full-scale range (±4 gauss). Conversely, if the tracer is too far from all magnetometers ($>$10 cm), the sample contains mostly noise with no useful information about the state of the tracer. Therefore, a magnetic sample is ignored if any of its values is higher than a saturation threshold, or all values are less than a noise threshold. The last step is to normalize the magnetic values (16-bit signed integer: ±32,768) to ±1.0 to be used as predictors.

To limit the number of combinations of hyper-parameters to tune for the model, some restrictions were imposed: fully-connected network, same activation function for all neurons (except for the output layer set to a linear function), same number of neurons per layer. We tested different architectures and found that a FNN composed of 3 layers with 100 neurons per layer and the hyperbolic tangent as the activation function offers the best trade-off between tracking accuracy and model complexity.

2.2 Hardware

2.2.1 Magnetic Capture Module

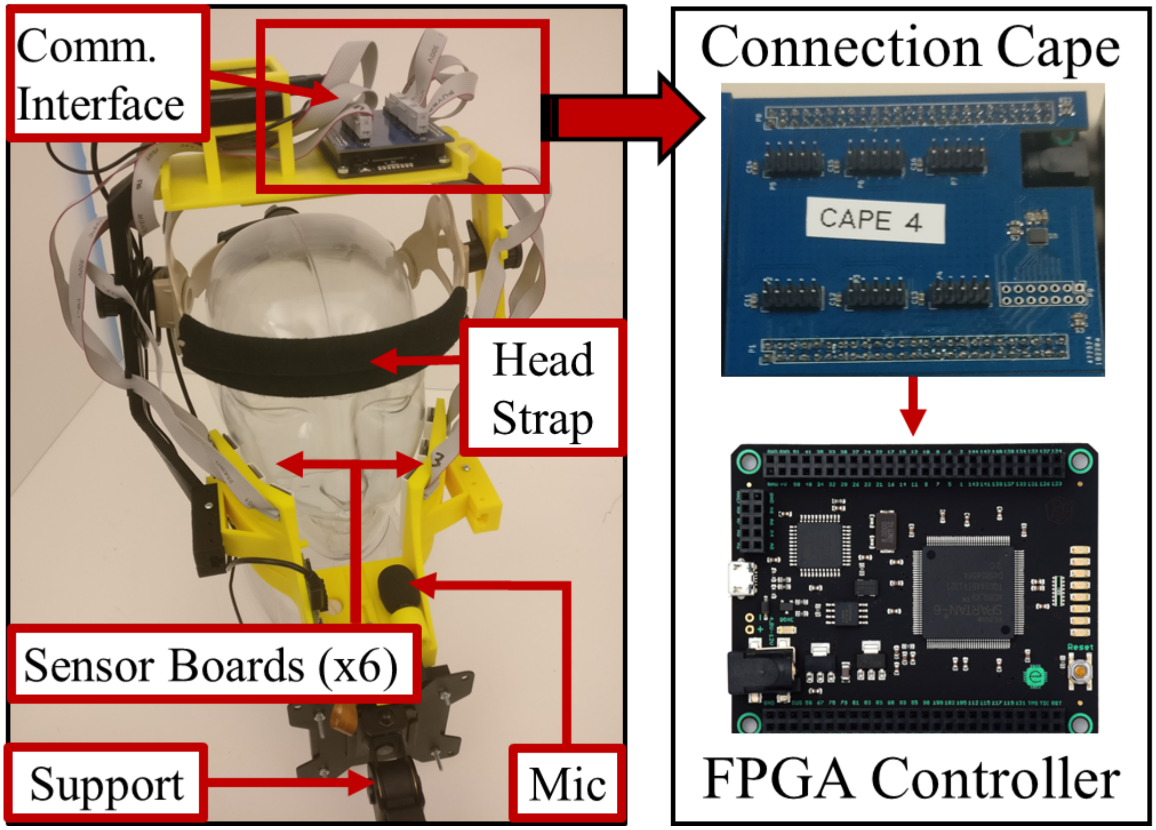


Figure 2.6: Magnetic capture module composed of six sensor boards connected to an FPGA controller through a custom-designed connection cape.

The changes in magnetic field induced by movements of the magnetic tracer are measured by 24 3-axial magnetometers. The choice of magnetometer is the LSM303D (STMicroelectronics, Geneva, Switzerland) because it provides one of the highest resolution in the market (~0.15 mgauss) thanks to a 16-bit analog-to-digital converter and a full-scale range of ±4 gauss. Also, the full-scale range can be changed to ±2/4/8/12 gauss, it can sample the magnetic field at a rate of 100 Hz, is compatible with both Serial Peripheral Interface (SPI) and Inter-Integrated Circuit (I2C) communication protocols, and is inexpensive since it is used in many mass-market electronics. To be modular, four magnetometers are housed in one custom-designed printed circuit board (PCB) which is referred as *sensor board* for brevity (Figure 2.7). All six sensor boards are connected via SPI to a field-programmable gate array (FPGA) (Spartan-6, Xilinx, San Jose, CA) embedded in a Mojo v3 board (Embedded Micro, Littleton, CO, USA) that also includes a USB interface to communicate with a computer. Figure 2.7 shows a high-level block diagram of the FPGA module. The Communication Manager block initializes and manages the SPI communication with the magnetometers through the SPI Controller module to poll all 72 magnetic field values. After receiving the digital magnetic field values from all the magnetometers, the Communication Manager generates a data packet with a total size of 153 bytes, composed of 144 bytes of magnetic field values (2 bytes per axis, 3 axes per magnetometer), 1-byte packet counter to verify if any packet loss has occurred, 4-byte header to identify the start of a new data packet, and 4-byte footer to signal the end of a data packet. The data packet is transmitted to the AVRI Controller block that delivers input packets to an Atmel AVR microcontroller on the Mojo board, which transmits the packet to the computer through an enumerated virtual COM port over USB at a baud rate of 500,000. The decision to use an FPGA to collect magnetic data in a high throughput parallel fashion stems from the fact that the initialization and sampling time of all magnetometers need to be accurately controlled and their output data to be concurrently polled to have an instantaneous snapshot of the tracer's magnetic field. However, the magnetometers in a sensor board are connected in an independent slave configuration with dedicated Cable Select pins (CS1-4), allowing only one magnetometer at a time to transmit data on the Master In/Slave Out (MISO) channel. This design of serial instead of parallel polling in a sensor board is due to the limited number of GPIOs available in the FPGA controller and to reduce the number of connections.

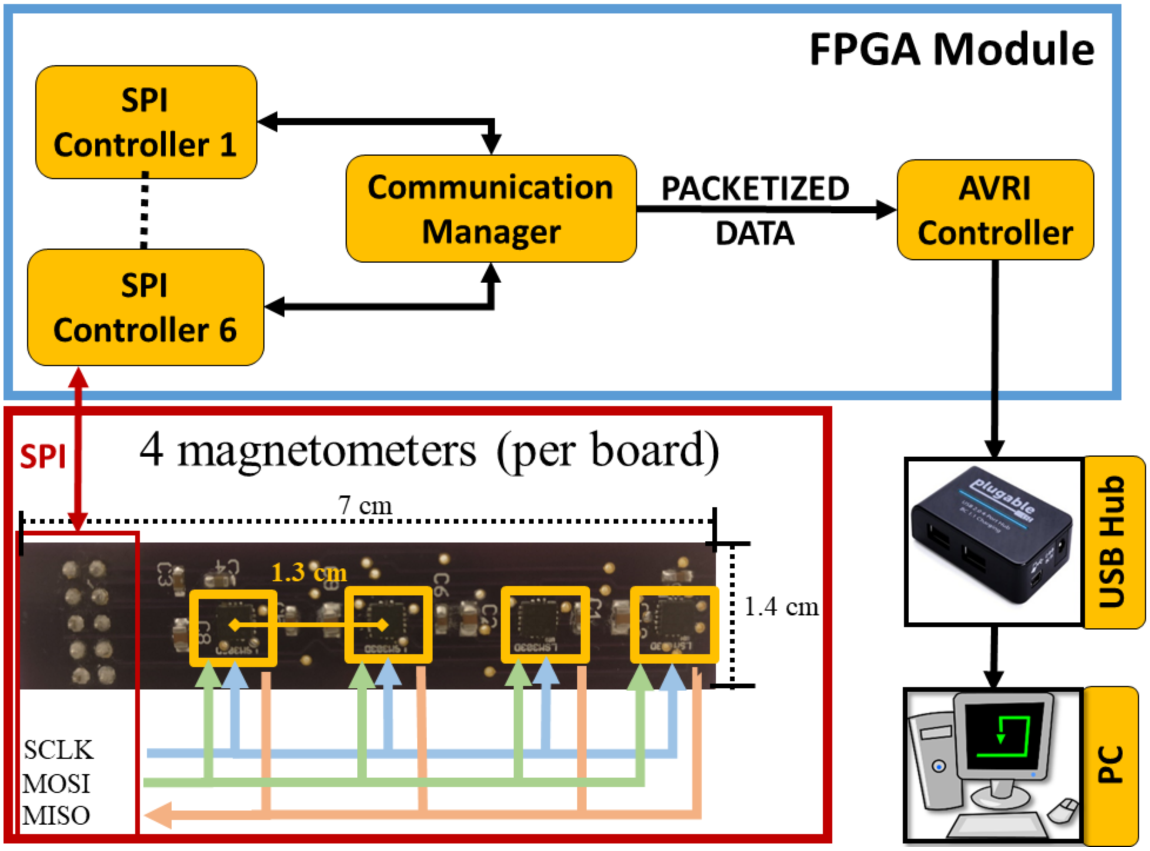


Figure 2.7: A high-level block diagram of the FPGA embedded in a Mojo board.

2.2.2 Voice Recording

The Mini-Akiro microphone (Kinobo, London, England) is used to record the user's voice at an adjustable sampling rate of 96 kHz. The choice for this model was due to a good trade-off between price and quality of recording.

2.2.3 Body Design

The design of the device's body has gone through a series of modifications to improve its functionality. There has been roughly five major versions with each one redesigned after a human study. Indeed, during each human study we conducted, participants were asked for their feedback about some specific aspects of using the device such as their level of comfort, fitness to their head/face, etc. Our team compiled the results and identified the most important design features to improve. Below is a list of each major version along with more details on the purpose and reasons for the redesign.

Benchtop V1

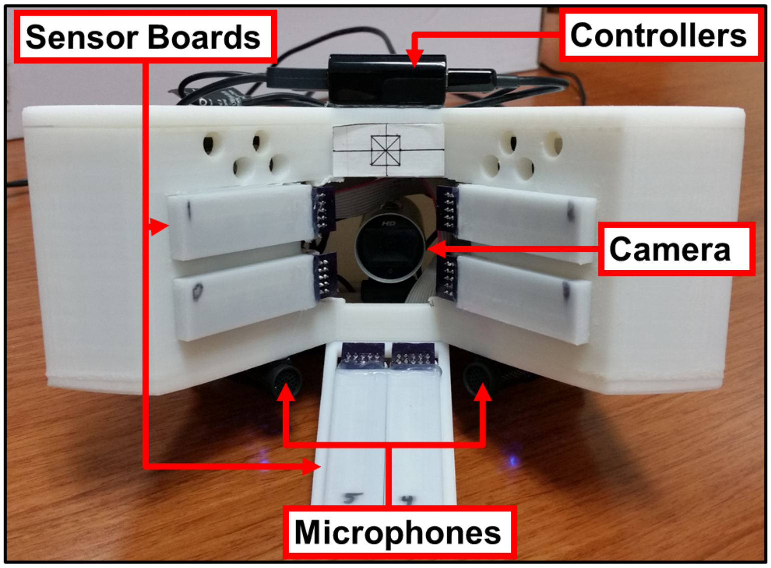


Figure 2.8: Benchtop design (version 1) of the device's body that includes six sensor boards, a pair of microphones, and a camera. The controllers are placed on top of the enclosure with the USB hub visible for easy connection to a computer. Two sensor boards are placed below the chin.

This is the initial design of the device. Its main feature is the placement of two sensor boards under the chin. The rationale for this choice of boards' placement is to have magnetometers all around the mouth to obtain a better measurement of the tracer's magnetic field. Although this was supposed to increase the tracking accuracy, we found that the accuracy was actually reduced because the user's jaw would bump against the under-the-chin boards when speaking and thus change the position of these magnetometers. A fixed and unchanged position of the magnetometers are indispensable for the localization to work properly since the models are calibrated and trained with a fixed set of positions. These boards could not be placed further away because the tracer's magnetic field would be too weak to be measured. This issue is actually a major technical challenge for tongue tracking using PML because it is hard to find a geometric arrangement of the magnetometers that would place them as close to the tracer as possible to increase the SNR but without impeding the motion of lips, jaw, and facial muscles.

Although less critical, another issue with this design is the placement of the controllers outside of the device. This version of the device's body was designed as a proof-of-concept, and thus its actual use in a human study was not a priority. An obvious issue is that they are in the user's field of view, and because neither the user nor the device can move once data collection starts, the controllers were blocking parts of the screen. Additionally, it was not uncommon for the sensor boards' cables to be disconnected from the FPGA or the headers to be damaged since they were not protected when the device was transported from one location to another.

As a side note, a camera is shown in most designs because lip tracking was thought to improve the usefulness of our system as a multimodal speech capture system. This system was previously called the MSCS as the acronym for Multimodal Speech Capture System. Recently, after determining that lip tracking was a challenging and multi-dimensional research problem in itself, we decided to focus on tongue tracking instead. This decision was also implemented because tongue tracking is not only the main innovation of this system but also is more practical and useful for its target users. Lip tracking might be added in the future but only after the tongue tracking technology is fully developed and widely used in the target applications.

Benchtop V2

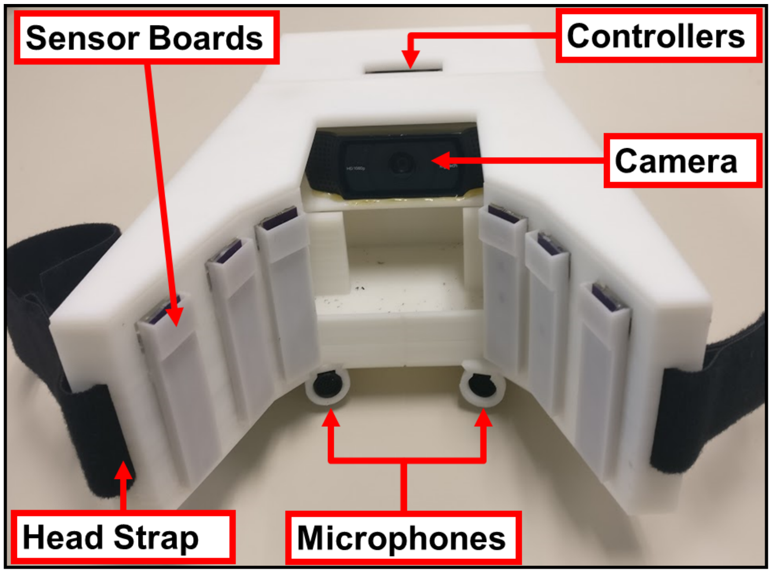


Figure 2.9: Benchtop design (version 2) of the device's body with the under-the-chin sensor boards moved to the side and all electronics placed inside the device's enclosure. Straps are also added to restrict head motion.

This is the first major redesign of the device. The main change was to move the under-the-chin sensor boards to each side of the device in order to allow for unhindered jaw motion. The sensor boards were also flipped vertically to improve the tracking accuracy on the vertical axis. Since more boards are placed laterally, we also increased the range of measurement on the anterior-posterior axis (i.e. forward/backward motion of the tracer). A back enclosure was added to house the controllers. A small aperture on the top was designed to enable the connection to a computer, therefore only two cables are connected to the device: one power and one USB cable. This not only protects all the electronics but also enable standard users to easily connect to a computer as a plug-and-play device.

Head straps were added to limit head motion. This is an important issue because the system has no means to discriminate between the tracer's motion and the head/body motion. Although participants are asked to remain still during data collection, it is impossible to fully prevent these motion artifacts from occurring because they are a natural motion that one cannot easily control. Therefore, straps were attached behind the head and gently fastened in an attempt to restrict head motion.

Benchtop V3

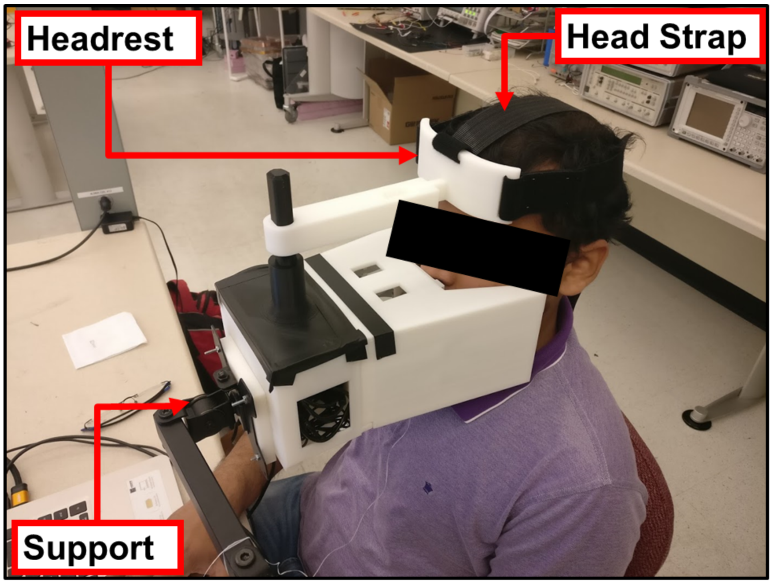


Figure 2.10: Benchtop design (version 3) of the device's body with a headrest added to better restrict head motion without worsening comfort.

The head strapping mechanism from the previous version was not efficient and substantial head motion was still found in the collected tongue trajectories. As a reminder, the difference in tongue placement for some phonemes, particularly for consonants, can be on the order of few millimeters [25]. Thus, even slight body/head motion can significantly affect the analysis of these tongue trajectories. In this version, a headrest was added to increase its restriction capability. The headrest is adjustable in both height and length to fit any head sizes. This had also the added benefit to increase comfort because the previous strapping method was pressing the user's cheeks against the device which was impeding their natural motion during speech. With this version, the head strapping is independent to the positioning of the main device's body which can be placed further away from the user's mouth.

Headset V1

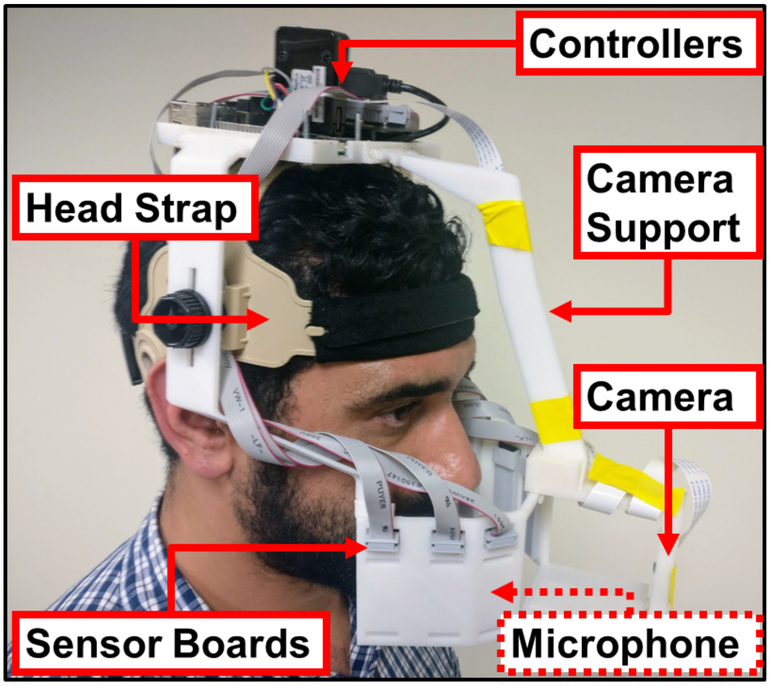


Figure 2.11: First attempt for a headset design. A support beam is placed over the mid-line section of the device to hold the camera. One microphone is placed at the bottom of the body (not visible in this picture).

Users of the last version of the benchtop device reported a satisfactory level of comfort and head restraint. However, because the final design of MagTrack's magnetic capture module must be wearable, it was decided to create a preliminary version of a headset. To bootstrap our design process, the base of the headset was a Browguard's headgear (Allegro Industries, Piedmont, South Carolina, USA) from which a custom-designed and 3D-printed structure is attached to using side screws. The structure houses all electronics, including the sensor boards placed around the mouth, a camera holder, a microphone placed underneath the structure, and a top plate to hold the controllers.

Current Headset Design

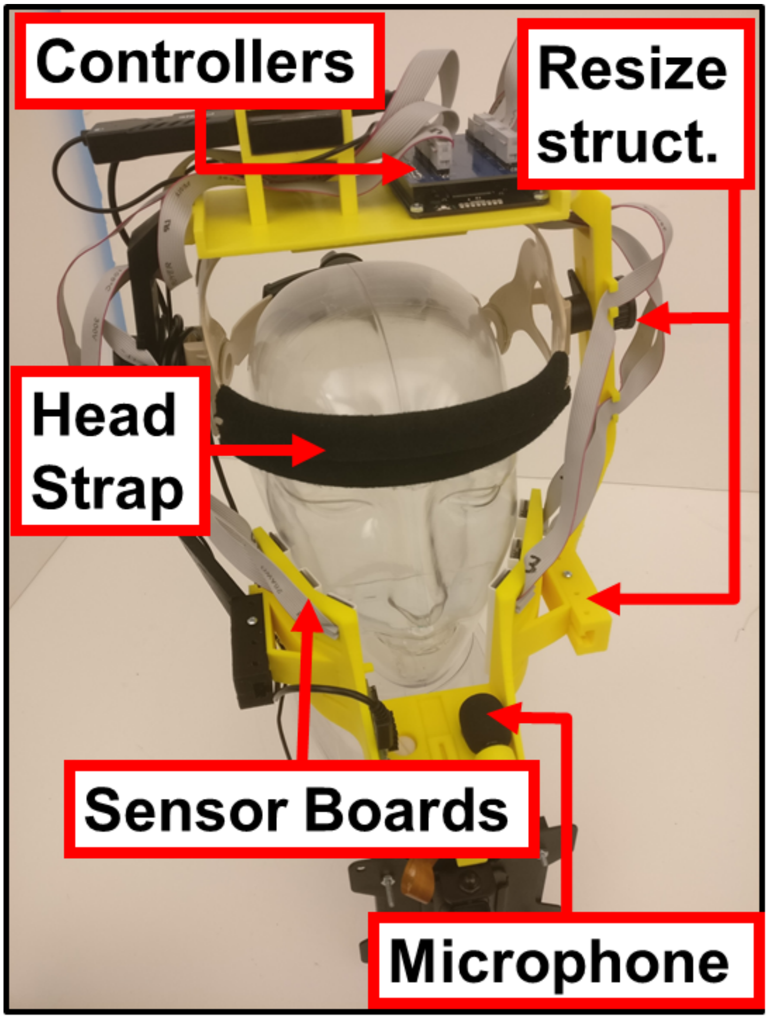


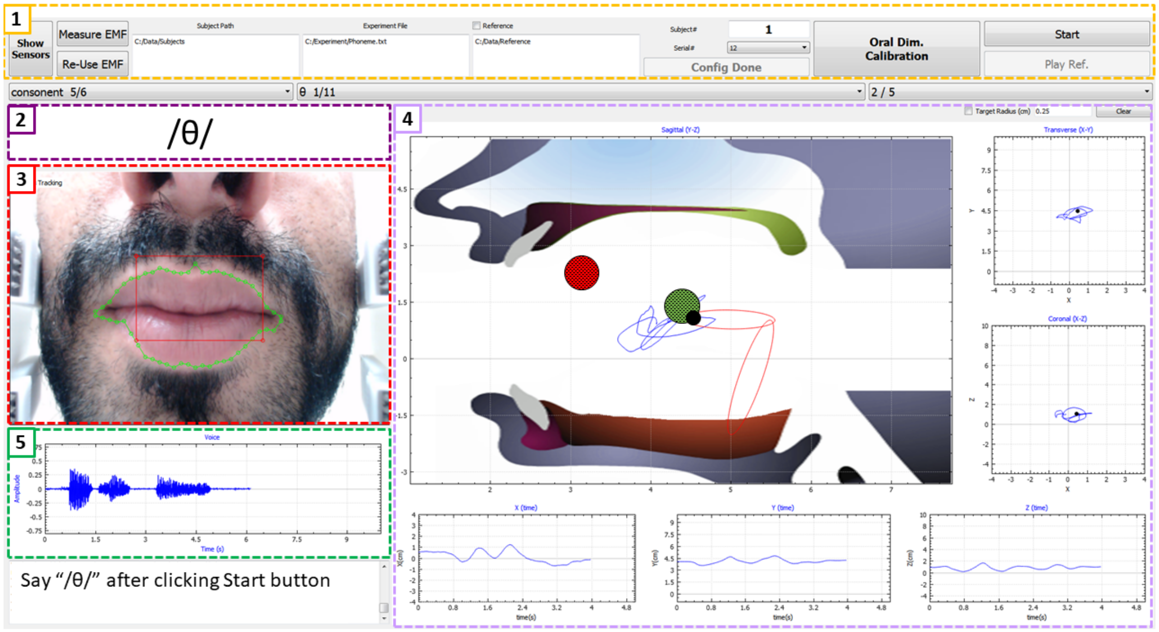
Figure 2.12: Current design of the headset with re-sizable structures to better fit users' head dimensions.

The latest design of the headset is shown in Figure 2.12. The headset's structure was made sturdier, lighter, easier to manufacture, and adjustable to better fit any head shapes and sizes. Also, the camera holder was removed since lip tracking is not implemented anymore. This enabled the design to be simplified by removing the mid-line camera support that was in the user's field of view, and placing the microphone closer to the mouth to increase the quality of recording.

More importantly, the bottom of the headset has a special design to anchor the whole device to an external support. As explained in section 2.1, the device has to remain fixed during data collection because our current algorithm lacks a dynamic BMF cancellation that will enable the localization to perform well even when the headset is moving. Since the BMF is recorded with the device being fixed in position, any deviation from this initial position will add errors to $\overrightarrow{B}^{meas}$ in equation 2.7 and, consequently, decrease tracking accuracy. This is the reason why the headset is not wearable in this version and it must be tightly anchored to a fixed support to prevent it from moving once data collection has started.

2.3 Software

2.3.1 User Interface



Subject Path: C:/Data/Subjects

Experiment File: C:/Experiment/Phoneme.txt

□ Reference: C:/Data/Reference

Subject#: 1

Serial#: 12

consonent 5/6

θ 1/11

2 / 5

□ Target Radius (cm) 0.25

Say "/θ/" after clicking Start button

Figure 2.13: User interface of the data collection program showing the following elements: (1) configuration parameters, (2) utterance, (3) video feedback of the camera, (4) tongue motion, and (5) voice signal waveform.

Figure 2.13 shows the current user interface developed specifically for human studies related to the treatment of SSD. The software was programmed in the C++ programming language using the Qt framework to facilitate the design of the user interface, and reduce complexity when developing a multi-thread program. This user interface is composed of the following parts:

1. ***Configuration***: This tab contains parameters that must be set before starting data collection:

    - Measure/Re-use BMF: These buttons are used to either record or load pre-recorded values of the background magnetic field. The BMF is recorded for 100 samples (1 second) and averaged to be used in the static BMF cancellation.
    - Subject Path: Various data are saved into files throughout the session: the raw magnetic recordings and the estimated tracer's states are saved as comma-separated value files (.csv), the voice as a waveform audio file (.wav), and logs of session activities as a text file. This field sets the root folder where these files will be saved to.
    - Experimental File: This field sets the path to the file that contains the utterances that must be spoken by the user.
    - Reference: This field sets the folder path where the reference audio files are located. These audio are the voice samples recorded for each utterance that were produced by a reference speaker (usually a speech-language pathologist).
    - Subject Number: A unique identifier for the participant.
    - Serial Number: Each MagTrack device has unique calibration and model parameters that are loaded by the program at startup to ensure the proper localization of the tracer.
    - Config done: this button must be clicked once the aforementioned fields are set. A sub-routine is then executed to ensure all parameters are properly set before starting the data collection.

- Oral Dim. Calibration: This button launches another user interface to obtain some information about the dimensions of the user's oral cavity. More details about this feature will be provided later in this section.
- Start: This button displays and saves the localized tracer.
- Play Ref: This button plays the reference audio for the current utterance.
- Show Sensors: This button opens the user interface shown in Figure 2.14 that displays the raw magnetic measurements. This visualization is used to verify whether the magnetometers are working properly.

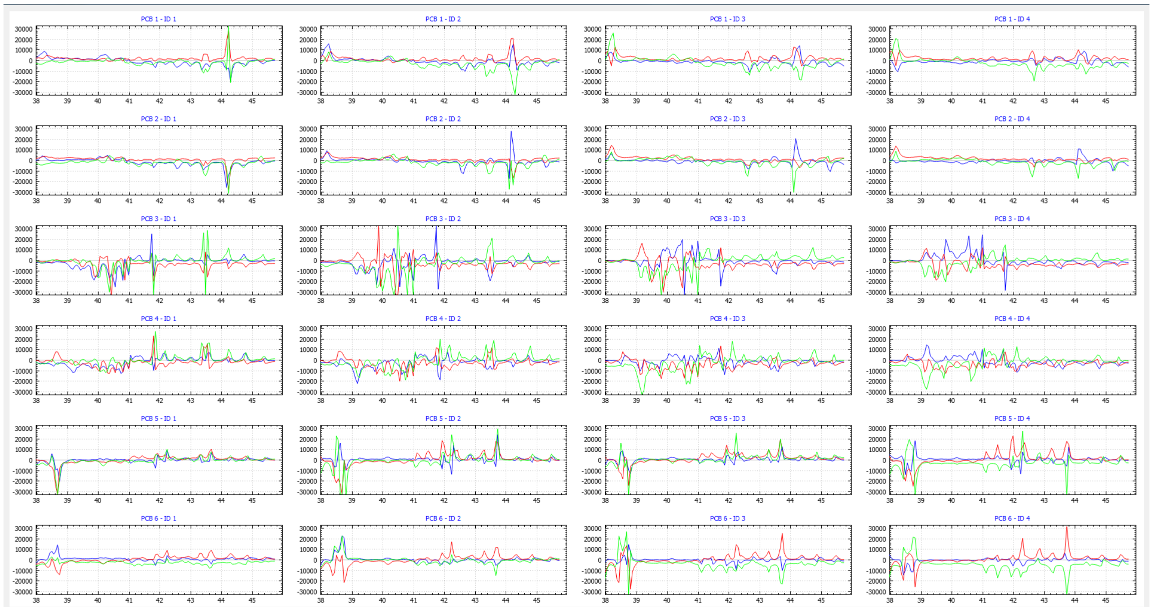


Figure 2.14: User interface showing the raw magnetic measurements split by magnetometer.

2. ***Utterance***: This text box displays the utterance to be spoken by the user. The utterance is automatically updated following the order in the list above it. This list contains all the utterances loaded from the experiment file and is split into three fields (from left to right): a category to provide a context (e.g. phoneme, word, object, month, color, question), the actual utterance, and the number of repetitions.

3. ***Video Feedback***: As mentioned before, the camera is no longer used in the current

version of MagTrack but is nonetheless shown here since it might be used again in the future. A centered red box provided a visual cue to ensure that the user's lips are properly positioned in front of the camera. Each frame was processed by a computer vision algorithm that detected the lip boundary which was then displayed over the original image as a green overlay.

4. ***Tongue Tracking***: The 3D position of the tracer is displayed in real-time and split into six graphs: the top graphs show the tracer's trajectory in transverse (X-Y), coronal (X-Z), and sagittal (Y-Z) planes, while the bottom graphs show the dynamic movement of the tracer along each axis vs. time. These spatio-temporal representations provide the SLP with valuable information about quality of speech and possible impairments both in terms of tongue placement (upper row) and tongue timing (lower row). The tracer's position is displayed as a black dot at the tip of an arbitrary delineated tongue in red that is shown solely to provide a visual context. The two ellipses that comprise this virtual tongue actually move with the tracer by stretching and folding. Furthermore, visual targets can be added to guide the user's tongue position. Two targets are shown with their color being either green if the tracer is inside or red otherwise. The diameter of these targets can be changed to adapt the difficulty of the speech exercise.

5. ***Voice Waveform***: The microphone's recordings are represented in real-time as an audio waveform. The amplitude of the audio data is normalized between ±1 and the waveform is downsampled by a factor of 10 to reduce unnecessary consumption of computer resources by the plotting library. This representation of audio data was selected because most SLPs are more familiar with voice waveforms.

2.3.2 Palatal Projection

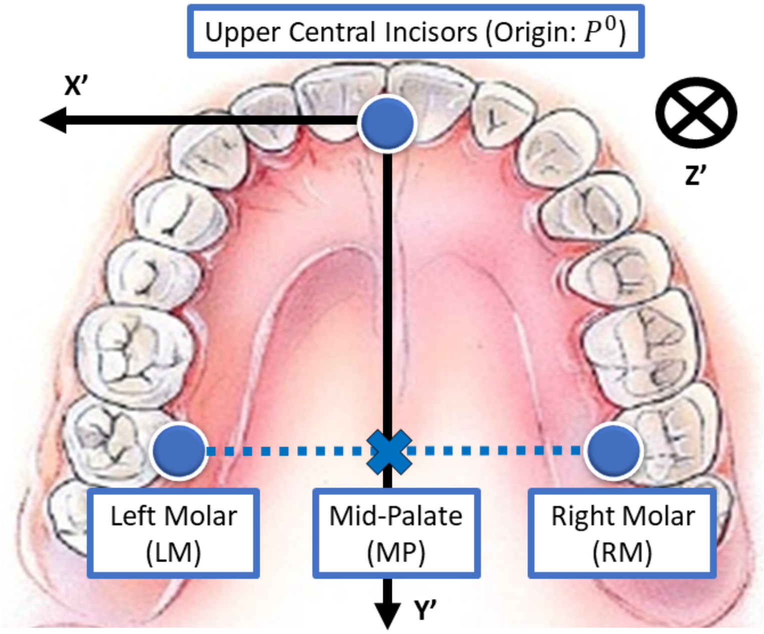


Figure 2.15: Illustration of the most important landmarks and axes used in the palatal projection.

When the headset is placed on the user's head, the orientation of the magnetometers relative to the user's mouth is not known. This is an issue because it creates difficulties when the tongue trajectory is analyzed, displayed, or just plainly compared to other trajectories. To illustrate this problem, let's consider the case of the tongue trajectory of a patient that is compared to a reference during speech therapy. The reference trajectory would be recorded from an SLP with the mouth being oriented at a certain angle, while the patient's mouth will be oriented at a different angle since there is no mechanism to place the headset at the same position and orientation each time it is placed on a user's head. Because there isn't any simple hardware/mechanical solution to this problem, a software solution was implemented. The idea is to project the coordinates of the tracer from the global FoR into a FoR that is always fixed relative to the user's mouth. Therefore, no matter the position and orientation of the headset, the tongue position will always be shown in the same FoR and thus facilitate the comparison and analysis of tongue trajectories between sessions. As shown in Figure 2.15, an obvious selection for this new FoR is the palate since it is a fixed part of the mouth. This palatal FoR is composed of three basis ($\overrightarrow{x}'$, $\overrightarrow{y}'$, $\overrightarrow{z}'$) fully described by these three landmarks: the origin (P^o) located between the upper central incisors, the left molar (LM), and the right molar (RM). The basis vector $\overrightarrow{x}'$ is set to be parallel to the line that passes through LM and RM, and arbitrarily directed positive towards the left side. The basis vector $\overrightarrow{y}'$ is set as the line passing through the origin and the mid-palate point (MP) which is placed at the center of the LM-RM line. This axis is oriented positively towards the inside of the mouth (i.e. the throat). The last basis $\overrightarrow{z}'$ is orthogonal to the $\overrightarrow{x}'$-$\overrightarrow{y}'$ plane and following the right-hand rule (here, directed upwards).

Every time the headset is placed on the user's head, the user is asked to place the tracer at each of these three palatal landmarks to locate their positions in the global FoR. Then, a matrix $Proj$ is calculated to project the position of the tracer from the global to the palatal FoR as follows:

$$\overrightarrow{X'} = \overrightarrow{LM} - \overrightarrow{RM};\ \ \overrightarrow{x}' = \frac{\overrightarrow{X'}}{\|X'\|}$$

$$\overrightarrow{Y'} = \frac{\overrightarrow{RM} + \overrightarrow{LM}}{2} - \overrightarrow{P^o};\ \ \overrightarrow{y}' = \frac{\overrightarrow{Y'}}{\|Y'\|}$$

$$\overrightarrow{z}' = \overrightarrow{x}' \times \overrightarrow{y}'$$

$$Proj = \begin{pmatrix} x'_0 & x'_1 & x'_2 & -P^o_x \\ y'_0 & y'_1 & y'_2 & -P^o_y \\ z'_0 & z'_1 & z'_2 & -P^o_z \end{pmatrix}$$

Therefore, for every new tracer's position ($x,\ y,\ z$), a set of coordinates ($x',\ y',\ z'$) is calculated using the following equation:

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = Proj \begin{pmatrix} x \\ y \\ z \\ 1 \end{pmatrix}$$

To reiterate the purpose of this palatal projection, if the tracer is placed at a same position in the mouth, the projected coordinates (x', y', z') should remain the same no matter the position and orientation of the user's head in the global FoR. Consequently, this projection should provide consistent trajectories between data collection sessions and thus facilitate their analysis and comparison.

2.3.3 Procedures for Data Collection

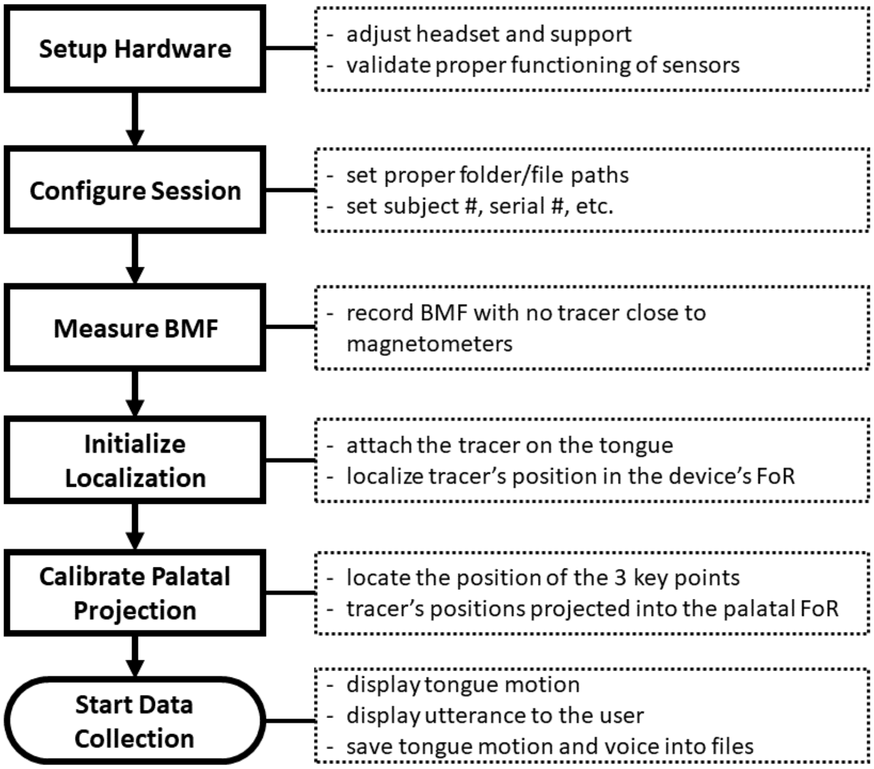

Setup Hardware
- adjust headset and support
- validate proper functioning of sensors
Configure Session
- set proper folder/file paths
- set subject #, serial #, etc.
Measure BMF
- record BMF with no tracer close to magnetometers
Initialize Localization
- attach the tracer on the tongue
- localize tracer's position in the device's FoR
Calibrate Palatal Projection
- locate the position of the 3 key points
- tracer's positions projected into the palatal FoR
Start Data Collection
- display tongue motion
- display utterance to the user
- save tongue motion and voice into files

Figure 2.16: Flow diagram of the procedures in a typical data collection.

Figure 2.16 provides an overview of the procedures that are followed in a typical data collection session. The first step is to setup the hardware by fastening the headset to the user's head and adjusting the support to ensure the user's posture is comfortable. The magnetometers are then tested for a signal response by waving a tracer in their vicinity. The next step is to configure the parameters for the session, and measure the background magnetic field to properly initialize the localization algorithm. Then, the tracer is attached on the part of the tongue whose motion is of interest, and the palatal projection matrices are calculated to convert the tracer's states from the global to the palatal frame of reference. Only after these steps are completed that the data collection can start. These procedures are typically completed in less than 5 minutes.

2.4 Future Work

Our current prototype is functional and can reach a satisfactory level of tracking accuracy as it will be shown in the next chapter. Furthermore, it has already been used in human studies to assess its capabilities in various applications, including in the creation of a baseline of phoneme landmarks (chapter 4) and in a silent speech interface (chapter 5). These studies enabled us to obtain quantitative and qualitative feedback about MagTrack's capabilities which are the basis of this discussion on the future improvement of the system.

2.4.1 Wearable Headset

As already mentioned, our current headset version is stationary and thus requires the user to remain still for the duration of the data collection session. However, natural body/head motion cannot be fully restricted and thus adds motion artifacts in the recorded trajectories that is not from the tongue. Having a wearable headset that remains fixed on the user's head but not bound to a fixed support will enable users to move freely without adding any body motion to the recorded trajectories. Additionally, a wearable design will drastically increase comfortability and ease of use. To reach this objective, the magnetic field produced by the tracer must be isolated from the background magnetic field that is a disturbance to be removed. In the current version, the BMF is simply measured for each tracking magnetometer before the data collection starts and with no tracer in the vicinity. These BMF values are then subtracted from any raw magnetic measurements to isolate the tracer's magnetic field (equation 2.7). This static BMF cancellation is possible because the magnetometers are fixed in space and thus, in theory, the BMF values remain unchanged.

In practice, this method is flawed for two main reasons. First, the headset does not remain absolutely fixed because the support is not perfect and allows small movements that are produced by natural body motion. Secondly, the BMF might vary during the data collection session because we are surrounded by dynamic magnetic field-generating devices (e.g. mobile devices) that adds magnetic disturbances not accounted for. More importantly, this method cannot be used when the headset is moving because the BMF values for each magnetometer will vary with its change in orientation.

As a possible solution, our last version of the FPGA's cape was designed with a LSM9DS1 reference sensor (STMicroelectronics, Geneva, Switzerland) which is an inertial measurement unit (IMU) composed of a magnetometer, accelerometer, and gyroscope. Similarly to the Tongue Drive System [39], the idea is to dynamically attenuate the BMF by projecting the magnetic measurement $\overrightarrow{\boldsymbol{B}}_{ref}$ from the LSM9DS1's magnetometer to each of the tracking magnetometers in the magnetic capture module. The assumption is that the BMF is affecting all magnetometers in the same way because its magnitude and direction are constant for a same location since its main component is the Earth's magnetic field. Therefore, its magnitude can be measured by the reference magnetometer of the IMU since it is placed far enough from the tracer that only the BMF is captured. However, to cancel the BMF at each of the tracking magnetometer, the BMF's direction must be rotated from the reference magnetometer's FoR to each of the tracking magnetometers' FoR. Furthermore, using the same rationale than in equation 2.7, a calibration must be performed to account for the nonidealities of the sensors. Therefore, a projection matrix P_i must be calculated to estimate the BMF B_i that should be read by the i^{th} tracking magnetometer from the B_ref

measured by the reference magnetometer:

$$\begin{pmatrix} B_x \\ B_y \\ B_z \end{pmatrix}_i = P_i \begin{pmatrix} B_x \\ B_y \\ B_z \\ 1 \end{pmatrix}_{ref}$$

$$P_i = \begin{pmatrix} g_{xx} & g_{xy} & g_{xz} & o_x \\ g_{yx} & g_{yy} & g_{yz} & o_y \\ g_{zx} & g_{zy} & g_{zz} & o_z \end{pmatrix}_i$$

with g and o being the gain and offset coefficients, respectively. The projection matrices are calculated using the least-square error method by collecting as many magnetic samples as possible while rotating the headset in all directions:

$$\widehat{P_i} = \left(B_{meas} B_{ref}^T\right) \left(B_{ref} B_{ref}^T\right)^{-1} \tag{2.8}$$

$$B_{meas} = \begin{pmatrix} B_x^1 & ... & B_x^N \\ B_y^1 & ... & B_y^N \\ B_z^1 & ... & B_z^N \end{pmatrix}_i$$

$$B_{ref} = \begin{pmatrix} B_x^1 & ... & B_x^N \\ B_y^1 & ... & B_y^N \\ B_z^1 & ... & B_z^N \\ 1 & ... & 1 \end{pmatrix}_{ref}$$

Although this method of dynamic BMF cancellation is fully known and reported to be performing well in the Tongue Drive System, the residual errors that remain after this cancellation are high enough to significantly decrease the tracking accuracy for MagTrack at a level that renders our PML unusable. The causes of this issue are not known at this moment but will be investigated in our next batch of improvements. One possible reason is that because the inverse matrix of $(B_{ref}B_{ref}^T)$ is actually a pseudo-inverse since it is not full ranked due to sensor noise, the method used to derive the pseudo-inverse might not provide a sufficiently close approximation. Thus, a better method to approximate the pseudo-inverse might be needed. Furthermore, it is also possible that there might have been enough number of magnetic measurements or rotations of the headset to generate a proper estimation of the pseudo-inverse.

2.4.2 Hardware Optimization

As can be seen in Figure 2.12, the headset is rather cumbersome and might not be practical to be used as a wearable system. Our current focus is to validate that the tracking can be performed with a satisfactory level of accuracy. Therefore, our design's objective was to maximize the number of magnetometers in order to increase the tracking accuracy. Indeed, increasing the coverage area around the mouth provides different ”viewing angles” for recording the tracer's magnetic field which should improve the overall tracking accuracy. In the current design, the magnetometers are place next to the cheek and as close to the tracer as possible while enabling the user to speak freely. However, it is possible that a more optimal configuration of the magnetometers, in terms of their position and orientation, could have a greater influence on the tracking accuracy than simply increasing their number. This study is challenging because it is difficult to build a physical setup that will enable the magnetometers to move in 6D (3D position and 3D orientation) in an efficient and automated manner. There are two different approaches that can be used to overcome this issue.

First, a top-down approach could be implemented with the development of a simulated environment of the physical space in which the magnetic measurements would be calculated from the point-source dipole equation with added gaussian white noise to simulate the magnetometers' noise, or even from a more realistic model that could be generated from real measurements. This simulated space would enable an automated method to move and rotate the magnetometers, with spatial constraints (e.g. remain outside of the oral cavity), while estimating the impact on the tracking accuracy. The results would then be validated in the physical space by placing the magnetometers at the optimal positions/orientations as estimated in the simulated space.

Secondly, a bottom-up approach could be implemented in which the optimal configuration for a small set of magnetometers will be studied. For example, one could start from an initial objective to find out the maximum distance between a pair of magnetometers, placed on each side of the mouth, that would still provide a satisfactory level of tracking accuracy. From that point, one pair could be added at each iteration and the position/orientation of all magnetometers could be re-evaluated to maximize tracking accuracy. Because of the inherent symmetry of the magnetic capture module, any change on the geometric arrangement of the magnetometers in one side can be mirrored to the other side, resulting in a significant reduction in the number of possible combinations of positions and orientations.

CHAPTER 3
ASSESSMENT OF TRACKING ACCURACY

This chapter will describe in details the setup, methods, and results of our assessment of MagTrack's tracking accuracy. As a brief overview, the tracking accuracy is defined as the distance error between the actual position of the tracer and its estimation generated by our localization algorithm. A tracer is placed at known positions along a predefined trajectory by a 5D positional stage that was designed and built in our lab. This positional stage has 5 motors that independently control each value of the tracer's state (x, y, z, θ, ϕ), thanks to a motor controller that transforms a desired state as input to the motors' angular rotation. A custom-designed program was developed to synchronize the placement of the tracer at the desired states and the magnetic recording from MagTrack. This program has historically been called the ”calibration” program due to its initial application to find the parameters needed to run the nonlinear optimization-based localization. This calibration program loads a trajectory file that contains all the desired states, transforms the states into a gcode format that is recognized by the motor controller, and saves all the magnetic measurements when the tracer reached the desired state. The first assessment of tracking accuracy was performed for the optimization-based PML using one reference trajectory to find the calibration parameters and three testing trajectories to measure the tracking error. This study was published in [32] and more details are provided in section 3.2.1. The second assessment was carried out after the development of the FNN-based PML in which three trajectories were used: a training set, a validation set, and a testing set. The training set is used to train a neural network by finding the weights of the edges, the validation set is used to attenuate the effect of over-fitting, and the testing set serves as a separate dataset unseen during the training process that provides a more realistic evaluation of the tracking error. This assessment was published in [40] and described in section 3.2.2.

Our assessment of tracking accuracy, described in more details in section 3.1, is more comprehensive that the methods that are used for EMA. Indeed, the typical assessment for EMA relies on an indirect tracking method that estimates the difference in the distances between probes (referred as *inter-distance*) placed at a known and fixed distance between each other. In [9], an evaluation of the tracking error of the AG500 (Carstens Medizinelektronik GmbH, Bovenden, Germany) was carried out using three methods. The first method relied on the calibration device provided with the AG500 in which a maximum of 12 probes can be placed inside a rotating disk. The placement of each probe within the disk is known, which means that their inter-distances are also known. The disk is rotated only in the horizontal plane, and the estimated positions of the probes are analyzed to calculate the deviation of their inter-distances to their actual values. In the second method, instead of an automated rotation in one plane, the disk was moved manually and randomly inside the EMA's working space. Therefore, the disk is rotated in different planes and positions. In the last method, two probes were attached on the jaw of a single subject that was asked to speak few utterances. Their inter-distances measured during speech are compared to its initial value. Overall, this study showed that the median error was around 0.5 mm. In [8], a similar but more elaborate apparatus was built by the researcher to conduct an assessment of tracking accuracy on another EMA, the Wave System (NDI, Waterloo, Ontario, Canada). Three methods were also used: a static tracking where the probes are placed at known locations, a dynamic tracking where the inter-distances are known while the probes are being moved using a crank-rocker four bars mechanism, and a speech task with two probes attached to the jaw in the similar manner than in [9]. The results are also similar to the study in [9] with a median of 0.5 mm. These indirect methods are suitable for a first approximation of the tracking accuracy, but they are far from being comprehensive. For instance, the automated motion of the probes are usually restricted to one or two planes at best, and their trajectory lacks a variation in their rotations. Furthermore, the motion encompasses a small portion of the working volume or a low density of locations inside a larger volume. Finally, some of the assessment methods relied on manual operations that are not easily reproducible.

3.1 Data Collection Setup

The objective of our assessment of tracking accuracy is to be (1) reproducible by automating the tracer's placement using a 5D positional stage (section 3.1.1), (2) consistent in the placement of the tracer by using predefined trajectories (section 3.1.2), and (3) easy to perform through an operator-friendly program (section 3.1.3).

3.1.1 5D Positional Stage

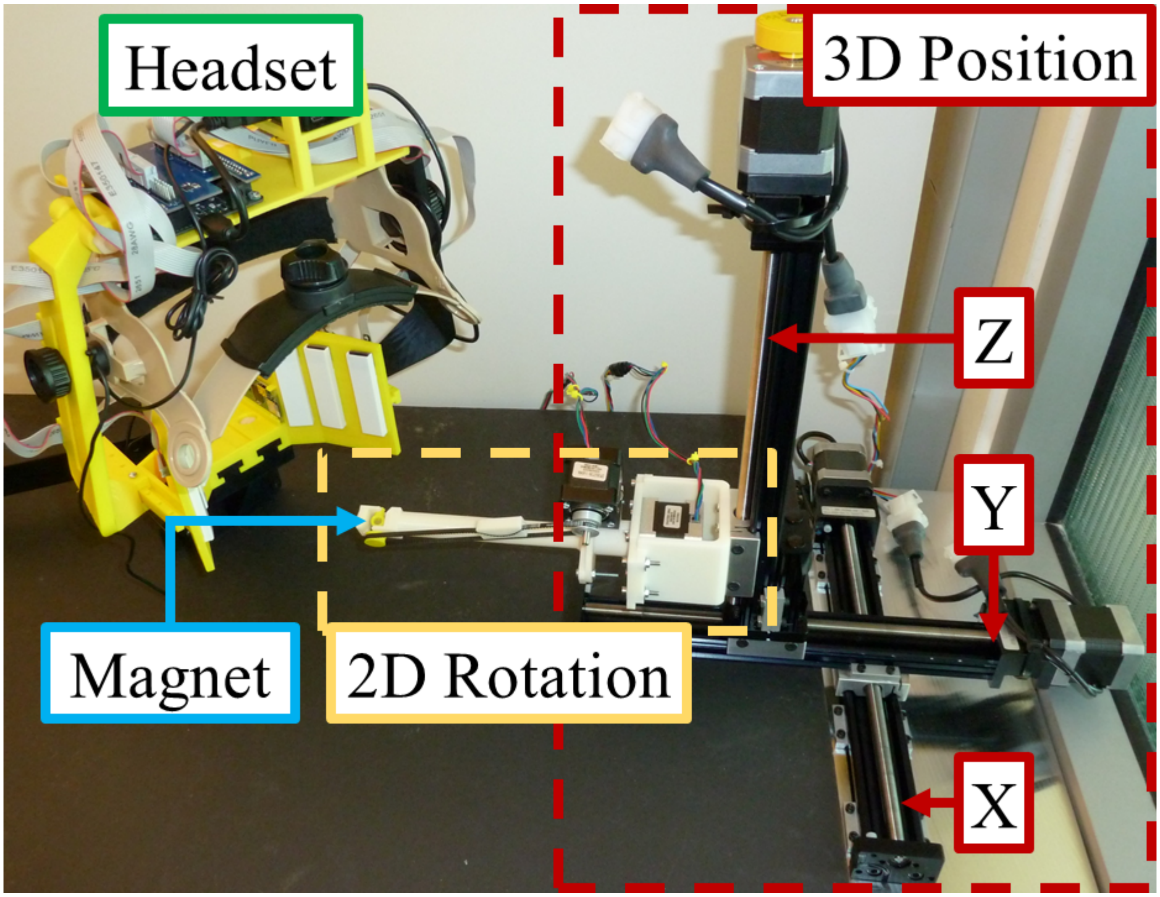


Figure 3.1: Automated 5D positioning setup composed of a 3D linear and a custom-designed 2D rotational stage to place the tracer to a 5D state with a precision of 76 µm and 1.8°.

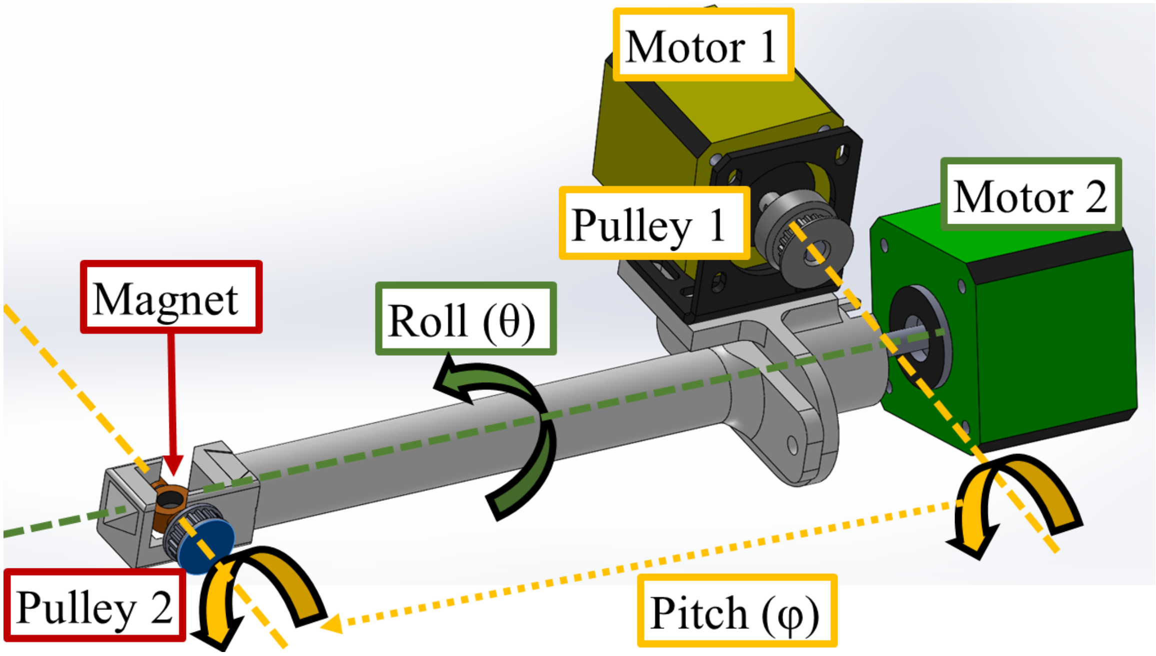


Figure 3.2: 2D rotational stage composed of two stepper motors to rotate a tracer along its pitch and roll (timed belt not shown), and a custom-designed and 3D-printed holder.

Placing the tracer at a desired 3D spatial (X, Y, Z) and 2D angular (roll: θ, pitch: ϕ) states in a controlled manner is a challenging problem. First, the positioning setup must be able to change any of these 5 parameters independently from each other. There are commercially-available robotic arms that can perform such tasks with great accuracy [38], but they are highly expensive, cumbersome, and will hardly be able to maneuver in the tight space inside our headset's working volume. Additionally, a critical constraint is that no magnetic materials should be in the vicinity of the magnetometers (~15 cm) to avoid corrupting their readings of the tracer. Therefore, we created our own 5D positioning system (Figure 3.1) composed of a 3D linear and a 2D rotational stage that are fully automated because the value of each dimension (X,Y,Z,θ,ϕ) is driven by a dedicated stepper motor that is controlled by a motion control system, sufficiently precise for our application, have a small footprint, are affordable (< $2,000 in total), and are specifically designed for use with a tracer.

The 3D linear stage is based on three motorized XSlides (Velmex Inc., Bloomfield, NY, USA) with an accuracy of 76 µm and repeatability of 0.25 µm. The 2D rotational stage is the main innovation of this setup (Figure 3.2) and enables the tracer to freely rotate about its two axes of rotation (roll and pitch). The third rotation (yaw) is not needed because the tongue cannot freely move around that axis and is the reason the tracer was chosen to be cylindrical. The tracer's pitch is controlled by a timed pulley (#2) driven by a matching pulley (#1) whose rotation is set by the stepper motor #1 (the belt is not shown). These parts are attached to a 3D-printed holder that can be directly rotated by the stepper motor #2, effectively controlling the roll angle of the tracer. All these parts are carefully designed to ensure that the tracer's center is always aligned with the two axes of rotation. The stepper motors have an angular accuracy of 1.8°, and the rotational stage is attached to the 3D linear stage through a custom-designed anchor (Figure 3.1). To avoid any magnetic disturbance to the magnetometer's readings, the parts in the rotational stage are made of non-magnetic materials (e.g. plastic, aluminum), except for the motors (that themselves contain a permanent magnet) which are placed far away from the magnetometers ($>$15cm). The tip of the stage, where the tracer can be found, has a width of 2.5 cm and a thickness of 1.6 cm, which is small enough to be easily maneuvered inside the headset.

The motors of this 5D positional stage are driven by the TinyG multi-axis motion control system (Synthetos, Washington D.C., USA). Because this motor controller can only drive 4 motors, two controllers are connected in parallel with one controller configured to drive 3 motors to control the position of the tracer, and the other controller configured to drive the remaining 2 motors to control the tracer's angular rotation. These controllers receive commands about the desired state values using G-Code commands, with G-code being the language that is the industry standard for automated motion control. A countless number of parameters can be configured in TinyG to control the motion of the stage. The most important parameter for our application is the motion velocity because it provides a trade-off between completion time to traverse the whole trajectory and precision in positioning the tracer. Indeed, a high velocity will speed up completion time but at the detriment of precision. More importantly, if the velocity surpasses a certain threshold, the motors might stall which will corrupt the collected data since there is no close-loop feedback to inform the controller to correct and recover from this problem. Empirically, we found that a velocity of 400 mm/min for the 3D linear stage and 10,000 deg/min for the 2D rotational stage are optimal values that provide the highest speed without stalling the motors.

3.1.2 Trajectories

For each type of PML, a set of three trajectories has been created to train, validate and test the resulting models. In the figures below, each state is represented in blue with the start and end positions in red as a star and triangle, respectively. Only the 3D positions are shown in the figures, but each position is actually composed of many states since the tracer is also rotated.

The objective is to assess the tracking accuracy in a volume broadly similar to the oral cavity. A reference volume for an oral cavity can be estimated to be around 40x40x70 mm^3 for the palatal width (X), palatal length (Y), and facial height (Z), respectively. These dimensions are actually set larger than an average oral cavity because each dimension is calculated using the 3rd quartile (Q3) of the largest reported values in [41] where anthropometric measures were collected on 21 adult speakers. Although the roll angle (θ) of the tracer does not vary much during speech, a range of ±50° is considered because it is difficult to ensure the tracer is attached flat on the tongue and that the user's head is perfectly aligned with the headset's frame of reference. This sets a reference for the following trajectories.

Nonlinear Optimization

The first set of trajectories were created to assess the tracking accuracy of the nonlinear optimization-based PML. At the time when this assessment was performed, the 5D positional stage was not operational yet and thus we were relying on a 4D positional stage which allowed the θ angle to rotate but not the ϕ angle. Therefore, ϕ was fixed at 0°while θ is rotated after each complete traverse in the range of ±50° with a 10° step to reflect a realistic range of tongue twist during speech.

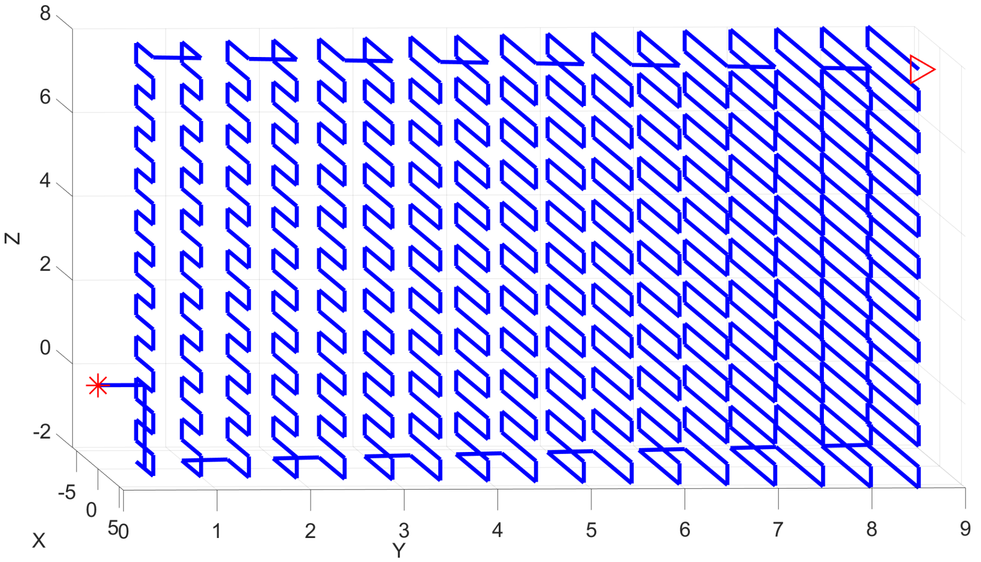


Figure 3.3: Trajectory designed to train a PML based on the nonlinear optimization method (all units in cm).

Figure 3.3 shows the training trajectory that is enclosed in a larger volume than in the reference with a volume of 4x8x10 cm^3 to account for head position offset. This trajectory is composed of 6,000 positions, per θ rotation, that are uniformly distributed with a constant distance of 5 mm between points. This sampling resolution is chosen to collect as many data in a reasonable amount of time.

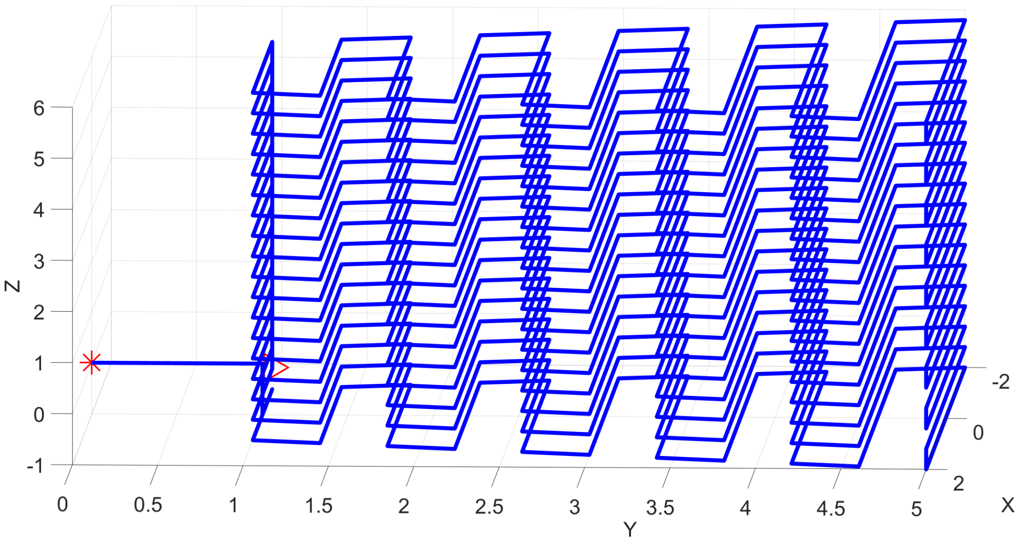


Figure 3.4: Trajectory designed to validate our optimization-based PML (all units in cm).

Figure 3.4 shows the validation trajectory that was designed to follow a similar trajectory than in the training but with the volume being traversed in a different manner. Therefore, the sequence of positions being traversed is different in order to validate the tracking performance when the initial state estimate is different in the optimization process.

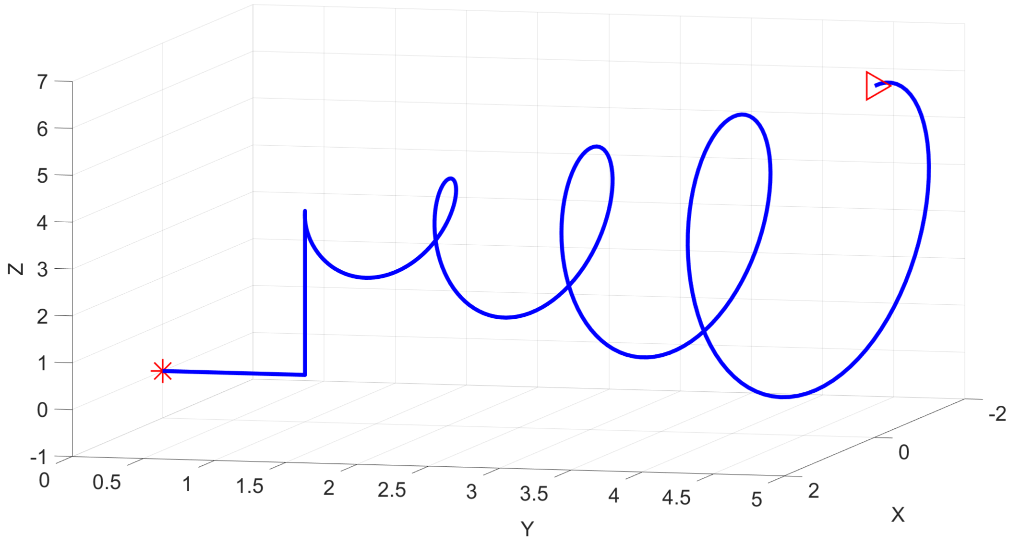


Figure 3.5: Trajectory designed to test our optimization-based PML (all units in cm).

The tongue usually moves in a curved motion during speech. To emulate this behavior, Figure 3.5 shows the testing trajectory as a damped helix (similar to other studies [29, 42, 26]) designed to assess the tracking accuracy with greater resolutions that can be as high as 0.1 mm for positional and 1° for rotational. Therefore, this trajectory emulates a more realistic tongue motion in speech with curves and twists.

Neural Network

The trajectories used for the assessment of the FNN-based PML are similar to the previous ones, with the exception that the pitch (ϕ) was also rotated since the 5D positional stage was fully operational. Because the tongue has more abilities to vary its angle, specially when the tracer is placed at the tip of the tongue, the range of pitch angles is set much wider at $\phi = \pm 90°$ to encompass all possibilities. A wider rotation can be reached by the tongue but is unnatural to its common uses, particularly during speech.

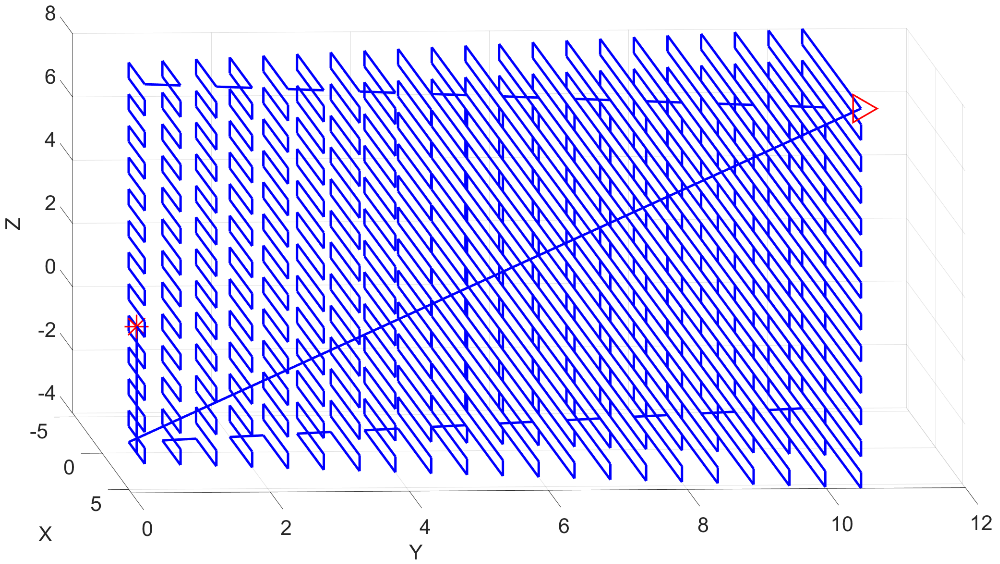


Figure 3.6: Training trajectory composed of 1.7M states traversed by the tracer to train our localization model (all units in cm).

The trajectory shown in Figure 3.6 is a prism that roughly follows the mandibular arc from the gnathion to the tragions [41] with a width (X) expanding from 3 cm to 10 cm, a length (Y) of 10 cm, and a height (Z) of 10 cm. This is a much larger volume than the reference to better train the localization model and to obtain a more comprehensive assessment of the tracking performance in the whole working volume of MagTrack's headset, with 8,100 positions uniformly distributed over the volume and separated by a constant distance of 5 mm. At each position, the tracer's pitch rotates at ϕ = ±90° with a 10° step, and this 4D trajectory is then repeated for the tracer's θ rotating for ±50° at a 10° step. This results in a trajectory composed of ~1,7 million states, which has a more dense and comprehensive coverage of the working volume than any methods used in EMA.

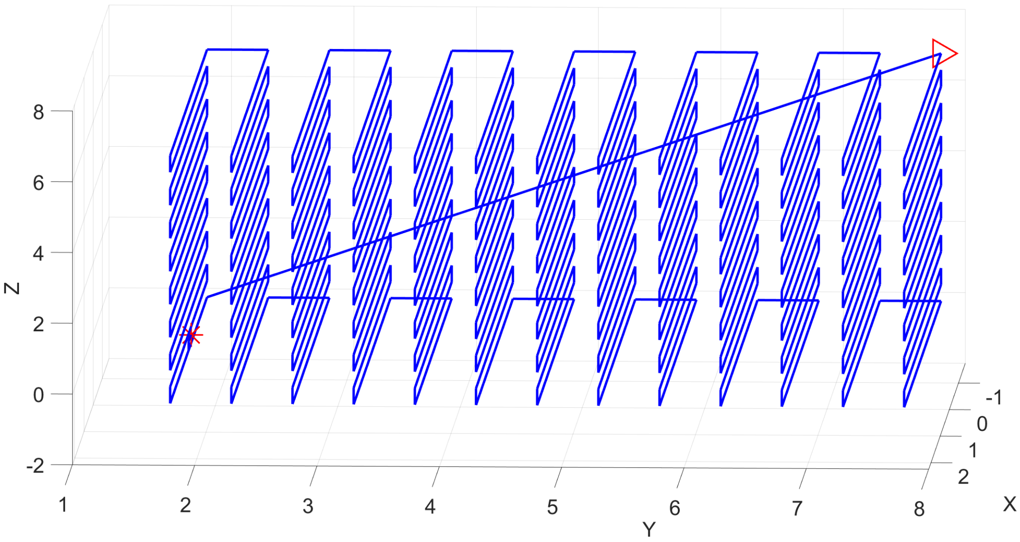


Figure 3.7: Validation trajectory composed of 337,000 states and used to avoid over-fitting during the training of the localization model (all units in cm).

The trajectory in Figure 3.7 covers a volume closer to our reference at 4x6x7cm^3 with the only difference that the length dimension (Y) is 1.5x larger to account for the extension/retraction of the tongue when producing some phonemes. This volume is also uniformly sampled with positions placed at every 5 mm, similarly to the training trajectory. However, to serve as a validation against over-fitting, none of its 5D states are equal to the ones in the training set by offsetting all positions by 2.5 mm (i.e. placed in between adjacent training positions) and rotational angles by 5° (ϕ: ±85°, θ: ±45°), resulting in a total of 337,000 states.

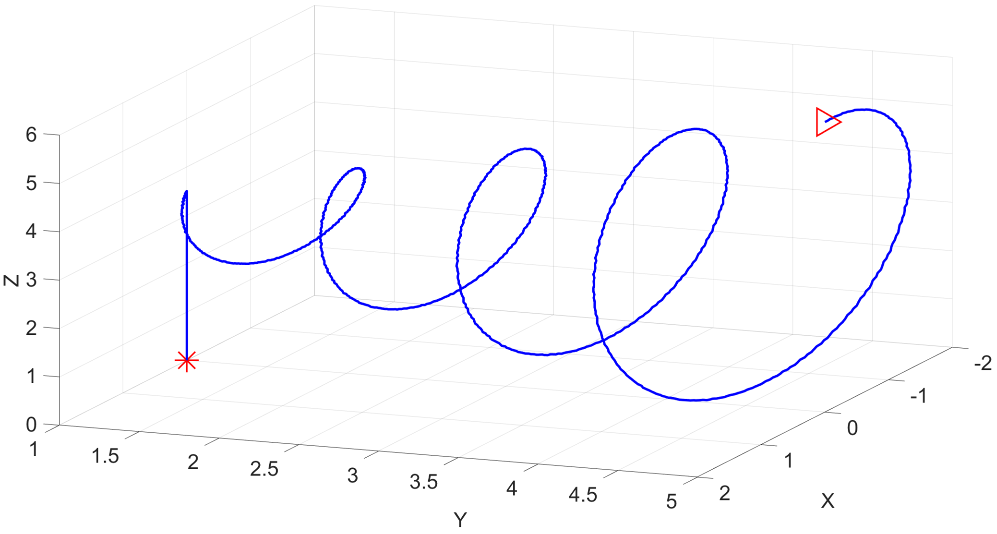
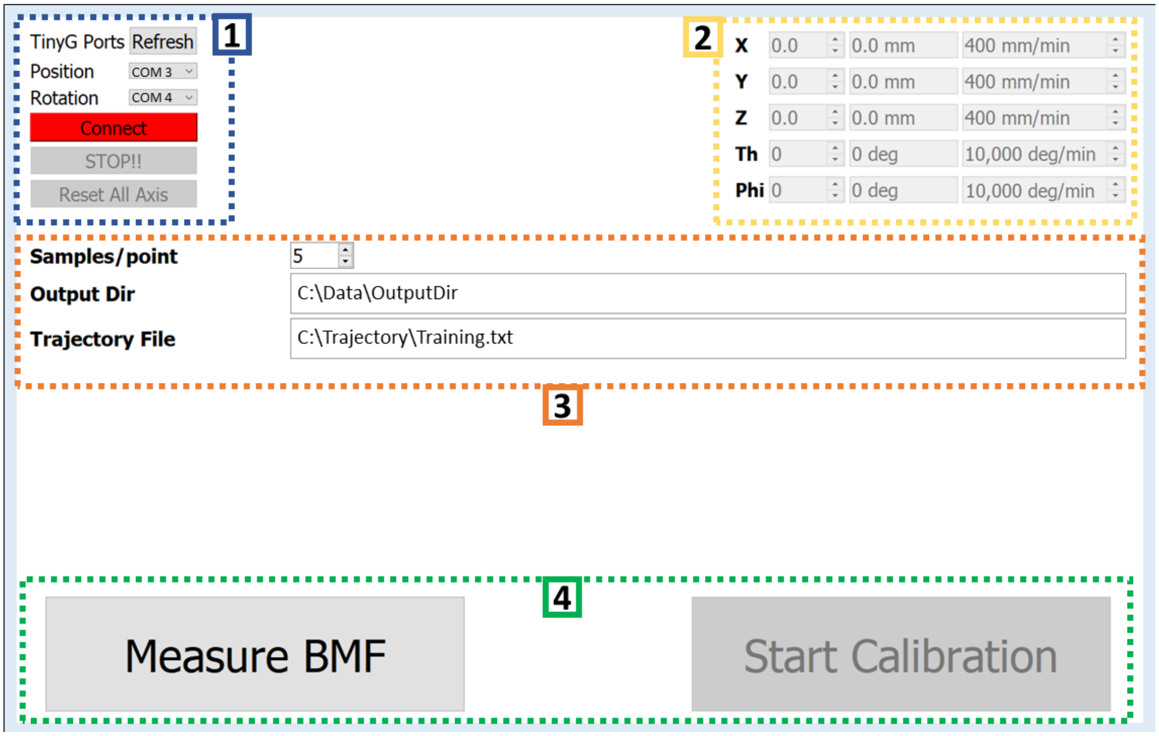


Figure 3.8: Test trajectory composed of 1,000 states used to assess the tracking accuracy on unseen states (all units in cm).

The testing trajectory shown in Figure 3.8 is similar than the one designed for the non-linear optimization-based PML. However, this testing set is composed of 1,000 states sampled in the following ranges (mm): X=[5, 20], Y=[10, 50], Z=[5, 30], θ=±50°, and ϕ=±90°. As opposed to the other trajectories, all parameters are changing between consecutive states and 100 magnetic samples are recorded for each state to better assess the robustness of our tracking against sensor noise. This results in a dataset of 100,000 samples.

3.1.3 Calibration Program



Position COM 3
Rotation COM 4

X	0.0	0.0 mm	400 mm/min
Y	0.0	0.0 mm	400 mm/min
Z	0.0	0.0 mm	400 mm/min
Th	0	0 deg	10,000 deg/min
Phi	0	0 deg	10,000 deg/min

Samples/point 5
Output Dir C:\Data\OutputDir
Trajectory File C:\Trajectory\Training.txt

Figure 3.9: User interface of the calibration program that is designed to load a trajectory, synchronize tracer placement and magnetic reading, and save data into a file for later use.

The calibration program is the main program that orchestrates all the different parts of the data collection setup. It is developed in C++ using the Qt framework to facilitate the design of its user interface (Figure 3.9). During boot up, the program attempts to find the communication ports where the magnetic reading module (i.e. the Mojo board) and the motor controllers are connected to. An error message is shown if any of them are not found, otherwise the user interface is displayed and the configuration of the data collection can performed through the following widgets:

- **Initialization of the motor controllers (1)**: Because the two controllers are exactly the same, the program cannot automatically determine which one is controlling the tracer's position as opposed to its rotation. Therefore, the COM ports can be manually switched in this widget if need be. The "Connect" button starts an initialization of the motor controllers where all important motion parameters are set (e.g. velocity, jerk, mapping between motion axis and motor). When the ”Stop” button is clicked, an emergency procedure is executed to stop the motion of the 5D positional stage and prevents any further movement until the program is restarted. The ”Reset All Axis” button sends an command to the controllers to consider the current state of the 5D stage as the origin. The need for such coordinate reset is explained below.

- **Manual control of the tracer's state (2)**: MagTrack's headset has its own FoR (referred as the global FoR) in which the coordinates of the tracer are described in. Although this FoR can be arbitrarily chosen, it must remain the same for the localization to provide correct estimations of the tracer's state. Therefore, before starting the data collection, the tracer must be placed at the origin of this FoR. This procedure has to be performed manually because the headset is not always placed at the same position from one session to another. This widget enables the operator to manually control the motion of the 5D stage by setting the desired position and orientation of the tracer until it reaches the origin of the device's FoR. Once done, the ”Reset All Axis” button is pressed to reset all values to zero. As a side note, the velocity of each motor can be dynamically changed in this widget. It can be useful to increase the velocities temporarily when the tracer is being moved to the origin since a lower precision, and even stalling, can be manually corrected while the advantage is a significant decrease in completion time for this procedure. The velocities are then reset back to their default values before starting the actual data collection.

- **Configuration of the data collection parameters (3)**: To increase the robustness of the localization model against signal noise, the number of samples that are recorded for each individual state can be set. This enables a sampling of the signal noise to be included during training and testing of the localization model. The output directory is the folder where all files pertaining to this session are saved. Lastly, the path to the trajectory file is provided in this widget.

- **Execution of the data collection (4)**: As explained in section 2.1, the background magnetic field must be measured and subtracted to the magnetic measurements. Thus, before starting the actual data collection, the tracer is removed from the 5D positional stage and the button ”Measure BMF” is selected to collect 100 magnetic samples (1 second). The BMF is then estimated for each axis of each magnetometer as the average value across all the recorded samples. Finally, the tracer is placed back into its holder and the operator selects ”Start Calibration” to begin the actual data collection in which the tracer traverses the input trajectory while its magnetic field is recorded by the magnetic capture module.

3.2 Results

The metrics selected as a measure of tracking accuracy is the positional error (E_p) which indicates the distance error, in the 3D space, between an estimated and a target position:

$$E_p = \sqrt{(x_e - x_t)^2 + (y_e - y_t)^2 + (z_e - z_t)^2}$$

with $(x, y, z)_e$ the estimated position by the PML, and $(x, y, z)_t$ a target position from the input trajectory.

3.2.1 Nonlinear Optimization

As explained in section 2.1.1, the objective of the training for this localization method is to find the calibration parameters shown in equation 2.5. There is a total of 432 parameters to estimate which are split into 18 parameters that are unique to each of the 24 magnetometers: 9 for the 3x3 gain matrix G_i, 3 for the offset O_i, 3 for the magnetometer's position s_p and 3 for its rotation angles s_a. The gain and offset are used to transform the raw magnetic recordings $\overrightarrow{\boldsymbol{B}}_i^{raw}$ into a measurement field $\overrightarrow{\boldsymbol{B}}_i^{meas}$ (equation 2.7) that should ideally be equal to the theoretical field (equation 2.4). Additionally, the position and orientation of the magnetometers are slightly different for each device and need to be estimated with great accuracy for the theoretical model to be valid (equation 2.6).

The calibration procedure resembles the one shown in the equations (2.5)-(2.6)-(2.7) with some important differences. First, the state of the tracer is known during calibration since it is accurately set by the 5D positional stage from the input trajectory. Second, an error function is defined for each magnetometer and optimized over the measurement samples:

$$E\ (G, O, s_p, s_a)_i = \sum_{j=1}^{N} \left\| \overrightarrow{\boldsymbol{B}}_j^{meas} (G, O)_i - \overrightarrow{\boldsymbol{B}}_j^{estim} (s_i, s_a)_i \right\|^2$$

where N is the number of magnetic samples collected along the trajectory and i is the magnetometer's index (1-24). The inputs to $\overrightarrow{\boldsymbol{B}}_j^{meas}$ and $\overrightarrow{\boldsymbol{B}}_j^{estim}$ are the calibration parameters whose optimal values are the ones that generates the lowest error E_i. Empirically, the Levenberg-Marquardt optimizer is found to generate lower errors, and thus more optimal values of the calibration parameters, than with Nelder-Mead.

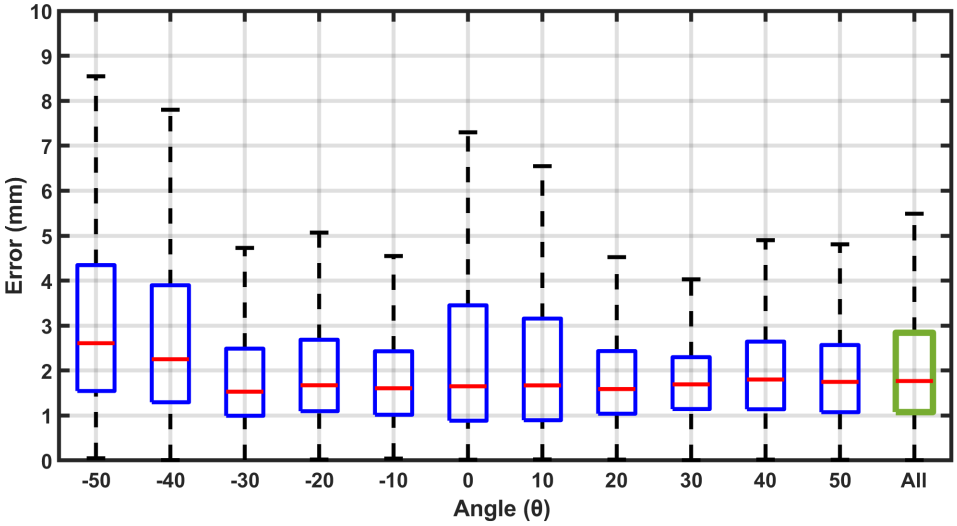


Figure 3.10: Box plots of the positional errors shown for the training set and split across each rotation of θ. The last box plot shows all the errors combined.

Figure 3.10 shows the positional errors for the training set. The errors are split by values of the rotation angle (θ), and we can observe that the median error is fairly similar at ~1.5 mm across all angles except for -40° and -50°. There is no obvious explanation because if there is an actual reason for the error to be higher with increased rotation, this should also be observed for 40° and 50° since the magnetic capture module is symmetrical. To be noted, these two rotations are the last ones to be traversed, thus, a disturbance might have occurred in the form of an added component to the BMF from a magnetic field-generating object placed near the system, and/or the headset was accidentally slightly moved. Fortunately, this issue had no significant impact on the overall distribution of the training errors as reported by the last box plot with a median error of 1.8 mm and a Q3 of 2.9 mm.

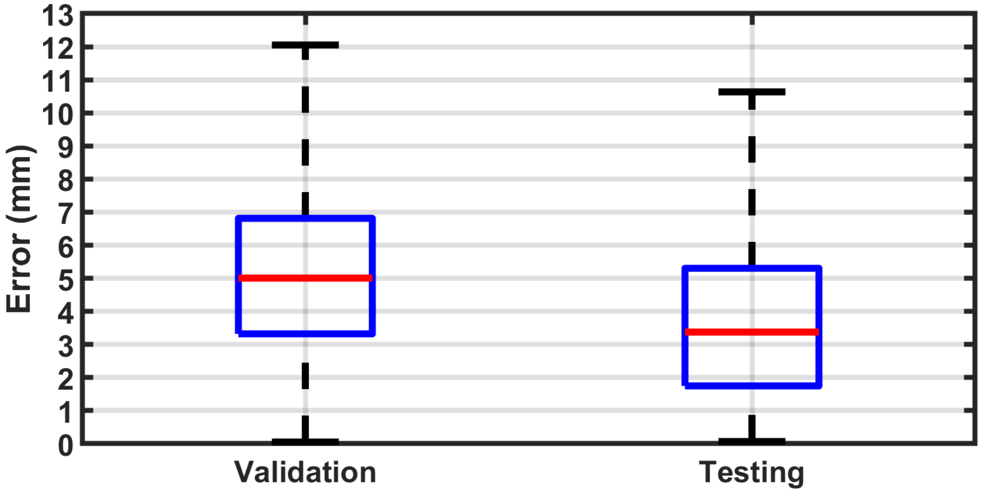


Figure 3.11: Box plots of the positional errors shown for the validation and testing sets.

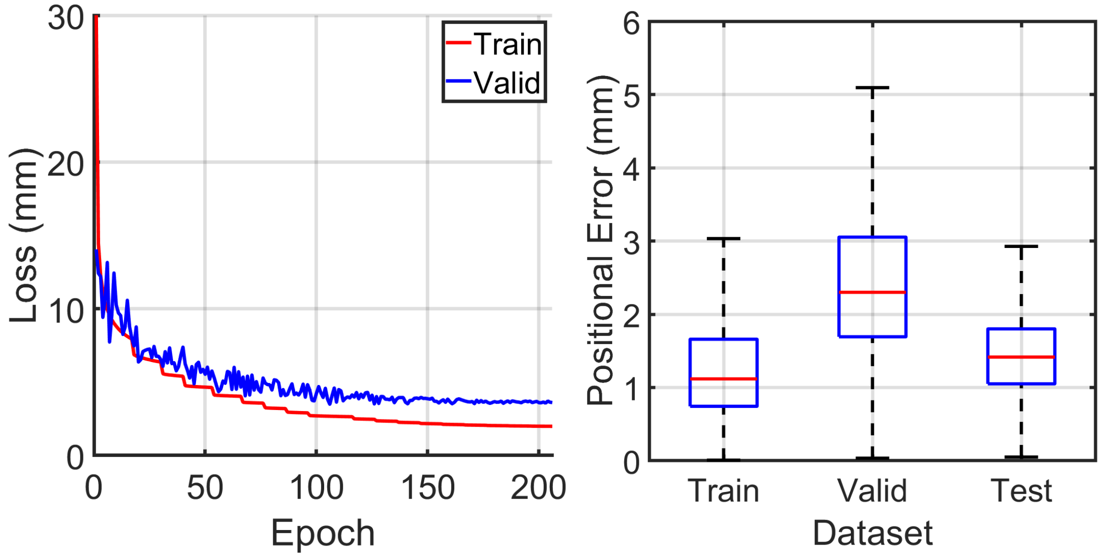
The positional errors for the validation and testing sets are shown in Figure 3.11. As expected, the errors are higher than in training with median errors of 5 mm and 3.4 mm as well as third quartiles valued at 6.8 mm and 5.3 mm for the validation and testing, respectively. The errors in the validation set are higher compared to the testing because the trajectory covers a wider volume with a denser coverage. Although over-fitting is observed since the validation errors are more than twice higher than the errors reported during training, the localization generalizes better on the testing set even though the trajectory traversed by the tracer is more complex with curves and twists.

3.2.2 Neural Network

The training of the FNN was performed using the Keras library [43] with TensorFlow as the backend [44]. The training set was shuffled before each epoch (i.e. one iteration over the entire dataset) using a repeatable sequence, and the hyper-parameters were found heuristically: RMSprop as the optimizer [45], a learning rate of 0.01, and a loss function set as the mean across all samples of the following error function (L) calculated for each sample (i):

$$L_i = |x_e - x_t| + |y_e - y_t| + |z_e - z_t|$$

with $(x, y, z)_e$ the estimated position by our model, and $(x, y, z)_t$ the target position. A model is selected by its ability to predict with the lowest 3rd quartile (Q3) validation error while being the simplest architecture. We chose to focus on Q3 as a metric because it sets a more realistic, though stricter, assessment of performance and robustness of our PML by ensuring that the majority of errors (75% by definition) are below that value.



(a) Loss

(b) Positional Error

Figure 3.12: Results of the training of our FNN-based PML model with a) learning behavior over the training and validation datasets, and b) positional errors for our three datasets.

Empirically, we found that a FNN with the following hyper-parameters provides the most optimal tracking accuracy: 3 hidden layers, 100 neurons per layer, and hyperbolic tangent (tanh) as the activation function. The values of the loss function over the training time (epoch) are provided in Figure 3.12a and show that the model is only slightly over-fitting the training set (red curve) since the gap with the validation set (blue curve) is small. The over-fitting can also be observed in the positional errors (Figure 3.12b) where Q3 for the validation set (3 mm) is higher that the training (1.7 mm). However, this has a reduced affect on the testing set with a Q3 of 1.8 mm and a median error of 1.4 mm.

3.2.3 Discussion

It is remarkable that our wireless localization method can reach a tracking accuracy that is similar to EMA which is considered the gold standard for tongue tracking. Indeed, a wired tracking method such as EMA is expected to perform much better than its wireless counterpart because the magnetic field intensities can be increased to improve the signal-to-noise ratio, resulting in a more accurate tracking. In contrast, EMA is much more cumbersome, expensive, and not practical for applications outside of a research lab. In [10], the research team has evaluated the tracking accuracy of many different models of EMA and showed that the positional errors fluctuates between 0.3 mm to 2.18 cm depending on the location of the probes in the measurement volume (30x30x30cm^3). For our FNN-based localization, our results show significantly more consistent errors between 0.05 mm and 2.9 mm for the testing set, and in the worse case, 0.03 mm to 5.1 mm for the validation set. Even the results of the optimization-based localization show errors that are more consistent with the worse case dataset (i.e. validation) having errors between 0.05 mm and 1.2 cm. These results show the potential of MagTrack to be a suitable alternative to EMA by providing a similar tracking accuracy while having the potential to be used by a wider population outside research labs.

Comparing our two methods of localization, it is clear that the FNN model performs significantly better than the nonlinear optimization method. Actually, the difference in tracking accuracy is even wider because the trajectories for the FNN-based localization include the rotation of the pitch angle (ϕ) while it remained constant at 0° during data collection for the nonlinear optimization method. The tracking errors increase with an added dimension of motion because the localization model becomes more complex and thus more difficult to approximate. Also, the FNN is faster to execute because it is non-iterative, which would allow our PML to run on devices that are restricted in terms of computing resources such as mobile devices. More importantly, the FNN model doesn't require the knowledge of the previous estimated state. Therefore, the accuracy remains the same regardless of the order of the sequence of positions traversed by the tracer. Indeed, in the nonlinear optimization assessment, the validation errors are significantly higher than in the training set. However, the main difference between the two trajectories is that the tracer follows a different path while traversing the same positions. Therefore, the FNN-based PML is not only more accurate but also more robust to variations in the tracer's motion.

3.3 Future Work

Although MagTrack's tracking accuracy is satisfactory since it is similar to that of EMA, theses results should only be seen as an indication of the potential of the PML method to provide suitable localization performance for tongue tracking applications rather than definite results. In this section, improvements to the current system are discussed.

3.3.1 Localization Method

The choice for a neural network architecture has been purposefully restricted to a fully-connected feedforward network, same activation function for all neurons, and same number of neurons per layer. This restriction was put in place to reduce the number of hyper-parameters to tune in order to obtain localization results faster. Indeed, the main objective was to use a rapid-iteration paradigm to validate as a proof-of-concept that a neural network could localize the tracer in our system. Therefore, many other hyper-parameters could be tuned and types of machine learning architectures could be implemented to find out if there is a more optimal model that can further reduce the positional error. For instance, more complex models could consider prior knowledge of the tracer's state to improve its current estimation (e.g. recurrent neural network). Also, ensemble models such as bagging or gradient boosting, could be studied to verify if using a combination of weaker models can result in a better overall estimation of the tracer's position. Additionally, the two PML methods assessed in this work could be combined by leveraging the domain knowledge provided by the PSDE and the unmatched capability of deep learning to model highly complex and nonlinear processes.

3.3.2 Trajectory

Increasing the resolution of positional sampling from 5 mm to 1 mm could help the PML training to better discriminate between the small differences in the magnetic field signature of adjacent points. But, this increase of resolution will also increase the duration of the data collection by 25, from 1 week to ~6 months for the training set alone. Thus, a smarter sampling of the headset's working volume must be established and will likely rely on a high resolution in the center, where the magnet can be mostly found, to a low resolution on the periphery.

3.3.3 Data Collection Setup

One of the most challenging tasks in collecting data for this assessment is to reach a perfect positioning of the tracer. Because our objective is to reach tracking errors in the range of few millimeters, the 5D positioning stage must then be capable of placing the tracer with a precision that should be much below these values. This is difficult to do because any mechanical system has imprecision. For instance, friction during movement creates a small error in placement that adds up over time. As a coarse assessment of the influence of friction, the tracer is placed back to its initial state after a data collection session is completed. It was not uncommon to observe errors in placement that can reach few millimeters and degrees. Although the manufacturer of the 3D linear stage states that the accuracy is 76 µm and repeatability is 0.25 µm, these values must have been found with no load present on the stage. In contrary, the 2D rotational stage in our system adds a load to the 3D stage that is not only adding significant weight but also the weight is not evenly distributed. This could generate a slight misalignment of the motion axes and thus could be responsible for lower accuracy and repeatibility.

In the current version of the 5D stage, the motion control is performed in open-loop, i.e. there is no feedback about the actual motion being taken into account by the controller to correct the actual position of the tracer compared to a target position. One possible solution to this problem is to add motion sensors on each motor's axis to enable a close-loop feedback mechanism for the controller to increase the accuracy in placement by having a better estimation of the actual tracer's state.

Another common concern with mechanical parts is that their dimensions are not exact, particularly for 3D-printed parts. Indeed, 3D printing is a complex mechanism that is highly sensitive to a variety of parameters such as the misalignment of the 3 axes of motion, bending and cleanliness of the printing bed, room temperature and humidity, quality of the plastic filament, temperature uniformity throughout the part while printing, vibration level of the printer, just to name a few. Even though these parameters can be better controlled with some professional 3D printers that typically have an enclosure to shield the printing environment from the outside, they are not perfect and our team does not have easy access to these type of printers. This issue of imprecision in the dimensions of the parts is not a problem in most applications, but in our case, this results in misalignment of the rotation axes of our 2D rotational stage. This misalignment adds more errors in the position of the tracer that are relative to its rotation angle position. Although this error can be measured to generate a model of the position offset that can be added to the original position, this creates more complexity to the system and is also not perfect because the rotation angle is not exactly known due to the same problems explained in the paragraphs above.

Lastly, the center of the tracer is not aligned with the rotational axis for θ because the belt-pulley system of the motor #1 that drives the ϕ angle (Figure 3.2) is pulling the tracer towards the pulley #1. This creates a tension force that bends the cylindrical rod that transfers the θ rotation from motor #2 to the tracer. This adds an offset to the tracer's position that is relative to the angular value of θ. Fortunately, there are few solutions that we can implement in the future to overcome this issue. One solution is to have a tensioner on the belt to adjust its tension in a way to prevent the rod from being bent while ensuring that enough tension is created to properly transfer the rotation from the pulley #1 to #2. Another solution is to print the 3D parts with a stronger material, e.g. carbon-fiber-reinforced Onyx, that can sustain such tension.

CHAPTER 4

TOWARDS THE CREATION OF A BASELINE OF PHONEME LANDMARKS

4.1 Motivation

Verbal communication is a key part of everyday activities such as working, learning, and socializing. Brain injuries, neuromuscular diseases, and developmental deficits can lead to speech sound disorders (SSD) which are responsible for the decreased intelligibility of an individual, resulting in a significant reduction in quality of life. In the United States alone, SSD affect 5 million adults, and nearly 1 in 12 children [46]. SSD are characterized by improper positioning and motion of the articulators due to reduction in the range of motion, lack of coordination, and strength (dysarthria), and/or their motion planning (apraxia of speech). Following diagnosis, to improve intelligibility, those with SSD undergo a series of therapy sessions, led by speech-language pathologists (SLP), who identify the root causes of errant speech and prepare a treatment plan. As already mentioned, the tongue is the most important articulator and plays a key role in producing most phonemes [1]. Thus, it is not surprising that some of the most common SSD are due to improper tongue motion [47]. A key challenge is that unlike lip gestures, which can be seen, and voice, the ultimate outcome of speech sound production, which can be heard, tongue motion is hidden from sight to both SLPs and their patients. Therefore, the ability to see and analyze tongue motion could make a significant difference.

In today's therapy sessions, SLPs are limited in the treatment tools available to them to guide patients towards proper placement of their tongue in target locations. These tools are usually tactile markers that include tongue depressors, a gloved finger of the SLP, tongue models, or even flavored lollipops in pediatric practices [48]. This is quite inefficient because patients still cannot practice while the target location is being marked, do not know how far they are from the target while speaking (also known as the *place of articulation*), and cannot coordinate proper timing of their tongue motion (known as *manner of articulation*). Similarly, the SLPs' assessment of improper tongue placement by their patients during diagnosis or attempted corrections during therapy sessions are inaccurate and subjective at best because they have no idea how the patients' tongue is moving during examination or while practicing.

Consequently, the lack of key salient features in the treatment tools used by virtually all SLPs has a negative impact on the therapy outcomes because these tools cannot track, record, or provide any kind of meaningful feedback about the tongue motion. SLPs are left with no choice but to rely on their perception of articulation and prior experience in their subjective assessment of speech performance, and subsequent plan for therapy. These are not only subject to internal biases but also inconsistent, and challenging to follow, document, and course-correct, if necessary. Additionally, the lack of feedback affects the quality, quantity, and efficacy of patients' practicing in the SLP office or at home because they are also unsure of their performance, any progress they are making, or lack thereof. This leads to a reduction in their compliance with the therapy, resulting in a diminishing rate and amount of recovery.

Therefore, there is a need for a tool that would provide access to tongue motion during speech to enable a more objective assessment of patients' progress by SLPs, and allow patients to better practice by providing quantifiable measures of their speech performance.

4.2 Objectives

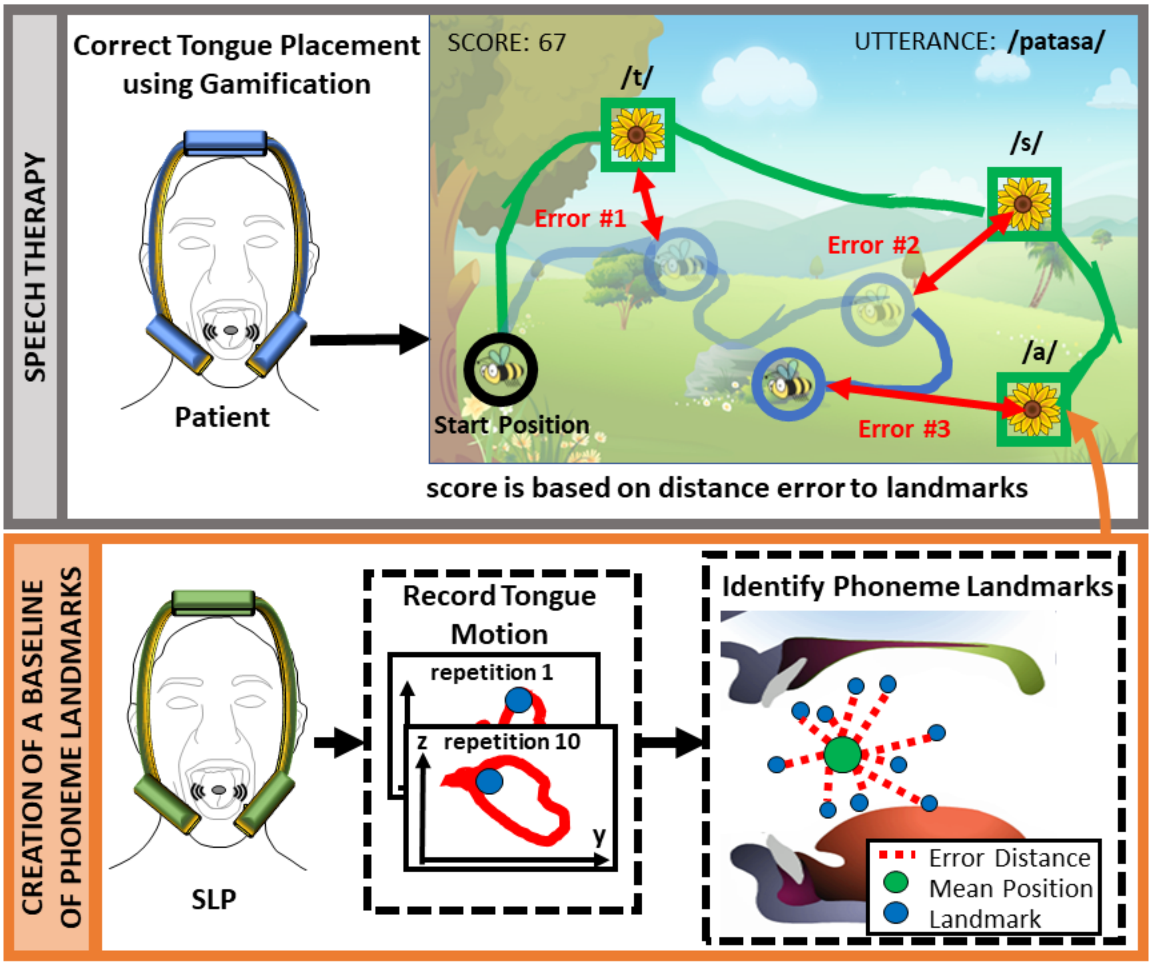
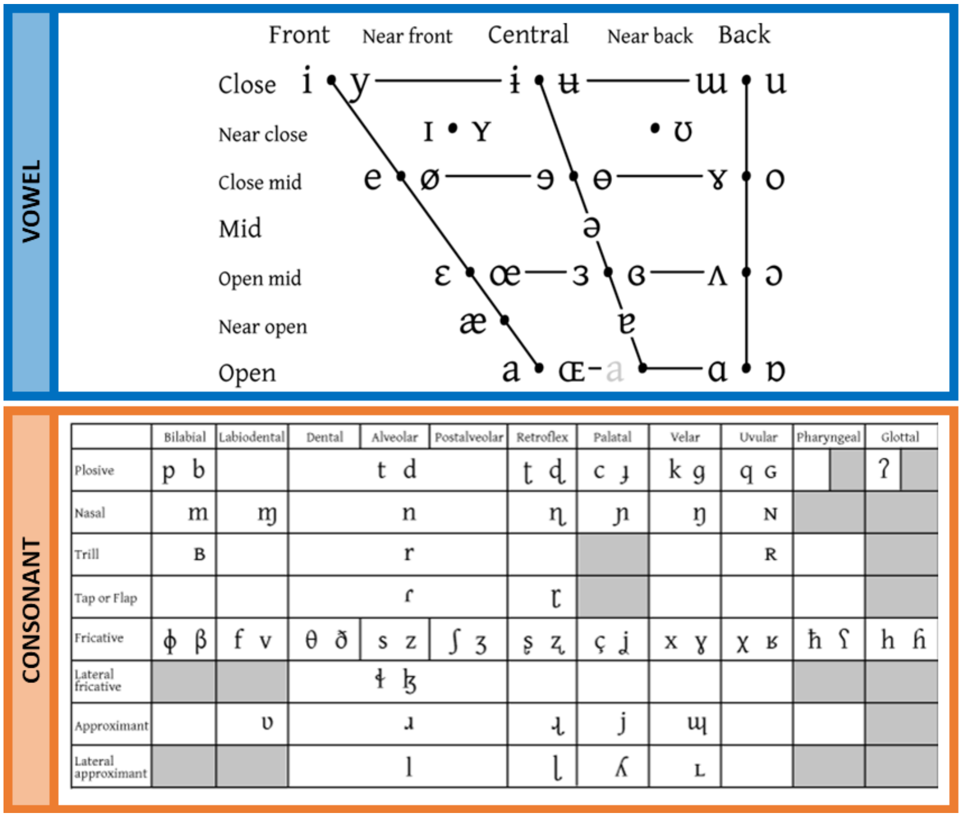


Figure 4.1: Overview of our research objective for the treatment of SSD. The correction of tongue placement is gamified thanks to visual targets (top). The patient would improve their speech production by increasing their score which is calculated from the distance errors between the patient's tongue placement and the visual targets. These targets are representative of the phoneme landmarks that were identified by analyzing the tongue motion of SLPs during their production of these phonemes (bottom).

Previous studies have shown that combining real-time visualization of tongue motion and overlaid visual targets for tongue placement can improve the quality of the produced sounds [49, 50]. Aligned with these findings, the overall objective of our research is to develop such visual feedback through the gamification of tongue placement. As illustrated in Figure 4.1 (top), the aim is to display the landmarks for the phonemes that the patient has difficulty producing (shown as flowers in our illustration). The patient will then be asked to place her tongue (shown as a bee) as close as possible to these landmarks during her attempt to produce the target utterance. A score will be generated based on the distance errors between the phoneme landmarks and the patient's current tongue position. This score will serve as an objective measure of speech performance which would reinforce good practice and result in improved recovery.

To develop this visual feedback, MagTrack is needed to not only provide the tongue position in real-time but also to find where the phoneme landmarks are located in the first place. Indeed, during the production of a phoneme, a landmark is the specific position of the tongue in the oral cavity that modulates the air to generate the intended sound. The problem boils down to identifying the location of the landmark for each phoneme of interest. In the current version of MagTrack, a phoneme landmark is placed manually by the SLPs as a circular mark whose color changes from red to green when the patient's tongue is within a set distance from the landmark's center (Figure 2.13). This manual marking of phoneme landmarks relies on the SLP's knowledge of proper tongue placement for each phoneme. This is not only subjective but will also create inconsistencies from one SLP to another because there is no realistic baseline of phoneme landmarks that has been produced so far. Indeed, SLPs rely on the International Phonetic Alphabet (IPA) chart as a reference (Figure 4.2) which only provides coarse and qualitative information about tongue placement. Therefore, researchers have been trying to determine more precisely where the actual landmarks for these phonemes are. Attempts to create such baseline were made previously for a limited set of lingual consonants with an electromagnetic articulograph as a the tongue tracking system [9, 25, 41, 51]. However, these studies were limited to few consonants only.

The objective of this research activity is to evaluate the feasibility of using MagTrack to generate a baseline of landmarks that includes vowels as well as consonants and for which tongue placement is the main contributor to their production. The key challenge of this



	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Figure 4.2: IPA charts for American-English with idealized tongue placement for the vowels (top) and coarse location for consonants (bottom).

research activity is to find a method to validate if MagTrack is capable to generate such baseline because there is no existing baseline of phoneme landmarks to compare against. However, it is assumed that the selected phoneme landmarks are fixed in position in the oral cavity because they were chosen due to the importance of tongue placement in their production. Therefore, if MagTrack is capable of generating such baseline, the positional variability of a landmark can be used as an indirect measure. To evaluate the positional variability, a human study was conducted to record the tongue motion of 10 SLP students that were asked to utter 25 phonemes with 10 repetitions. As illustrated in Figure 4.1 (bottom), a landmark is identified for each repetition, and the final landmark for a phoneme is selected as their mean position. Positional errors are calculated between each of the 10 landmarks (one per repetition) and the final phoneme landmark. More details about our methods are provided in the next section.

4.3 Methods

4.3.1 Data Collection

Ten SLP students at the University of Georgia were recruited based on the following criteria: passed a Phonetic class, have no history of speech disorders, and have no intra-oral magnetic device that would interfere with the magnetometers. To reduce positional variability due to different methods of articulation, all subjects were female and recruited from the Atlanta metropolitan area. The subjects were between the age of 21-36 y.o, and grew up in the state of Georgia (USA) except for subject #10 in Chicago (Illinois, USA). This study was approved by the Georgia Tech Institutional Review Board and carried out at the University of Georgia under the supervision of Dr. Nina Santus. A new and sterilized tracer was used for each subject and placed mid-line on the blade of the tongue using the adhesive (~1 cm from the tip). The subjects were first asked to read the Grandfather Passage [52] to be accustomed to speak with the tracer, and then asked to repeat 10 times each phoneme in the following lists:

- r sound: [ɔr, aɪr, ar, aʊr, ɪr, ɛr]
- vowel: [æ, ɔ, e, ə, ɚ, ɛ, i, ɪ, u,ʊ]
- consonant: [d, l, n, s, t, z, ʃ, ʒ, θ]

The inclusion of 6 different type of /r/ sounds stems from the fact that the produced sounds differ depending on the phonemic context. Therefore, it was decided to include these /r/ variants in order to capture some of these differences. The phonemes were produced in isolation except for the consonants that were followed by a vowel. The subject's voice was also recorded by the headset's built-in microphone, and a clicker was used by the subject to start/stop the recording of each repetition. The subjects were instructed to elongate and articulate their speech, and were allowed to record back any repetition. In an attempt to minimize variation in phoneme production between subjects, a reference audio was played before each new phoneme and a word was displayed to provide a context in which the target phoneme is used. At the end of the recording session, the tracer was detached from the tongue and disposed of.

4.3.2 Data Analysis

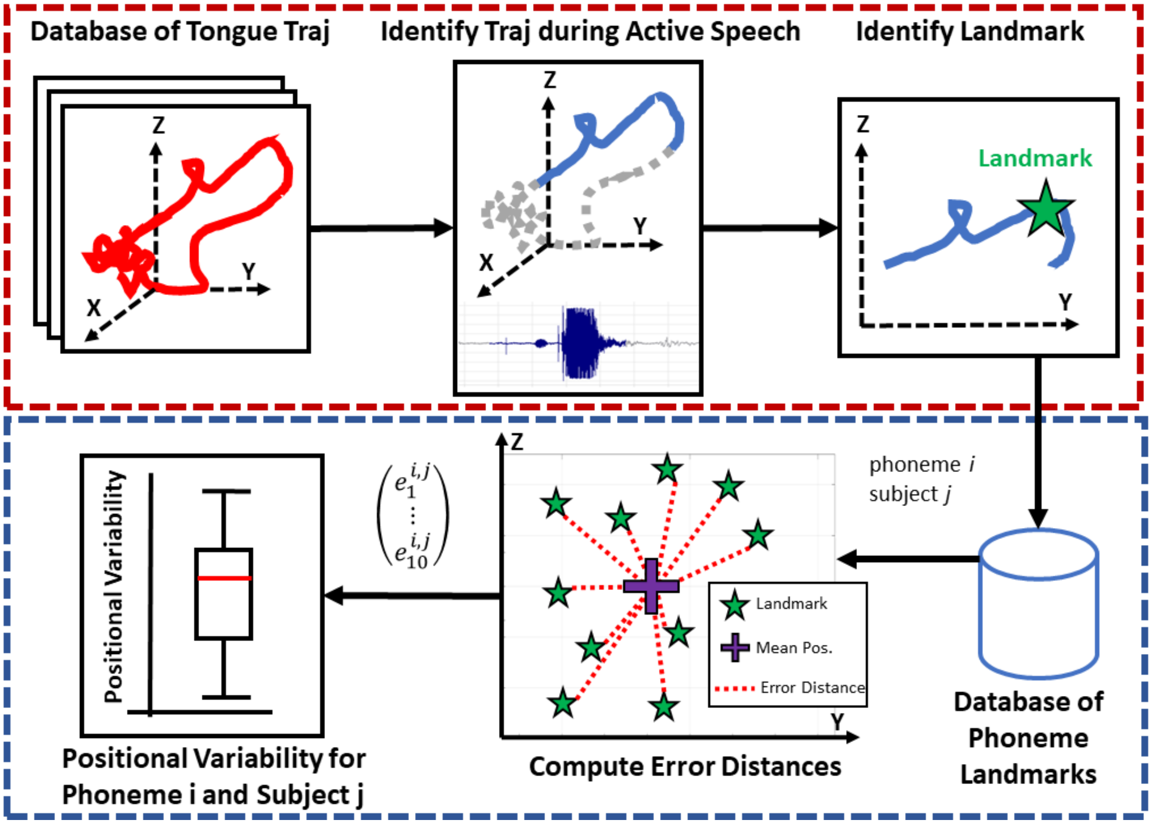
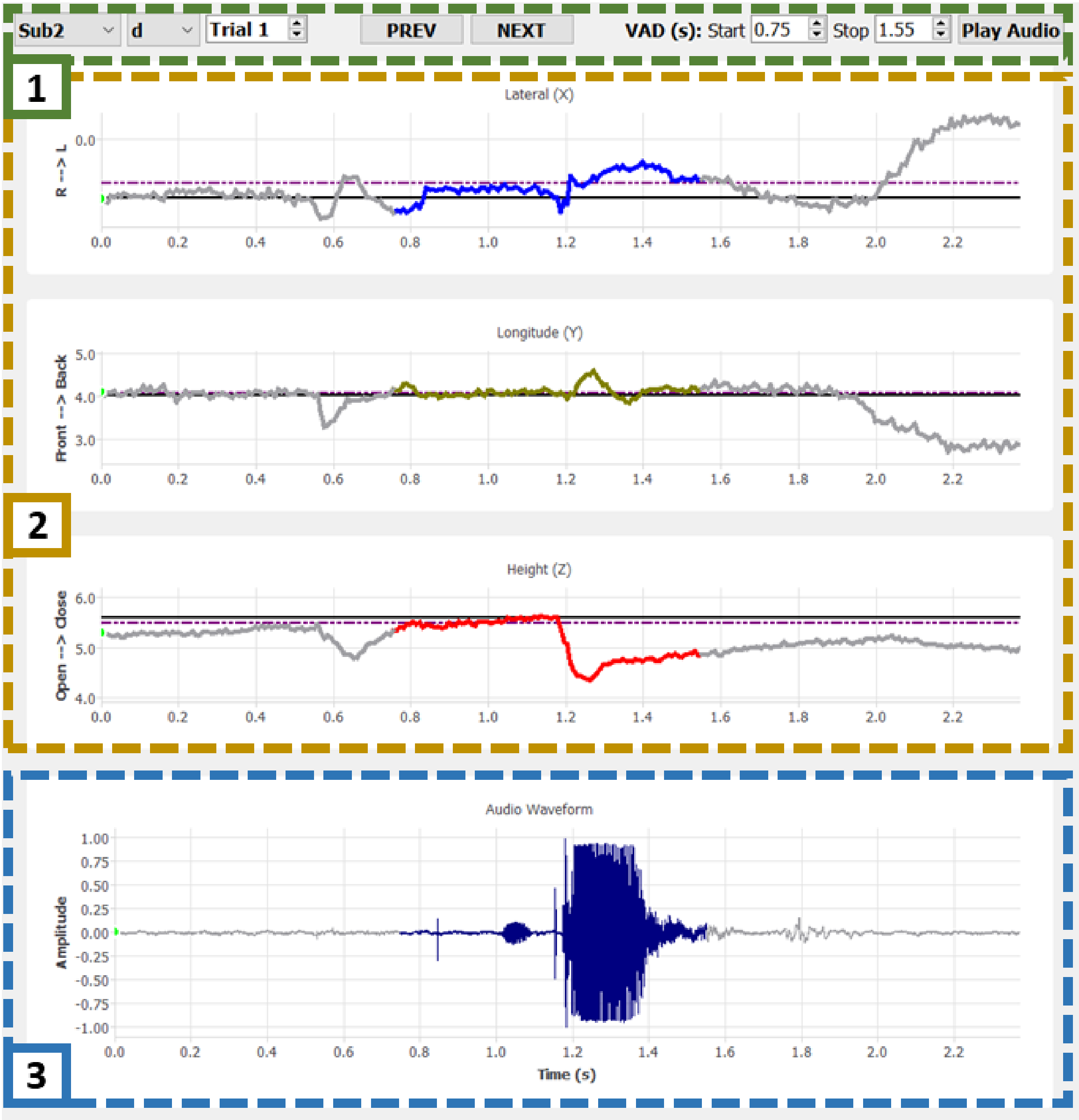

Database of Tongue Traj
Identify Traj during Active Speech
Identify Landmark
Z
Y
X
Landmark
phoneme i
subject j
$\begin{pmatrix} e_1^{i,j} \\ \vdots \\ e_{10}^{i,j} \end{pmatrix}$
Positional Variability
Landmark
Mean Pos.
Error Distance
Database of Phoneme Landmarks
Compute Error Distances
Positional Variability for Phoneme i and Subject j

Figure 4.3: Overview of our data analysis in which the phoneme landmarks are identified from the tongue trajectories that are pre-processed by trimming out its components not recorded during active speech. The positional variability is then calculated for each phoneme as the error distance between a landmark and the mean position estimated from all the 10 landmarks of that phoneme and per subject.

An overview of our data analysis is shown in Figure 4.3. The first step is to identify the landmarks from the 2,500 tongue trajectories (10 subjects x 25 phonemes x 10 repetitions) collected during our human study. In [25, 41], the position of the highest elevation in the tongue trajectory was selected as the landmark. However, this identification can be corrupted in our system by the unintentional head/body motion that can generate a higher elevation in a section of the trajectory not related to the phoneme production. Therefore, another phoneme landmark identification method was selected and is based on the assumption that the phoneme landmark is the position at which the tongue is dwelling. To increase the validity of this assumption, the subjects are asked to elongate their phoneme production in order to obtain a higher density of tongue positions at that phoneme landmark. Consequently, the objective of this first step is to identify the location at which the highest density



Sub2 | d | Trial 1 | VAD (s): Start 0.75 Stop 1.55

Figure 4.4: User interface of the phoneme landmark identification program that displays the tongue trajectory as three time series, one per axis, and its associated voice waveform.

of positions can be found.

Because a large number of trajectories must be analyzed, a program was developed to assist the process of identifying landmarks. Its user interface is shown in Figure 4.4 and is composed of 3 parts: (1) a settings bar where the operator can select a particular trajectory from the database, play the recorded voice, and set the start/stop time markers of the active period of speech; (2) tongue trajectory displayed as three time series (one per axis) with the positions recorded during active speech highlighted in color; (3) voice waveform with active speech highlighted in blue. The period of active speech is highlighted in order to facilitate the visual identification of a landmark by the operator. The 3D position of a landmark is set by clicking on a target point on any of the time series, typically the longitude (Y) or height (Z), where the operator estimates that it is representative of that phoneme (the other 2 values are automatically set based on the time stamp of the selected point). To further assist the operator, a candidate landmark is automatically selected by an algorithm following these steps for each axis: generate a histogram of the positions, and then select the bin with the highest distribution. The operator is free to choose this candidate or select another position as the landmark.

The second step has for objective to generate the positional variability from the database of 2,500 phoneme landmarks (Figure 4.3, bottom). For each phoneme i and subject j, the mean position across the 10 landmarks (one per repetition) is selected as the landmark that is characteristic of that phoneme and subject. Its positional variability is then defined as the distribution of the error e_k^{ij} calculated as the euclidean distance between the final phoneme landmark and each $k \in [1, 10]$ repetition. Specifically, we will focus on the 3rd quartile (Q3) of the distribution since it provides a strict but realistic measure of the capability of MagTrack to generate a baseline of phoneme landmarks.

4.4 Results

There are many different ways to visualize the 250 positional variability (10 subjects x 25 phonemes). As a high-level view of the results, the variability will be aggregated in the sections below and shown by subject, by phoneme, and with all results combined as the final measure of MagTrack's capability to track tongue motion for the purpose of generating a baseline of phoneme landmarks.

4.4.1 Split by Subject

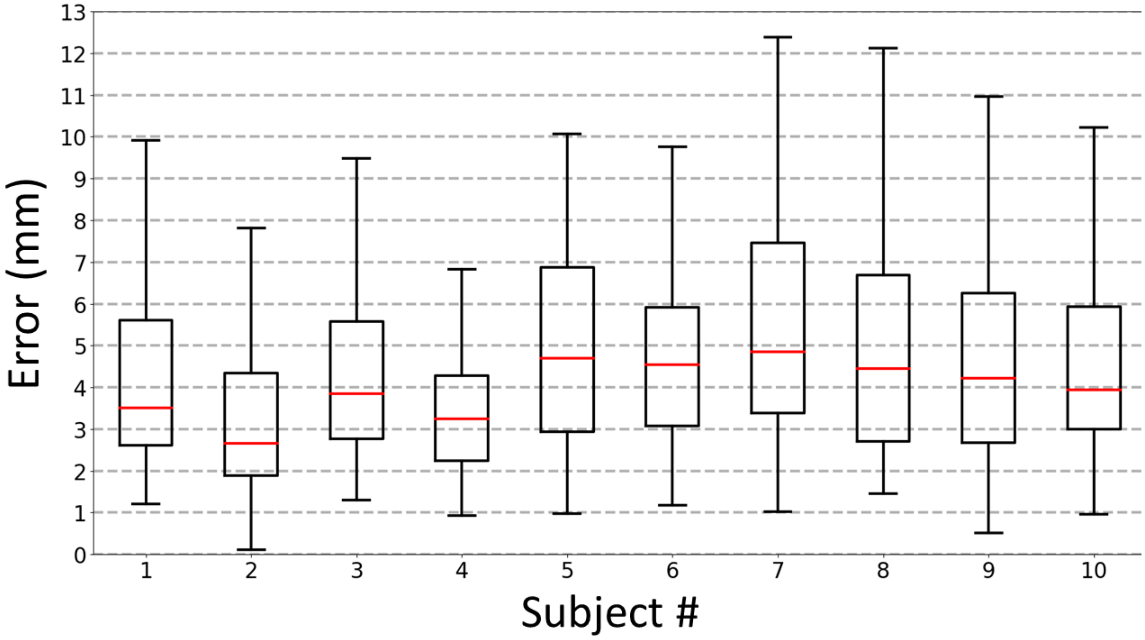


Figure 4.5: Positional variability of phoneme landmarks split by subject.

Figure 4.5 shows the positional variability across all phonemes for each subject. The highest errors are found for subject #7 with a Q3 of 7.4 mm, and the lowest for subject #2 with a Q3 of 4.2 mm. Interestingly, there is no demographically distinctive features that were found between these two subjects: both were born and raised in state of Georgia, are exactly the same age (22 y.o), same ethnicity, and have similar accent (slightly southern). Overall, the first 4 subjects have lower errors and variability than the other subjects. Although one major difference between these two groups is that their data collection sessions were performed in two separate days, with a duration of one week between them, the subject #5 belongs to the first group while her variability is more similar to the second group. There is no apparent distinction between the group composed of subjects #1-4 and #5-10 that would explain such difference in variability since the same system was used, the data collection occurred at the same location, and the same researchers were conducting the recording session. However, a smaller difference can be observed on the median values of the errors with the lowest being at 2.8 mm and highest at 4.9 mm.

4.4.2 Split by Phoneme

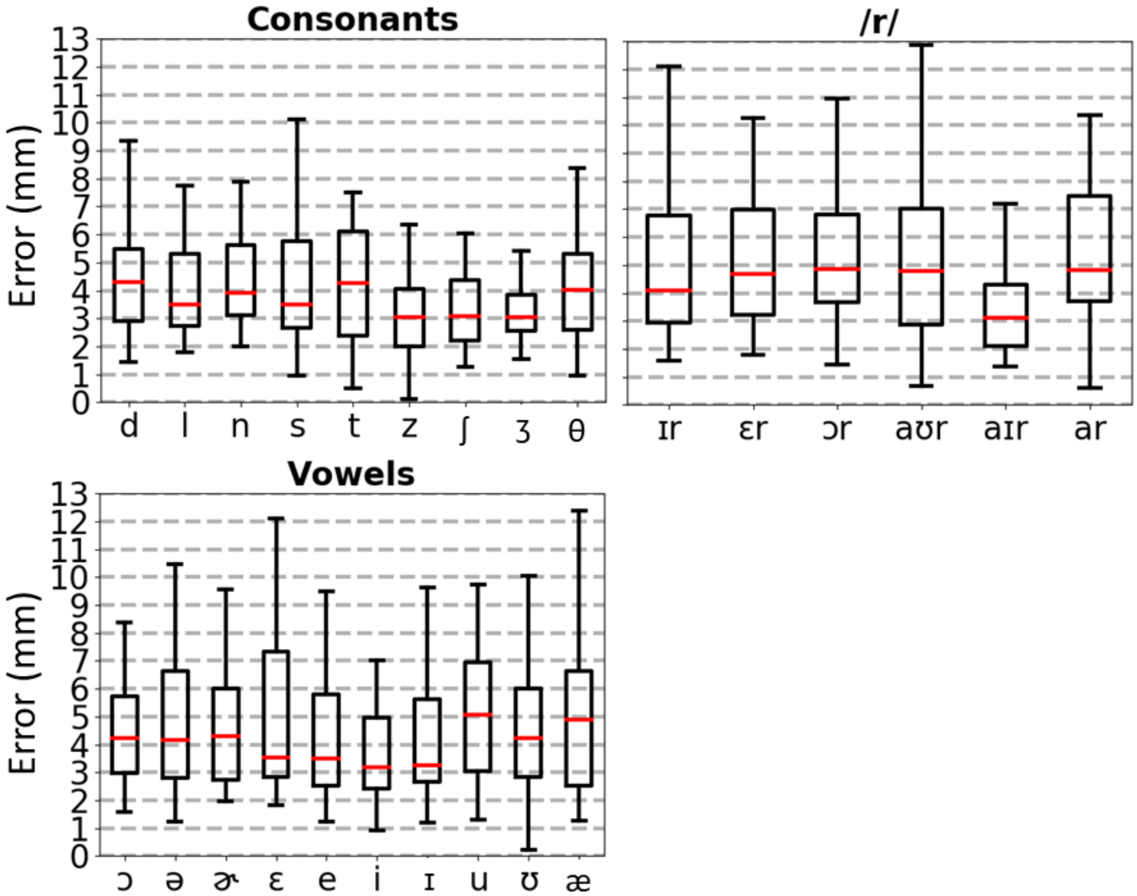


Figure 4.6: Positional variability of phoneme landmarks split by phoneme.

Regarding the variability by phoneme (Figure 4.6), the consonants have generally lower errors than observed for the vowels and the /r/ sounds. This could be explained by the fact that the placement of the tongue's blade has a more important impact on the production of these consonants. Indeed, the tongue's blade is in contact with the palate and/or front teeth, which increases the proprioception of the position of the tongue, and thus improves our ability to accurately place the tongue in a same location.

The higher errors reported for the set of /r/ sounds could be a consequence of the fact that two different manners to produce an /r/ sound exist: retroflexed and bunched. Figure 4.7 shows an ultrasound image of the oral cavity of an individual producing the /r/ sound, with the left picture being the retroflexed shape that is characterized by the tongue tip being raised and curled back, and the right picture depicting the bunched production in which the tongue tip is down and the tongue forms a bunch near the back of the palate. The landmarks are in utterly different place depending on the produced shape, and because the subjects were not asked which shape they used to produce /r/, the variability might be artificially higher that it is supposed to be.

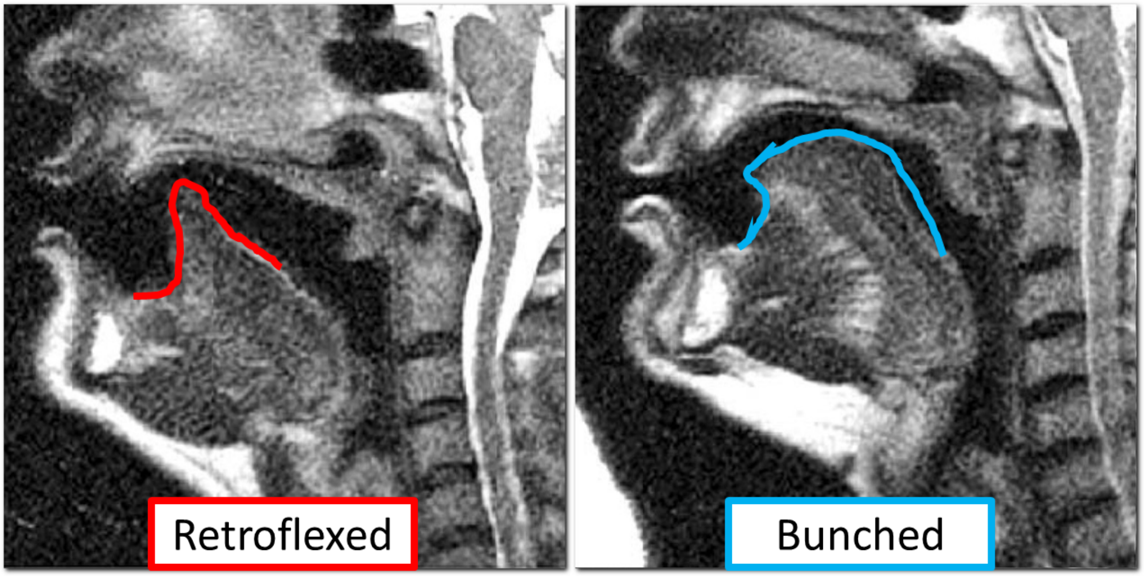


Figure 4.7: Difference between retroflexed and bunched shapes of the tongue for the production of the /r/ sound [53].

4.4.3 Summary of All Variability

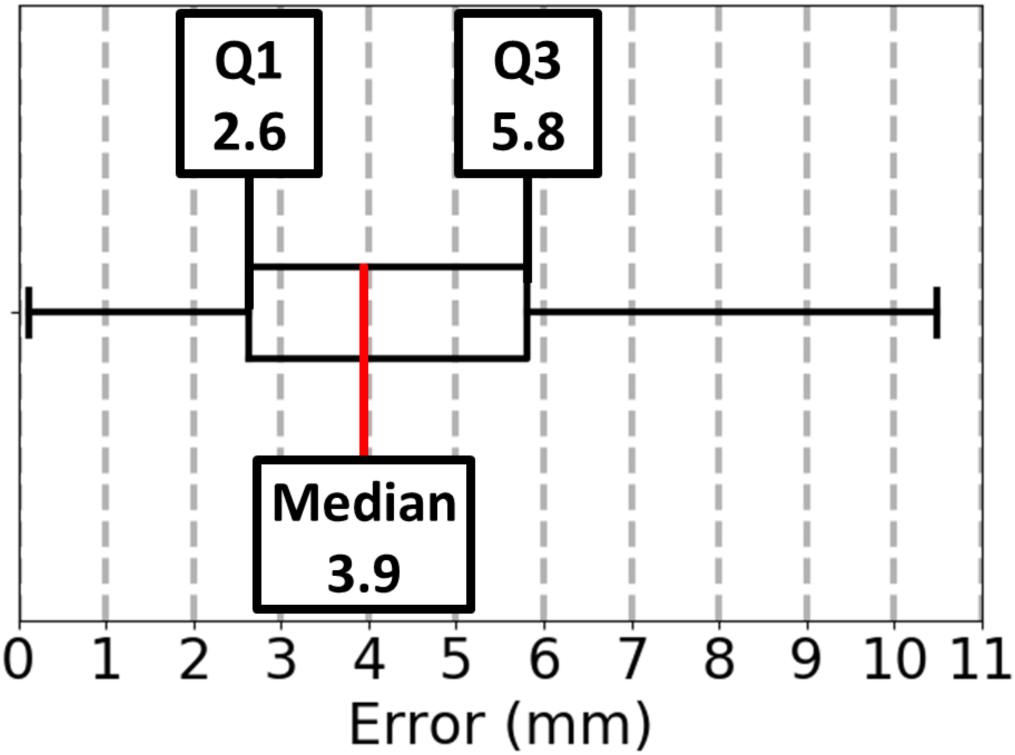


Figure 4.8: Positional variability across all phoneme landmarks.

Figure 4.8 shows a summarized view of all the 2,500 distance errors combined. Overall, 75% of the phoneme landmarks identified from the tongue motion are within 5.8 mm of their estimated true position, with a median error of 3.9 mm and an inter-quartile range of 3.2 mm.

4.5 Discussion

Although MagTrack is still an early prototype and our wireless tracking method in its infancy, the result of this study shows that MagTrack is capable of tracking actual tongue motion with the majority of positional errors within 5.8 mm. Actually, this value is an upper bound on the tracking accuracy because it also includes a natural positional variability that is unavoidable when producing a phoneme, as shown in [25, 41]. Although unknown, this natural tongue placement variability can be roughly estimated to be within a range of 1-3 mm [25] which would significantly reduce the positional variability found in this study. In our previous work, although the methods to assess the tracking accuracy between EMA and MagTrack were different, there were at least some objective and quantitative reference values to compare against. In this study, even though it is difficult to conclude on the feasibility of MagTrack to generate a baseline of phoneme landmarks, these results provide reference values of errors that future improved versions of MagTrack can be compared against. Once the results will be found satisfactory, an objective validation will be made by evaluating the impact of our gamification of tongue placement on recovery time for people with SSD.

At this stage of the research, a deeper analysis of the results should not be made because the science in speech sound production is still in its infancy with only few studies that attempted to record tongue motion during speech production including the studies in [11, 12, 50, 22, 25]. There are many reasons for the lack of research in that field such as the fact that tracking the tongue during speech and without impeding its natural motion is difficult to achieve, and because of the inherent complexity in articulation, motor control, voicing, among other mechanisms involved in speech production. For instance, the majority of the phonemes were produced in isolation while some consonants were followed by a vowel (C-V sequence). We found that it was more challenging to identify a characteristic landmark when a phoneme was produced in isolation and it might be more more difficult for the subjects to place their tongue in a consistent location when producing the phonemes in that manner instead of a C-V-C or V-C-V sequence as done in other studies [41]. Additionally, it was observed that the subjects were not consistent between each other in how they uttered some phonemes even though our data collection was designed to reduce such variability. Indeed, a reference sound was played when a new phoneme was displayed, a word was shown alongside the target phoneme to provide a context for its pronunciation, and the subjects were students that were not only trained as SLPs in the same program but also recently passed a Phonetics class where they learned how to properly produce these phonemes.

4.6 Future work

Before conducting a follow-up study, the following technical challenges must be overcome. First, the tracer is not placed at the exact same position on the tongue between subjects. This creates differences in the landmark positions that are not due to differences in tongue placement between speakers but to this tracer offset. There is a need to find a method to either place the tracer at the same position and in a consistent manner or to account for this offset and post-process the trajectories. Secondly, studies such as [41] show that differences in the dimension of the oral cavity and the tongue must be taken into account when comparing the landmark positions between subjects. Third, the current headset version is stationary and thus forces the user to remain still for the duration of the data collection session. However, natural body/head motion cannot be fully restricted and thus adds motion artifacts in the recorded trajectories that are not from the tongue. Our future plans to develop a wearable headset might solve this problem by enabling the user to move freely without adding any body motion to the recorded trajectories. Finally, it might be beneficial to restrict the set of target phonemes to the ones that most patients with SSD have difficulty producing. In this study, we selected a large sample of phonemes which, as far as we know, is the largest database of tongue motion recorded for the purpose of identifying phoneme landmarks. Most studies from speech researchers focus on few consonants, and as was shown in Figure 4.6, these are the ones with lower variability, and thus have the higher chance to generate a more accurate baseline of phoneme landmarks.

CHAPTER 5
ADDITIONAL APPLICATIONS FOR MAGTRACK

The main focus of our research has been the development of MagTrack for the treatment of SSD, and more specifically, the development of the visual feedback to show preliminary evidence of the potential of MagTrack as a tool for assisting therapy. However, the ability to track tongue motion wirelessly can be used in a variety of applications that widely differ in scope. For instance, instead of correcting speech, a silent speech interface tries to recognize speech from tongue motion only. In section 5.1, the results of the preliminary studies that show evidence that MagTrack can be used for such application will be discussed. Another important application is in assistive technology in which tongue motion can help quadriplegics becoming more autonomous for some activities of daily living. More details about this application will be provided in section 5.2. Lastly, far-fetched but nonetheless realistic applications can be found in consumer markets in which the tongue could be used as an added controller in games, or to help architects and designers to create and manipulate objects in 3D using a specialized stylus pen whose position is tracked in 3D using our permanent magnet localization method.

5.1 Silent Speech Interface

5.1.1 Motivation

Laryngectomy is a surgical procedure that removes the larynx of an individual affected by laryngeal cancer [54]. As part of the larynx, the vocal folds are at the source of phonation. As a consequence of the removal of the vocal folds, an individual is not capable of producing audible sounds which, similarly to SSD, affects their ability to communicate with others. As reported by the American Cancer Society in 2018, there is an estimated 16,000 individuals every year that are diagnosed with a type of cancer (e.g. laryngeal, hypopharyngeal) that necessitates a laryngectomy.

In order to regain the ability to communicate verbally, these individuals have limited options. The electrolarynx is one of the commonly known methods (Figure 5.1) by generating vibrations that emulate the functioning of the vocal folds.

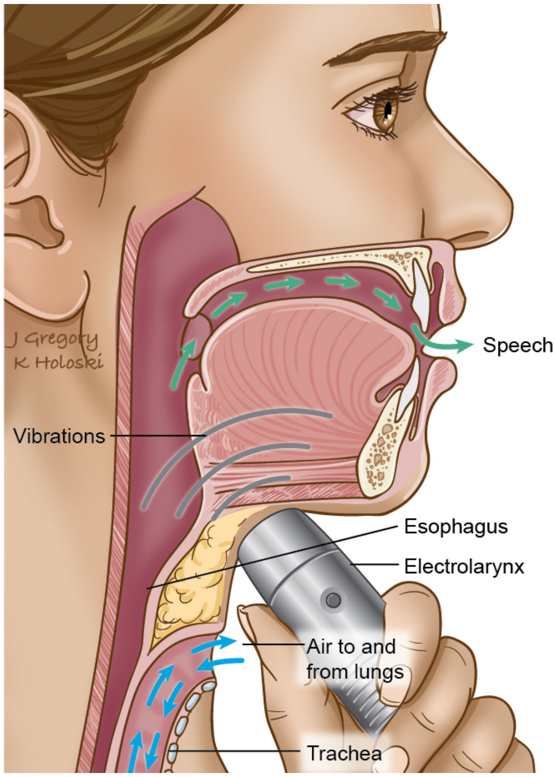


Figure 5.1: Illustration of a hand-held electrolarynx [55].

Other methods include tracheoesophageal puncture (TEP) [56], and esophageal speech [57]. Although these methods allow an individual to recover a certain ability to speak, the actual produced sounds are typically perceived as hoarse or even robotic [58, 59]. Consequently, it was reported that the individuals may develop symptoms of depression and anxiety during social interactions since they may be cognizant that others may perceive them as being abnormal [60]. Therefore, there is a need for a technology that would generate a more normal-sounding voice.

In recent decades, researchers have been working on the development of a silent speech interface (SSI) that can convert the motion of the articulators to synthesized speech [61]. The major advantage of SSI is that the synthesized speech can be processed to sound more natural for the user. VocaliD, a company created by the speech researcher Dr. Rupal Patel, has been able to collect millions of voice samples from more than 25,000 individuals over the world. From these samples, a unique voice can be tailored to an individual by modulating the input audio to match a respiratory drive, vocal pitch, breathiness, and resonance. Thus, there has been significant progress on the output of SSI. Now, the main challenge is on the input of SSI, and more precisely, recognizing speech from the articulatory motion. This process can be done in two different ways: (1) *recognition-and-synthesis* in which articulatory motion is first translated into text which is then fed into a text-to-speech synthesizer [62]; (2) *articulation-to-speech* in which the motion is transformed into audio features that can be directly fed into a speech synthesizer. The first method might be easier to implement because a large body of work and software already exist on processing, correcting, and synthesizing text. These tools can help attenuate some errors in speech prediction from the articulation-to-text module, and it would be easier to assess the performance of the system by comparing the predicted text to the one uttered by the speaker. The second method is more complex to implement but its main advantage is a lower delay between produced and synthesized speech because the model by-passes any intermediate processing steps, such as text-to-speech, and allows the speech delivery to be perceived as more ”real-time” which is important for verbal communication.

Regardless of the SSI method, it remains difficult to capture and process the motion of all articulators. Therefore, researchers are focused on the tongue since, as it was already mentioned before, it is the most important articulator for speech production. Naturally, EMA is the tongue tracking system being the most used by researchers for this purpose [7, 63, 64]. However, the same issues reported earlier for speech therapy are also applicable to SSI. For instance, EMA cannot be a wearable device which is critical for the use of SSI in daily activities. Additionally, EMA’s reliance on wired probes would prevent users to eat and drink without having to remove the probes because of the wires. Considering these issues, MagTrack has been seen as a suitable alternative that would allow SSI to become a practical system for the end-users by being portable, light-weight, affordable, and its wirelessly tracking method allows users to not only speak without hindrance but also to eat/drink without the need to detach the tracer.

5.1.2 Preliminary Results

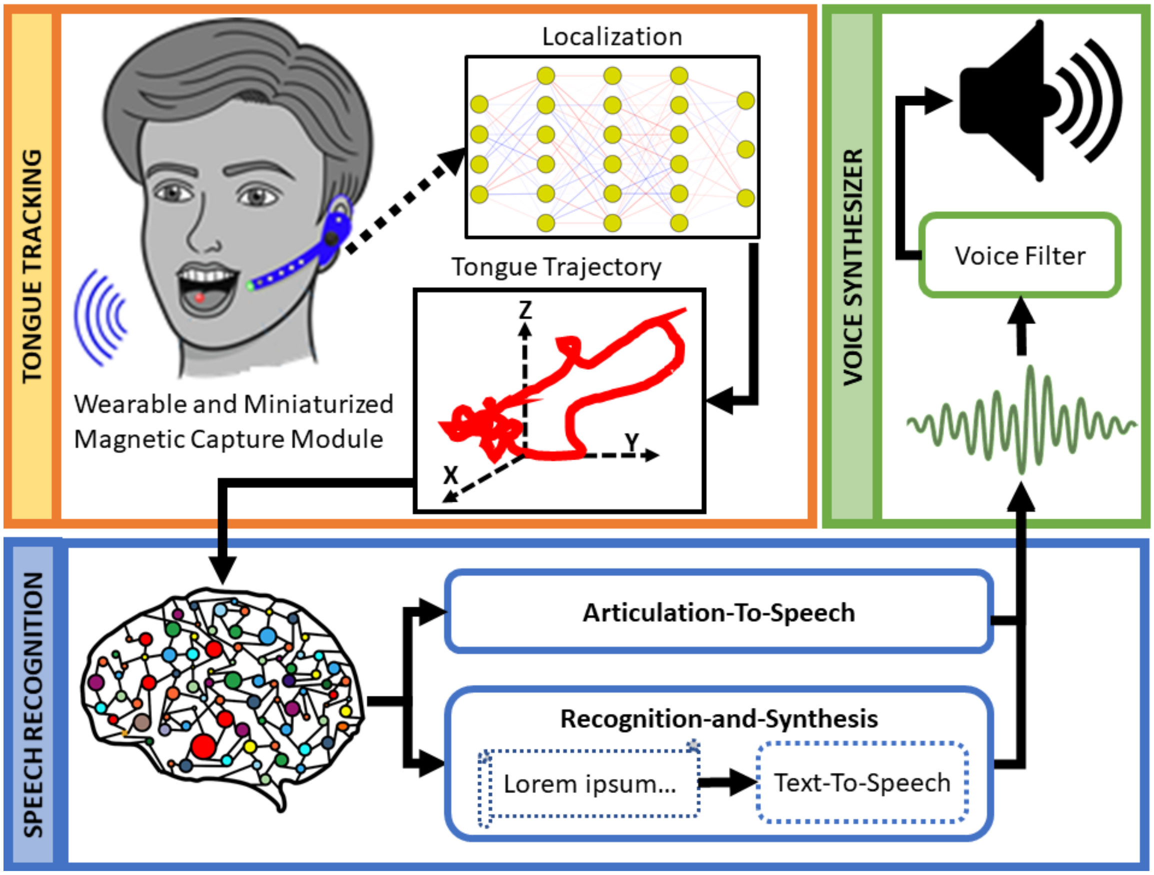


Figure 5.2: Overview of the proposed silent speech interface using MagTrack as the tongue tracking system, and a speech recognition algorithm that generates a base voice that is then filtered to produce a more normal-sounding voice.

As part of an ongoing collaboration with Dr. Jun Wang's lab at the University of Texas - Austin, our research objective is to develop a silent speech interface whose system overview is shown in Figure 5.2. MagTrack will be used as the tongue tracking system to capture the tongue trajectories, but our team will be working on redesigning the headset to be not only wearable but also miniaturized. Dr. Wang's team will be responsible for the development of the speech recognition algorithm that will transform the input trajectories into a base voice. Finally, the base voice will be modulated by an audio filter that will generate a voice that sounds closer to the one that the user had before the laryngectomy. As mentioned earlier, there are two methods to process the trajectories, therefore, our team conducted a human study as a proof-of-concept in our lab and for each method.

Recognition-and-Synthesis

In this study [65], six native American speakers were recruited among the Georgia Tech student body with the approval of the Georgia Tech Institutional Review Board. The participants were gender-balanced (3 females, 3 males), with an average age of 20.5 ±1.7, and with no history of speech disorder. MagTrack was used to collect their tongue trajectories while they were asked to speak 10 times a list of 90 utterances split between 65 words and 25 phrases. These utterances were selected by their common usage in daily life (e.g. "*okay*", "*hungry*", "*how are you?*", "*do you want to?*"). The total number of utterances, words, and phonemes that were recorded in this data collection are 5,400, 7,560, and 35,760, respectively. Figure 5.3 shows the example of tongue trajectories collected for a same speaker and for two different words. We can observe that the repetitions of a same utterance are well aligned which shows that MagTrack can deliver an adequate repeatability of tracking which is crucial for this application.

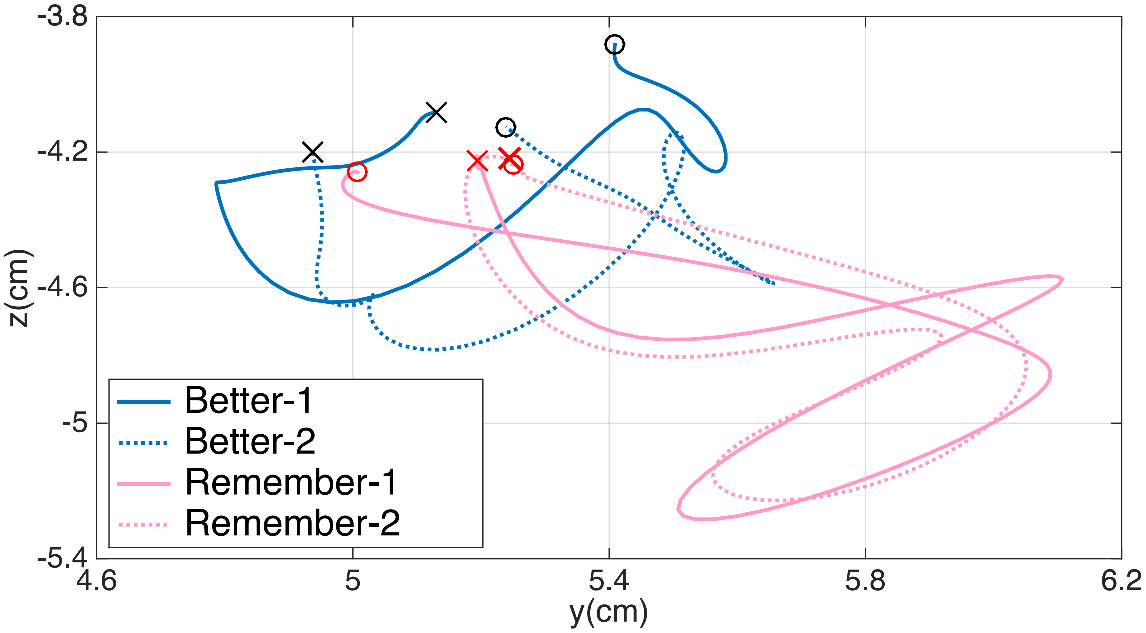


Figure 5.3: Example of tongue trajectories projected on the sagittal plane (Y-Z). These trajectories were collected for one speaker and for two repetitions of ”*better*” and ”*remember*”.

The objective of this study is to assess the classification accuracy of selected models in predicting words and phonemes. Predicting words is easier than phonemes because predictions can be improved with existing algorithms that provide additional predictive capability using grammar rules and the context of the word in the sentence to reduce the number of possible choices. However, this method is ”*closed-vocabulary*”, i.e. the model can only predict a finite set of words. Conversely, predicting phonemes is more challenging because there is not only less information to be extracted from the tongue trajectories to discriminate between phonemes, but also no simple set of rules can be used to assist in the prediction. Historically, a Gaussian Mixture Model (GMM)-Hidden Markov Model (HMM) has been used as a classifier [66, 67], but researchers are recently leveraging the capabilities of deep neural networks (DNN) [62]. Therefore, in our system, classifiers models based on DNN-HMM were implemented and compared against GMM-HMM. More details about this study can be found in [65]. Figures 5.4 and 5.5 show that our best classifier reached a word error rate (WER) of 32.1% and a phoneme error rate (PER) of 37.3%. Our results are similar than reported in the literature for EMA [68] with a PER of 35.5% for a male speaker and 37.1% for a female speaker. Although these results cannot be directly compared because the datasets and the data collection protocols are different, it shows nonetheless the potential for MagTrack to be used in SSI.

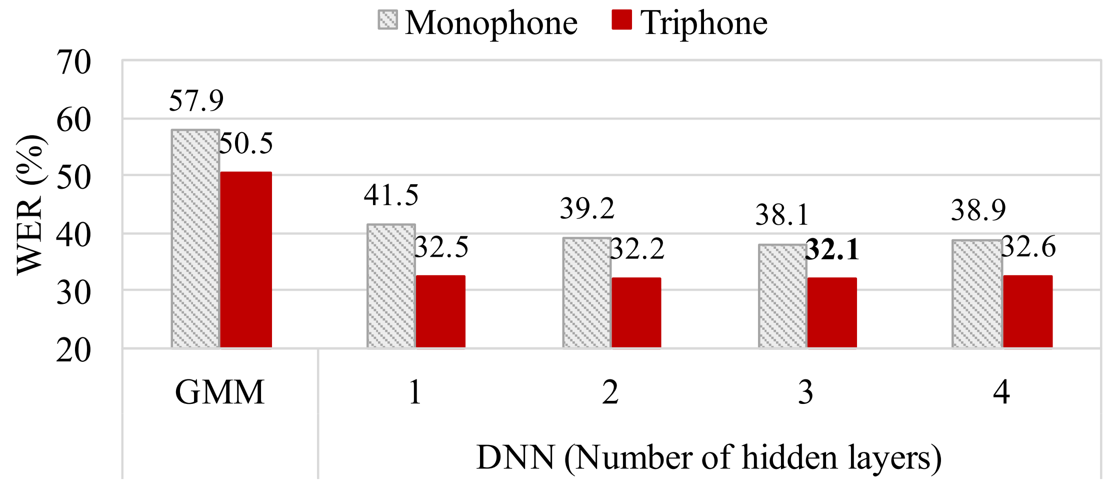


Figure 5.4: Comparison of the word error rates (WER) between GMM-HMM and DNN-HMM models in recognition-and-synthesis [65].

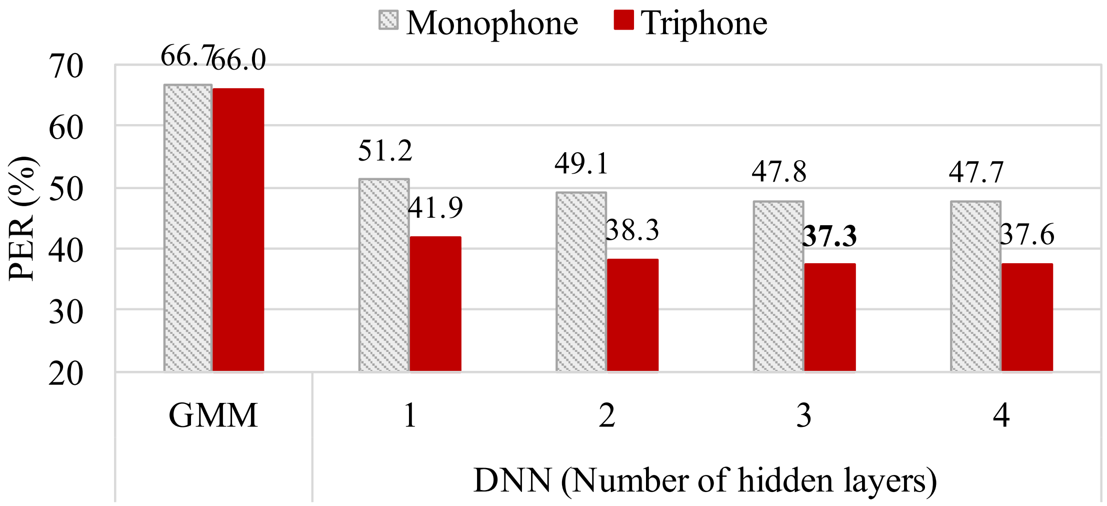


Figure 5.5: Comparison of the phoneme error rates (PER) in recognition-and-synthesis [65].

Articulation-to-Speech

In this study [69], the objective was to compare the capabilities of MagTrack and EMA for the purpose of SSI. Similarly to the previous study, two groups of 10 participants (one for each system) were tasked to speak 132 utterances with 2 repetitions. The rationale for more utterances but less repetitions is to collect a wider variety of context for the usage of each word. The first repetition of an utterance was uttered in a normal voice, while the second was spoken ”silently” (i.e. no voice was produced, only the articulatory motion was performed). Because the articulation-to-speech (ATS) model outputs 180 audio features based on mel-cepstral coefficients, there is not an intuitive and simple method to assess the accuracy of prediction. The mel-cepstral distortion (MCD) is the measure that most researchers rely on, and it is used in this study to estimate the quality of the generated speech as compared to the recorded voice, with lower values representing a better audio quality. More details can be found in [69], but the results are summarized in Figure 5.6.

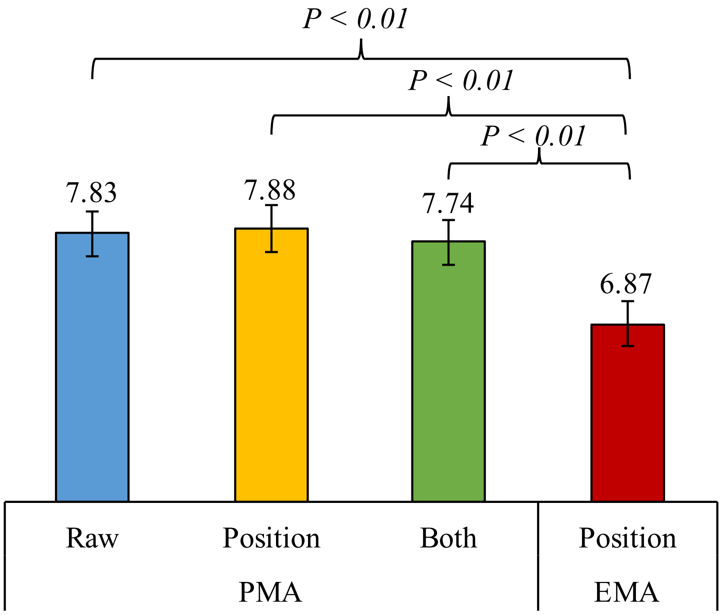


Figure 5.6: Comparison of the MCD between MagTrack (PMA) and EMA systems [69].

A first comparison was performed on the impact of using only the raw magnetic values of MagTrack (refer in this paper as a permanent magnetic articulograph, or PMA) as opposed to solely the localized state of the tracer, and to both set of values. Since no statistically significant differences were observed, this result shows that despite the additional errors added by the localization of the tracer, the generated speech are of similar quality. The speech researchers in our team were eager to perform this first evaluation with MagTrack because they don't have access to the raw data with EMA but only the localized states of the probes. This result is significant because using the localized state as input to the model not only reduces its complexity due to a significantly lower number of inputs to process (here, 3 positional values instead of 72 magnetic measurements) but will also enable simpler pre-processing of the inputs to improve the performance of the model, including linear and/or geometric transformations such as scaling/normalization or rotation, instead of complex and non-intuitive nonlinear transformations for the magnetic measurements. Furthermore, using the state of the tracer as input enables the ATS model to be decoupled from the device configuration. Therefore, the number of magnetometers, their position, their model type, and their settings can be changed without having to train a new ATS model as long as the localization accuracy of the new device remains at least the same.

The second comparison carried out in this study concerns the MCD values between MagTrack and EMA. It is shown that the difference is statistically significant ($p < 0.01$) in favor of EMA. This result is not surprising because EMA is reportedly a more precise motion tracking technology than the current version of MagTrack. However, this study fails to show whether the value of MagTrack's MCD (7.88) means that the generated audio is comprehensible enough for a listener to understand what the original utterance was supposed to be. Furthermore, although the two groups of participants were roughly age and gender-matched, they are nonetheless different individuals residing in two different states (Georgia vs. Texas) and whose data collection sessions were conducted by two different teams. Finally, MCD is not the only measure of quality of speech since other studies have used ban aperiodicity distortion (BAP) [70], root mean square error of fundamental frequencies (F0-RMSE), and voiced/unvoiced (V/UV) error rate. Therefore, a listening evaluation with text transcription of the generated speech would provide more conclusive results for SSI.

5.1.3 Future Work

Our team was awarded a R01 research grant from the National Institutes of Health to fund this research for five years. Led by Dr. Wang, the team will be composed of three units: (1) our team at the Inan Research Lab at Georgia Tech (Dr. Omer Inan) to continue improve MagTrack to be wearable and have an increased tracking accuracy; (2) the Speech Disorders and Technology Lab at the University of Texas - Austin (Dr. Jun Wang) to develop the speech recognition model and interface with a voice filter to generate the synthesized speech; (3) clinicians at the University of Texas - Southwestern Medical Center (Dr. Ted Mau) to provide access to patients with laryngectomy. Additionally, two consultants will be part of the team to add further experience in speech-language pathology, voice disorder, and advanced signal processing for speech.

5.2 Assistive Technology for Quadriplegics

MagTrack is actually an offshoot of the Tongue Drive System (TDS) which has been the flagship project of the GT-Bionics lab at Georgia Tech. This project has been led by Dr. Nazmus Sahadat and has for objective to use the tongue as a controller to help people with quadriplegia to regain autonomy for some activities of daily life. More details about the TDS can be found in [4, 5, 39, 71, 72, 73, 74, 75, 76] and in the Ph.D thesis of Dr. Nazmus Sahadat [77]. In a nutshell, the TDS tracks the tongue in a similar manner than MagTrack by relying on the same magnetic tracer attached on the tongue whose magnetic field is captured by an array of magnetometers. In the TDS, the main difference with MagTrack is the fact that no localization of the tracer is performed. Indeed, the objective of the TDS is to generate 7 commands with each command being associated to an area of the oral cavity. Any area of the mouth can be selected as a command, but it was found that the ones shown in Figure 5.7 provide the best performance because they are well separated and easy to locate: the four canines, left and right cheeks, and the natural resting position of the tongue.

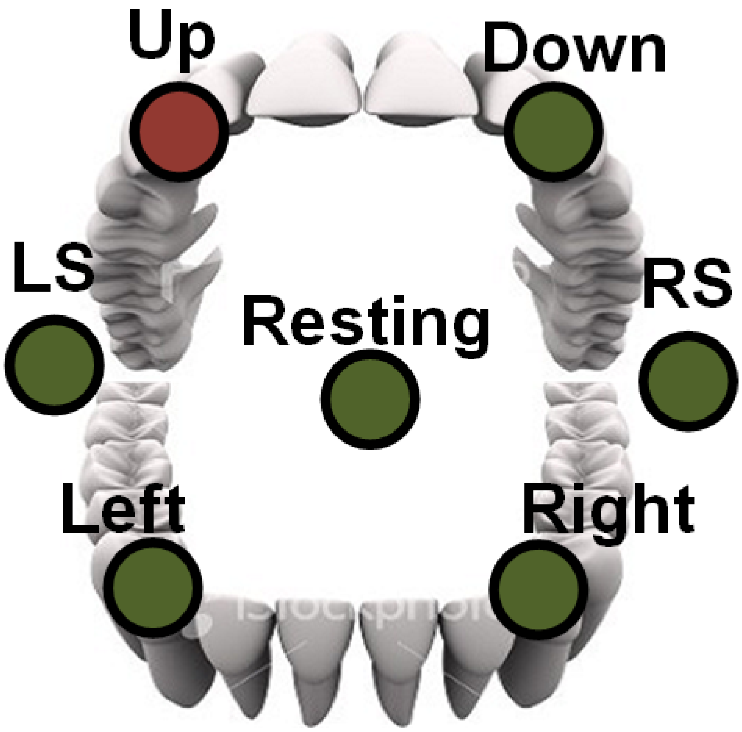


Figure 5.7: Example of a mapping of tongue positions and commands issued by the TDS [39].

Each of the 7 locations is then mapped into a unique command. The main application of the TDS is to help quadriplegics to drive their powered wheelchair using the tongue as the controller. In Figure 5.7, four commands are mapped to a direction of motion (i.e. up, down, left, right). In our illustration, if the user touches her upper-right canine, the wheelchair would move forward (UP). Another application for the TDS is the emulation of a mouse to control a computer. Here, six commands are used since a typical mouse has not only four directions of motion (up, down, left, right), but also a left and right click that are mapped to the left select (LS) and right select (RS) commands in Figure 5.7.

Because the exact location of the tracer is not needed, the TDS relies on a classifier to convert the magnetic measurements to a command with a reported accuracy above 95% [39]. Additionally, the TDS is wearable by implementing a dynamic background cancellation that is similar than the method described in 2.4.1, and the headset has only four magnetometers which makes the device light-weight. A human study is being conducted at the Brooks Rehabilitation Center (Jacksonville, Florida, USA) in which the TDS is being used by people with quadriplegia to drive a powered wheelchair on a custom-designed race track with difficult motion maneuvers such as U-turns, backing, and roundabouts.

However, there are some limitations of the TDS. Chief among them is the fact that the classifier must be trained at each session, i.e. every time the headset is placed on the user's head. More details about why the training must be performed at each use can be found in [77] but, in summary, it is due to the fact that the placement of the magnetometers relatively to the tracer changes every time the headset is worn and thus the previous classifier model is not validate any longer. Therefore, the user must performed this training, which lasts about 5 minutes, by placing the tracer at the 7 positions with a certain number of repetitions. Additionally, a calibration of the dynamic BMF cancellation must also be performed to estimate the projection matrices (refer to equation 2.8). This calibration requires the user to move their head in a circular fashion for about 1 minute. The problem is that the performance of the dynamic cancellation depends on the user's ability to rotate her head as wide as possible which cannot be done by all users since their spinal cord injury might restrict their neck's range of motion. In summary, these two calibrations are not just an inconvenience but limit the classification accuracy of the TDS when used by actual people with quadriplegia.

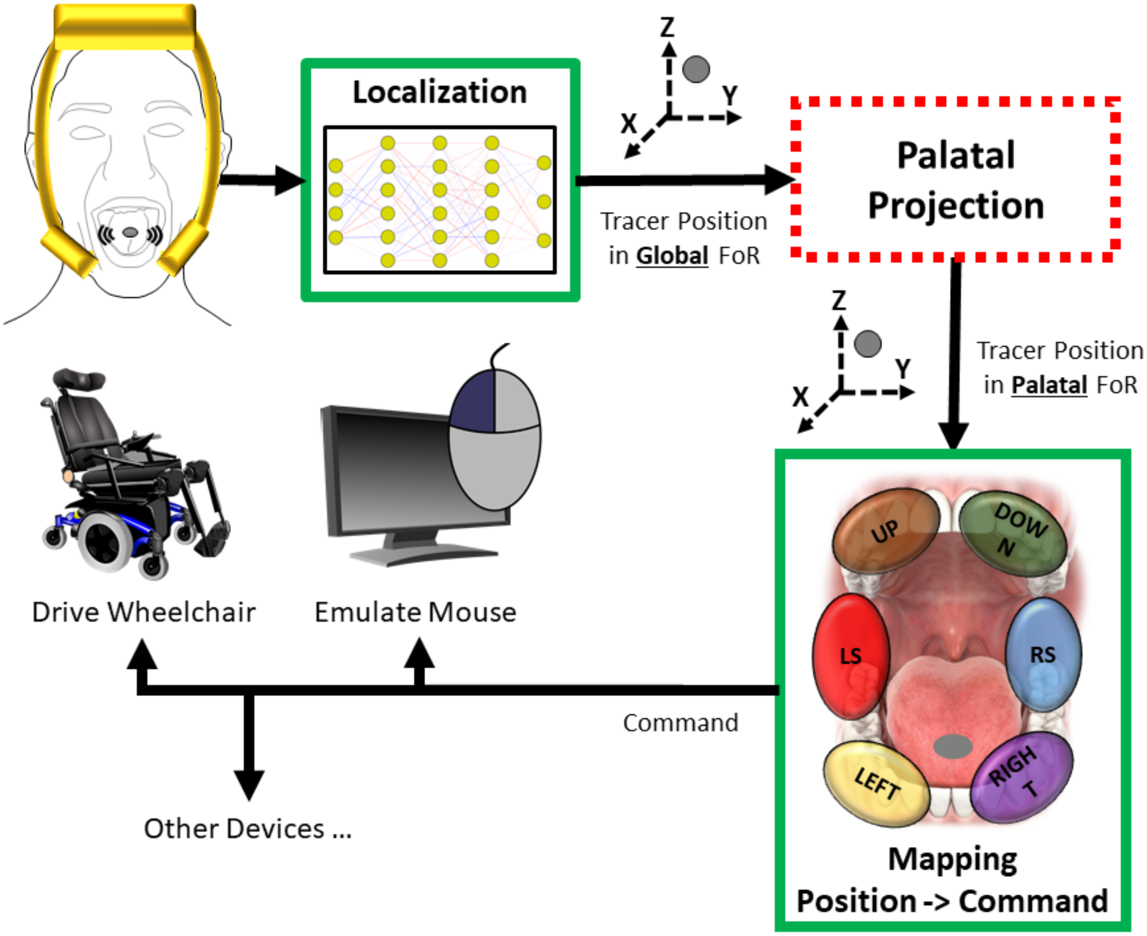


Figure 5.8: Improvement of the TDS by reducing the need for calibration. Algorithms that require a once-a-lifetime calibration are emphasized with a solid green rectangle, while the ones with calibration that must be performed at every use are shown in dashed red.

Figure 5.8 shows how MagTrack could solve this issue by reducing the need for calibration. First of all, the localization model is only trained once after a headset is built and the training is performed in the lab/manufacturing site. Once the model is trained, it remains unchanged for the entire lifetime of the device. Then, the tracer position in the global FoR that are generated by the localization are projected into the palatal FoR (more details in section 2.3.2). This is the only step that requires a calibration at every use since the headset will never be placed on the user's head at the exact same position and orientation. However, this step requires only a few seconds to complete since the tracer must only be placed once at three locations on the palate.

The final step is to generate the desired command by finding the labeled area associated to the current position of the tracer in the palatal FoR. These labeled areas must only be set once by the user to create a mapping between an area in the mouth and a command. A re-mapping is only required if the dimensions of the oral cavity has changed significantly. The mapping is easy and quick to perform by setting the position of each area's boundaries. As an added advantage, the number of areas, therefore commands, are not restricted to seven and can be set to virtually any number. The only restriction is that there should not exist any overlap between the areas in order to assign a unique command per area.

Finally, because the tracking and reference magnetometers are fixed in position in MagTrack, the calibration of the dynamic BMF cancellation will be done only once after a headset is built since the projection matrices should remain unchanged for the lifetime of the device. Therefore, there is no need for the user to rotate their head for this calibration at each session. We believe that these improvements will make the system more practical for the users, and consequently, enable MagTrack to have a better chance to be widely adopted by this community.

5.3 Conclusion

Below are the main intellectual contributions of this thesis:

- Created a wearable tongue tracking system based on a wireless permanent magnet localization algorithm
- Quantified the accuracy of a permanent magnet localization algorithm based on a neural network and from the largest dataset of 5D states of a magnetic tracer
- Quantified the positional variability of tongue placement from the largest dataset of tongue motion recorded by a tongue tracking system for American-English phonemes
- Predicted speech with a practical tongue tracking system as input to a silent speech interface

This work has produced the first practical tongue tracking system that allows its users to move their tongue without hindrance thanks to our implementation of a real-time and wireless permanent magnet localization algorithm based on a neural network. For the first time, the largest dataset of 5D states of a magnetic tracer was recorded to train our models and quantify the tracking accuracy with the most comprehensive set of tongue trajectories. This novel tracking system was tested on real tongue motion to quantify the positional variability during the production of the largest set of phonemes recorded to date and to predict speech solely from tongue motion.

Looking into the future, the next major milestone in this project is to allow the headset to be untethered, and thus, fully wearable. This will enable MagTrack to be used outside of the lab and in the hands of early adopters that will help us refine the system for its target applications.

REFERENCES

[1] W. R. Zemlin, *Speech and Hearing Science: Anatomy and Physiology*, 4th ed. Boston: Allyn and Bacon, 1998.

[2] R. D. Kent, “Research on speech motor control and its disorders: A review and prospective,” *Journal of Communication Disorders*, vol. 33, no. 5, pp. 391–427, 2000.

[3] E. Maas, D. A. Robin, S. N. Austermann Hula, S. E. Freedman, G. Wulf, K. J. Ballard, and R. A. Schmidt, “Principles of motor learning in treatment of motor speech disorders,” *American Journal of Speech-Language Pathology*, vol. 17, no. 3, pp. 277–98, 2008.

[4] M. N. Sahadat, S. Dighe, F. Islam, and M. Ghovanloo, “An independent tongue-operated assistive system for both access and mobility,” *IEEE Sensors J.*, vol. 18, no. 22, pp. 9401–9409, 2018.

[5] M. N. Sahadat, A. Alreja, and M. Ghovanloo, “Simultaneous multimodal pc access for people with disabilities by integrating head tracking, speech recognition, and tongue motion,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 192–201, 2018.

[6] T. Kaburagi, K. Wakamiya, and M. Honda, “Three-dimensional electromagnetic articulography: A measurement principle,” *Journal of the Acoustical Society of America*, vol. 118, no. 1, pp. 428–443, 2005.

[7] P. Schoenle, K. Grbe, P. Wenig, J. Hhne, J. Schrader, and B. Conrad, “Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract,” *Brain and Language*, vol. 31, no. 1, pp. 26–35, 1987.

[8] J. J. Berry, “Accuracy of the ndi wave speech research system,” *Journal of Speech, Language, and Hearing Research*, vol. 54, no. 5, pp. 1295–1301, 2011.

[9] Y. Yunusova, J. R. Green, and A. Mefferd, “Accuracy assessment for ag500, electromagnetic articulograph,” *Journal of Speech, Language, and Hearing Research*, vol. 52, no. 2, pp. 547–555, 2009.

[10] C. Savariaux, P. Badin, A. Samson, and S. Gerber, “A comparative study of the precision of carstens and northern digital instruments electromagnetic articulographs,”

Journal of Speech, Language, and Hearing Research, vol. 60, no. 2, pp. 322–340, 2017.

[11] W. F. Katz, S. V. Bharadwaj, and M. P. Stettler, "Influences of electromagnetic articulography sensors on speech produced by healthy adults and individuals with aphasia and apraxia," *Journal of Speech, Language, and Hearing Research*, vol. 49, no. 3, pp. 645–659, 2006.

[12] W. F. Katz and M. R. McNeil, "Studies of articulatory feedback treatment for apraxia of speech based on electromagnetic articulography," *Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, vol. 20, no. 3, pp. 73–80, 2010.

[13] Speech Disorders and Technology Lab. (2019). Major equipment, [Online]. Available: `https://csd.utexas.edu/research/wang-lab/lab` (visited on 09/29/2019).

[14] Speech Production Lab. (2019). Electromagnetic articulography, [Online]. Available: `https://www.utdallas.edu/speech-production-lab/research/previous-ema-visual-feedback-studies/about-ema` (visited on 09/29/2019).

[15] S. Kelly, A. Main, G. Manley, and C. McLean, "Electropalatography and the linguagraph system," *Medical Engineering & Physics*, vol. 22, no. 1, pp. 47–58, 2000.

[16] F. Gibbon and A. Lee, "Electropalatography for older children and adults with residual speech errors," *Semin Speech Lang*, vol. 36, no. 04, pp. 271–282, 2015.

[17] F. E. Gibbon and L. Paterson, "A survey of speech and language therapists views on electropalatography therapy outcomes in scotland," *Child Language Teaching and Therapy*, vol. 22, no. 3, pp. 275–292, 2006.

[18] CompleteSpeech. (2019). Smartpalate, [Online]. Available: `http://completespeech.com/smartpalate` (visited on 09/29/2019).

[19] A+ Speech Therapy. (2019). Speech programs, [Online]. Available: `https://palatometertherapy.blogspot.com/p/programs-available.html` (visited on 09/29/2019).

[20] L. Mnard, J. Aubin, M. Thibeault, and G. Richard, "Measuring tongue shapes and positions with ultrasound imaging: A validation experiment using an articulatory model," *Folia Phoniatrica et Logopaedica*, vol. 64, no. 2, pp. 64–72, 2012.

[21] M. Grimaldi, B Gili Fivela, F. Sigona, M. Tavella, P. Fitzpatrick, L. Craighero, L. Fadiga, G. Sandini, and G. Metta, "New technologies for simultaneous acquisition

of speech articulatory data: 3d articulograph, ultrasound and electroglottograph," *Proceedings of LangTech*, pp. 1–5, 2008.

[22] J. Wang, A. Samal, P. Rong, and J. R. Green, "An optimal set of flesh points on tongue and lips for speech-movement classification," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 1, pp. 15–26, 2016.

[23] Haskins Lab. (2017). Ultrasound visual feedback, [Online]. Available: `http://www.haskins.yale.edu/uvf` (visited on 09/29/2019).

[24] L. Davidson, "Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance.," *The Journal of the Acoustical Society of America*, vol. 120, no. 1, pp. 407–415, 2006.

[25] Y. Yunusova, J. S. Rosenthal, K. Rudy, M. Baljko, and J. Daskalogiannakis, "Positional targets for lingual consonants defined using electromagnetic articulography," *Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1027–1038, 2012.

[26] N. Sebkhi, D. Desai, M. Islam, J. Lu, K. Wilson, and M. Ghovanloo, "Multimodal speech capture system for speech rehabilitation and learning," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 11, pp. 2639–2649, 2017.

[27] C. Hu, M. Q.-H. Meng, and M. Mandal, "A linear algorithm for tracing magnet position and orientation by using three-axis magnetic sensors," *IEEE Trans. Magn.*, vol. 43, no. 12, pp. 4096–4101, 2007.

[28] W. Yang, C. Hu, M. Q. Meng, S. Song, and H. Dai, "A six-dimensional magnetic localization algorithm for a rectangular magnet objective based on a particle swarm optimizer," *IEEE Trans. Magn.*, vol. 45, no. 8, pp. 3092–3099, 2009.

[29] C. Hu, M. Li, S. Song, R. Zhang, and M. Q.-H. Meng, "A cubic 3-axis magnetic sensor array for wirelessly tracking magnet position and orientation," *IEEE Sensors J.*, vol. 10, no. 5, pp. 903–913, 2010.

[30] S. Song, B. Li, W. Qiao, C. Hu, H. Ren, H. Yu, Q. Zhang, M. Q.-H. Meng, and G. Xu, "6-d magnetic localization and orientation method for an annular magnet based on a closed-form analytical model," *IEEE Trans. Magn.*, vol. 50, no. 9, pp. 1–11, 2014.

[31] S. Song, C. Hu, and M. Q.-H. Meng, "Multiple objects positioning and identification method based on magnetic localization system," *IEEE Trans. Magn.*, vol. 52, no. 10, pp. 1–4, 2016.

[32] N. Sebkhi, D. Desai, A. Khan, N. Prasad, S. Banerjee, J. Eng, K. R. Wilson, and M. Ghovanloo, "Towards a wireless multimodal speech capture system," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2016, pp. 82–85.

[33] G. Yin, Y. Zhang, H. Fan, and Z. Li, "Magnetic dipole localization based on magnetic gradient tensor data at a single point," *Journal of Applied Remote Sensing*, vol. 8, no. 1, 2014.

[34] T. Nara, S. Suzuki, and S. Ando, "A closed-form formula for magnetic dipole localization by measurement of its magnetic field and spatial gradients," *IEEE Trans. Magn.*, vol. 42, no. 10, pp. 3291–3293, 2006.

[35] C. Cheng, X. Huo, and M. Ghovanloo, "Towards a magnetic localization system for 3-d tracking of tongue movements in speech-language therapy," in *IEEE Engineering in Medicine and Biology Society Conf. (EMBC)*, 2009, pp. 563–566.

[36] A. Farajidavar, J. M. Block, and M. Ghovanloo, "A comprehensive method for magnetic sensor calibration: A precise system for 3-d tracking of the tongue movements," in *IEEE Engineering in Medicine and Biology Society Conf. (EMBC)*, 2012, pp. 1153–1156.

[37] S. Foong and Z. Sun, "High accuracy passive magnetic field-based localization for feedback control using principal component analysis," *MDPI Sensors*, vol. 16, no. 8, p. 1280, 2016.

[38] L. Marchal, S. Foong, Z. Sun, and K. L. Wood, "Design optimization of the sensor spatial arrangement in a direct magnetic field-based localization system for medical applications," in *IEEE Engineering in Medicine and Biology Society Conf. (EMBC)*, Aug. 2015, pp. 897–900.

[39] M. N. Sahadat, N. Sebkhi, D. Anderson, and M. Ghovanloo, "Optimization of tongue gesture processing algorithm for standalone multimodal tongue drive system," *IEEE Sensors J.*, vol. 19, no. 7, pp. 2704–2712, 2018.

[40] N. Sebkhi, N. Sahadat, S. Hersek, A. Bhavsar, S. Siahpoushan, M. Ghovanloo, and O. Inan, "A deep neural network-based permanent magnet localization for tongue tracking," *IEEE Sensors J.*, 2019.

[41] K. Rudy and Y. Yunusova, "The effect of anatomic factors on tongue position variability during consonants," *Journal of Speech, Language, and Hearing Research*, vol. 56, no. 1, pp. 137–149, Feb. 2013.

[42] H.-M. Shen, Y. Yue, C. Lian, D. Ge, and G. Yang, "Tongue computer interface prototype design based on t-type magnet localization for smart environment control," *Applied Sciences*, vol. 8, no. 12, p. 2498, Dec. 2018.

[43] F. Chollet *et al.*, *Keras*, `https://keras.io`, 2015.

[44] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, *et al.*, *TensorFlow: Large-scale machine learning on heterogeneous systems*, 2015. [Online]. Available: `https://www.tensorflow.org/`.

[45] T Tieleman and G Hinton, "Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning," in *Technical report*, 2017.

[46] NIDCD, "Statistics on voice, speech, and language," National Institute on Deafness and Other Communication Disorders, Report, 2016. [Online]. Available: `https://www.nidcd.nih.gov/health/statistics/statistics-voice-speech-and-language`.

[47] S. Fuchs and P. Perrier, "On the complex nature of speech kinematics," *ZAS Papers in Linguistics*, pp. 137–165, 2005.

[48] P. Marshalla, "Horns, whistles, bite blocks, and straws: A review of tools/objects used in articulation therapy by van riper and other traditional therapists," *International Journal of Orofacial Myology*, vol. 37, pp. 69–96, 2011.

[49] W. Katz, T. F. Campbell, J. Wang, E. Farrar, J. C. Eubanks, A. Balasubramanian, B. Prabhakaran, and R. Rennaker, "Opti-speech: A real-time, 3d visual feedback system for speech training," in *Annual Conference of the International Speech Communication Association*, 2014, pp. 1174–1178.

[50] W. F. Katz and S. Mehta, "Visual feedback of tongue movement for novel speech sound learning," *Frontiers in human neuroscience*, vol. 9, p. 612, 2015.

[51] M. McNeil, T. Fossett, W. Katz, D. Garst, G. Carter, N. Szuminsky, and P. Doyle, "Effects of visually augmented kinematic feedback with constant practice for the treatment of apraxia of speech: A single subject experiment," in *Clinical Aphasiology Conference*, 2007.

[52] J. Reilly and J. L. Fisher, "Sherlock holmes and the strange case of the missing attribution: A historical note on the grandfather passage," *Journal of Speech, Language, and Hearing Research*, 2012.

[53] X. Zhou, C. Y. Espy-Wilson, M. Tiede, and S. Boyce, "An articulatory and acoustic study of retroflex and bunched american english rhotic sound based on mri," in *Annual Conference of the International Speech Communication Association*, 2007.

[54] B. J. Bailey, J. T. Johnson, and S. D. Newlands, *Head & Neck Surgery–Otolaryngology*, 1st ed. Lippincott Williams & Wilkins, 2006.

[55] Head & Neck Cancer Guide. (2019). Speech and swallowing rehabilitation, [Online]. Available: https://headandneckcancerguide.org/adults/cancer-diagnosis-treatments/surgery-and-rehabilitation/surgeries-to-aid-breathing-and-eating/speech-and-swallowing-rehabilitation/ (visited on 09/29/2019).

[56] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production," *Journal of Speech and Hearing disorders*, vol. 49, no. 2, pp. 202–210, 1984.

[57] M. Hyman, "An experimental study of artificial-larynx and esophageal speech," *Journal of Speech and Hearing disorders*, vol. 20, no. 3, pp. 291–299, 1955.

[58] T. Mau, "Diagnostic evaluation and management of hoarseness," *Medical Clinics*, vol. 94, no. 5, pp. 945–960, 2010.

[59] T. Mau, J. Muhlestein, S. Callahan, and R. W. Chan, "Modulating phonation through alteration of vocal fold medial surface contour," *The Laryngoscope*, vol. 122, no. 9, pp. 2005–2014, 2012.

[60] J. Mertl, E. Zakova, and B. Repova, "Quality of life of patients after total laryngectomy: The struggle against stigmatization and social exclusion using speech synthesis," *Disability and Rehabilitation: Assistive Technology*, vol. 13, no. 4, pp. 342–352, 2018.

[61] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[62] M. Kim, B. Cao, T. Mau, and J. Wang, "Speaker-independent silent speech recognition from flesh-point articulatory movements using an lstm neural network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2323–2336, 2017.

[63] B. Cao, M. Kim, J. V. Santen, T Mau, and J Wang, "Articulation-to-speech synthesis using articulatory flesh point sensors orientation information," in *Interspeech*, 2018, pp. 3152–3156.

[64] F. Bocquelet, T. Hueber, L. Girin, C. Savariaux, and B. Yvert, "Real-time control of an articulatory-based speech synthesizer for brain computer interfaces," *PLoS computational biology*, vol. 12, no. 11, e1005119, 2016.

[65] M. Kim, N. Sebkhi, B. Cao, M. Ghovanloo, and J. Wang, "Preliminary test of a wireless magnetic tongue tracking system for silent speech interface," in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, 2018, pp. 1–4.

[66] Y. Deng, J. T. Heaton, and G. S. Meltzner, “Towards a practical silent speech recognition system,” in *Annual Conference of the International Speech Communication Association*, 2014.

[67] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, “Silent speech recognition as an alternative communication device for persons with laryngectomy,” *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[68] S. Hahm and J. Wang, “Silent speech recognition from articulatory movements using deep neural network.,” in *ICPhS*, 2015.

[69] B. Cao, N. Sebkhi, T. Mau, O. T. Inan, and J. Wang, “Permanent magnetic articulograph (pma) vs electromagnetic articulograph (ema) in articulation-to-speech synthesis for silent speech interface,” in *Proceedings of the Eighth Workshop on Speech and Language Processing for Assistive Technologies*, 2019, pp. 17–23.

[70] M. Morise, “D4c, a band-aperiodicity estimator for high-quality speech synthesis,” *Speech Communication*, vol. 84, pp. 57–65, 2016.

[71] M. N. Sahadat, Z. Zhang, A. Alreja, P. Srikrishnan, S. Ostadabbas, N. Sebkhi, and M. Ghovanloo, “Live demonstration: A tongue-operated multimodal human computer interface and robotic rehabilitation system,” in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2015.

[72] M. N. Sahadat, A. Alreja, P. Srikrishnan, and M. Ghovanloo, “A multimodal human computer interface combining head movement, speech and tongue motion for people with severe disabilities,” in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Oct. 2015, pp. 1–4.

[73] M. N. Sahadat, A. Alreja, N. Mikail, and M. Ghovanloo, “Comparing the use of single versus multiple combined abilities in conducting complex computer tasks hands-free,” *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 26, no. 9, pp. 1868–1877, 2018.

[74] M. N. Sahadat, N. Sebkhi, F. Kong, and M. Ghovanloo, “Standalone assistive system to employ multiple remaining abilities in people with tetraplegia,” in *IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–4.

[75] F. Kong, M. Sahadat, M. Ghovanloo, and G. Durgin, “A standalone intraoral tongue-controlled computer interface for people with tetraplegia,” *IEEE Trans. Biomed. Circuits Syst.*, 2019.

[76] M. Ghovanloo, M. N. Sahadat, Z. Zhang, F. Kong, and N. Sebkhi, “Tapping into tongue motion to substitute or augment upper limbs,” in *Micro-and Nanotechnology*

Sensors, Systems, and Applications IX, International Society for Optics and Photonics, vol. 10194, 2017, p. 1 019 413.

[77] M. N. Sahadat, “Design and evaluation of a multimodal assistive technology using tongue commands, head movements, and speech recognition for people with tetraplegia,” PhD thesis, Georgia Institute of Technology, 2019.