COMPUTER GAMES FOR MOTOR SPEECH REHABILITATION

MICHAEL BRANDON HAWORTH

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF SCIENCE

GRADUATE PROGRAM IN DEPARTMENT OF ELECTRICAL ENGINEERING
AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

JANUARY 2016

Abstract

This research investigates the problem of creating a system for interactive digital visual feedback of articulator kinematics measures for speech rehabilitation. Recent technology provides precise non-line-of-sight positional tracking of small sensors which affords exploration into the motion of articulators such as the tongue. By utilizing recent game development technology, articulation kinematics can be visualized in realtime. Using these technologies the basis for an interactive rehabilitation system is formed. The system is posed as both a research apparatus and a potential clinical rehabilitation delivery system. As such, this system provides an extensible software and design architecture for the creation of interactive feedback visualizations and kinematic speech metrics as well as a clinical research front end for the creation and delivery of speech motor rehabilitation protocols.

Acknowledgements

I would like to express my deepest appreciation to my supervisors, Professors Petros Faloutsos and Melanie Baljko. Their expertise, patience, and support has made my graduate experience a rewarding and transformative adventure. Professor Faloutsos' encouragement led to my pursuit of graduate studies, and it was by his invitation that I came to be a part of this exciting research project. I am appreciative of his immense knowledge in so many areas of our research and community, as well as his continued guidance and collaboration. Professor Baljko has been a tremendous source of knowledge, perspective, and support. I am grateful for her immensely valuable revisions and standpoints that have been so transformative. I must also acknowledge her ability to mediate, transcribe, and recapitulate knowledge in so many ways.

I would also like to thank Professor Sotiris Liaskos for accepting the position as my external examiner, as well as Matthew Kyan for accepting the position as my examination chair.

My sincerest appreciation goes to the members of the Speech Production Lab in the Department of Speech-Language Pathology at the University of Toronto. I would like to

thank Dr. Yana Yunusova whose guidance, motivation, and support in pursuit of speech research has been so valuable. She has provided direction, insight, and resources towards the successful completion of this thesis and the ongoing research project. I would also like to thank Elaine Kearney, who, amongst many research achievements, has spent so much time placing small sensors in the mouths of our study participants and running data collection sessions. Her guidance, insight, and research efforts have been immensely valuable in our collaborations. Her feedback and willingness to operate so many versions of software I wrote for our research has been paramount to the project.

In conclusion, I wish recognize the various sources of financial support which made this project possible: the Department of Electrical Engineering and Computer Science at York University; the Department of Speech-Language Pathology at the University of Toronto; the NSERC Discovery Grant Program; the ASHA Foundation New Investigator Award; the University Health Network – Toronto Rehabilitation Institute; the Parkinson Society of Canada Pilot Project Grant Program; and the Centre for Innovation in Information Visualization and Data-Driven Design (CIV-DDD).

Table of Contents

List of Tables

List of Figures

Publications from the Thesis

All data collection and testing presented in this thesis was conducted at UHN: Toronto Rehabilitation Institute. This work is covered under the University Health Network Research Ethics Board (Certificate: 13-6235-DE Visual Feedback Systems in Speech Rehabilitation).

An overview of the work has been published in workshop proceedings (Haworth et al. 2014b). A portion of Chapter 4 has been published (Shtern et al. 2012). Figure 4.4 and portions of text from Section 4.2 are reused with permission. Section 5.3 was presented as a poster at the Biennial Motor Speech Conference (Haworth et al. 2014a). Chapter 7 is based on collaborative work conducted at UHN: Toronto Rehabilitation Institute in the Vocal Tract Visualization Lab and at the University of Toronto Speech Production Lab. Analysis of preliminary data for Figure 7.5 provided by Elaine Kearney.

Chapter 1

Introduction

This thesis describes the development of a computer-based speech therapy system (CBST) for augmented kinematic visual feedback. This pilot CBST system is deployed as a speech research tool in a rehabilitation research context. The driving motivation here is the need to examine the effects of interesting visualized feedback in the field of speech motor rehabilitation therapy. This approach is prompted by speech and motor rehabilitation practices, a field that is now in the initial stages of moving towards the visualization and gamification of therapy processes. Researchers in this field require a system that affords the principles of motor learning, the practices of speech therapy, and the artistry and interaction of digital visualization, while providing the underlying computational requirements of a digitally driven framework.

The development of such a system should be a needs driven design task prompted by user, data, and hardware requirements. These requirements must be derived from a user-centric process of development, and lead towards an architecture for hardware

that delivers data requirements and for software that affords required functionalities and extensibility in the face of iterative and incremental processes of research. Thus, the research question examined by this thesis is as follows: What are the underlying architectural and computational requirements of a CBST for delivering augmented kinematic visual feedback?

1.1 Computational speech rehabilitation

Speech rehabilitation therapy encompasses a wide range of services provided by Speech-Language Pathologists (SLP) to optimize communication and swallowing in an effort increase an individual's quality of life (Bankson et al. 2002). Current clinical practices in speech intervention and rehabilitation rely on a wide range of techniques, highly dependent on the conditions being treated. For instance, evaluation of intelligibility and/or progress may be based on the SLP's auditory and visual perceptions of the client's speech, automated audio analysis, or family reported and self-reported improvements in communication. Other "low-tech" visual methods, such as the use of mirrors may be incorporated into clinical and at-home practice methods to increase efficacy or ease of treatment. State-of-the art technologies in speech research are now taking advantage of 3D electromagnetic articulography (EMA). EMA is a recently developed sensor technology that gathers large volumes of real-time data about the movement of the tongue and other articulators. Without such sensor technologies, these movements can not be easily studied, as most of the articulators are hidden from view and entail millisecond-duration movements.

EMA represents a leap forward in information accessibility over earlier technologies that have been employed by speech scientists, such as x-ray micro-beam, in terms of logistic feasibility—specifically cost and the need for highly specialized expertise, staff and infrastructure. Thus, for speech rehabilitation, the application of new technologies such as EMA presents a major step forward in clinical practice. To harness and take advantage of these technologies in the clinical setting, frameworks and systems must be put into place which afford the study of speech kinematics in the context of new clinical practices, and efficacy of treatment. These systems would provide the means to empirically evaluate new directions in motor speech rehabilitation, which is especially important in the case of neurodegenerative diseases, such as Parkinson's and amyotrophic lateral sclerosis (ALS), that have known speech motor symptoms.

1.2 Visual feedback

Visual information plays an important role in speech development, perception, and intelligibility (Grant and Seitz 2000, Kuhl and Meltzoff 1982, McGurk and MacDonald 1976). Enhancing visual information might benefit individuals with proprioceptive and tactile deficit symptoms of neurodegenerative diseases, such as Parkinson's disease (PD) (Marchese et al. 2000, Seiss et al. 2003). Since clinicians believe visualizations are important and provide a more motivating and memorable learning experience, we seek to drive visual feedback with EMA technologies. The primary motivation of this thesis is providing a system which affords the further study of this approach to speech motor rehabilitation.

The Vocal Tract Visualization (VTV) project, of which the system presented here is a subset of, is aimed at creating new clinical practices for speech interventions. The short-term objectives of this project are to describe the methodologies, architecture, and processes of the aforementioned system. The long term goal of the project is to devise novel empirically validated clinical speech rehabilitation practices that utilize state-of-the-art tracking and visual feedback technologies. In particular, this project is positioned at the intersection of many fields of study, namely speech-language pathology, human-computer interaction, computer graphics, computational linguistics, and computational geometry. As such, the research is driven by various intertwining goals: empirical speech research, user experience, rehabilitation practice, efficacy, and discovery.

1.3 Thesis objectives

This research project seeks to develop and to evaluate the steps taken in generating a preliminary system which affords the study and delivery of kinematic visual feedback for speech rehabilitation. In this thesis, I will describe the development of such a system, its software and hardware architecture, and underlying computational processes. The first aim is to break down the design of the system in a top-down approach from a set of design parameters and user needs. Careful needs analysis exposes a requirement for extensibility in the overall software architecture and flexibility in the front end which is addressed throughout. The second aim is to describe the necessary steps to making such a system evolve into a working tool mapping from architecture to computational processes. The

data control paradigm developed and applied in the context of filtering and metrics that make the system perform in production environments spanning the research and clinical. The final objective is to validate the system. This entails the design and computational processes of a data analysis suite defined by prior speech research and the needs of an active SLP research lab. This data processing and analysis pipeline consists of a series of computational processes that clean and measure data derived from an EMA system. This suite is then used as validation of the second aim, presenting both as the offline and online processes for kinematic speech data respectively and the convergence of their outputs as a measure of success.

1.4 Thesis overview

Chapter 1 is the introduction, expressing the need for evolving the current clinical speech practices with the latest speech research technologies, and rehabilitation practices. The need, and challenges therein, for a system to conduct such preliminary investigations into augmented kinematic visual speech feedback is expressed. The objectives of the thesis are introduced as a response to the preliminary needs of a long term project.

Chapter 2 is the literature review and theoretical background pertaining to motor learning practices, visual feedback and gamification of rehabilitation practices. This review supports the need, approach, and delivery of the system described in the thesis.

Chapter 3 describes the design space, development process, requirements, and evaluation strategies. These first steps and specifications are set in the context of the challenges

of an interdisciplinary project with numerous interacting goals. The Agile development methodology is described in response to these challenges, as well as the need for empirical evaluation of the system's modules in relation to background speech science, user experience, and compliance to stakeholder specifications.

Chapter 4 describes the hardware components, system architecture, software subsystems in detail. These architectures are explored in the context of requirements set out in Chapter 3.

Chapter 5 describes the online data process that produces accurate data driven visualization. These processes are the real-time filtering and transformations required to isolate articulator kinematics. An experimental method for producing absolute articulator motions in an arbitrary game space and its foundation in basic speech research is described.

Chapter 6 describes the offline data process for cleaning kinematic speech data and producing experimental measures for research. This process is developed in the framework of both carrying out research and as a ground truth for system validation presented in Chapter 7.

Chapter 7 describes the validation of the system in the framework of the requirements set out in Chapter 3. Two validation studies are described and there results detailed. The completion of tasks and status of ongoing tasks with respect to validation of the system as a whole are discussed.

Chapter 8 concludes the thesis reviewing the body of work, contributions, and outputs

of the research project. Future work regarding the overarching research project and the system is considered and discussed.

Chapter 2

Background and literature review

2.1 Introduction

This chapter summarizes the research literature relevant to the objectives of this thesis. In particular, this review addresses the following question, what is needed in a computational system which affords the study and delivery of kinematic visual feedback for speech rehabilitation?

There is a large body of results in speech science research literature pertaining to the focuses of this thesis. This review will delineate this body of work addressing a number of pertinent questions. These questions are: what evidence exists that visualization of the speech sound articulator can lead to improvements in speech? what are the clinical techniques for bringing about improvements to speech? how are these improvements actually accomplished? what is the retention of this learning?

This thesis focuses on the application of a certain mode of speech rehabilitation in a general software and hardware framework. As such, this literature review will high-

light the pertinent results from speech and motor rehabilitation science, which serve as the bases for the key design decisions that will underlie the software and hardware framework developed in subsequent chapters. The structure of the literature review is as follows the: principles of motor rehabilitation 2.2.1; the use of EMA as a step towards information accessibility 2.2.4; application of EMA in motor speech rehabilitation 2.2.4.1; derivation of intervention targets given background speech science and EMA afforded information 2.2.4.2; a look at speaker variation in terms of targets 2.2.4.3; and motivational and guiding work in the gamification of motor rehabilitation 2.3.

2.2 Application domain

In speech rehabilitation, clinical goals are typically derived from modifiable speech parameters which lead to increases in intelligibility. With the inclusion of EMA as a mode of data acquisition, speech rehabilitation is afforded the view of kinematics, or motor movement outputs, of hidden articulators, such as the tongue. This is a massive leap in accessibility of information regarding these articulators and has led to the expanded exploration of kinematic speech parameters. Furthermore, this has naturally led to the exploration of augmented kinematic feedback of articulators.

2.2.1 Overview of motor speech therapy

The clinical speech rehabilitation field aims to produce therapeutic outcomes. In particular, the primary desired outcome of speech rehabilitation is intelligibility. This outcome

is typically achieved through the modification of speech parameters. Speech parameters may be acoustic in nature such as volume or pitch, or kinematic in nature such as articulator speed or position. These articulators include the passive such as alveolar ridge, teeth, and hard palate or the active, which move relative to the passive, such as the tongue and lower lip (see Figure 2.1).

Methods for eliciting therapeutic outcomes in terms of speech parameters can be complex. Outcomes are predicated upon the successful positive modification of motor parameters and the sustainability of those parameters in the face of the motivation and engagement of the client. Thus, speech motor learning draws upon many general motor learning principles to build towards successful speech motor rehabilitation protocols. These motor learning principles are reviewed here.

Augmented feedback is supplemental information to a person's sensory feedback, such as sight or proprioception (Schmidt and Lee 2011). Feedback in motor learning practice has been divided in the literature into two categories, knowledge or results (KR) and knowledge of performance (KP). Both KP and KR are subsets of augmented feedback. This approach has found success in conditions where visualizations or supplemental information takes the place of a lack of information, such a hearing in deaf children or motor rehabilitation of hidden articulators.

With KP, feedback is supplied on the patterning and quality of the motor action in progress, or kinematics (Gentile 1972). Historically, this is delivered verbally via coaching during the performance of the motor action, e.g., feedback on the motion such as speed

of repetition or range of motion.

With KR, feedback is supplied on the parameters of the performance relating to the outcome (Salmoni et al. 1984). Historically, this is delivered via coaching on the result of the movement, e.g. feedback on a measure of the performance such as distance from the goal or time elapsed during the action. The benefit of KR comes with the inherit goal reaching and setting feedback (i.e. attaining goals and setting new ones), and has been known in motor learning for some time (Locke et al. 1968).

The benefit of augmented feedback in motor learning has been demonstrated repeatedly, however numerous studies have shown that control over feedback parameters such as frequency of the feedback delivered, bandwidth of threshold for providing feedback, and delay in feedback delivery has clear impact on the results of learning and must be taken into account (Winstein 1991). In particular, it has been shown that reducing the frequency of KR feedback improves retention (Winstein and Schmidt 1990). Furthermore, too frequent augmented feedback may even reduce learning or produce maladaptive behaviours or dependencies on feedback (Schmidt 1991). Reduction in feedback has also shown efficacious in speech motor learning (Hula et al. 2008). To measure the effects of KR feedback without confounding learning with performance it is necessary to provide no-KR retention and transferability design (Schmidt and Lee 2011).

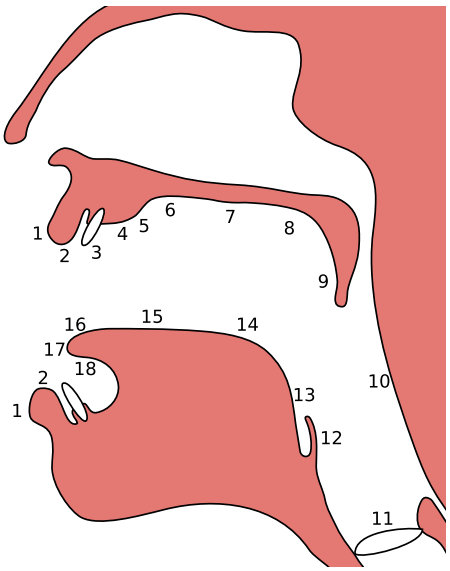


Figure 2.1: Cross-sectional overview of the active and passive articulators. 1. Exo-labial (outer part of lip), 2. Endo-labial (inner part of lip), 3. Dental (teeth), 4. Alveolar (front part of alveolar ridge), 5. Post-alveolar (rear part of alveolar ridge & slightly behind it), 6. Pre-palatal (front part of hard palate that arches upward), 7. Palatal (hard palate), 8. Velar (soft palate), 9. Uvular (or Post-velar; uvula), 10. Pharyngeal (pharyngeal wall), 11. Glottal (or Laryngeal; vocal folds), 12. Epiglottal (epiglottis), 13. Radical (tongue root), 14. Postero-dorsal (back of tongue body), 15. Antero-dorsal (front of tongue body), 16. Laminal (tongue blade), 17. Apical (apex or tongue tip), 18. Sub-laminal (underside of tongue) (Catford 1977). Image:Ishwar/Rohieb / CC BY-SA 3.0 - sagittal section image based on (Minifie et al. 1973).

2.2.2 Computer-based speech therapy

Computer-Based Speech Therapy (CBST) systems aim to provide therapeutic outcomes through computationally-based means, typically via automatic signal acquisition, analysis, and feedback design. Thus, CBST can provide a computational means to deliver speech therapy protocols and feedback in the framework of motor learning. To align with the motor learning literature, we distinguish between *product-oriented* and *process-oriented* approaches. Methods that focus on the final output of speech (i.e., the resultant sound) are said to be *product-oriented*, whereas methods that focus on the actions or motor programs associated with the production of speech are *process-oriented* (Povel and Arends 1991). Process-oriented CBST systems focus on articulators, deriving visualizations about some parameter of articulation during speech production.

2.2.3 Acoustic-based feedback

Product-oriented CBST systems focus on the final output of speech and, thus, depend on high-quality acoustic data acquisition. Historically, CBST systems tended to be product-oriented as opposed to *process-oriented*, since acoustic information is readily available at low cost. In both cases, the CBST systems are intended to provide visual feedback on some goal or target of the rehabilitation practice. These may be either speaker dependent or independent, providing goals which are individually defined or normalized respectively. In some cases these goals or targets are abstract and/or gamified, in the sense that they afford increasing difficulty, awards, and exploration of the feedback. In this section we

provide a number of examples of both types of systems, their feedback, and means of goal setting.

2.2.4 EMA-based feedback

Electromagnetic Articulography (EMA) is a sensor tracking technology based on the principles of electromagnetic induction, wherein induction coils induce current in the sensors. It provides a powerful alternative to tracking articulators that reduces exposure to harmful ionizing radiation such as in x-ray techniques (Westbury et al. 1990). The creation of EMA stems from a long history of needs to accurately track articulators, such as the jaw (Hixon 1971). Early methods for tracking hidden articulators were primarily limited to two-dimensional (2D) data outputs, and the devices required significant calibration and attention to detail (Perkell et al. 1992, Schönle et al. 1987). Later methods afforded full three-dimensional (3D) tracking of position and rotation (Kaburagi et al. 2005, Zierdt 1993). Methods and principles of measurement such as these led to commercially available speech research products, such as the Carsten ag (Carstens Medizinelektronik GmbH, Bovenden) line and the Wave Speech Research (NDI, Waterloo) systems. These commercial systems have been tested for their accuracy and shown adequate in the application of speech science (Berry 2011, Kroos 2008, Yunusova et al. 2009). Current commercial systems such as the Wave Speech Research System produce audio aligned and timestamped six-dimensional (6D) kinematic data, covering full positional and rotational data within their generator fields.

The need to track articulators, especially the tongue, comes with certain constraints and measurement effects. The design of flesh point, or point parametrized, tracking systems such as EMA requires that a sensor be placed at the point where data is to be collected. In practice, this means that sensors are fixed directly to the articulator, and as such may introduce reductions in intelligibility or alter articulator kinematics (Katz 2006). Furthermore, sensor placements may be consider uncomfortable, and must be taken into account. In particular, placement of a tongue sensor has been found uncomfortable if nearer than 1cm to the tip of the tongue (Hoole and Nguyen 1999, Perkell et al. 1992). The field based nature of these types of systems also introduces a need to account for spurious head motions, whether the head is free to move or not (Perkell et al. 1992, Westbury et al. 1990)

2.2.4.1 Visual speech therapy

Visual speech therapy is a form of augmented feedback in which successful production of speech parameters are displayed rather than coached. There is a rich body of literature regarding the use of visual speech therapy in deaf children, in which visuals help reinforce feedback with respect to hearing (Öster 2006).

In the case of speech disorders such as aphasia, apraxia, or dysarthria, which can cause articulation errors, EMA affords visual augmentation of proprioception during articulator movements. The literature describes efficacious application of EMA-provided feedback in visual speech therapy for motor speech disorders (Katz and McNeil 2010). This work

supports the direction of EMA supplied augmented feedback in the framework of prior motor learning principles, including combined KP ad KR in reduced feedback frequencies. In particular, it has been shown that EMA supplied augmented feedback may lead to a improvement of accuracy, or articulation placement, in speakers with apraxia of speech (AOS) and aphasia, specifically Broca's aphasia (Katz et al. 1999, 2007).

2.2.4.2 Speech intervention targets

Rehabilitation goals must first be defined on the basis of the nature of the speech 'signal' and then through production parameters that can be modified via motor learning. The selection of these rehabilitation goals with respect to their measurement should also be theoretically motivated and empirically validated.

Empirical evidence shows that somatosensory information is principal to speech production (Tremblay et al. 2003). Furthermore, this information is principal to speech movement precision (Nasir and Ostry 2006). Successful target production can be derived from absolute movement precision or relative measures. These goals may relate to anatomically spatial targets. It has been shown such that spatially derived measures may need anatomically based coordinate systems (Westbury 1994). On the other hand, studies have shown that some lingual consonants have distinct hard palate contact regions (Yunusova et al. 2012).

The current state of the literature on augmented kinematic visual feedback provides insight on clinician derived motor targets for EMA produced feedback. These augmented

kinematics are based on head corrected articulator positions. That is, current speech intervention targets delivered via EMA feedback are mainly positional in nature. In the training of the Japanese flap, a consonant sound produced by alveolar tap, the target contact point was defined as a region near the alveolar ridge. By eliciting speech kinematics which are produced by making contact with the alveolar ridge, the nearby posterior target was found. A participant repeatedly aimed for this positional target region using an EMA sensor position feedback screen and an experimenter marked target region during repetitions of the speech stimuli (Levitt and Katz 2010). In the treatment of apraxia of speech using EMA feedback, a target region was also defined by experimenter positioning. The region was fixed in size and based on the placement of the tongue tip for a given speech motor target after several baseline recordings of correct responses. During feedback trials the participant was shown the the EMA sensor position and the target region prompting the participant to guide their tongue to the place of articulation (Katz et al. 2010). A system has been engineered for delivering these types of investigator defined positional target regions in EMA supplied feedback called OptiSpeech (Katz et al. 2014). This system defines the target region as a fixed dimension sphere hand positioned at the alveolar place-of-articulation for English consonants. In this system the feedback is a virtual tongue model and the goal is to enter the target region with the tongue sensor point.

2.2.4.3 Speech variation

Interspeaker variability may be an inherit issue in the design and operationalization of kinematic motor speech intervention targets since individuals may adopt new lingual articulation patterns to overcome the affects of palate morphology on speech acoustics (Lammert et al. 2013). The definition of a kinematic goal for one speaker may not generalize to other speakers. In fact, speakers may produce entirely different gestural targets and global kinematics (Johnson et al. 1993, Westbury et al. 1998). This variability could arise from a number of conditions including disease, emotional, or morphological differences. This is in contrast to intraspeasker variability which in turn may present difficulties in goal setting during the continued treatment of an individual. It is important to note that many studies focusing on interspeaker variability draw from small populations and in many cases are two speaker comparisons.

At the macro level, speakers may produce different articulatory kinematics because of the type and expressive power of emotional states (Kim et al. 2011, Lee et al. 2005). Since speakers may enter a treatment session and react to clinical environments or approach the process of speech therapy differently, this may produce variability in articulatory kinematics across speakers. This points to potential intraspeaker differences as well, which may arise as a difference in emotional state at the time of treatment.

A body of research has shown that anatomical features, especially hard palate morphology, play a significant role in motor control strategies and result in different articulator kinematics. A biomechanical tongue model for studying interspeaker variability

was constructed (Winkler et al. 2011). These speaker-specific models, constructed as Finite Element Meshes (FEM), provide a means of studying motor control and articulatory variability. The study showed that individual articulatory strategies may relate to vocal tract morphology. An EMA and EPG based study of German alveolar obstruents related speaker-specific palatal contact points and general tongue positioning to differences in palate shape (Fuchs et al. 2006). The positional targets of consonant stops for english speakers has also shown a high degree of variability attributable, in part, to hard palate morphology (Rudy and Yunusova 2013, Yunusova et al. 2012).

2.3 Gamification of motor rehabilitation

Digital games can be described as goal driven interactive visualizations that embed artistry, desire, compulsion, and narrative into a user driven framework. They can provide abstract feedback on skill performance and deliver engaging narrative and visuals. The effect of games on learning, in many realms is a well known and active field of research. There is strong evidence for the use of games in rehabilitation practices (Lohse et al. 2013). Here the focus is on motor rehabilitation practices and their gamification, or the use of game design elements in non-game contexts (Deterding et al. 2011).

Examining and developing design processes for motor rehabilitation games has been identified as an important step in the following literature. The development of two childhood gait rehabilitation games, Gabriello v1.0 and v2.0, were documented as a means of supporting explorative design in this domain (Martin et al. 2014). The authors identify

a need to measure emotional response and physiological effort when designing motor exercise games. The iStoppFalls exergames was used to generate design principles and directions towards engaging user driven digital game aesthetics (Marston et al. 2014).

Gamification is not a simple process and there is a clear need for careful, user-driven design processes. Increasing engagement and enjoyment of a system is of particular importance. A series of multi-modal games for stroke rehabilitation were surveyed, which showed that various incremental design changes involving the feedback and challenges present in a "serious game" impact user engagement and enjoyment (Shah et al. 2014). These changes included increasing graphical fidelity, and making goals clearer and easier. It has been shown that feedback fidelity impacts engagement for older adults as well (Smeddinck et al. 2013). It is clear that, in gamifying rehabilitation processes, design which engages the user is important both for clinical outcomes and the user. Specifically, feedback fidelity, awards, and meaningful, clear goals appear to impact engagement and enjoyment of any gamified rehabilitation process.

2.4 Conclusion

This chapter presented a strategic review of the background theory and literature involved in the generation of a CBST for augmented kinematic visual feedback. Motor learning theory and application is reviewed and provides background on the need for control of feedback parameters in motor rehabilitation practice. This leads to a functional requirement in the system, as the clinician must be able to control feedback scheduling. The

application of CBSTs and EMA in speech therapy is reviewed providing motivation for this direction in development. There is preliminary evidence of efficacy for the approach of EMA-supplied feedback for speech motor learning. The background on developing kinematic targets and interspeaker variations in kinematics is reviewed providing insight into the difficulties and potential approaches to devising such targets. This too leads to functional requirements, since the system will be deployed to multiple users. Finally, the application of visualization and gamification in motor rehabilitation is reviewed motivating the use of interactive feedback and visualization for motor speech therapy. This highlights the need to take a user-centric approach to design and development, which is paramount to the system's successful deployment.

Chapter 3

Methodology

This chapter describes the methodology for developing and generating a system to deliver augmented kinematic visual speech protocols. This system requires the engineering of software modules that afford different functionalities for proper delivery. These functionalities are a product of clinical research desires to deploy EMA in a user-centric application domain.

3.1 Agile development

To address the needs of this project I chose to apply a loose adaptation of the agile development methodology (Beck et al. 2001, Cockburn 2006). In particular, I employed the methods of the SCRUM development process (Schwaber 1995). Key tenets of the agile development methodology, and in particular SCRUM, are preparedness for emergent requirements, quick cycles of identification and implementation small production units, empirical basis of units, and co-location of key participants (Schwaber 2004). The

SCRUM process is well suited to handle the modification and turnover of requirements.

SCRUM addresses requirements and development churn through a series a best-practices regarding meetings and environment. The focus of the process is an iterative and incremental development of backlogged features amongst a co-located team. The sprint is a regular cycle of development that begins with the sprint planning meeting, or weekly scrum. This meeting allows the key stakeholders (typically the product owner) to define what are the current priority needs, features, or functionalities—moving items from the product backlog to the sprint backlog. During a sprint, regular meetings, or daily scrum, are held to capture the progress and current impediments of contributors on the team. The sprint ends with a sprint review meeting to exhibit the products, or deliverables, and a retrospective meeting to reflect on process.

In interdisciplinary, cross-institution research projects, the nature of the contributors limits the use of some of features of SCRUM and requires adaptation of the process. To address these limitations: daily scrums are adapted into email updates; weekly scrums include the sprint review with presentation of empirical results and end with the sprint retrospective; and co-location of contributors is at a common institution where development, analysis, and trials can take place. For example, in this project, the process was driven by regularly held bi-weekly scrums to address theoretical concepts and massage them into potentially viable future clinical and research concepts. These meetings served as a platform for presenting production units and defining next steps in terms of empirical analysis and new units. Meetings, development, and analysis were held in a common lo-

cation, the UHN: Toronto Rehabilitation Institute, to encourage strong interdisciplinary exchange and promote progress on parallel goals.

SCRUM is particularly useful for the developing and emergent requirements found in research projects such as this one. This project is driven by the desire for novel clinical practices centred on visual feedback of articulator kinematics. The SLP stakeholders knew that they needed a general framework for visualizing EMA-based feedback relative to clinical goals, but not the nature of the feedback or goals. The emergent requirements of this project were, necessarily, empirically-based research results. This required iterative and incremental deliverables involving quick cycles of refinement. These increments would result in working prototypes which could be used and tested to define next steps and functionalities. To support these increments, the development process needed to include repeated pilot testing with different populations including healthy and treatment populations.

3.2 Requirements: Iteration #1

The research needs of the stakeholders generated the following requirements

R1: Use of the Wave system

Flesh-point hidden articulator tracking systems have evolved from ionized radiation producing x-ray techniques to less intrusive and dangerous commercial EMA systems. In particular, the most current iterations of commercially available EMA systems are highly accurate and produce quality data. Speech research stakeholders wished to make use

of the Wave Speech Research System (Northern Digital, Waterloo) a high precision, high frequency, 6 Degree of Freedom (DoF) electromagnetic articulograph. The system is capable of sampling articulator kinematics data at up to 400Hz with an accuracy of $< 0.5mm$ (Berry 2011). This error size is within acceptable limits for this context (Yunusova et al. 2009). The device is comprised of several components both hardware and software. The hardware components are: sensors (either 6 DoF or 5 DoF); sensor interfaces units; a control unit; a standard computer; and an electromagnetic field generator. The sensors are connected to the control unit which is in turn connected to the computer running software for receiving and presenting the data stream. Remote real-time data streaming is afforded by the TCP based Real-Time Application Program Interface (RTAPI). The motions of the sensors are valid within a certain selectable cubic field size. In the lab configuration that is relevant to this research project the valid field is $300mm^3$ to the side of the field generator, with its origin at the centre.

R2: Multi-user sytem which implements clinical protocols

The needs of the the user are split between two user classes, the participant and the clinician. The clinician must be able to deliver speech motor therapy protocols. The participant must be able to interact with the system, receiving visual feedback and driving exercise progress.

R2a: Clinical protocol follows the principles of motor learning

Clinical speech protocol typically revolve around the repetition of some speech stimuli. These stimuli are reading passages usually loaded with some linguistic or phonemic ar-

tifacts designed to elicit certain utterances. In this system, these utterances produce kinematics data which ultimately drive visualizations. The visualizations are digital visual feedback related to the kinematics.

1. construct exercises of both clinical and research protocols, allowing a clinician to set up repetitions of of stimuli matched with visualizations driven by incoming articulator kinematics. These exercises have parameters and processes that apply to different experimental controls in the research. These include:
   (a) goal setting based on pre-exercise calibrations to define client capabilities
   (b) feedback scheduling for visual feedback frequency, as a process of motor learning (see: Section 2.2.1)
   (c) withheld, or no feedback, for learning retention (see: Section 2.2.1)
   (d) visual feedback driven by goals and metrics
   (e) success feedback driven by progress and achievement of goals
2. deliver these exercises, prompting the participant with stimuli in order to produce feedback
3. provide visual feedback, affording Knowledge of Results (KR) and Knowledge of Performance (KP) (see: Section 2.2.1)
4. provide results
5. drive feedback and results with kinematics

R2b: System is useable for participants

The system is delivering the clinician defined protocols to the participant. The participant-involved portions of the protocols must be clear and easy to learn, or recognize, from the visual interface. The user must be able to follow the course of the exercise mainly from the on-screen prompting with minimal input from the clinician, outside of basic instructions.

R3: System robustness

The system functions robustly with respect to the goals of **R2**. This includes generation of an exercise by the clinician and successful deliver to the participant. These are predicated on the functionality of the system architecture.

R4: Establish system fidelity

The system produces measures from the incoming articulator kinematics. These measures are accurately represented in the visualizations.

3.3 Requirements: Iteration #2

Subsequent scrums served to refine and development several of the requirements.

R2a.1a: Clinical goals

In order to identify a suitable clinical goal, the SLP stakeholders decided to focus on a particular population: people with Parkinson's Disease (PD). In PD, motor control can be strongly affected. In particular, symptoms concerning speech motor control are related

to a reduction in the range of motion across all articulators (Forrest et al. 1989, Yunusova et al. 2008). As such, the first targets or goals of the system were those associated with the expansion of range of motion.

R2a.5 Drive feedback and measures with kinematics

Online processes are defined by the need to produce real-time feedback of measures. The online data process is primarily dictated by the live data streaming of the Wave system. It was found that the Wave does not perform online head, or reference sensor, correction for the recorded sensors. There was also no guarantee of recording stimuli length or content for future feedback scenarios. There is also the desire to "future-proof" the system in the sense of affording different modes of game play by supplying methods for both absolute and relative articulator based game controllers.

1. **Head correction**

   the Wave does not include an anchor system for the head and relies on 6 DoF reference sensors to afford free head movement. Thus the system requires an automatic, live method for removing head motions from desire articulator kinematics.

2. **Data filtering**

   given the unknown nature of future use and data artifacts found during preliminary tests, there was an expectation of data artifacts such as spikes and gaps in the system. This coupled with the realtime requirement dictates a requirement for online filtering, or smoothing.

3. **Controller processing**

   to afford the creation of various types of games within the framework, it was necessary to define different controllers or processes that transform articulator movements to game design spaces.

R2a.2: Speech stimuli

A set of loaded sentences was chosen to elicit speech. Loaded sentences are tongue-twister-like stimuli that focus on a single phoneme or sound which is distributed regularly throughout the sentence. These need to be delivered in sets of linguistically related stimuli (3 randomly ordered in each set of repetitions) and individual stimuli (1 for each set of repetitions). These repetitions of stimuli need to be captured correctly between onset and offset of the speaker's productions.

R3: System robustness

The coupling of the Wave system with an external visualization platform will require a series of interconnecting and interfacing components. This system of coupled components will transfer data across their various links. This need requires verification of the component links and their data transfers.

1. verify channel functionality, it is necessary to verify the individual component connections. No data can be lost or spurious artifacts added between the component connections.

2. the data rate at the remote connection and the game system must match the RTAPI requested data sampling rate .

R4: Establish system fidelity

The use of the Wave in a novel real-time manner differs from default data recording process. That is, the online and offline process which handle the transformations and filtering of incoming live data during treatment sessions, and the processing of speech kinematics for research respectively are fundamentally different. The key difference between the two being that online processes are necessarily applied to the local signal due to the real-time requirements, whereas offline processes are applied to the global signal as defined by speech research needs. The Wave typically produces head-corrected data during an offline blackbox export process. To drive the augmented kinematic visual speech rehabilitation system accurately there is need of an online data process that mimics the output of the offline recording process in data transformation and quality as well as output measures. The offline process is poised as the ground truth for which the online process is evaluated. The divergence of these two streams is an indication of infidelity.

1. **Offline processing**

   to establish a ground truth requires a pipeline capable of processing speech kinematics as driven by the needs of the speech lab's research goals. These needs diverge into two-types of data: short recordings containing repetitions of a single speech stimulus; and long recordings containing single repetitions of long form speech stimulus.

   (a) short recordings are typically less than a minute and contain multiple repetitions of loaded sentences. These require the parsing of repetitions into single

recordings to derive metrics for individual utterances

(b) long recordings are typically between one and three minutes and contain an entire reading passage. The passages require the removal of long pauses that introduce unwanted non-speech kinematics. Preliminary analysis also dictated a need to remove spurious data artifacts such as spikes in sensor position data an gaps, or loss of sensor data.

3.4 Summary

This chapter sought to devolve the development methodology of the system at hand. This included the development process methodology, limits imposed by the nature of stakeholder involvements, subsequent process adaptations, and the development of two set of long term requirements. SCRUM and its adapted use as this project's development process was detailed to give the reader and idea of the challenges presented by generating such a system and how one might tackle them effectively. The iteration of requirements presented here represents a move from theory and need to prototype in iteration #1 to working production system in iteration #2. These requirements drive the rest of the thesis and are summarized as follows. The requirement to drive the entire system using the Wave Speech Research System as the primary sensor device (**R1**). The requirement to deliver visualized feedback of articulator kinematics in the framework of motor speech protocols (**R2**). The requirement to verify system robustness as a function of a working interconnected architecture (**R3**). The requirement to establish system fidelity as a func-

tion of convergence between offline kinematic speech research data processing and the online real-time data processing (**R4**).

These requirements can be operationalized as a list of final functional criteria the system must meet:

1. Incremental packet numbers at the end point in the system
2. Received packet frequency at the end point matches requested
3. Convergence of online and offline AWS volume outputs
4. Online head correction
5. Online data smoothing/filtering
6. Implementation of new visualizations and metrics possible

Chapter 4

System architecture

This chapter presents a hierarchical view of the entire system architecture from its hardware components down to the individual software subsystems at the leaves of the hierarchy. At each level, the architecture is described in terms of the requirements it addresses.

4.1 Overview of system hardware structure

The system is comprised of a series of digital hardware components interconnected to support the robustness of the system. These hardware components are the sensor device, the data collection and sensor control system, and the visualization system, see Figure 4.1.

The data sensor device is the NDI Wave Speech Research System (see: **R1**). The data collection device is a desktop PC (Microsoft Windows 7 Professional, Intel® Core™ i3-2100 CPU @ 3100MHz, 2 Cores, 4 Logical Processors, and 4.00 GB RAM) running the supporting WaveFront and WaveProxy softwares supplied by NDI (data collection control, and remote data streaming support respectively). The visualization system is a desktop

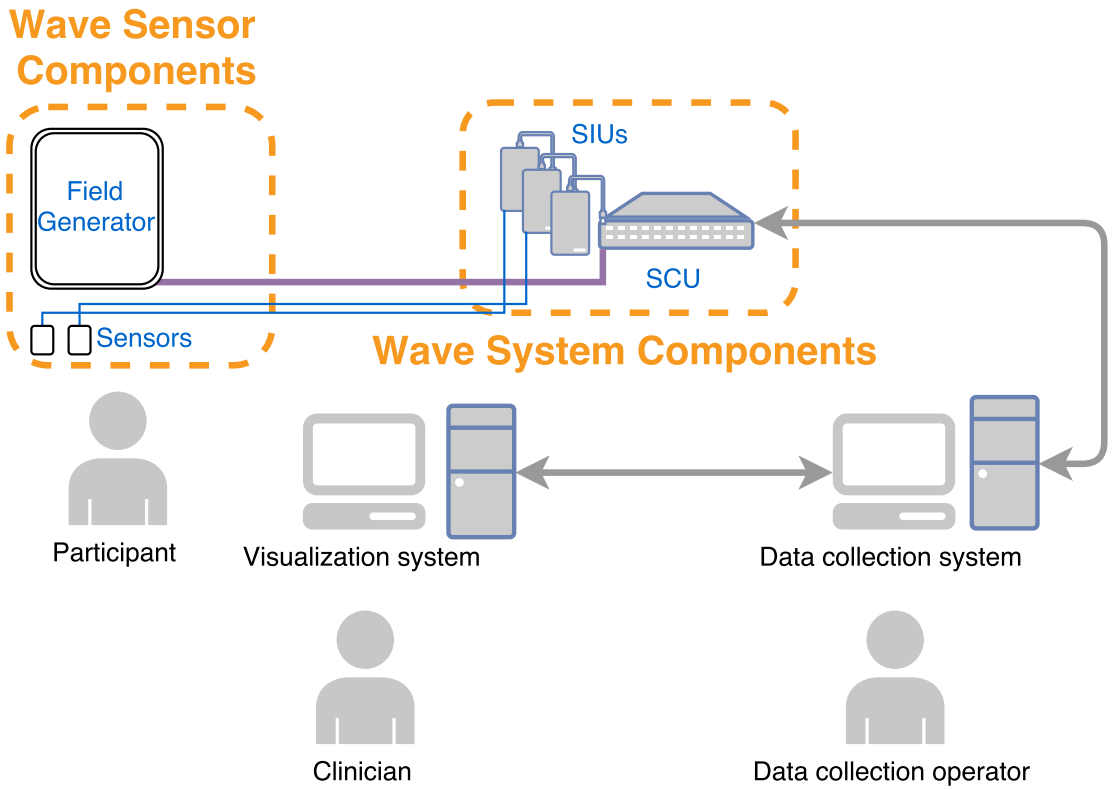


Figure 4.1: Overview of the system hardware including the Wave system components and control machines.

PC (Windows 8.1, Intel® Core™ i7-4790 CPU @ 3600MHz, 4 Cores, and 8.00 GB RAM) with a graphics card (NVIDIA GeForce 720 2.00 GB) for rendering visualizations, game engine software, and a remote data streaming application. The details of this application will be described in Section 4.3.1. The Wave device connects to the collections systems via proprietary hardware comprised of the following: sensors to sensor interface unit (SIU) (Figure 4.3(a)); SIU to sensor control unit (SCU) (Figure 4.3(b)); SCU to collection system via serial bus and serial to usb connections. The collections device connects to

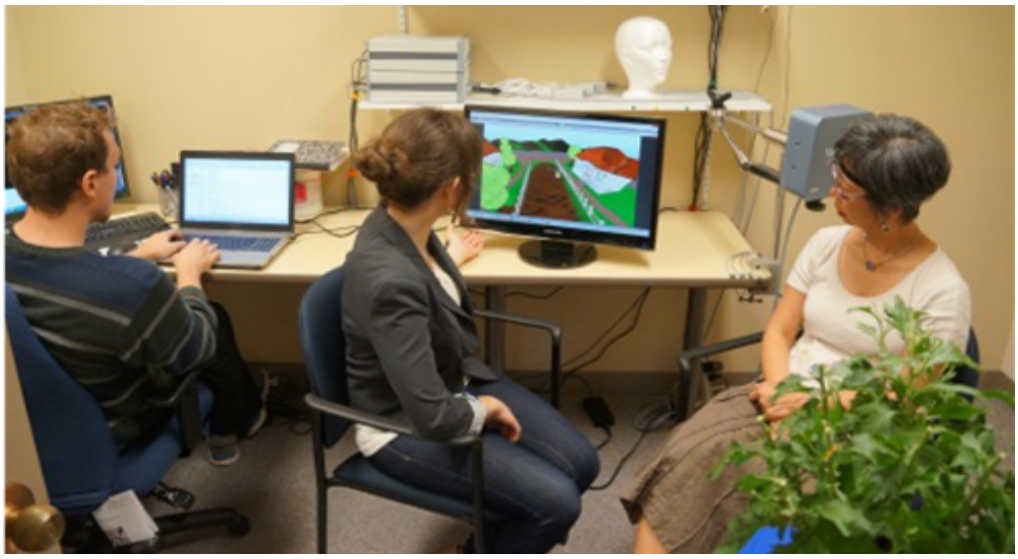


Figure 4.2: An overview of the system in practice.

the visualization system via an ethernet crossover cable to enable direct two-way TCP communication.

4.2 Overview of system components

The architecture of the system components an overview of the various working parts and their symbolic or direct relationships as seen in figure 4.4. These parts connect three principal pieces of the clients. These are: 1) the clinician 2) the patient, or participant, and 3) the data collection device. The system, users included, function as a whole to produce motor learning through visual feedback of articulator kinematics. The clinician is an SLP, the participant takes part in clinical and/or research based exercise protocols, and the data collection device serves to drive the system's data stream. These clients,

(a) (b) (c)

Figure 4.3: The primary components of the NDI Wave Speech Research System: (a) Sensor Interface Units (SIU); (b) the Sensor Control Unit (SCU); and (c) the field generator.

operating each module respectively, can be seen in Figure 4.2.

The rest of the system is comprised of software controlling the data channels as well as the delivery of and interfaces for clinical and research protocols. These parts are represented abstractly as follows:

- **Clinician User Interface (UI)**: A set of Unity modules that allows a clinician to configure a training session, associate games or visualizations with speech stimuli, input patient information, and drive protocols.
- **Middle Layer**:
  - **Unity Plug-in**: The game engine driver and the main interface between the middle-layer and the Unity game engine serving as a client to the data server.
  - **Wave Filter**: A series of filters that transform and clean the raw motion data into usable control signals.
  - **Dispatcher**: A module that is responsible for distributing high level data

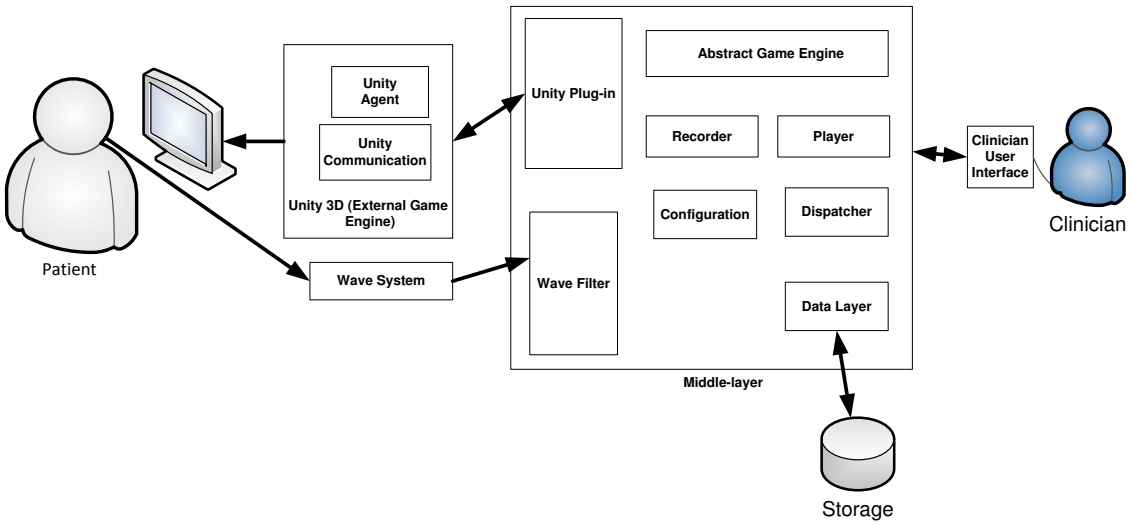


Figure 4.4: Overview of the system components including abstract constructs.

between the middle-layer components and the external network client (and locally the Unity Plug-in).

- **Data Layer**: A module that implements the main database operations for storing kinematic traces, recorded sessions, and patient-related information.
- **Abstract Game Engine**: A layer that serves to make the framework game engine independent from the other middle-layer modules, affording use with other interactive visualization engines.
- **Configuration**: A module that configures each session on the basis on the parameters set by the clinician.
- **Recorder**: A module that records each training session and stores its relevant data in the main database.
- **Player**: A module that affords replay of any previously recorded session that

is stored in the main database.

- **Unity 3D (External Game Engine)**: Commercially provided game engine.
  - **Unity Communication**: An abstraction of the communication with our middle layer.
  - **Unity Agent**: A module that translates between the gaming instructions in the abstract game engine's format and the Unity3D scripts.

These abstract modules take shape in the form of different softwares. Data collection and transmission starts at the proprietary software set WaveFront and WaveProxy. WaveFront provides both a recording interface, most notably for speech research, and live visualization of sensor positions and orientations within the field generator's valid field. The WaveProxy is a background process initiated by WaveFront that affords access to remote streaming and control of data channels via standard network protocols (TCP).

A central piece of software, dubbed the DataMuffin, handles much of the middle layer in the system architecture. It streams data from WaveProxy to the Unity game engine via client server relationships, provides a pipes and filter architecture for pre-filtering data, and provides secondary interfaces for storing data in SQL databases (as opposed to the standard text file provided by WaveFront).

Finally, an architecture in the unity game engine provides several of the exercise, or trial, specific functions—these are explored in detail in Section 4.3. These include: exercise creation; visualization and stimuli selection; visualization and interactive control

(feedback, or gaming scenarios); and exercise data collection.

The architecture presented here addresses a number of the requirements set out in Chapter 3. The use of the Wave system, **R1**, is embedded in the hardware and components architecture. The construction and delivery of motor learning exercises, **R2**, are afforded by the clinician's user interface and the combined processes of the middle layer.

This architecture was a partial input to the design process. The system as a whole makes use of the architecture as a core streaming process that connects the Wave data stream to the game system. Later additions and changes to this architecture included restructuring the TCP connections to handle queueing of incoming data packets from the Wave system, 6 DoF data packets for the streams and database storage, and binary data streaming (originally ASCII-based). The game system (see Section 4.3.2) makes use of the Unity Communication layer as an external dynamic link library (ClientLib) to facilitate Wave data access.

4.3 Software subsystems

4.3.1 DataMuffin

The DataMuffin is the central piece of software that enables the remote connection to WaveProxy via RTAPI and the local connection to the game engine via the client library. These connections are modelled as client-server relationships and build upon the reliability of the TCP network protocol. The remote connection with the data collection machine is produced by connecting as a client to the data collection system running the WaveProxy

sever and issuing RTAPI commands to open a TCP-enabled data stream. Locally the game system connects as a client to the DataMuffin server and opens a TCP-enabled data stream.

The DataMuffin provides a user interface for controlling the connections of the data streams, monitoring the status of the connections, and recording streaming data to a database. This interface is build in the Windows Presentation Foundation graphical subsystem and takes advantage of the command pattern and extended command delegates in the model-view-viewmodel pattern (MVVM).

The DataMuffin provides both filtering and storage capabilities. These are afforded through a pipes and filters architecture. In this architecture filter, data filters or storage procedures, can be implemented as standalone objects in pipeline. Thus, the DataMuffin has an extensible framework for filtering data and pushing it to databases. The database objects are generic such that different types of database storage methods can be used. For the purposes of this research project a MSSQL database is used.

4.3.2 Game system

The software architecture described here enables the flexibility and extensibility required of such a system. The game system must deliver motor learning protocols (MLP) as experimental trials. This includes both the visualized (KR) and non-visualized (retention) MLP, see Section 2.2.1. The experimental trials consist of exercises wherein participants repeat speech stimuli and receive feedback on their success and progress. These exercises

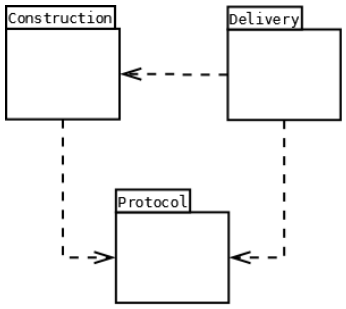


Figure 4.5: Overview of the subsystems in the System Architecture.

include stimulus prompting, recording of speech data, selection of feedback (or not), visualization (or not), and summary feedback. These constitute the exercise parameters, which the clinician must be able to select in order to produce a desired protocol.

From an abstract point of view, the architecture is composed of three subsystems: definition of protocol; construction of exercises within protocol; and delivery of exercises. While, the subsystems are interconnected by various components and te game engine itself, the general architecture has a dependency structure loosely isomorphic to the Model-View-Controller (MVC) pattern. That is, the delivery of exercises is dependant on both protocol and construction, while construction is dependant on only protocol, and protocol is not dependant on either as seen in Figure 4.5.

Subsystems in the game system architecture are broken into classes that implement the necessary features. The core features serve as pillar for the extensible features. The extensible features are associated with the needs to support various protocols, metrics related to treatment conditions, and visual feedback. The primary goal of the system is to deliver visual feedback and provide interfaces for constructing protocols. These

subsystems are as follows:

- **ExerciseBuilder**: An interface for selecting the parameters of the exercise protocol to be delivered. This interface implements the front-end for **R2a.1**.
- **ExerciseAgents**: A generic object that delivers exercises to the participant. This can be inherited from to support fundamentally different exercise protocols, such as live visualized feedback versus KR based feedback exercises.
- **ProtocolExercises**: A generic exercise representation that encapsulates the material and structure of an exercise for a given protocol. The ExerciseBuilder constructs this from the clinician-selected parameters and delivers it to the ExerciseAgent.
- **SummaryFeedback**: A generic summary KR feedback class. This can be inherited from for end-of-block summary feedback, end-of-session summary feedback, or even post-session feedback incorporating take home results, such as an HTML formatted report for display or printed out in hard-copy format.
- **WaveController**: The main data driving controller. This class implements the relationships between sensors, such as head reference and tongue, supports filters and transformations of the sensor data. Ultimately it drives a publicly accessible in game scene object which other classes can access to gather kinematics or visualize articulators.
- **ParameterController**: The controller for updating metrics based on articulator kinematics coming out of the WaveController.

- **SensorMetric**: A generic SensorMetric class affords the implementation of desired parameters, which can be stored for research or used to drive visualization. Each SensorMetric child implements a specific parameter computation and stores timestamps snapshots of that parameter, such as distance, position, or kinematic speech motor measures. These parameters are accessed and controlled via the ParameterController.

- **Sensors**: The main access point in the game to kinematic data. This class connects to the DataMuffin via the supplied client library, and enables the streaming of a single sensor's kinematic data.

- **VisualizaitionBehaviour**: A generic class implementing the basic need of a visualization. It is a render-able game object connected to the game engine's Update loop. This allows all manner of visualization to be produced and easily added to the system. These are selectable in the ExerciseBuilder, and delivered by the ExerciseAgent.

4.3.3 Design summary

The designs presented in this section support flexibility and extensibility of an ongoing research system. In this project, there is a need to support the research of different motor speech disorders, populations (ages in particular), and protocols. Different disorders may require new and interesting metrics requiring new modes of control. Different populations may require artistic approaches to feedback and visualization designs. Finally, clinicians

may wish to design new and interesting protocols that take advantage of augmented kinematic visual feedback which will require new types of exercises and modes of delivery. This ongoing, iterative, and incremental approach to research without the need to re-engineer in the face of new requirements is afforded by the designs presented here.

4.4 Summary

This chapter described the resultant design of the system, addressing the requirements set out in Chapter 3. The system is broken down into its constituents parts by first starting a birds-eye view of the architecture as a whole. The hardware architecture was described in detail to give an overview of the experimental set up. The abstract system architecture was described as a series of interacting subsystems that afford different functionalities for the user population. Finally, the software subsystems implementing the various processes and connections required by the system were described.

Chapter 5

Online processing

This chapter describes the underlying computational techniques of the live game system. First, the head correction process that isolates the articulator kinematics is described. Second, data filtering for smoothing potential spikes and replication of the offline low-pass filtering is described. Then an experimental general normalization technique is described as a means to position an articulator sensor avatar absolutely in an arbitrary game space. Finally, the metrics derived from the online kinematics data stream are described. These are used for speech research and to drive the visualization feedback central to the system design.

5.1 Head correction

To isolate articulator motions from the combined articulator+head motions, the system must account for the free moving head (see **R2a.5.1**). The 6D reference sensor is placed on a static position on the head. For the purposes of this research project and experimental

Figure 5.1: Tongue and head sensors attached to a participant sitting within the active field of the Wave system's field generator.

configurations, the sensor is attached to a headband and placed high on the forehead, see Figure 5.1. The implementation of this transformation required an extension of the core streaming system to included 6 DoF sensor data. It also required the implementation of an online process for performing the head correction over the desired sensors. An example of this sensor configuration, the field generator space, and the head sensor orientation can be seen in Figure 5.2.

Since the head is free moving, a frame of reference needs to be chosen for which the head moves relative to. This can be a fixed position defined by the game designer or defined empirically. For the purposes of this research the first good initial head sensor sample is considered the reference. The relative difference between the current head position and orientation and the reference dictates a transformation matrix which can be applied to articulator sensors. This process is illustrated in Figure 5.3.

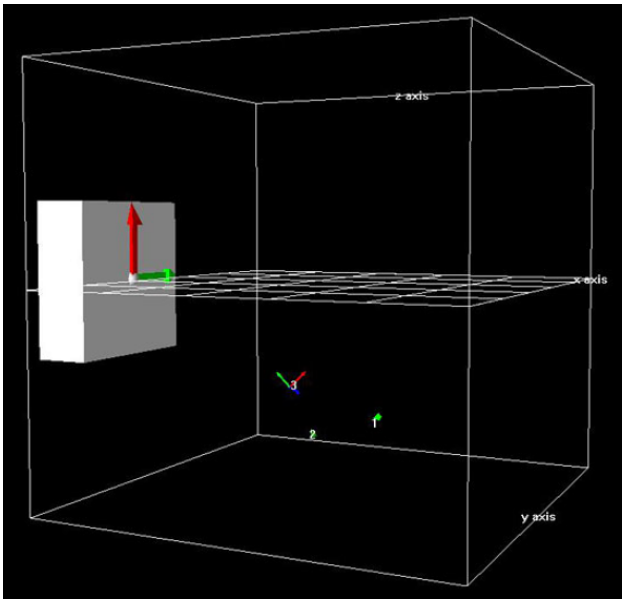
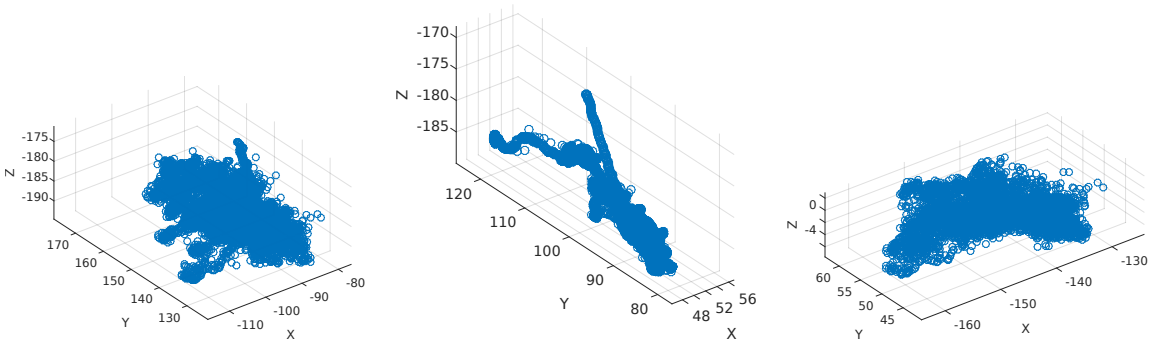


Figure 5.2: Tongue and head sensors shown within the coordinate system of the Wave field generator.



(a) (b) (c)

Figure 5.3: Head correction illustrated on a recording including repetitions of sentences. (a) Raw articulator (tongue tip) data (b) head movement isolated (c) articulator data with head movement removed. Note the range and swapping of the axes in the final product (c).

In more detail, the initial or reference head position vector and orientation quaternion are p_i, q_i, likewise for the current head sensor are p_h, q_h respectively. The relative position vector and orientation quaternion of the head sensor are p_r, q_r respectively, where $p_r = p_h - p_i$ and $q_r = q_h^{-1} * q_i$. The tongue sensor position vector and orientation quaternion are p_t, q_t respectively, and finally the head corrected tongue parameters p_t^h, q_t^h are:

$$p_t^h = Im(q_r * (0, p_t - p_r)), \qquad q_t^h = q_r * q_t \tag{5.1}$$

where Im is the imaginary part of a quaternion, and '*' indicates quaternion multiplication. In practice, to subtract head rotations, an angle axis rotation is formed by the quaternion and the sensor is rotated around the point p_i after being translated by $p_t - p_r$.

5.2 Data filtering

Positional time series data recorded with the Wave system is subject to artifacts. First, the sensor system is subject to high frequency submillimeter noise, as shown in Figure 5.4. In preliminary tests, data artifacts such as spikes were found in some recordings, as shown in Figure 5.5. In many cases, especially with long recordings, these spikes are accompanied by gaps, or missing data, as shown in Figure 5.6. To address these spikes in real-time, a moving median filter is used. The window width of this median filter is defined empirically. From preliminary tests, it was found that a window size of 3-5 samples works best for recordings at 100Hz and of 6-21 samples for recordings at 400Hz. An example of median filtering on data with spikes and gaps is shown in Figure 5.7.

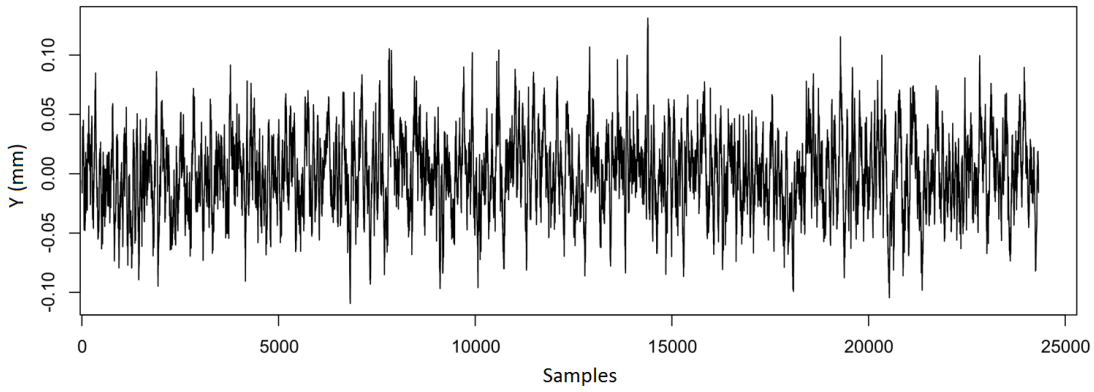


Figure 5.4: An example of high frequency noise in an articulator sensor position time series captured using EMA, shown in one dimension for clarity.

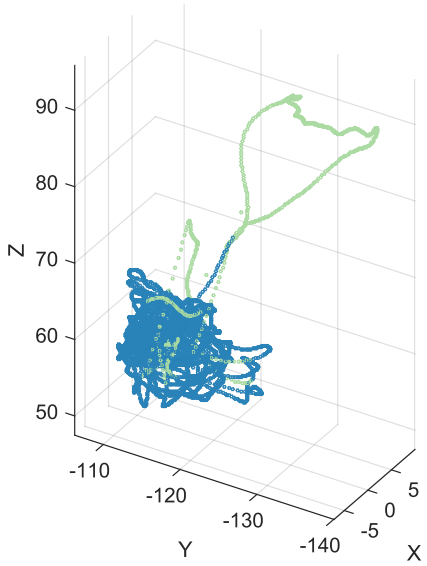


Figure 5.5: An example of a labelled spike (green/lighter) in 3D articulator trajectory data.

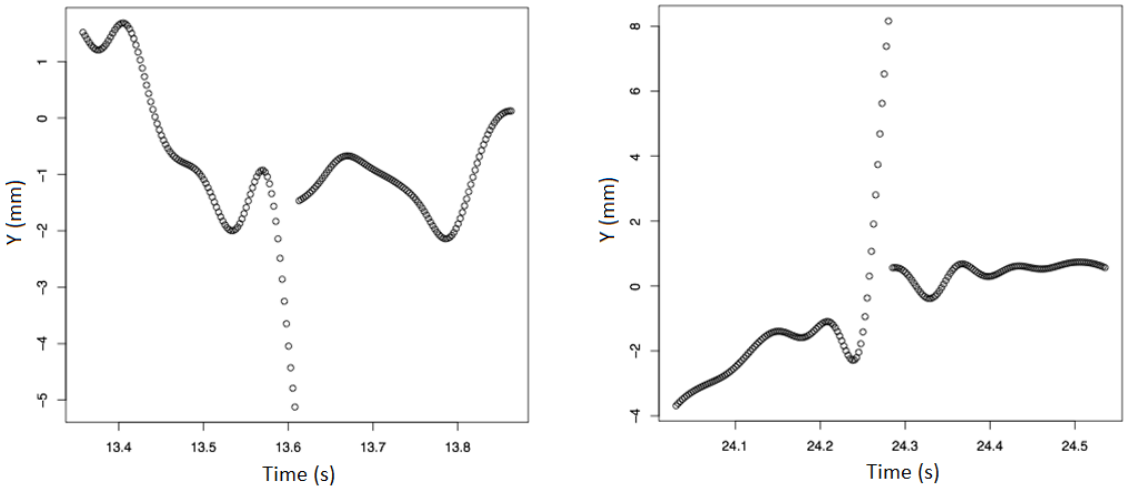


Figure 5.6: Two examples of spikes accompanied with gaps in an articulator sensor position time series, shown in one dimension for clarity.

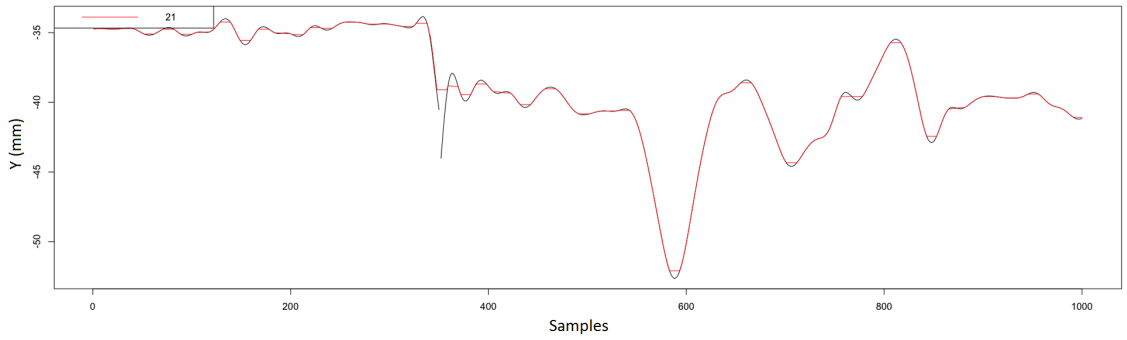


Figure 5.7: The effects of median filtering on data with spike and gap artifacts in an articulator sensor position time series, shown in one dimension for clarity. The data shown here is recorded at 400Hz, the black data is the raw signal and the red (lighter) data is median filtered with a window size of 21 samples.

5.3 Normalization

In game play, there are innumerable modes of input which afford interaction with game elements across genres. In particular, in motion-based games there are absolute and relative controllers. In this research project, the Wave is appropriated as a type of game controller. The Wave records motions in the space of the field generator and further, as per Section 5.1, the final articulator motions take place in head space. The head reference is placed on the participant in a consistent way, but between sessions and participants this is not guaranteed to be exactly the same. Thus, for games that require absolute positioning of the articulator sensor, a transformation is needed which brings the arbitrarily placed articulator motions into the visible usable game space. The process by which this transformation is formed and the final online transformation is referred to here as normalization.

For the first and primary treatment population of this research project, users with PD, the metrics driving the game are independent of position and orientation of the visible game volume, and the visualization, or feedback, was a measure of progress defined by the metric and calibration recordings. This means that for the first population, the system only required head correction and filtering to produce a usable signal, see Section 5.4.1 for calibration based derivation of signal and targets. However, future treatment populations and their needs were considered. It was found that, in order to provide feedback on real-time absolute articulator positions within a the game design space, processes were needed to define and transform articulator movements into visible, or usable, game space. These

processes would afford the control of an avatar in the game space in both an absolute and relative fashion - controller modes common to games. To achieve this controller flexibility, a two-fold process is required for forming a transformation,or normalization, that encompasses all articulator movements in a game space. First, a centre is defined. This is possible through process known as biteplate correction, which provides an anatomical centre consistent in definition across participants. Second, the bounds or range of motion of the articulation space must be defined. The range of motion is partially defined by biomechanical constraints, i.e. the hard palate, but has no strict bounds in all other directions. This means the process for defining bounds would need to be a novel empirically defined calibration process that encompasses the entire range of motion of the desired articulator.

5.3.1 Biteplate correction

The need for an anatomical basis in speech measures is known (see Section 2.2.1). Biteplate correction provides such a basis by giving an anatomically based position and orientation for use in correction. A biteplate consists of sensor(s) attached to a rigid plate which is marked with a centre line and tooth limit. The plate is placed in the participant's mouth by aligning the centre line markings with the participants midsagittal plane and their front teeth with the tooth limit marking.

This basis is static with respect to the head reference position and thus can be achieved by taking a snapshot of a biteplate position and orientation and using this to later trans-

form articulator sensors in the same manner as head correction by performing the added translation and rotation to bring the biteplate to the origin of the space. This effectively provides a cross-speaker anatomical basis for all tongue motions, further simplifying the transformation of articulator based controls to game spaces.

5.3.2 Range of motion

For the purpose of this research project, the primary articulator is the tongue. The human tongue is a flexible muscular organ with an ephemeral deformable shape. Thus, the range of motion of an articulator, especially the tongue, is difficult to define without calibration, or testing of limits. In the case of articulator motions the range of motion during articulation is desired. To capture the range of motion for game calibration use, an experimental calibration method was proposed on the basis of speech research and tested experimentally. This experiment is described in the following Section 5.3.2.1.

5.3.2.1 Experimental characterization of articulator range of motion

SLP practitioners make use of speech passages to elicit the range of linguistic and phonemic characteristics of a language. In the case of the English language a number of these passages exists, such as *Grandfather Passage* and *Rainbow Passage*. An experiment was conducted to see if these passages could elicit kinematics that encapsulate the range of motions of articulation during speech. The hypothesis was that if the passages elicit the range of linguistic patterns in English, then they may also elicit the range of kinematic

patterns and thus the range of motion of a given articulator.

Methodology For this experiment, the population was 15 healthy adults (7 male, 8 female) aged > 60 years (mean: 69.7 ± 2.96). These participants were native speakers of Canadian-English, had no reported history of neurological impairment, speech and/or language difficulties. The participants were also tested for normal hearing, vision, and cognition using a standard Snellen chart vision test, tonal range hearing test, and the Montreal Cognitive Assessment (MoCA) (Nasreddine et al. 2005).

The speech stimuli in this experiment were two repetitions of the *Grandfather Passage* and *Rainbow Passage* each and a series of randomized sentence stimuli including various loaded sentences—all elicited at a normal rate and style. Participants were asked to utter these stimuli while sitting the working field of the Wave Speech Research System. Each participant had one 6 DoF head reference sensor attached to a head band, one 5 DoF articulator sensor attached 10mm posterior to the tongue tip, and one 5 DoF sensor attached at tongue back 30mm posterior to the tongue tip. These sensors were used to capture the articulator kinematics of the tongue during the speech stimuli utterances.

All tongue kinematics were low-pass filtered using a 5th order Butterworth filter with a cutoff frequency of $15Hz$. The kinematics were then filtered for spikes and data artefacts using a 2SD ellipsoid point test, as described in Section 6.5.2.

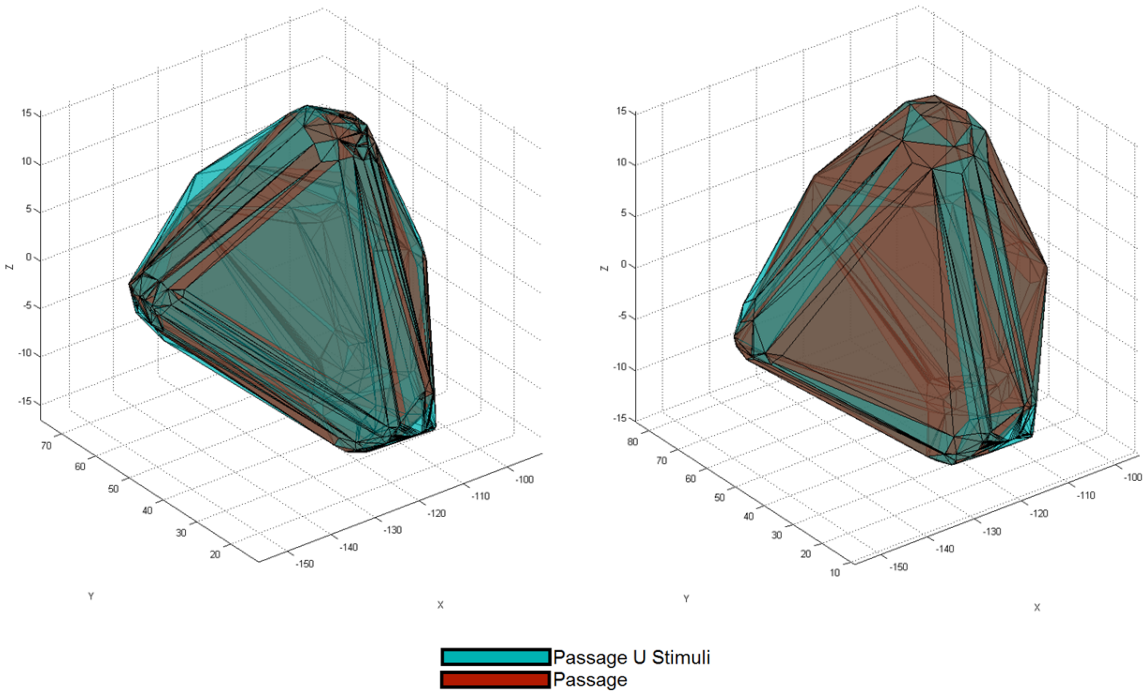
Results The speech utterances produced the following 3D point parametrized trajectories: *Grandfather Passage*, $k_{gf1}, k_{gf2} \in K_{gf}$, the *Rainbow Passage*, $k_{rb1}, k_{rb2} \in K_{rb}$, and the randomized sentence stimuli, $k_n \in K_{stim}$.

The spaces of the speech kinematics are compared using AWS volume, or the volume of the convex hull encapsulating the speech kinematics. The volume values were computed for the kinematics of the stimuli. The resulting volume values are denoted as: *Grandfather Passage*, $v_{gf1}, v_{gf2} \in V_{gf}$, the *Rainbow Passage*, $v_{rb1}, v_{rb2} \in V_{rb}$, and the randomized sentence stimuli, $v_n \in V_{stim}$.

To motivate the comparison of volume results, a preliminary visual investigation of the convex hulls was performed. This visual analysis was generated by rendering overlay of the 3D convex hulls of V_{gf} with $V_{gf} \cup V_{stim}$, see Figure 5.8(a), and V_{rb} with $V_{rb} \cup V_{stim}$, see Figure 5.8(b). The ratios of the total stimuli with the passage stimuli was also computed for comparison as $V_{rb}^R = V_{rb}/V_{stim}$ and $V_{gf}^R = V_{gf}/V_{stim}$, these were 97.29% and 94.52% respectively.

To examine the relationship between volume ratios and passage repetition, a within subject AWS across repetition test was performed, see Figure 5.9. Summary statistics for these ratios are shown in table 5.1. A Paired Wilcoxon Signed Rank test was conducted to compare the volume ratios of the first and second repetitions of both passages. The volume ratios of the two repetitions did not differ significantly, $v_{gf1} <> v_{gf2}$ p-value: 0.169 and $v_{rb1} <> v_{rb2}$ p-value: 0.208.

To examine the relationship between volume ratios and passage stimulus, a within subject AWS across passages test was performed, see Figure 5.10. Summary statistics for these ratios are shown in table 5.2. A Paired Wilcoxon Signed Rank test was conducted to compare the volume ratios of the two passages. The volume ratios of the two passages



(a) (b)

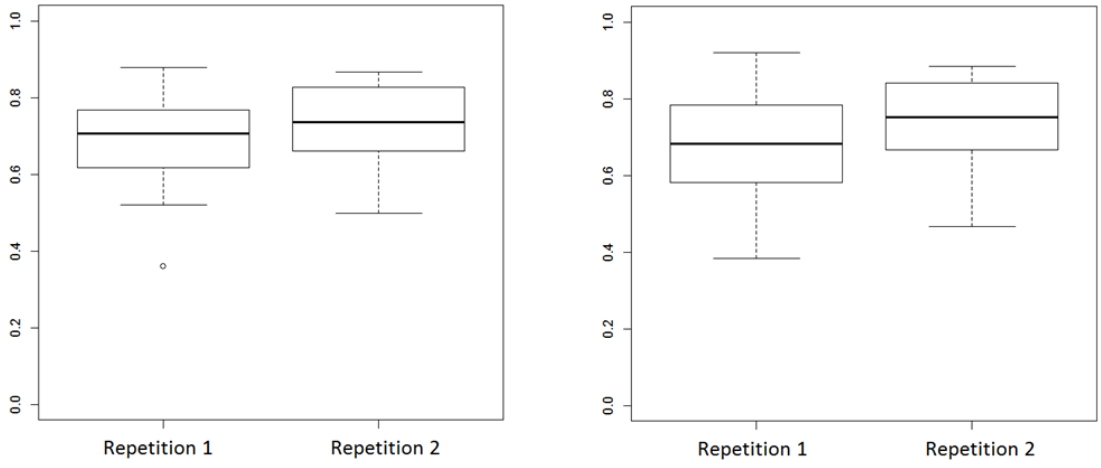
Figure 5.8: Visual overlay of convex hulls for the randomized speech stimuli with (a) *Grandfather Passage* and (b) *Rainbow Passage*.

did not differ significantly, $V_{gf} <> V_{rb}$ p-value: 0.670.

This experiment shows that it may be possible to use as little as two repetitions of either the *Grandfather Passage* or *Rainbow Passage* to derive a characterization of the participant's AWS for general speech.

5.3.3 Procedure

The procedure for this form of normalization is three-fold. The method described here is experimental and based on the aforementioned means of deriving data. To form the



(a) (b)

Figure 5.9: Passage repetition and stimuli volume ratios for (a) *Grandfather Passage* and (b) *Rainbow Passage*.

Rep. Volume	v_{gf1}	v_{gf2}	v_{rb1}	v_{rb2}
Mean	68.1%	73.4%	68.2%	74.3%
Median	70.6%	73.6%	68.3%	75.2%

Table 5.1: Passage repetition and stimuli volume ratio summary statistics for *Grandfather Passage* and *Rainbow Passage*.

Pass. Volume	V_{gf}	V_{rb}
Mean	70.8%	71.2%
Median	72.3%	71.8%

Table 5.2: Passage and stimuli volume ratio summary statistics for *Grandfather Passage* and *Rainbow Passage*.

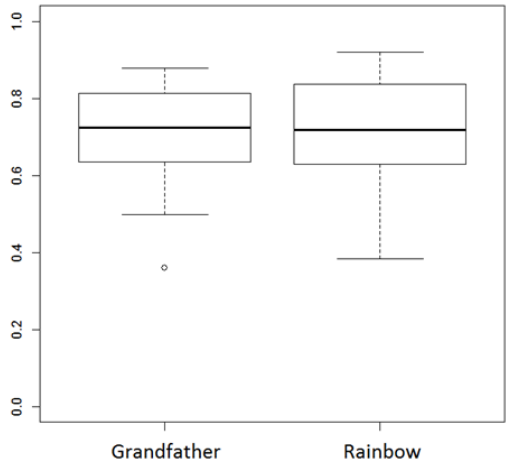


Figure 5.10: Passage volume ratios for (a) *Grandfather Passage* and (b) *Rainbow Passage*.

final step it is necessary to know the game design space parameters beforehand, and are entirely dependent on the type and design of the game.

1. **Biteplate basis** The biteplate is configured as described in 5.3.1. A recording is taken for 5 seconds. The position and orientation of the biteplate, post head correction, is averaged to get a good basis.

2. **Range of motion** A characterization of the participant's AWS is formed based on an elicitation of a speech passage as described in 5.3.2. The extents of the AWS values, or extrema, for each desired dimension are referenced. This characterization is transformed to biteplate space to centre the derived AWS around the origin.

3. **Normalization** Using the relative difference between the orientation and position of the biteplate and game space origins as well as the relative difference between the AWS extents and game space extents a scale, rotation, translation matrix is

formed. This matrix is used online to transform all incoming points into the game space.

5.4 Measures

In this research project, several speech kinematic measures are recorded, as per **R2a.1d**. Furthermore, visualizations are driven by measures, or metrics, as required by experimental speech rehabilitation protocol needs (see **R2a.4-5**). In this research project, for the application to the user population with PD, the AWS measure drives the feedback visualizations, other measures are recorded for future speech research analysis.

5.4.1 Articulatory Working Space (AWS)

The AWS measure relates tongue kinematics to the space they traverse in the vocal tract. The definition of an AWS is the convex hull surrounding the trajectory traversed during a speech production. The AWS was chosen by clinicians for driving visualizations in the current research protocol. The target population, elderly adults with PD, show a reduction in articulatory movements (Forrest et al. 1989, Yunusova et al. 2008). This reduction of movement is reflected in the individual's AWS (Weismer et al. 2012).

The 3D AWS volume (mm^3) is used to characterize the participant's overall AWS. This volume is computed over a speech sample, typically a single utterance of a loaded sentence. The available articulatory volume is limited to the biomechanical constraints of the vocal tract such as Temporomandibular Joint (TMJ) angle and hard palate.

To discretize the space of an AWS and compute its volume, the process starts with delaunay triangulation in three dimensions, or tetrahedralization. This algorithm generates space-filling tetrahedrons whose combined free surface forms a convex hull as seen in Figure 5.12.

This delaunay triangulation algorithm in three dimensions is sensitive to degenerate regular point configurations (colinear, cospherical, and coplanar) which must be taken into account. For real-time purposes, this is limited to two techniques: identical pairwise point removal; and input joggling. The pairwise removal operation checks if the next incoming data point matches with the previous, in practice the distance between the two is less than some epsilon. The input joggling process is carried out by point vector addition defined by a random point within a sphere of a radius small enough not to impact measures. Further the resulting volume can be computed as follows:

$$V_{AWS} = \sum_{t \in T} \frac{|(a_t - d_t) \cdot ((b_t - d_t) \times (c_t - d_t))|}{6} \tag{5.2}$$

where $\cdot$ and $\times$ are the dot and cross product respectively, a_t, b_t, c_t, d_t are the vertices of the tetrahedron t, and T is the set of all tetrahedrons.

To make use of the AWS volume as a clinical target, goals are set such that participants are motivated to increase the measure from their personal current best, see 2.2.1. The *target* AWS is derived on a speaker-specific basis and is a proportional enlargement of the *baseline* AWS as defined by prior analysis. A calibration stage provides a *baseline* target by finding the median of 5 repetitions of speech stimuli in both normal and clear styles

of speech. The *baseline*, is defined as 95% of the midpoint between the AWS volume of normal and clear speech

$$t_b = (v_n + (v_c - v_n)) * 0.95 \tag{5.3}$$

where t_b is the *baseline* target and v_n, v_c are the spatial volumes of the median AWS for normal and clear productions of the stimulus. An example visualization of this target type being visualized using the system can be seen in Figure 5.11.

This primary measure, driving visualizations in the system, illustrates the connection between disorder and motor effects. This measure also illustrates a particular approach to behavioural based PD treatment. However, this is one approach to treatment of speech disorders related to progressing PD which varies from drug-based to behavioural therapies (Ramig et al. 2008). The most well known behavioural approach, the *Lee Silverman speech therapy (LSVT) LOUD* technique (Fox et al. 2012), approaches rehabilitation by focusing partially on the sensory awareness of speech loudness. A recent augmentation for treating limb motor control in PD, *LSVT BIG*, focuses on large limb motions. The AWS volume increase approach is analogous to these approaches with respect to articulators. The spacial AWS volume is currently subject to experimentation and validation of efficacy as a modifiable speech parameter that produces increased intelligibility as judged by Speech Language Pathologists.

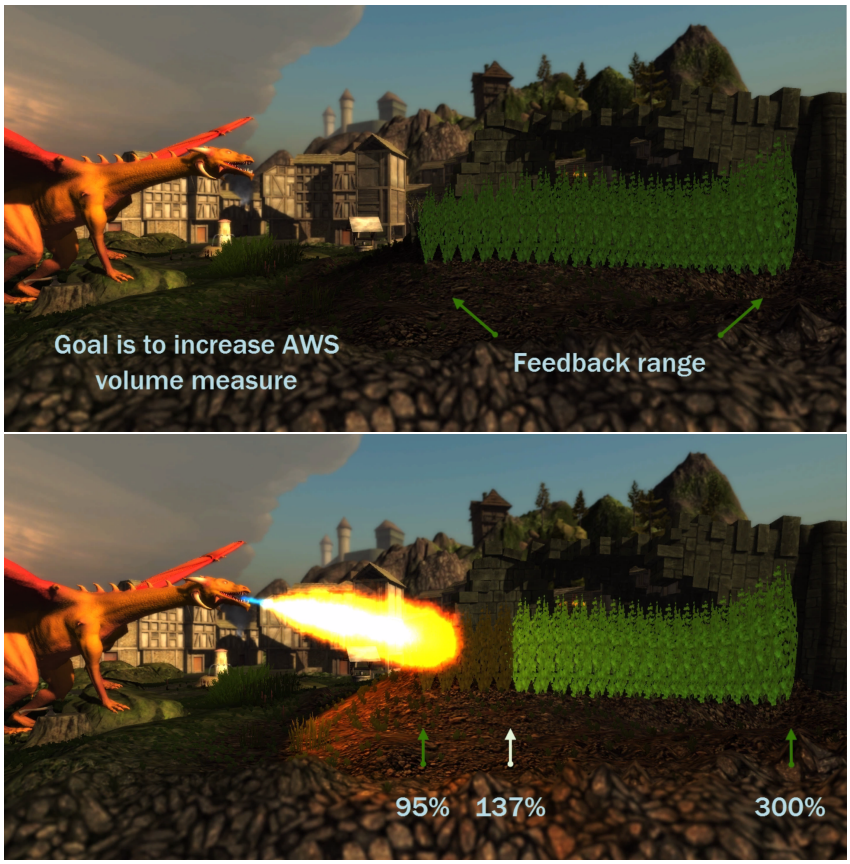
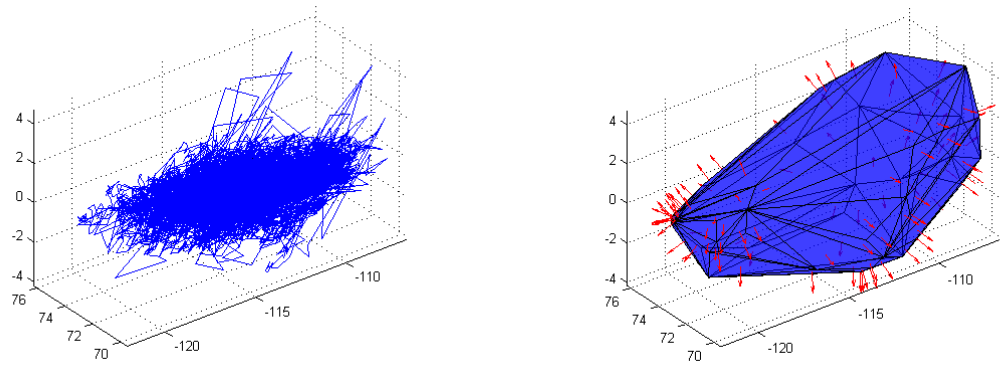


Figure 5.11: An example feedback visualization demonstrating the AWS driven visual goals. The previously achieved AWS volume is shown as burnt/brownish hedges. That target space ranges from the *baseline* to 300% of the normal utterance AWS volume.



(a) (b)

Figure 5.12: Delaunay triangulation in 3D, or tetrahedralization, is performed on (a) a noisy input signal producing (b) a convex hull at the free surface (surface normals shown).

5.4.2 Distance

The distance the sensor has travelled is an empirical measure of potential interest to speech science researchers. To compute this measure, a pairwise distance is taken between the last point and the new incoming data point. This is accumulated over time such that the final output is the total distance of the trajectory and each timestamped sample contains the distance up to that point

$$D = \sum_{i=2}^{n} dist(p_i, p_{i-1}) \tag{5.4}$$

where n is the number of points in the current recording, p_i is the current 3D (x, y, z) sensor position, and $dist(p_a, p_b)$ is the 3D Euclidean distance between points p_a and p_a.

5.4.3 Duration

The duration of an utterance is another empirical measure of potential interest to speech science researchers. To compute the duration, T the difference between the timestamp of the last incoming data point and the first is taken

$$T = t_n - t_1 \tag{5.5}$$

where n is the number of points in the current recording and t_i is a timestamp associated with a data point.

5.5 Summary

This section has described the main portions of the online data process that transforms kinematic data points for use in the game system. The first two processes, head correction and filtering, are necessary for isolating articulator kinematics and removing potential data artifacts that affect measures. The third process, normalization, is an experimental calibration and transformation process designed to enable absolute positioning of articulator sensors in arbitrary game spaces. Finally, a set of measures derived for each utterance is described. This includes the main experimental measure for this research project the AWS, and how targets are calibrated and can be visualized online.

Chapter 6

Offline processing

Speech research focused on articulator kinematics generally requires data processing. As with most basic research, the data derived from the sensor system is subject to noise. In particular, electromagnetic sensor systems are sensitive to magnetic disturbances and high frequency sensor noise.

6.1 Context of offline processing

Offline processes are defined by the needs of speech researchers and their methods of deriving kinematic speech measures. In speech processing there is a general need to isolate desired speech data. This entails removing, or reducing, noise in the data capture/measurement system and maintaining a global signal—examples of these artifacts are discussed in Section 5.2. This data can then be used to compute useful measures, or for general study. A general overview of the pipeline described in this chapter is rendered in Figure 6.1.

In this research project, the offline process serves two purposes. The first purpose is to afford study of speech kinematics, while the second is to provide a means of functional testing of the online process the game system is subject to, or the system fidelity (see **R4**).

6.2 Pipeline input

The lab making use of this pipeline studies many types of speech recordings. These speech recordings are derived from reading passages called speech stimuli. Participants are asked to read the stimulus while the Wave system and an audio microphone records data. This results in a proprietary raw file and audio file.

In order to process the raw recording output, the first step in the process is to use the NDI WaveFront export process. This process corrects sensors to the 6D reference (typically head), separates the two channel audio file with timing signal (SMTPE time coding) and recorded speech, matches data points to audio timing (using SMTPE time coding signal), and produces two output files for each recording. The first is a kinematics file with timestamps relating audio timing with position and orientation data points for each sensor. The second file is simply the recorded speech audio.

6.2.1 Recording specific processing

Various types of speech recordings are typically studied in this lab. For simplicity, these can be broken into two groups, short recording and long recordings. Short recordings

are usually comprised of sentence length material or shorter. This could be anything from Vowel-Consonant-Vowel (VCV) productions to full length sentences. Depending on purpose, these could potentially be batched in repetitions for a single stimulus such that a single recording contains up to four repetitions. Longer recordings can be anything which take more than a minute to produce, and are typically reading passages commonly used by SLP researchers and linguists. The processing of these two recording types is described in detail in Sections 6.4 and 6.5

6.3 General processing

The WaveFront export produces the two aforementioned files per recording. In the kinematics file timestamped samples of position and orientation data for all sensors are stored. The first of these sensors is the 6 DoF reference sensor. All other sensors have been "corrected" to remove the motions of this first sensor.

The motion data stored in the kinematics file requires processing to denoise the signals and fill data loss gaps (general processing area in Figure 6.1). The steps are as follows: trim gaps to reduce potential noise spikes; fix sampling errors inherit in the Wave; fill gaps using an interpolating spline; and remove high frequency sensor noise. An example of the effects of this process on the input signal is show in Figure 6.2. This process is augmented by a basic data quality analysis.

6.3.1 Data quality

The sensor trajectories are checked for drops, or sudden losses of data. These appear as Not-a-Number (NaN) entries in the kinematics file for the rows corresponding to the position and orientation data of a given sensor. If sufficiently large ($> 100ms$), the onset and offset of the gap is stored and the file marked as a potentially a bad recording. It is possible that the gaps occur outside of speech material and thus SLPs use the data quality output to identify usable and unusable recordings.

6.3.2 Global signal

The Wave is subject to both the aforementioned data loss gaps and issues with regular data sampling. At higher frequency recording rates (400Hz), timestamps may fluctuate around the desired sampling period. Empirically these fluctuations occur for approximately $16 - 20\%$ of the data and are within $\pm 0.0002s$ of the desired sampling period. To produce a global signal that is regularly sampled, a cubic interpolating spline is fit to the data. This spline is sampled regularly using the included timestamps to interpolate for desired positional data.

6.3.3 Denoising

Given a global and regularly sampled signal, the motion data can now be filtered to remove high frequency sensor noise. This is accomplished with a standard low-pass filtering procedure. In this case, a common 5th order Butterworth filter with a cutoff frequency

of 15Hz is used. This cutoff frequency is a product of prior research which shows that the frequency of tongue motions associated with speech production is typically lower than 15Hz (Gracco 1992).

6.3.4 Output

The output of the general processing procedure is a new processed data file for each input recording file. There is also a data quality and flag file produced. The structure of this output can be seen in Section 6.6.

6.4 Short recording processing

For convenience, short recordings typically contain multiple repetitions of short stimulus productions batched in one recording file. These types of recordings must be segmented such that repetitions can be processed separately (short recording processing area in Figure 6.1). This process requires an SLP to listen to the audio recording and identify onset and offset times of the repetitions. These onsets/offsets are stored in a consistent format which is used as input to the pipeline to segment both kinematic and audio data.

6.4.1 Output

The output of the sentence processing procedure is n new kinematic and audio files corresponding to the n repetitions as defined by the input onset/offset file. The resulting measures for the individual repetitions are also produced. The structure of this output

and the measures produced can be seen in Section 6.6.

6.5 Long recording processing

Long recordings, typically containing a single long production of a speech stimulus, require preprocessing and noise removal (Long recording processing area in Figure 6.1). Preliminary experiments revealed spurious data artifacts associated with long recordings. First, it was found that participants produced long pauses during the production of long stimuli. Second, it was found that long recordings with the Wave system contained spikes. The spikes appear to correlate with the duration of the recording, the longer the recording the higher probability of spikes appearing.

6.5.1 Pause removal

In speech science, long pauses may introduce spurious kinematics artifacts, which are not important and may skew measures in the study. The artifacts derive from human movements such as yawning, licking of the lips, deep breaths, or simply resting between sentences. To remove these, an automated software identifies pauses by applying a threshold to the audio signal amplitude and producing onset and offset times for the pauses. These onset/offset times are then used in the pipeline to remove these pause segments before producing measures. This is a simple process of matching the nearest timestamped data points corresponding with the onset and offset time and simply removing the data.

6.5.2 Spike culling

In preliminary tests it was found that kinematic spikes, i.e. spurious high speed jumps in position, and gaps, i.e. missing data, are common in long recordings with the Wave. It was empirically found that these spikes are not speech data by identifying that they: (1) violate biomechanical constraints of the hard palate; (2) violate the upper limits of tongue speed and distance in English; and (3) are typically accompanied by gaps or loss of data.

The third stage of the pipeline characterizes the long recording speech data with a 2σ error ellipsoid. Intuitively, this means that all positional data within the ellipsoid are classified as speech data, whereas data outside the ellipsoid are considered non-speech. This ellipsoid is defined by the squared Mahalanobis distance, or generalized squared inter-point distance (Gnanadesikan and Kettenring 1972), and a constant threshold K:

$$(x - m_x)^T \Sigma_x^{-1} (x - m_x) \leq K \tag{6.1}$$

where x is a data point, m_x is the mean of x, and Σ_x is the covariance matrix. Setting $K = 7.84$ corresponds to $95\%CI$ or 2σ (Ribeiro 2004). All points with distance equal to or less than K are on the surface or within the ellipsoid respectively. The assumption, verified with preliminary tests, is that long recordings contain normally distributed positional data.

6.6 Pipeline output

6.6.1 Measures

The pipeline produces two separate output measure files for the long and short recording data processes. Both files contain basic summary statistics such as the min, max, mean, median, and modes of the following: x, y, z, x speed, y speed, z speed, 3D speed. They also contain the AWS volume, cumulative distance, and duration. In addition, the long recording results include the volume of the 2SD ellipsoid and the principle components of the data.

6.6.2 File structure

The output of the pipeline involves a well structured directory of the segmented and processed files as well as data measures and data quality files (final step in Figure 6.1). This directory has the following structure as needed by this research and automation of the pipeline:

- Results
  - [participant ID]_data_log.csv
  - [participant ID]_sentence_results.csv
  - [participant ID]_passage_results.csv
- SentencesProcessed

  - [original filename]_P_[repetition number].csv
  - [original filename]_P_[repetition number].wav
  - ...
- PassagessProcessed
  - [original filename]_P_PP.csv
  - ...
- [original filename].tsv
- [original filename].wav
- [original filename]_P.csv
- ...

where "..." means this structure is repeated for all recordings.

In this structure, the Results directory stores the data log for the participant's data files, and the files for passage and sentence measures. The *SentencesProcessed* directory stores the segmented repetition sentence kinematics (*csv*) and audio files (*wav*) with the repetition number appended to the file names. The *PassageProcessed* directory contains the kinematics without pauses and culled by the ellipsoid characterization process (*csv*). The main directory contains these directories, the original kinematics (*tsv*) and audio files (*wav*) and the general processed kinematics files appended with *'_P'*. Here the *csv* and *tsv* filenames denote comma and tab delimited files respectively.

6.7 Summary

This chapter covered the offline processing pipeline for all kinematic and acoustic data related to this research project. The inputs are the raw Wave recording files, and the outputs are well-organized and useful data files ready to be analysed. This processing pipeline serves a portion of the system testing in Chapter 7 and the processing of future speech kinematics data for other studies. In addition, the pipeline is available as open source to other speech research labs in North America.

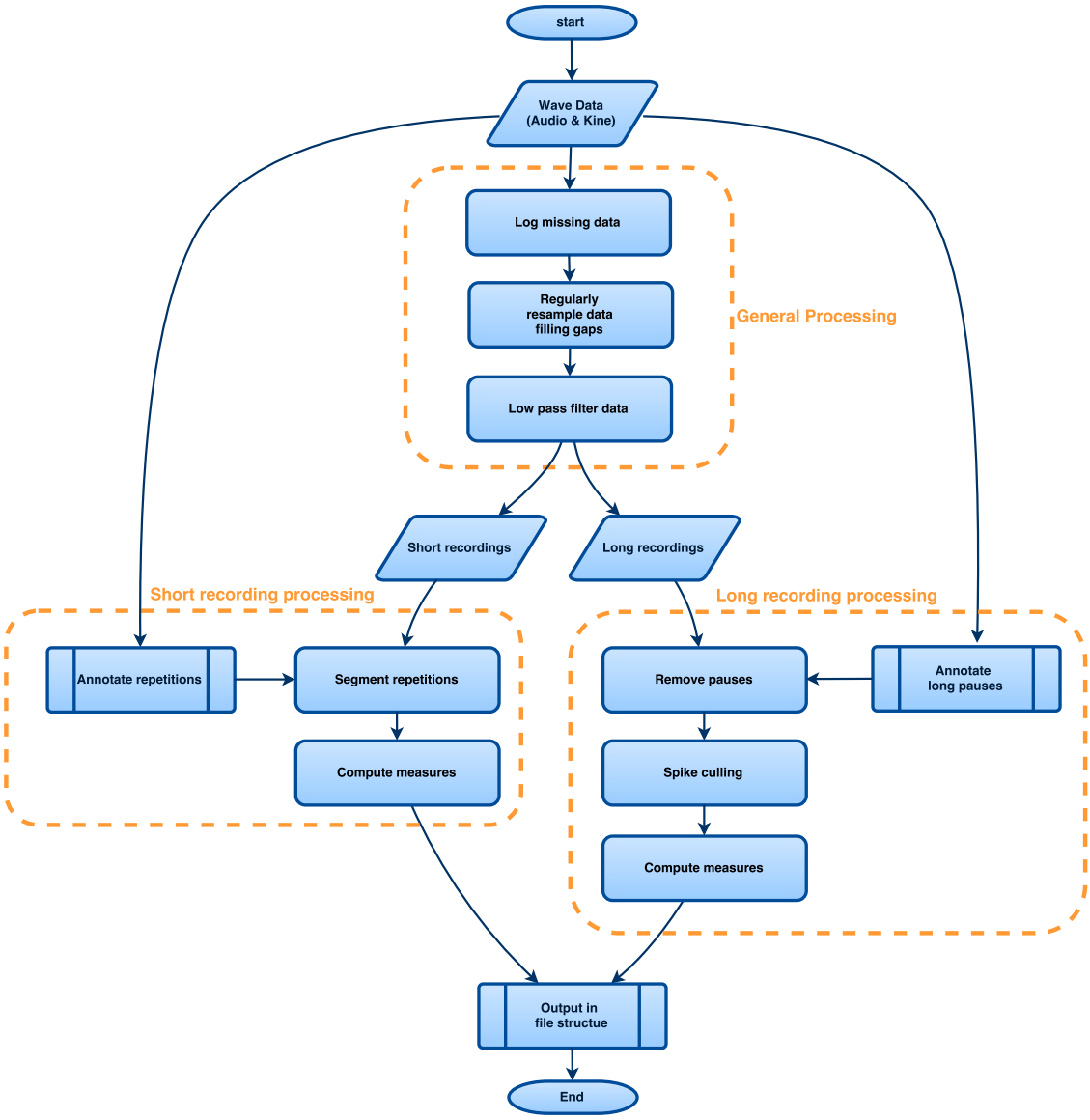
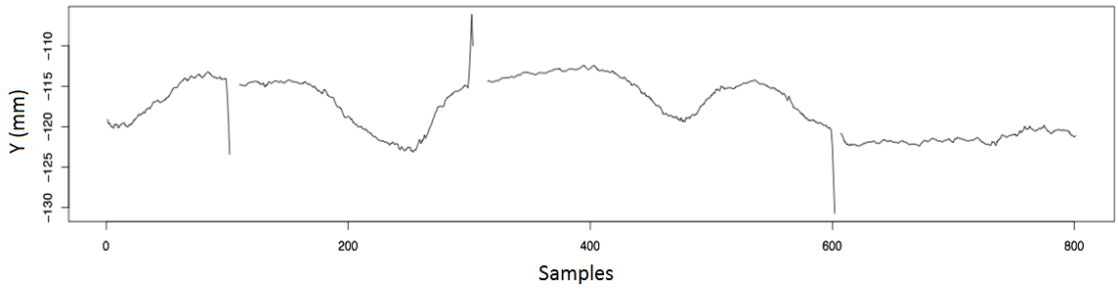
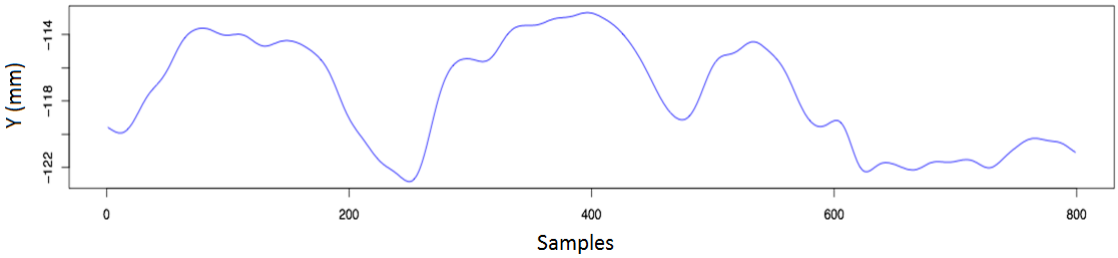


Figure 6.1: Overview of the offline data processing pipeline.



(a)



(b)

Figure 6.2: (a) An example input and (b) output of the general processing portion of the offline pipeline for an articulator sensor position time series, shown in one dimension and scaled for clarity.

Chapter 7

Functional Testing

This section describes the fulfilment of the requirements to validate the system. The first requirement **R3** (system robustness) is satisfied through functional testing study #1 (described in Section 7.1). The second requirement, **R4** (establish system fidelity), is satisfied through functional testing study #2 (described in Section 7.2).

7.1 Study #1

Purpose The purpose of this study is to validate the network data channels, and meet the requirement **R3** (system robustness)

Method A three step process is followed to validate the individual connections. Validation of the individual steps of data transfer between the interconnected components implies the working order of the networked nature of the system.

A simplified data collection session is used for this experiment. The components are the Wave Speech Research system, the data collections computer, and the visualization

computer (see Section 4.1 for details). The Wave to data collection computer connection is facilitated via a proprietary serial based cabling, and the data collections computer is connected to the visualization computer via crossover Ethernet cable (see Section 4.1 for details). The data transfers are facilitated by proprietary data and control signals and finally TCP protocol respectively, as facilitated by the various control softwares (see Section 4.2 for details). The Wave sensors recorded for these experiments are one 6 DoF reference sensor and one 5 DoF articulator sensor fixed with respect to each other.

Step 1: Data recording from the Wave. The Wave must successfully produce raw kinematic and audio files on the data collections computer when a recording is initiated via WaveFront software. A number of recordings are executed using default settings. The sensors are left stationary and then moved them for a small amount of time (approx. 5 seconds each). We examine the exported files (see Section 6.2), to ensure sensor data matches stationary and induced movements—accuracy is ignored and assumed to be within system limits, see Section 3.2.

Step 2: Remote connection to data stream and reception of TCP packets. A remote connection to the Wave data source, the data collections computer software WaveProxy, is established using the Real-Time Application Programmers Interface (RTAPI) by requesting a stream at $100Hz$. This connections is left running for approximately 10 seconds to accumulate data. The remote connection (DataMuffin) is examined for a data stream (TCP packets carrying articulator kinematics data) and the average frequency at which they are received

$$f_a = \frac{\sum_{i=2}^{n} \frac{1}{t_i - t_{i-1}}}{n} \tag{7.1}$$

where f_a is the average frequency, n the number of packets, and t_i the received time for packet i in seconds.

Step 3: Game system connects to and receives remote data stream. A connection is established between the game system and the DataMuffin using the ClientLib. The packets from *Step 2* are numbered before being sent to the game system. The packet numbers received at the game system are examined to ensure they are equivalent and ordered as they are at DataMuffin. The frequency of the packets received at the game system is examined in the same manner as *Step 2* using equation 7.1.

Results

Step 1: Results were positive. The Wave successfully produce recordings on the data collections machine

Step 2: Results were positive. The remote connection successfully receives a TCP packet stream from the data collection computer at the requested frequency, within $\pm 0.02\%$.

Step 3: Results were positive. The game system receives all packets in order, and at the receiving frequency of the DataMuffin (*Step 2*)—noting that the connection between the game system and DataMuffin is local (on the same machine).

Conclusions This study showed that the chain of connected components, of which the system hardware architecture is comprised, successfully transfers the desired data

within operating parameters. It should be noted that, during development logical and buffering errors were fixed until the connections produced desired results. There are cases when the connection between one of the components is lost, i.e. system errors, Wave software errors. In this case the game system waits, and the connection can be established again by sending another RTAPI data stream request from the DataMuffin software to the WaveProxy software.

7.2 Study #2

Purpose The purpose of this study is to validate the online data processes

Method A ground truth for AWS derivation was constructed as a data processing pipeline that outputs SLP verified speech kinematic data measures (see **R4.1**). This pipeline is described in detail Chapter 6. The results are examined for the online process relative to ground truth.

The online and offline processes are fundamentally different because of there real-time and post approaches respectively. Thus, the two processes can be considered two different means of measurement for the same value (AWS volume). As such, the values are not expected to be exactly the same, but must be in convergence of less than 1% relative error as per the needs of the stakeholders.

A standard data collection session is used for this experiment. The Wave sensors recorded for these experiments are one 6 DoF reference sensor attached to the head and one 5 DoF articulator sensor fixed to 1cm posterior the tongue tip. The participant was

tested for normal hearing, vision, and cognition using a standard Snellen chart vision test, tonal range hearing test, and the Montreal Cognitive Assessment (MoCA) (Nasreddine et al. 2005). The participant was asked to repeat load sentence stimuli (/sh/ and /t/) in two styles of speech, *Normal* (regular everyday speech) and *Clear* (exaggerated enunciation of speech) (see Section 5.4.1 for how these are used in the game system). There are five repetitions of each stimuli for each style, resulting in ten repetitions for each style. The output AWS volume (online and ground truth) for the two different styles is examined.

Results

Normal: Results are positive. The AWS volumes computed by the online and ground truth processes are highly correlated for the *Normal* style of speech, as seen in Figure 7.1. The distribution of relative error can be seen in Figure 7.3

Clear: Results are positive. The AWS volumes computed by the online and ground truth processes are highly correlated for the *Clear* style of speech, as seen in Figure 7.2. The distribution of relative error can be seen in Figure 7.4.

Conclusions Since the percent error of the AWS volume measures was $< 1\%$ at and average of $\pm 0.36\%$, it can be concluded that the online measurements are succeeding in reproducing the measurements of the ground truth, or offline process.

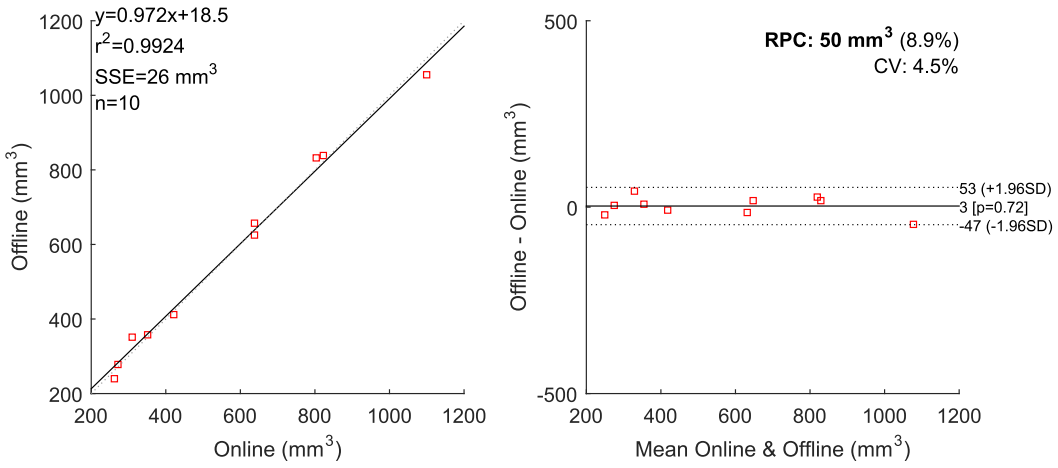


Figure 7.1: Bland-Altman figures for the 'Normal' style stimuli productions—online relative to ground truth (offline) processes.

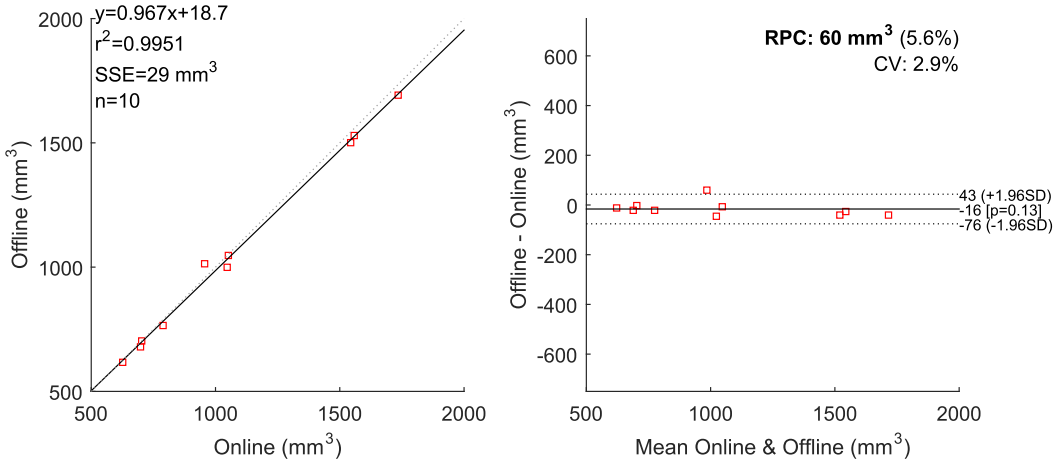


Figure 7.2: Bland-Altman figures for the 'Clear' style stimuli productions—online relative to ground truth (offline) processes.

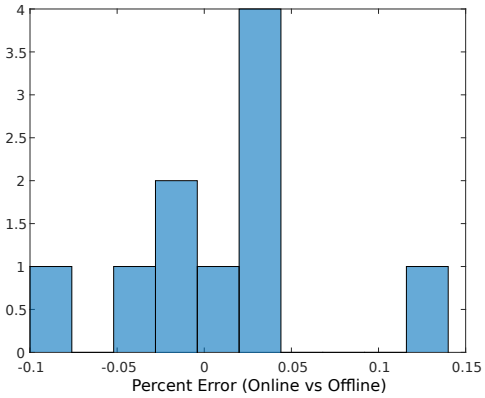


Figure 7.3: The percent error distribution for the 10 'Normal' style trials shown in Figure 7.1.

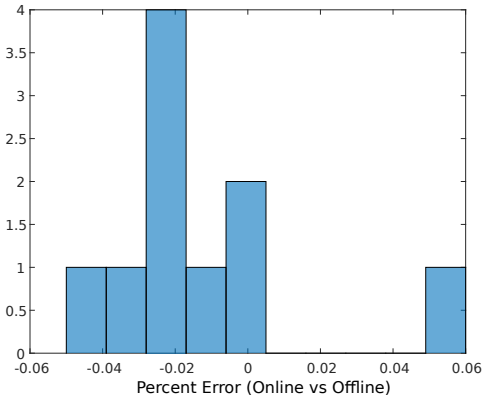


Figure 7.4: The percent error distribution for the 10 'Clear' style trials shown in Figure 7.2.

7.3 Discussion

The functional testing of the system robustness and fidelity was described in studies presented in this chapter. These studies fulfil requirements **R3** and **R4** respectively. These requirements pertain mainly to the functionality of the system. Other, requirements are met and validated through use and further development of the system.

The following success criteria were used to judge the functionality of the system (see Section 3.4):

1. Incremental packet numbers at the end point in the system [*Completed*]
2. Received packet frequency at the end point matches requested [*Completed*]
3. Convergence of online and offline AWS volume outputs [*Completed*]
4. Online head correction [*Completed*]
5. Online data smoothing/filtering [*Completed*]
6. Implementation of new visualizations and metrics [*Completed*]

7.3.1 Future work

Further work in validating the system will take on a user centric approach. This will involve investigating the usability across the two user classes (clinicians and participants).

7.3.1.1 Usability

Purpose The need to validate usability (see **R2b**) is an ongoing process. The system currently delivers speech research protocols to participants in a lab setting. It is used successfully by clinicians to produce exercises that fit protocols. These exercises are delivered to participants with instructions from the experimenters. However, the usability of the system for participants requires user experience studies.

Method This study makes use of multiple methodological tools to asses the usability of the system from different perspectives.

A set of preliminary training and retention sessions are performed. These sessions are 48 hours apart and consist of one visualized feedback training session and one non-visualized retention session. The training session consists of three sets of ten repetitions for two different visualization scenarios. Each visualization is assigned a group of two linguistically similar loaded sentences, which are randomized of the ten repetitions. The participant have control over feedback, using tokens, up to a feedback frequency of 50%. The retention session has no assigned visualizations and participants are shown a black screen during the regular feedback phase.

The Wave sensors recorded for these experiments are one 6 DoF reference sensor attached to the head, one 5 DoF articulator sensor fixed to 1cm posterior the tongue tip, and one 5 DoF articulator sensor fixed to 2cm posterior the tongue tip. The participants are tested for normal hearing, vision, and cognition using a standard Snellen chart vision test, tonal range hearing test, and the Montreal Cognitive Assessment (MoCA) (Nasreddine

et al. 2005).

System usability: To asses usability in a preliminary way, a scheme was developed to identify critical incidents (Flanagan 1954). In this case, critical incidents are significant situations in which interaction is abruptly stopped or can not continue because of a mismatch between user understanding or critical failure of some feature or procedure. Video coding provides recordings of both the system apparatus and the participant during the session. Our method was to document critical incidents as they occurred during the prototyping process, discuss in next relevant scrum, and address through system revisions by identifying a product and placing it in the appropriate backlog.

Visualization fidelity impact: To asses the impact of visual fidelity on th participant two self reporting mechanism are used. The Borg Rating of Perceived Exertion Scale *BORG CR-10* to assess localized exertion (Borg 1998). The second tool is the Self-Assessment Manikin (SAM) to assess affective stimulus (Bradley and Lang 1994). Both are given after each set of visualizations in the training session.

Conclusions Pilot studies have thus far driven several iterations of the visualizations, system interaction methods, and general interface and feedback design. While conclusions on motor learning can not be made at this preliminary stage, retention data shows that training and retention sessions produce different output values (AWS volume), see Figure 7.5. These early conclusions motivate further examination of the usability of the system.

Figure 7.5: Preliminary retention experiment data highlighting differences between training and retention sessions.

7.4 Conclusion

This chapter has described the various approaches to the functional testing of the system. Two studies were conducted to ensure the meeting of the numeric requirements set out in Chapter 3. These test verified the interconnected functionality of the system as well as the correctness of the output measures in the training system. Furthermore, the current and future work regarding the usability of the system was discussed. The ongoing usability experiment was described, and early results show that the system is operating as per the current requirements.

Chapter 8

Conclusion

This thesis set out to describe the development of an experimental augmented kinematic speech feedback system for Computer Based Speech Therapy. The main objective of the thesis was to provide insight into this development while answering questions such as: what are the development requirements of a real-time speech kinematics based CBST for speech motor rehabilitation?; what systems and subsystems are required to support a functioning architecture?; what data processes are necessary for providing real-time feedback?; and finally, what are the tasks involved in validating such a system? To answer these questions, the thesis included a number of chapters describing: the background literature motivating the pursuit of this research project and the theory from which some of the fundamental system requirements are derived; the design process and generation of the requirements for the system; the system architecture in a top-down approach delineating the system from its hardware to its software subsystems; the online processes of the system which afford real-time use of articulator kinematics data; the offline data

processes which serve as a ground truth for validation of the system outputs; and finally the approach to validation as a function of acceptable data outputs. This chapter will review a synthesis of the aforementioned chapters and research questions, a discussion on the implications of the research project outcomes, and end with a discussion of the limitations and subsequent future directions and research areas of the project.

8.1 Findings

The primary objective of the thesis was to explore the requirements of developing a CBST within given parameters. First, a literature review provided the theoretical background and motivating work for the CBST. From the theory review in motor learning it was clear that control over feedback parameters, head movement correction, and affording normalization techniques were important foundational requirements. Further, the Agile development framework and SCRUM processes revealed several user-centric requirements leading to a robust final product. The application of this development framework suited the research project particularly well and lead to a system grounded theoretical background.

The development process also provided framework on addressing expanding and shifting system requirements. This led to robust subsystems in a flexible architecture. These designs not only support the current application of the system to research in a specific treatment population but further research into other treatment populations.

Thorough review of background work in CBSTs as well as the co-located development

of the system revealed several data processes necessary to achieve desired functioning. These data processes, related to the research requirements of speech kinematics, provided insight into how offline global data processes can be recreated using real-time local online processes.

It was clear, from background literature in the gamification of motor rehabilitation and the subsequent generation of system requirements through user-centric design, that the user is core to the system outcomes and development. Current validation of the system focused on verification of data processes to ensure proper measurements. However, validation of such a system is multifaceted and must include, or lead to, the user perspective (usability) as well as the clinical application perspective (efficacy).

8.2 Implications

This thesis presented a strategic description of the research and development of a CBST for augmented kinematic speech feedback. The thesis describes several contributions to the field of CBST systems in speech rehabilitation research.

A description of the application of an approach for developing an active clinical research software system that has flexible requirements for an emerging clinical process was given. This has implications for future approaches to development of CBSTs. This approach can be further applied to the clinical product CBST this system is the precursor to.

A robust game system architecture was described. It is capable of: visualizing articu-

lators and their kinematic parameters; producing experimental motor therapy exercises; and flexible development of speech parameters, visualizations or game scenarios, and feedback. This architecture provides evidence towards extensible system designs that support a wide range of augmented feedback and applications to different treatment populations, in contrast to past CBST systems that typically focus on singular modes of visual feedback or treatment populations.

A method for normalizing articulator motions in an arbitrary game space for absolute articulator feedback was described. Empirical evidence was provided for a means of operationalizing the method as a calibration process. Further, the process affords a level of flexibility in future games design for augmented kinematic feedback. Future games can take advantage of both the real-time transformations derived from the process and automatic calibration for arbitrary game spaces, in contrast to past CBSTs that require head movement constraints or experimenter calibration.

Finally, the design of a ground truth offline data process afforded a means for validation of the system. Comparing the offline and online data processes outputs gives insight into the functioning of the online CBST system. Furthermore, this approach provides insight into how needs driven design may benefit multiple applications. That is, the offline process serves a dual purpose as validation ground truth and as a tool in the future research of speech kinematics.

8.3 Limitations and future work

The development of the system described in this thesis has generated many answers for the overarching research objectives as well as several questions and new directions, while highlighting limitations in the current approach. This section seeks to describe future work in reference to these questions and limitations.

8.3.1 User experience

Future works will take a user-centric approach to validation by investigating usability from the perspective of the user. Section 7.3.1.1 presents the current methodology applied in the research project for examining user experience. The next goal in this ongoing study is to complete data collection and analysis of the results gathered through surveys and self-assessment devices. Future work in this direction involves applying these method to different treatment populations which may involve new speech disorders and/or age groups—potentially defining new usability parameters.

Preliminary work in the gamification of motor learning systems identified the potential affects of feedback fidelity on user experience (Section 2.3). Thus, another step in this research project includes the investigation of visual feedback fidelity in the application of a CBST. This is pertinent since this system supports a range of feedback from textual to minimal graphical shapes to high fidelity rendered scenarios. The system is also flexible with respect to treatment populations with different speech disorders which may respond differently to feedback fidelity.

Potential future work, with respect to usability, may include new modes of feedback and interaction with the user. An interesting direction is the application of co-play, or multiplayer, for participants or between clinician and participant. This direction promises to deliver interesting interactions that may be more engaging, moving therapy from a directed process to a cooperative one. There are innumerable potential modes of visualization and that span game genres. This space of potential visualizations is only partially constrained by the user-based input (interaction) to the gamin system. We plan to explore this space in the application of the system to various speech disorders and user demographics.

8.3.2 Efficacy

Novel clinical practices are typically judged on their efficacy and more specifically the retention of the therapeutic outcomes they produce. Thus, an important next step for the application of this CBST system is to ensure that this approach produces sustainable positive change in the treatment populations. This system supports this form of investigation and it is a current goal of the speech science portion of the research project. This is currently afforded by visualized feedback session outcomes with respect to no feedback retention session. In Section 7.3.1.1, early preliminary retention results were given as validation of usability of the system (delivery of training versus retention sessions), however they also motivate further long term retention experiments.

8.3.3 New treatment populations

The treatment population motivating the creation of the system (PD) is only one of many the system could potentially support. In the case of people post-stroke, the next population target, these visualizations would address inaccuracies in articulator movements, in contrast to reduction of articulator movements in PD. A potential measure for this is accuracy of articulator placement during consonant stops which translate into palatal contact points. This would be similar to the targets devised in the current literature covered in Section 2.2.4.2, but would be automatically defined through calibration processes supported by the system. This may take the form of automatically defining the area around repeated alveolar ridge contacts. This can be done by forming a 2SD ellipsoid about the max height position of trajectories during repeated speech stimulus productions (accepted as successfully intelligible by the operating clinician). Further study in this area may lead to articulatory trajectory matching with respect to successful trajectories, feedback using the absolute articulator motion mapping described in Section 5.3, or mapping hard palate contact regions to controls mimicking electropalatography (EPG) devices. All of these approaches are examples of possible directions in treating articulator movement inaccuracies and would require their own efficacy measuring experiments to validate—further adding to the body of future work.

8.4 Conclusion

This thesis presents the successful application of a user-centric design process, founded in the framework of Agile development and SCRUM processes, to the development of a CBST. The thesis has delivered the theoretical background and description of practical work related to the generation of the CBST, as set out by the research question. The deliverables of this work are a research system (the CBST) actively in use at a world-class rehabilitation institute (UHN; Toronto Rehabilitation Institute), and a data processing pipeline currently being distributed to research labs using the Wave Speech Research System to gather speech kinematics data. Furthermore, the CBST supports future work in the field and the ongoing research project.

Bibliography

Nicholas Bankson, Allan Diefendorf, Roberta Elman, Susan Forsythe, et al. Scope of practice in speech-language pathology. *Communication Disorders Quarterly*, 23(2):77, 2002.

Kent Beck, Mike Beedle, Arie Van Bennekum, Alistair Cockburn, Ward Cunningham, Martin Fowler, James Grenning, Jim Highsmith, Andrew Hunt, Ron Jeffries, et al. Manifesto for agile software development. 2001.

Jeffrey J Berry. Accuracy of the ndi wave speech research system. *Journal of Speech, Language, and Hearing Research*, 54(5):1295–1301, 2011.

Gunnar Borg. *Borg's perceived exertion and pain scales.* Human kinetics, 1998.

Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

John Cunnison Catford. *Fundamental problems in phonetics.* Midland Books, 1977.

Alistair Cockburn. *Agile software development: the cooperative game.* Pearson Education, 2006.

Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM, 2011.

John C Flanagan. The critical incident technique. *Psychological bulletin*, 51(4):327, 1954.

Karen Forrest, Gary Weismer, and Greg S Turner. Kinematic, acoustic, and perceptual analyses of connected speech produced by parkinsonian and normal geriatric adults. *The Journal of the Acoustical Society of America*, 85(6):2608–2622, 1989.

Cynthia Fox, Georg Ebersbach, Lorraine Ramig, and Shimon Sapir. Lsvt loud and lsvt big: behavioral treatment programs for speech and body movement in parkinson disease. *Parkinsons Disease*, 2012, 2012.

Susanne Fuchs, Pascal Perrier, Christian Geng, and Christine Mooshammer. What role does the palate play in speech motor control? insights from tongue kinematics for german alveolar obstruents. *Speech production: Models, phonetic processes, and techniques*, pages 149–164, 2006.

Ann M Gentile. A working model of skill acquisition with application to teaching. *Quest*, 17(1):3–23, 1972.

R. Gnanadesikan and J. R. Kettenring. Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, 28(1):pp. 81–124, 1972. ISSN 0006341X. URL `http://www.jstor.org/stable/2528963`.

Vincent L Gracco. Analysis of speech movements: practical considerations and clinical application. *Haskins Laboratories status report on speech research SR-109/110*, pages 45–58, 1992.

Ken W Grant and Philip-Franz Seitz. The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, 108(3):1197–1208, 2000.

M B Haworth, E Kearney, Y Yunusova, P Faloutsos, and M Baljko. Rehabilitative speech computer game calibration using empirical characterizations of articulatory working space (aws), March 2014a. Poster presented at 17th Biennial Motor Speech Conference.

M Brandon Haworth, Elaine Kearney, Melanie Baljko, Petros Faloutsos, and Yana Yunusova. Electromagnetic articulography in the development of serious games for speech rehabilitation, 2014b.

Thomas J Hixon. An electromagnetic method for transducing jaw movements during speech. *The Journal of the Acoustical Society of America*, 49(2B):603–606, 1971.

Philip Hoole and Noel Nguyen. Electromagnetic articulography. *Coarticulation–Theory, Data and Techniques, Cambridge Studies in Speech Science and Communication*, pages 260–269, 1999.

Shannon N Austermann Hula, Donald A Robin, Edwin Maas, Kirrie J Ballard, and Richard A Schmidt. Effects of feedback frequency and timing on acquisition, retention, and transfer of speech skills in acquired apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 51(5):1088–1113, 2008.

Keith Johnson, Peter Ladefoged, and Mona Lindau. Individual differences in vowel production. *The Journal of the Acoustical Society of America*, 94(2):701–714, 1993.

Tokihiko Kaburagi, Kohei Wakamiya, and Masaaki Honda. Three-dimensional electromagnetic articulography: A measurement principle. *The Journal of the Acoustical Society of America*, 118(1):428–443, 2005.

William F Katz. Influences of electromagnetic articulography sensors on speech produced by healthy adults and individuals with aphasia and apraxia. *Journal of Speech, Language, and Hearing Research*, 49(3):645–659, 2006.

William F Katz and Malcolm R McNeil. Studies of articulatory feedback treatment for apraxia of speech based on electromagnetic articulography. *SIG 2 Perspectives on Neurophysiology and Neurogenic Speech and Language Disorders*, 20(3):73–79, 2010.

William F Katz, Sneha V Bharadwaj, and Burkhard Carstens. Electromagnetic articulography treatment for an adult with broca's aphasia and apraxia of speech. *Journal of Speech, Language, and Hearing Research*, 42(6):1355–1366, 1999.

William F Katz, Gregory C Carter, and June S Levitt. Treating buccofacial apraxia using augmented kinematic feedback. *Aphasiology*, 21(12):1230–1247, 2007.

William F Katz, Malcolm R McNeil, and Diane M Garst. Treating apraxia of speech (aos) with ema-supplied visual augmented feedback. *Aphasiology*, 24(6-8):826–837, 2010.

William F. Katz, Thomas F Campbell, Jun Wang, Eric Farrar, J Coleman Eubanks, Arvind Balasubramanian, Balakrishnan Prabhakaran, and Rob Rennaker. Opti-speech: A real-time, 3d visual feedback system for speech training. In *Proc. Interspeech*, 2014.

Jangwon Kim, Sungbok Lee, and Shrikanth Narayanan. An exploratory study of the relations between perceived emotion strength and articulatory kinematics. In *INTERSPEECH*, pages 2961–2964, 2011.

Christian Kroos. Measurement accuracy in 3d electromagnetic articulography (carstens ag500). In *Proceedings of the 8th international seminar on speech production*, pages 61–64, 2008.

Patricia K Kuhl and Andrew N Meltzoff. The bimodal perception of speech in infancy. *Science*, 218:1138 – 1141, December 1982.

Adam Lammert, Michael Proctor, and Shrikanth Narayanan. Interspeaker variability in hard palate morphology and vowel production. *Journal of Speech, Language, and Hearing Research*, 56(6):S1924–S1933, 2013.

Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan. An articulatory study of emotional speech production. In *INTERSPEECH*, pages 497–500, 2005.

June S Levitt and William F Katz. The effects of ema-based augmented visual feedback on the english speakers acquisition of the japanese flap: a perceptual study. *Stroke*, 4: 5, 2010.

Edwin A Locke, Norman Cartledge, and Jeffrey Koeppel. Motivational effects of knowledge of results: A goal-setting phenomenon? *Psychological bulletin*, 70(6p1):474, 1968.

Keith Lohse, Navid Shirzad, Alida Verster, Nicola Hodges, and HF Machiel Van der Loos. Video games and rehabilitation: using design principles to enhance engagement in physical therapy. *Journal of Neurologic Physical Therapy*, 37(4):166–175, 2013.

Roberta Marchese, Manuela Diverio, Francesca Zucchi, Carmelo Lentino, and Giovanni Abbruzzese. The role of sensory cues in the rehabilitation of parkinsonian patients: a comparison of two physical therapy protocols. *Movement Disorders*, 15(5):879–883, 2000.

Hannah R Marston, Michael Kroll, Dennis Fink, and Sabine Eichberg. Digital game aesthetics of the istoppfalls exergame. In *Games for Health 2014*, pages 89–100. Springer, 2014.

Anna Lisa Martin, Ulrich Götz, Cornelius Müller, and René Bauer. gabarello v. 1.0 and gabarello v. 2.0: Development of motivating rehabilitation games for robot-assisted locomotion therapy in childhood. In *Games for Health 2014*, pages 101–104. Springer, 2014.

Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 1976.

Fred Minifie et al. *Normal Aspects of Speech, Hearing, and Language*. Prentice-Hall, Inc., 1973.

Sazzad M Nasir and David J Ostry. Somatosensory precision in speech production. *Current Biology*, 16(19):1918–1923, 2006.

Ziad S Nasreddine, Natalie A Phillips, Valérie Bédirian, Simon Charbonneau, Victor Whitehead, Isabelle Collin, Jeffrey L Cummings, and Howard Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

Anne-Marie Öster. *Computer-based speech therapy using visual feedback with focus on children with profound hearing impairments*. PhD thesis, KTH, School of Computer Science and Communication (CSC), Speech, Music and Hearing, 2006.

Joseph S Perkell, Marc H Cohen, Mario A Svirsky, Melanie L Matthies, Iñaki Garabieta, and Michel TT Jackson. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, 92(6):3078–3096, 1992.

Dirk-Jan Povel and Nico Arends. The visual speech apparatus: Theoretical and practical aspects. *Speech communication*, 10(1):59–80, 1991.

Lorraine O Ramig, Cynthia Fox, and Shimon Sapir. Speech treatment for parkinson's disease. *Expert Review of Neurotherapeutics*, 8:297–309, 2008.

Maria Isabel Ribeiro. Gaussian probability density functions: Properties and error characterization. *Instituto Superior Tcnico, Lisboa, Portugal, Technical Report*, 2004.

Krista Rudy and Yana Yunusova. The effect of anatomic factors on tongue position variability during consonants. *Journal of Speech, Language, and Hearing Research*, 56 (1):137–149, 2013.

Alan W Salmoni, Richard A Schmidt, and Charles B Walter. Knowledge of results and motor learning: a review and critical reappraisal. *Psychological bulletin*, 95(3):355, 1984.

Richard A Schmidt. Frequent augmented feedback can degrade learning: Evidence and interpretations. In *Tutorials in motor neuroscience*, pages 59–75. Springer, 1991.

Richard A Schmidt and Tim Lee. *Motor control and learning : a behavioral emphasis - 5th Edition*. Human kinetics, 2011.

Paul W Schönle, Klaus Gräbe, Peter Wenig, Jörg Höhne, Jörg Schrader, and Bastian Conrad. Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31(1):26–35, 1987.

Ken Schwaber. Scrum development process. In Jeffrey V Sutherland, Dilip Patel, Cory Casanave, Joaquin Miller, and Glenn Hollowell, editors, *Business Object Design and Implementation: OOPSLA95 Workshop Proceedings*, pages 117–134. Springer, April 1995. ISBN 978-1-4471-0947-1.

Ken Schwaber. *Agile project management with Scrum*. Microsoft Press, 2004.

E Seiss, P Praamstra, C Hesse, and H Rickards. Proprioceptive sensory function in parkinson's disease and huntington's disease: evidence from proprioception-related eeg potentials. *Experimental Brain Research*, 148(3):308–319, 2003.

Neil Shah, Farshid Amirabdollahian, and Angelo Basteris. Designing motivational games for stroke rehabilitation. In *Human System Interactions (HSI), 2014 7th International Conference on*, pages 166–171. IEEE, 2014.

Mark Shtern, M Brandon Haworth, Yana Yunusova, Melanie Baljko, and Petros Faloutsos. A game system for speech rehabilitation. In *Motion in Games*, pages 43–54. Springer, 2012.

Jan Smeddinck, Kathrin M. Gerling, and Saranat Tiemkeo. Visual complexity, player experience, performance and physical exertion in motion-based games for older adults. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 25:1–25:8, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2405-2.

Stéphanie Tremblay, Douglas M Shiller, and David J Ostry. Somatosensory basis of speech production. *Nature*, 423(6942):866–869, 2003.

Gary Weismer, Yana Yunusova, and Kate Bunton. Measures to evaluate the effects of dbs on speech production. *Journal of Neurolinguistics*, 25(2):74–94, 2012.

John Westbury, Paul Milenkovic, Gary Weismer, and Raymond Kent. X-ray microbeam speech production database. *The Journal of the Acoustical Society of America*, 88(S1): S56–S56, 1990.

John R Westbury. On coordinate systems and the representation of articulatory movements. *The Journal of the Acoustical Society of America*, 95(4):2271–2273, 1994.

John R Westbury, Michiko Hashi, and Mary J Lindstrom. Differences among speakers in lingual articulation for american english/. *Speech Communication*, 26(3):203–226, 1998.

Ralf Winkler, Susanne Fuchs, Pascal Perrier, and Mark Tiede. Biomechanical tongue models: An approach to studying inter-speaker variability. In *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, pages 273–276, 2011.

Carolee J Winstein. Knowledge of results and motor learningimplications for physical therapy. *Physical therapy*, 71(2):140–149, 1991.

Carolee J Winstein and Richard A Schmidt. Reduced frequency of knowledge of results enhances motor skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(4):677, 1990.

Yana Yunusova, Gary Weismer, John R Westbury, and Mary J Lindstrom. Articulatory movements during vowels in speakers with dysarthria and healthy controls. *Journal of Speech, Language, and Hearing Research*, 51(3):596–611, 2008.

Yana Yunusova, Jordan R. Green, and Antje Mefferd. Accuracy assessment for AG500, electromagnetic articulograph. *Speech Language, and Hearing Research*, 52:547–555, 2009.

Yana Yunusova, Jeffrey S Rosenthal, Krista Rudy, Melanie Baljko, and John Daskalogiannakis. Positional targets for lingual consonants defined using electromagnetic articulography. *The Journal of the Acoustical Society of America*, 132(2):1027–1038, 2012.

Andreas Zierdt. Problems of electromagnetic position transduction for a three-dimensional articulographic measurement system. *Forschungsberichte-Institut für Phonetik und Sprachliche Kommunikation der Universität München*, 31:137–141, 1993.