TRACKING VISIBLE FEATURES OF SPEECH FOR COMPUTER-BASED SPEECH THERAPY FOR CHILDHOOD APRAXIA OF SPEECH

MEHRNAZ ZHIAN

A THESIS SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

GRADUATE PROGRAM IN COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO

May 2017

Abstract

At present, there are few, if any, effective computer-based speech therapy systems (CBSTs) that support the at-home component for clinical interventions for Childhood Apraxia of Speech (CAS). PROMPT, an established speech therapy intervention for CAS, has the potential to be supported via a CBST, which could increase engagement and provide valuable feedback to the child. However, the necessary computational techniques have not yet been developed and evaluated. In this thesis, I will describe the development of some of the key underlying computational components that are required for the development of such a system. These components concern camera-based tracking of visible features of speech which concern jaw kinematics. These components would also be necessary for the serious game that we have envisioned.

To My Parents,

For their endless love, support and encouragement

Acknowledgements

I would like to thank my supervisor Prof. Melanie Baljko also Prof. Petros Faloutsos for all their support and inspirations. The implementation of this research project could not be possible without their kind advice. Also, I would like to thank Dr. Aravind Namasivayam, Dr. Yana Yunusova, clinical and research speech language pathologists for their valuable input, comments, and recommendations. Many thanks go to Speech Production Lab at the University of Toronto and the Toronto Rehabilitation Institute – the University Health Network (TRI-UHN) for providing invaluable research support. I also would like to thank the members of my supervisory and examination committees for their time and effort in reviewing my thesis. I must thank all of the participants who helped me in user studies for their invaluable participation. I also want to thank my family for encouraging me in all of my pursuits and inspiring me to follow my dreams. I am especially grateful to my parents, who supported me emotionally and financially.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

At present, there are few, if any, effective computer-based speech therapy systems that provide the at-home component for clinical interventions for Childhood Apraxia of Speech (CAS). PROMPT, an established speech-therapy intervention for CAS that is non-computer-based, has both in-clinic and an at-home components. The at-home components presently consist of assigned articulation exercises performed by the child. It has been suggested that the at-home component has the potential to be supported via a computer-based system, which could increase engagement and provide valuable feedback to the child. However, the necessary computational techniques has not yet been developed and evaluated. In this thesis, I will describe the development of some of the key underlying components that are required for the development of such a system. It is important during the development of these components to keep the needs of the eventual end users in mind throughout the development process.

1.1 Background

1.1.1 *Apraxia of speech*

Apraxia of speech is a form of oral motor speech disorder. It is thought to be caused by a form of impairment to the cerebrum. Apraxia causes complications in performing movements tasks and thus disturbs a person's ability to articulate speech using the correct motor plans. It does not affect sensory or comprehension impairment. Apraxia can cause

people to have difficulty in articulation, which, in turn, creates problems for those who seek to understand what the speaker is trying to convey.

Apraxia of speech can occur in both children and adults. In adults, who have already attained speaking ability, it is usually an acquired condition, the cause of a progressive genetic illness, or a stroke. It comprises a high demographic in young children (Maassen & van Lieshout, 2010).

In children, the disorder is known as Childhood Apraxia of Speech (CAS). Children with CAS have reduced ability to produce syllables or units of words in a structured fashion. There is no clear reason as to how a child develops this motor disability, but studies have shown that it relates to the neurological system and the ability of the brain to process signals related to speech (Beukelman & Mirenda, 2005; Ziegler, 2008). By and large, CAS has no impact on the natural intelligence or the comprehension abilities of a child.

For a child with CAS, some improvement can be achieved by practicing speech exercises under supervision of a speech therapist (Ballard, Maas et al., 2007). Some of such treatment practices are as follows:

- **Motor-programming approaches**: these approaches employ motor-learning techniques such as continues word repetitions and are shown to be helpful in improving the ability to speak, repeat, and articulate.

- **Linguistic approaches**: these approaches are one of the primary methods in dealing with CAS, which aim to helping the child to make different sounds in their language.
- **Grouping methodologies:** these approaches are hybrid methodologies that incorporate techniques from both linguistic and motor-programming approaches.
- **Sensory prompting methodologies:** these approaches employ the child's own sense to practice the words and constructing sentences. PROMPT, described below, is one such approach.

1.1.2 *Speech Therapy for CAS*

Speech therapy involves the variety of services provided by a speech therapist, typically with a professional designation as Speech Language Pathologist (SLP). Today, clinical speech therapists often make use of different techniques when dealing with individuals who are impacted by speech disorders. In order to tackle and exploit different therapeutic advances in clinical settings, structures and frameworks must be established which afford the investigation of articulatory kinematics with regards to new clinical practices. In a speech disorder such as CAS, these frameworks are very important (Bartle-Meyer, Goozée et al., 2009).

1.1.3 ***PROMPT***

The approach ‘Prompts for Restructuring Oral Muscular Phonetic Targets,’ also known as PROMPT, is a clinical intervention method used to treat people with speech disorders such as CAS (D. A. Hayden & Square, 1994). PROMPT delivers therapy for speech disorders such as aphasia, apraxia/dyspraxia, dysarthria, pervasive development disorders, cerebral palsy, acquired brain injuries and autism spectrum disorders.

The focus of intervention for CAS is on improving the planning, sequencing, and coordination of muscle movements for speech production. Isolated exercises designed to "strengthen" the oral muscles alone are not effective, due to the fact that CAS is a disorder of speech coordination and not a disorder of strength; therefore, treatment approaches should take other factors such as ability to coordinate muscle movements and planning into account.

While there is a lot of variation within speech production and incomprehension for children following typical development mile stones, there are clinically established standards for the diagnosis of disorders of speech (ASHA, 2016).

The PROMPT treatment approach has both a clinical component as well as homework component. The efficacy of the at-home component PROMPT relies heavily on the patients’ behaviour. The clinical component entails one-on-one sessions with a speech pathologist (Bose, Square et al., 2001). Research has shown that the children with CAS have more success when they receive frequent exercises (e.g., 3-5 times per week)

(Murphy & Carbone, 2008). During a session, the speech language pathology prompts the child to produce certain words or sentences according to the stage of the PROMPT therapy, and also provides feedback specific to each child (Hodge, 1998). This clinical component can be coupled with at-home practice with parents and guardians. Concerning the at-home component, children with CAS may benefit from working with rhythms, frequent practice of the pronunciation of different sounds and words, and by practicing to string the different sounds together (Case-Smith & Bryan, 1999). In current practice, children do not receive corrective feedback during homework, but do gain valuable practice.

There are many factors that impact treatment efficacy. For instance, children seen alone for treatment tend to do better than children seen in groups. As the child improves, they may need treatment less often, and group therapy may be a better alternative. As well, a child with guided feedback during speech exercises via different modes of feedback, such as tactile "touch" cues and visual cues (such as looking at her/himself in the mirror or receiving other forms of visual feedback), can be expected to make better progress than those without this feedback (ASHA, 2016). With multi-sensory feedback such as this, the child can more effectively modify their speech behaviour to repeat syllables, words, sentences and longer utterances to improve muscle coordination and sequencing for speech.

Technology-based interventions, such as computer-based speech therapy systems (CBSTs), have incredible potential to improve the delivery of speech therapy. PROMPT

is one of the various speech therapy approaches that involve numerous repetitions of articulation exercises. The basic framework for a CBST is: (i) to support the user in undertaking the speech exercises specified within a given clinical speech rehabilitation protocol, (ii) to perform formative assessment of the speech elicited from the user, and (iii) to provide feedback to the user that supports the objectives of the clinical protocol. There is an opportunity to introduce a CBST system that can support the at-home component of PROMPT, and as well, an opportunity to introduce a CBST system to complement to in-clinic therapy.

Such a CBST system, before clinical deployment, must be thoroughly evaluated and its positive contribution to the therapeutic intervention needs to be validated.

1.2 **Problem Statement**

This thesis is concerned with a number of issues.

The first issue concerns identification of the sub-components of PROMPT which are potentially suitable for support via a CBST system. As described earlier, a CBST needs to, among other things, to provide feedback to the user that supports the objectives of the clinical protocol, on the basis of speech elicited in the context of a particular therapeutic protocol. Thus, this issue entails identification of clinical targets within the PROMPT approach that a CBST could feasibly support. The thesis entails an investigation of the different clinical targets within the PROMPT approach, which results in the identification of two clinical targets: correct degree and plane of jaw opening. These targets correspond

to two sub-stages within the protocol: phonatory and mandibular control. Support of these stages requires effective corrective feedback, which is given on the basis of the child's actual articulation relative to the target articulation. This feedback depends on the correct identification of the actual-vs-required degree of jaw opening and lack of lateral jaw movement.

The next issue concerns the identification of the main requirements of a CBST system to support delivery of PROMPT. In addition to the requirements that the system correctly track the targeted features of speech and provide effective feedback, there are a number of other considerations such as price, computational power, and sensor availability. The use of a User Centered Design (UCD) methodology affords the opportunity to identify these crucial stakeholder requirements at an early stage of research, when accommodations can and should be taken, rather than much later. The thesis entails undertaking an early iteration of UCD, which identifies, among other requirements, that a commodity-grade tablet-based camera should be the basis for tracking the clinical targets of correct degree and plane of jaw opening.

Given the need to provide effective corrective feedback concerning degree of jaw opening and jaw sliding, a research problem is to identify an inventory of visible facial features that have the potential to have a high degree of correlation with these clinical targets. By virtue of their correlation, these facial features become 'features of speech'.

A further task is to develop a camera-based computational technique that could potentially be the basis for the automatic identification of the facial features that are

correlated to the clinical targets. My conjecture is that these techniques can be developed using already-developed camera-based tracking techniques, with some additional software augmentation.

Given a technique which performs camera-based tracking of features of speech, a further task is to determine its accuracy. Determination of the fidelity of a tracking technique requires an appropriate methodology and the use of a secondary sensing system to provide the basis of comparison. The issue of fidelity is crucial, as it determines the viability of a CBST system to support the particular clinical targets that have been selected.

Last, once a technique has been developed which performs camera-based tracking of features of speech with an acceptable degree of accuracy, the next research problem is to investigate the degree to which this technique can serve as the basis of the design of corrective feedback. Can a module be developed that, on the basis of camera tracking, distinguishes between clinically-relevant categories of levels of jaw opening and jaw sliding?

Once these issues have been addressed, subsequent stages of research will be to use the techniques developed here in the design of a game-like scenario which delivers stage 2 and 3 PROMPT therapy and which provides accurate corrective feedback. Once the corrective feedback is validated, then the game-like scenario can be further augmented into a fully-fledged computer-based therapy system for at-home support of

Stage 2 and 3 PROMPT therapies for Childhood Apraxia of Speech that uses game-like scenarios.

1.3 Thesis Structure

This thesis presents the results concerning the development of several key antecedent components which are necessary for the development of a home-based CBST system that supports the delivery of PROMPT therapy for CAS.

In chapter 2, an overview of speech rehabilitation for CAS is provided. This chapter also describes the need for the CBST support for the at-home component of PROMPT. The design methodology is described and the stakeholders' needs are investigated. An overview of the relevant facial feature tracking techniques, including those based on camera input as well as other sensor systems, is provided.

Chapter 3 describes the two phases of this research (1) the identification of visible features of speech from a video corpus; (2) the development of relevant computational techniques. It also provides a report on the fidelity study.

Chapter 4 describes the work towards a computational module that, on the basis of camera tracking, is able to distinguish between clinically-relevant categories of levels of jaw opening and jaw sliding. This work focuses on the use of off-line classifiers. This chapter also present an investigation of the degree to which CVC productions can be classified using visible features of speech.

Finally, chapter 5 states the conclusion and future work.

Chapter 2
Design Methodology and Stakeholder Needs

2.1 Introduction

This chapter covers a number of topics. First, an overview is provided of PROMPT speech therapy. Next, the stakeholder needs are identified and an initial requirements analysis is performed, followed by a discussion of the user center design methodology as applied to this project. Further, we performed a review of literature concerning computer-based speech therapy (CBST). Additionally, a review of speech articulator tracking and face tracking techniques is provided, with an emphasis on techniques for tracking visible facial features of speech articulation, especially the key speech articulators, such as jaw, lips, and mouth.

2.2 PROMPT Speech Therapy

Prompts for Restructuring Oral Muscular Phonetic Targets, also known as PROMPT, is a clinical intervention method used to treat people with speech disorders (Chumpelik, 1984). PROMPT delivers therapy to individuals affected by a range of conditions, including apraxia, dyspraxia, aphasia, dysarthria, pervasive development disorders, Cerebral Palsy, and acquired brain injuries (Bose, Square et al., 2001; Chumpelik, 1984; Ward, Leitão et al., 2014). Children on the Autism Spectrum may benefit from working with rhythms, and from practice in the pronunciation of different sounds and words and in stringing different sounds together (Case-Smith & Bryan, 1999). Many patients with

childhood apraxia of speech have benefited from PROMPT therapy (Cumley & Swanson, 1999).

PROMPT's conceptual framework is based on the assumption that the process of speech production is the result of external and internal interaction factors (D. Hayden & Stockman, 2004) such as physical, cultural and social entities. The internal factors are specifically physical-sensory (facial structures, neuromuscular integrity), cognitive-linguistic (perception, sensation) and social-emotional (trust, willingness) entities. Considering these factors, the PROMPT approach takes a direct approach to evaluating and to facilitating improvement in each of these factors individually and in correlation with one another. PROMPT is based on a tactile-kinaesthetic approach to speech motor treatment, which is an approach that uses touch cues to a patient's articulators (jaw, tongue, lips) to help them with manual feedback throughout a targeted word, phrase or even a sentence. PROMPT develops motor control and the development of proper oral muscular movements while eliminating unnecessary muscle movements such as jaw sliding and inadequate lip rounding. Therefore, PROMPT therapy is a systematic approach to treat children with CAS as a movement disorder (Strand, 1995).

Most of the time, children with Childhood Apraxia of Speech (CAS) will benefit greatly from speech therapy (Murray & Parker, 2004). CAS is mostly a sensory-motor disorder which usually affects the articulatory restrictions of speech production. Children with CAS most likely have difficulties with repeating words or deliberate movements necessary to produce speech.

A goal of this project is to determine what is needed to create an effective computer-based system of PROMPT therapy for Childhood Apraxia of Speech in the home setting. To this end, I review the main component of the PROMPT therapy below.

In PROMPT therapy, the clinician provides tactile 'prompts' to provide children with feedback about their motor control while articulating words (Square, Chumpelik et al., 1986). PROMPT contrasts with traditional intervention methods which focus solely on auditory and visual feedback. PROMPT therapy is based on the 7 stages of the motor speech hierarchy and includes the following sequence of intervention components (D. A. Hayden & Square, 1994):

1. Tonal control
2. Phonatory control
3. Mandibular control such as jaw control (vertical plane of movements)
4. Labial or facial control (horizontal plane of movement)
5. Lingual control (anterior-posterior plane of movement)
6. Sequenced movements (co-articulated multiple planes)
7. Prosody

These components guide clinicians through the intervention procedures, beginning with the first stage of tonal control and ultimately finishing on the prosody stage. The above seven components are independent of each other. The first and second components (tonal control, phonatory control) mostly concentrate on the child's tone and

articulatory postures. The third to sixth components (vertical, horizontal, anterior-posterior, and co-articulated planes of movements, respectively) focus on the conception of "plane of movement" of the speech articulators. The final component of the intervention (prosody) focuses on the important factors for speech production and stress alterations which is responsible for creating different meanings in words and utterances.

The determination of whether a child is an appropriate candidate for PROMPT therapy is usually carried out by a PROMPT-trained speech therapist (Baranek, 2002). Speech Pathologists are the only professionals with the prerequisite knowledge to learn and apply PROMPT in the holistic manner in which it is intended to be used. Children who are impacted from disorders of speech should undergo extensive evaluations conducted by an expert Speech Language Pathologist (SLP). As part of the assessment process, the SLP will evaluate the child's oral abilities in order to indicate if the child is experiencing CAS, and devise an ideal approach to assist the child. Assessment includes listening to words spoken by the child and assessing whether appropriate stress on the syllables is being made during articulation. Children with CAS normally mispronounce the two similar words in consecutive repetition. To assess this, the SLP asks the child with speech disorder to repeat words which are close in pronunciation (e.g. "beam" and "bead") and determines whether the child articulates the words correctly (Takada, Miyawaki et al., 1994). Another important motor assessment for CAS concerns about the jaw movement displacement. As part of the diagnosis procedure, the SLP checks the child's face movements while assessing their articulation. For instance, the jaw sliding

from one side to another side of the mouth once the child starts to speak is a clear sign of this speech disorder (Geng, 2012).

As part of the assessment process, an expert SLP should also assess for the presence of additional motor disorder, such as dysarthria (Forrest, 2003). Even though a child might not seem to have CAS, they still might have very subtle signs that only an expert SLP can detect. The CAS signs might not be detectable to a layperson just by looking at the face, it is necessary to have their speech activities assessed by an expert SLP. In order to make an assessment, the SLP should listen to their speech production and asses their muscle control.

As part of the assessment process, an expert SLP should assess inability to pronounce under certain conditions. For instance, the SLP may ask the child to imagine eating an apple and then to articulate a series of words.

The PROMPT intervention has both a clinical component and an at-home component. The clinical component entails one-on-one sessions with a speech language pathologist (SLP), typically three to five times per week (Bose, Square et al., 2001). During the session, the clinician prompts the child to produce certain words or sentences according to the stage of the PROMPT therapy and provides feedback (Hodge, 1998). This clinical component is coupled with at-home practice, undertaken under the supervision of parents or guardians. Unlike the clinical component, children typically do not receive corrective feedback during their homework component, but nonetheless, at-home therapy is known to be an important part of PROMPT therapy.

2.3 Stakeholder Needs and Initial Requirements Analysis

At an early stage of this research project, a SLP clinician approached the research team with the idea to use a computer-based support for the homework component of PROMPT therapy. In response, a design process was initiated. The first step was to conduct an initial system 'needs and analysis' phase. This phase was supported by a series of weekly meeting among the research team members and the SLP, held in the VTV research lab in TRI-UHN (Toronto Rehabilitation Institute – University Health Network). A long term objective was developed, which is to create an effective computer-based system for at-home support of PROMPT therapy for CAS.

2.3.1 *Requirements*

Early requirements analysis revealed the following:

- R1: system should be effective in terms of providing accurate feedback and supporting clinical objectives more broadly
- R2: system should be low cost (for family member requirements)
- R3: system should be highly engaging (for client requirements)
- R4: system should be tablet-based
- R5: system should make use of device-embedded camera for facial feature tracking (for clinical requirements)
- R6: system should also function on traditional computing platforms (e.g. desktop workstation) (for parents and clinical requirements)

- R7: system should be usable in a home setting (for parent and clinical requirements)

R1: Efficacy (requirement for: client, family member, and clinician)

First and foremost, the system must properly and correctly support the delivery of the PROMPT therapy. Specifically, the system should elicit the necessary practise articulations and also provide useful, corrective feedback. Incorrect feedback, such as a false positives (system assess production as 'correct' even when it was incorrect) and false negatives (system assess production as 'incorrect' even when it was correct), was felt to be potentially very harmful. False positives could serve to reinforce incorrect productions and false negatives could be discouraging and disheartening. Thus, even no feedback would be better than incorrect feedback. It can be difficult to determine minimum accuracy levels, but we consider 90% to be a bare minimum.

It was felt that the phonatory control (Stage 2) and the mandibular control (Stage 3) components of PROMPT would be the best PROMPT stages to target as a first step. These stages offer the best prospect of being able to generate correct computer-based feedback, as compared to the type of corrective feedback that would be required of other PROMPT stages. Phonatory control is one of the essential steps of speech production, since its main focus is on the child's posture. Mandibular control mostly concentrates on the theory of "planes of movement". A focus on Stage 2 and 3 of PROMPT provides the basis for a specific target for clinical intervention.

R2: Low cost (requirement for: family member)

Cost was identified as a key factor. High cost would be a barrier to many families and would reduce the likelihood of the system being adopted by the target user groups. Thus, the system should be delivered via a common, relatively low-cost computational platform, such as a tablet device. Moreover, free software platforms and libraries should be used, to avoid the expense of licensing costs.

R3: High degree of engagement (requirement for: client, family member)

Boredom and monotony was identified as a problem with the current delivery of PROMPT. This is to be expected, given the large number of repetitions that the therapy requires. In fact, demotivation was described by the SLP as a key reason for lack of compliance with the 'at-home' component of PROMPT therapy. Thus, it will be important that the system be highly engaging to children and the system should be learning. A game-like experience was felt to be important to the design of the system.

R4: System should be tablet-based (requirement for: clinician)

Since the envisioned context is home use, it is important to identify devices that may already be available, as this would reduce barriers to adoption. Thus, tablet-based devices, which are already very common, would be ideal. Since the application is meant to be used by young children in a home setting, a tablet platform would be the best alternative. Due to its worldwide recognition and ease of implementation, iPad has

proved to be a very user-friendly platform in which the application can be efficiently developed. Using iPad's built-in camera; facial tracking algorithms can be used to correctly position and trace a user's facial structure when the application is in use. However, other camera-equipped tablets could also serve as the platform.

R5: System should operate using the device-embedded camera for facial feature tracking (requirement for: clinician)

While it is true that the performance of high-end camera rigs are superior to tablet-embedded cameras, they are also highly specialized, expensive, and unlikely to be adopted by the target user groups. As well, Kinect-based systems with 3D motion capture also represent potentially useful sensing technologies, but are not yet widely found in most residences. Thus, it was felt that the tablet-embedded camera should be used, if possible, as the source of input to the system. Moreover, it was felt that marker-based facial feature tracking techniques would present considerable usability obstacles (disinclination to attach markers and potential for incorrect placement). Instead, it was identified that marker-less facial feature tracking would be needed in the system.

R6: Ability to be used in home settings (requirement for: all stakeholders)

Children should be able to use a system that supports the therapeutic outcomes in home settings. Considering the stakeholders' needs, the SLP's focus was on providing a compelling context for children to eagerly perform the given speech therapy exercises at

home. Traditional computer-workstation applications may be suitable in clinical settings; however, a more cost-effective platform is desired with regards to mobility for conducting in-home practices. There are many advantages of practicing the speech therapies in the home setting. Home practice could increase the efficacy of the treatment.

2.3.2 *User Groups*

The design session discussions revealed different categories of users:

1. Clinicians
2. Children (clients)
3. Family members (support network around children)

There is potentially a large population of children who would benefit from a system delivers an effective at-home computer-based system of PROMPT or other types of speech therapy. According to Canada's 2001 Participation and Activity Limitations Survey, approximately 155,000 children between 5 to 14 years old have some form of speech impairment such as apraxia or dysarthria (Behnia & Duclos, 2003) In Canada, there are large numbers of applications to access special education services, such as services for children with speech disorders. One study reported that more than 29,000 children receive special education services from their speech therapies centers (Uppal, Kohen et al., 2008). And while speech disorders in children usually do not create insurmountable obstacles for interacting with others, a major issue for these children can be seen in the effect on their behaviour and social skills (Svirsky, Robbins et al., 2000).

2.3.3 *Research Objectives*

On the basis of this initial requirements analysis, objectives along three timeframes were identified.

- Short term:
    - to identify an inventory of visible facial features that have the potential to have a high degree of correlation with jaw opening and with jaw sliding,
    - (ii) to identify a slate of ready-made camera-based face tracking libraries that could potentially be the basis for the automatic identification of facial features, and
    - (iii) to determine the accuracy of the camera-based marker-less facial feature tracking techniques
- Mid-term: to design a game-like scenario which delivers Stage 2 and 3 PROMPT therapy and which provides accurate corrective feedback
- Long-term: to design an effective computer-based system for at-home support of Stage 2 and 3 PROMPT therapies for Childhood Apraxia of Speech that uses game-like scenarios.

2.4 **User Centered Design**

Upon completion of the first phase, a User Centered Design (UCD) methodology was employed. UCD is used in Human-Computer Interaction design projects as an approach that involves users during the development process. A main goal in UCD is to enhance

the user experience of the system (Dabbs, Myers et al., 2009). The UCD methodology stands in contrast to Technology-Centered Design, which is typically adopted in the design of software systems.

UCD involves the users from an early stage of the design throughout the development process, so that the usability of the system will be increased (Gabbard, Hix et al., 1999; Van Velsen, Van Der Geest et al., 2008). A main focus of UCD is on the design process which allows users to have an impact on the formation of a design (Abras, Maloney-Krichmar et al., 2004). UCD involves focusing on the participants' needs, such as in a general requirements analysis. Some variants of UCD involve users in their design at specific times, such as at information gathering or testing phases. In other variants, users are involved throughout; they have influence the design process at all stages and become partners in design. Figure 1 provides a general overview of the UCD process.

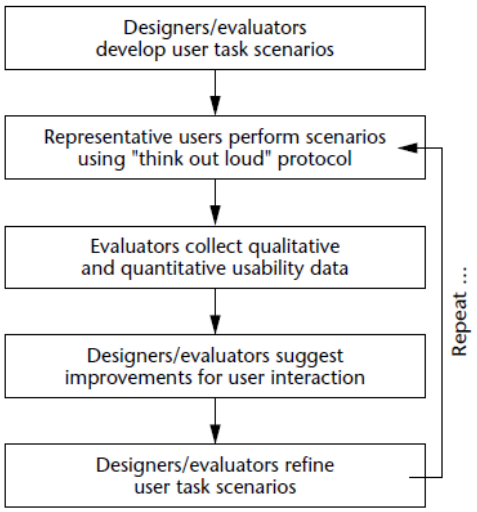

Figure 1. User-Centered Design Gabbard, Hix et al. (1999)

In UCD, “plans are just the beginning of the process, but the main mission is not conforming to the plan; rather, it is responding to changes throughout the life cycle of the project” (Baek, Cagiltay et al., 2008). By employing user center design, stakeholders are becoming the center of the design.

Research through design (RTD) is another relevant methodology. RTD supports empirical questions and stakeholder contribution and is mainly suitable for explorative inquiries in the HCI field of study (Zimmerman, Forlizzi et al., 2007). There is strong evidence for the use of the RTD research methodology for game design for children who experience different type of disabilities (Frauenberger, Good et al., 2011; Hamidi, 2016).

For the reasons described above, UCD is a methodology that improves the likelihood of a successful outcome. Therefore, the approach adopted in this research project is based on UCD methodology.

2.5 **Relevant Background and Literature Review**

This section provides a review of the prior work related to this research project. With respect to identifying an inventory of visible facial features that have the potential to have a high degree of correlation with jaw opening and with jaw sliding, we look to prior research concerning Computer-Based Speech Therapy Systems (CBSTs), both for CAS and for articulatory and language disorders. We also examine prior research on ‘Serious’ games and speech therapy.

With respect to identifying a slate of ready-made camera-based face tracking libraries that could potentially be the basis for the automatic identification of facial

features (and their degree of accuracy), we look to prior research on tracking visible features of speech and on facial feature detection relevant to movements of speech articulators.

2.5.1 *CBSTs for Childhood Apraxia of Speech*

Computer-Based Speech Therapy systems (CBSTs) are computational hardware and software systems that support speech therapy (Bälter, Engwall et al., 2005). Many CBSTs described in the research literature focus on particular speech and language disorders and only a small number have specifically targeted delivery of therapy to children with Childhood Apraxia of Speech (CAS).

In a recent pilot study on children with apraxia of speech (Parnandi, Karappa et al., 2013), children with CAS were asked to complete a series of exercises under SLP supervision. The exercises were performed on a tablet in at in-home setting, delivering the Nuffield Centre Dyspraxia Programme (NDP3) (Williams, McLeod et al., 2010). The tablet was running the "Brick Wall" application, which provided visual aids such as video, audio, and animations to elicit articulations by the child. The SLP remotely assigned tablet-based speech production exercises, and in response, the children recorded their own speech in response to the prompts shown on the tablet's screen. The SLP then reviewed all the recordings and provided feedback to the children and their parents on the basis of the acoustic data that was recorded. The speech therapy practice exercises were given to each child remotely, and the child could take his/her own time to practice the

exercises. This pilot study demonstrated positive outcomes on the participants' learning process and the utility of the remote therapy.

Shahin, Ahmed et al. (2015) conducted a study of computer-based speech therapy for children with CAS. A goal of this research was to find the way to help the SLPs, as well as the parents, to reduce time and cost in the delivery of speech therapies. The system "Tabby Talk" was developed, which is a multi-layer remote system which involves the parents, clinician, and the child. In this system, the clinician builds exercises and remotely assigns them to the child. Tabby Talk uses a combination of modules for all speech therapy activities. The modules were trained to derive representative acoustic models of "correct" productions by identifying any delay or error in the children's' voice while pronouncing a word. These modules use a database of US-American speech samples of children with CAS. The dataset was collected through a speech therapy clinic.

When a child performs exercises using Tabby Talk, their articulations are recorded and sent back to the clinician. Through this system, the clinician monitors the child's progress. The clinician analyses the recorded exercises using the speech recognition engine which runs on the Tabby Talk system. The engine identifies certain classes of errors that are associated with CAS, such as voice delays, articulation errors within words, and incorrect syllabus stress patterns on the words. The SLP assigns speech therapy exercises to each child that is tailored to address the identified errors.

To assess the accuracy of the error identifications made by Tabby Talk, an evaluation study using two speech datasets was performed (the training and the testing

dataset). Speech data in the training dataset was collected from 1100 children between the ages of 4 and 16. In the training dataset, all participants performed the same task, which was the articulation of 100 long sentences which included 205 unique words. The testing dataset contains the articulations of two children with CAS and four typically-developing children. The children were asked to produce the 205 unique words. In order to create the ground truth classifications, all the articulations in the training and testing datasets were manually labelled by expert SLPs (according to correct or incorrect articulation and, if incorrect, then labeled according to error type). The error types are: voice delays errors, pronunciation errors, and stress errors. An evaluation in a testing dataset, including two children with CAS and four typically-developing children, has shown the accuracy of 88.2% of voice delays errors, 80.7% of pronunciation errors, and 83.3% of stress errors.

2.5.2 *CBST for Articulatory Disorders*

Although only a small number of studies have examined CBSTs that have specifically targeted delivery of therapy to children with Childhood Apraxia of Speech (CAS), other CBSTs have targeted delivery of therapy to speech articulation disorders more generally. An articulatory disorder is one that involves the production of the sounds of a language, oftentimes due to problems with the movements of the articulators (including the jaw, lips, tongue, and mouth).

CBSTs that have been developed for use in clinical settings are designed to support speech language pathologists and are often based on motor learning theory which put an emphasis on repetition with feedback (Wiepert & Mercer, 2002). CBSTs that have

been designed and developed for home use are based on the enforcement theory and employ protocols that put an emphasis on the target performance (Fell, MacAuslan et al., 2006; Ferster, 1964; Koegel, O'dell et al., 1987; Whalen & Schreibman, 2003). These systems are designed to encourage the participants to interact with the system in a way that creates the desired target behaviour.

CBSTs have been utilized in two primary modes (Bälter, Engwall et al., 2005): (1) as a platform to help the child to practice the speech therapy exercises when the SLP is not present (Bälter, Engwall et al., 2005; Vicsi, Roach et al., 2000), (2) as learning software to help the SLP and parents in order to interact with the child (Öster, House et al., 2003).

Murray and Parker (2004) have conducted a research study to determine the effectiveness of the Sound Therapy Lite software application toward meeting speech-therapy goals. In this study, a set of first-and second-grade students who had previously-defined articulation goals were chosen to partake in the study. These students were then introduced on an individual basis to the Sound Therapy Lite program in order to allow for them to familiarize themselves with the software. After a three-week period, during which each student used the software four times a week for 60 minutes each time, 85 percent of these students were observed to have improvements in their articulation of speech. This study showed that the use of technology for articulation therapy can be beneficial.

Brain Computer Interfaces (BCIs) provide a mode of interaction system in which no muscle activity is involved (Wolpaw, Birbaumer et al., 2002). The core concept of BCI is to translate the brain activities into useful input signals (Vaughan, Heetderks et al., 2003). BCIs have been demonstrated as an effective mode of feedback in treatment of patients with physical disabilities (Rao, 2013), and with Alzheimer and Parkinson disease (Rupp, Kleih et al., 2014).

A group of researchers (Al-Nafjan, Al-Wabil et al., 2015) designed and developed a BCI-based system to deliver speech-language rehabilitation in clinical settings. In this research, BCI was used to recognize the limitation of a system designed to help patients in their speech therapy programs. Signals from the brain were collected via an Electroencephalography (EEG) device. The EEG signals collected electrical activity in the analytical cortex and were also used as the basis for observations about the participant's emotional state. A study was conducted to assess how BCI improved the speech therapy learning process. Participants in this study were 7 people, including children and adults. Due to the fact that some participants were not comfortable with the use of the EEG device, they decided not to participate in the study. The EEG device was attached using a headband on each participant's head. The EEG device comprised of 14 electrodes, and each electrode was attached to a different location on the scalp. In general, setting up the BCI system ranged from 2-10 minutes per participant. In total, two sessions were conducted for the study: (1) participants were asked to perform a reading (read a story of a few sentences length in Arabic aloud, during which they could hear

themselves), for the duration of 3 minutes; (2) participants were asked to summarize and to discuss the story they had read earlier. Data about the individual participant's emotions was collected both from the EEG and from the SLP's perceptions with respect to different major categories such as engagement, enjoyment, and frustration. Moreover, during the experiments, the SLP sat beside the participants to observe the participant's emotion and cognitive status from the EEG device. The result from the BCI study has shown that different participants took different amount of time to finish each session. The reading and talking tasks ranged from 1-2:05 minutes. Further, the EEG device results was compared to the SLP assessments results and found out that there is a difference between the values of the four major categories. The results from this study have shown that BCI system can increase the speech therapy session in clinical settings.

2.5.3 *Computer-Based Approaches for Language Disorders*

Related to systems that focus on disorders of speech are other systems that focus on disorders of language. Speech language disorders are a class of disorders that consist both of low level speech disorders (of which Childhood Apraxia of Speech is one) and of high level language disorders. Language disorders can affect different parts of language such as content of language, function of language or a combination of both of them (ASHA, 1993). Thus, high level language disorders can also impact on speech.

Specific language impairment (SLI) refers to a type of language disorder which can be difficult to identify. There has been some previous research that concerned the development of computer-based means to diagnose SLI. Diagnosis of SLI was

accomplished by using a new approach based on a fuzzy cognitive map (FCM) (Georgopoulos, Malandraki et al., 2003; Stylios, Georgopoulos et al., 2008). FCM is a methodology in which is used in complex system modeling (Craiger, Goodman et al., 1996). The FCM model has been used in many different areas such as Electrical Engineering (Styblinski & Meyer, 1991), and medicine (Stylios, Georgopoulos et al., 2008). FCM has previously been used for models for radiotherapy treatment (E. I. Papageorgiou, Stylios et al., 2003), for brain tumour detection (E. Papageorgiou, Spyridonos et al., 2008), and for urinary infection identification (E. I. Papageorgiou, Papadimitriou et al., 2009).

Yet other systems target pragmatic aspects of language use and the social use of language. Beneficiaries of these systems can include individuals with Autism Spectrum Disorder (ASD). Computer-based systems represent an alternative to traditional approaches for children on the Autism spectrum (Heimann, Nelson et al., 1995). Traditionally, modeling instructions via video was employed as a way to improve these children's social skills (Bellini & Akullian, 2007). Empirical evidence shows that new technologies can be beneficial in the development of vocabulary for children with ASD (Moore & Calvert, 2000). The effectiveness of new technology was compared with traditional, instructor-led language instructions. The study was conducted on 14 children between the ages 3 to 6 (12 boys and 2 girls). Participants were randomly asked to work either with a teacher instructor or with a software-based learning application. Some of the participants who were not familiar how to use computer were provided with a practice

session. The researchers introduced the children to the click button function in the practice sessions. Both the instructor–led and the learning software introduced the children to new vocabulary words by an exercise that required the labeling of a target object. In each exercise, participants were asked to choose an object and name it verbally. In the instructor-led condition, verbal responses were provided to the participants if they did not respond to the correct name. In a software-based condition, verbal responses were motivated and generated using visual aids such as music, color, and animation. The results of this study indicated that verbal responses in computer application helped children to name more vocabulary words than in the instructor-led condition.

2.5.4 *'Serious' Games and Speech Therapy*

Computer games have been shown to provide positive learning circumstances and therapeutic outcomes for their users (Baker & Uhlig, 2011; Pinnell, 2015; Prensky & Prensky, 2007; Squire & Jenkins, 2003). *Serious games* are considered games with a primary purpose to support learning (Boyle, Hainey et al., 2016; Johnson, Vilhjálmsson et al., 2005; Zyda, 2005). In order to create a serious game, a combination is needed of learning, of educational tasks, and of game characteristics, such as rewards and increasing the difficulty of each stage of the game. This combination in speech therapy is called *gamification* (DomíNguez, Saenz-De-Navarrete et al., 2013).

Speech and language therapies also been "gamified" with many beneficial effects (Gros, 2003). Studies have shown that computer games can provide rich learning contexts to

support speech therapies in children by encouraging and engaging them in the games (Frauenberger, Good et al., 2012; Lohse, Shirzad et al., 2013; Shibata, Mitsui et al., 2001). By engaging patients in computer games, there is a possibility that patients could forget that they are involved in speech therapy exercises (Flores, Tobon et al., 2008). Means of engagement in computer games for supporting speech therapy have entailed: (i) gamification (Brederode, Markopoulos et al., 2005; Bunnell, Yarrington et al., 2000; Lan, Aryal et al., 2014; Vicsi, Roach et al., 2000), (ii) movements (Hummels, Van der Helm et al., 2006; Zakari, Ma et al., 2014), and (iii) tabletop games (Bakker, Vorstenbosch et al., 2007; Magerkurth, Stenzel et al., 2003; Piper, O'Brien et al., 2006).

The benefits from game-based approaches has been also extended to other types of language therapies (Maas, Robin et al., 2008). These serious games could be helpful in speech rehabilitation for getting feedback on each child's performance. As an example, Hoque, Lane et al. (2009) conducted a study to present an intervention for modifying speech-enabled games to help the participants produce intelligible speech. Participants were divided into two groups, and both groups first underwent two weeks of traditional speech intervention. Then, the first group went through an additional two weeks of traditional speech intervention, whereas the second group underwent two weeks of computerized speech intervention. The computerized speech intervention started with a therapist asking the participants to repeat a number of sentences; however, instead of getting direct feedback from the therapist, the participants received the feedback from interactive games. These interactive games were developed at MIT's research labs and

are freely available to the public (Kortemeyer, Fish et al., 2013). Results showed that participants enjoyed interacting with the games so much that they normally continued to play the games even after their time was up, whereas participants in the traditional mode of speech therapy were often distracted, annoyed, or bored. This demonstrates that the use of games and digital media offers promising solutions to address speech disorders.

A challenge for children with ASD is the development of appropriate facial expression (Adolphs, Sears et al., 2001). A study has been conducted of the SmileMaze game, created in order to help children to produce a greater range of appropriate facial expressions (Cockburn, Bartlett et al., 2008). In order to detect the child's face via a webcam, this game uses the computer expression recognition toolbox (CERT). The objective of this game is to collect candies while moving within the maze. Some part of the maze are blocked with a yellow smile symbol, and in order to continue through the maze path, the child needs to smile for a certain amount of time. The results of this study show that children were motivated to continue developing facial expressions, since smiling was one of the essential parts of the game. Figure 2 shows a child playing the game.

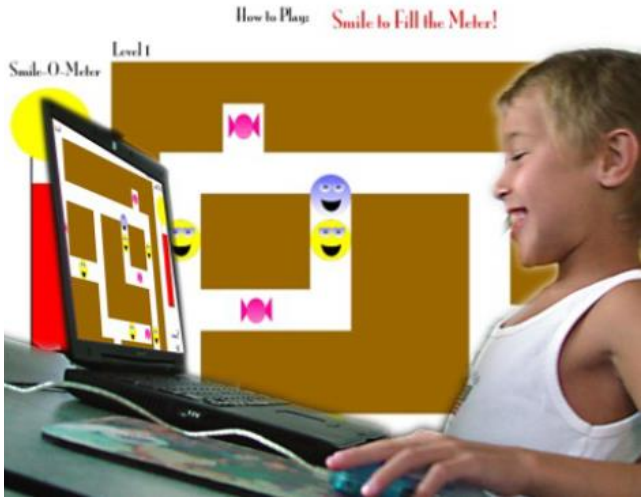

Figure 2. Children with autism playing SmileMaze (Cockburn, Bartlett et al., 2008)

In addition to the applications that support language intervention, many other computer-based applications also encourage patients in their speech and language therapies. A game called pOwerball is one such example, a comprehensive and flexible articulation program (Brederode, Markopoulos et al., 2005). This tabletop game was created by a certified SLP for children ages 8-14, with physical and/or learning disorders. The main goal of this game was to encourage children with disabilities to interact and improve their social skills. In this evaluation study, children were divided into 6 teams of 3 players each. Each group took 40 minutes to complete a session. To play this game, the team members sit together around a table where the game is illustrated. The goal of this game is to free creatures by moving virtual balls on the game table. The movement of each ball is done through the use of some switches. At the end of each game, the winner is the child who has freed the most creatures on the table. At the end of each session, an interview was conducted to gather children's feedback. According to the feedback from

children, the results of this study showed that the game was both engaging and easy to follow. Figure 3 shows children while playing the pOwerball game.

Figure 3. Children playing pOwerball (Brederode, Markopoulos et al., 2005)

2.5.5 *Tracking the Movements of Speech Articulation*

There are many non-camera based approaches for recognizing the movements of speech articulation, such as magnetic resonance imagining (MRI), tMRI imaging, x-ray micro-beam systems, and 3D electromagnetic articulography (EMA) (Earnest & Max, 2003). In general, EMA was introduced into the market to track the articulators of speech, such as the lips and the jaw (Hixon, 1971). Some drawbacks of the non-camera based techniques are that they are usually expensive and may cause discomfort.

Over the past few decades, various camera-based approaches have been developed for facial feature detection, which can be applied specifically to tracking the movements of the speech articulators, such as lips, jaw, and tongue.

A markerless, low-cost face tracking technique has been developed, making use of a Kinect sensor and a Vicon motion system (Bandini, Ouni et al., 2015). This system makes use of four different cameras to capture and to track the speech articulators, including the jaw, tongue, and lips. An evaluation study was performed to compare this technique to a marker-based technique called *Interface.* The *Interface* system, which operates on 2D images, fits a model consisting of 49 landmark positions on a face within the scene (10 landmarks on the eyebrow, 12 on the eyes, 9 on the nose, and 18 on the lips). Participants in this study were two people without any history of speech disorder. Each participant was asked to repeat 100 sentences and 200 words. In order to perform a comparison, the 2D coordinates derived from the Interface system need to be extrapolated to 3D coordinates for each landmark. Moreover, some articulatory parameters, such as distance between the corner of the lips, and distances on the frontal axis between the midpoints were computed. Once the relevant comparison points from each of the two techniques were extracted, the root mean square error (RMSE) was derived. The best RMSE values obtained were in the range of 1-3 millimetres. A recommendation arising from the study results was that, for camera-based studies making use of images with image resolution greater than 640*480 pixel, the distance between the

camera and the face should be around 0.5 meters. This study demonstrated the viability of low-cost marker-less approaches for facial movements in speech therapy practices.

2.5.6 *Face detection*

Face detection is a well-defined technique in computer vision to locate and to detect the presence of a face within an image or video stream. In order to detect the face, the face detection algorithm should be able to identify the face location, and possibly additional features, such as facial posture, facial expression, and facial features such as lips, nose, eyebrows, and eyes. Each algorithm for face detection has its own advantages and disadvantages. Some of these algorithms use contour models, neural networks, and flesh tones to detect the facial features. Being highly computational expensive and time-intensive are two significant disadvantages of some face detection algorithms (Bradski, 1998).

Viola and Jones (2001) presented a technique for detecting objects, including faces, within an image and proposed the AdaBoost classifier algorithm to detect any faces in frontal view. Nowadays this algorithm is used in many different application domains due to its performance at high speed. AdaBoost uses a small number of features among a potentially large set of features. Therefore, the accuracy of the classification is generally more efficient (Freund & Schapire, 1995). It uses Haar-like feature detection based on detecting features instead of pixels. The bottom line for detecting features rather than pixels is to increase the ability of the algorithm to detect faces. The features employed by the detection framework involve calculation over the image pixels within a given

rectangular region. In order to define and to detect an object, the Haar-like features are employed to detect the presence of three different types of features: lines, edges, and these in combination. For face detection specifically, a cascade with thousands of rectangular features is used in order to successful identify the presence of a face (Viola & Jones, 2001).

Haar cascade classification is a machine learning method where a cascade object is trained from different images (image of faces), and which is then used to detect the objects in the image (Viola & Jones, 2001). The Haar Cascade classifier must be trained with a large number of sample images of faces (or the particular object which is the target of detection). These training images should consist of positive and negative images. Positive images are those images that include the face; negative images are those which do not include any faces. After the classifier is trained, it is then applied to a region of interest within an image or video (which can be set to be the whole input image or video). The classifier provides binary outputs, depending on whether the region shows the specific target object or not. The region of interest can be set to be the whole input image or video or a specific scalable region. This method is known to be an effective object detection method (Viola & Jones, 2001).

Once a face can be detected within an image, the problem then turns to the detection of the internal features of the face.

2.5.7 *Detection of Facial Features*

The detection and tracking the facial features plays a significant role in many applications (Amberg & Vetter, 2011; Belhumeur, Jacobs et al., 2013; Rapp, Senechal et al., 2011). Facial feature points, also known as facial landmarks, can be used as reference points for the detection of basic features, such as eyes, nose, and lips (Gu, Su et al., 2003; Wu, 2013). Tracking specific features on the face can be challenging due to problems such as obstructions of the face, pose changing, and head swivelling (Gu, Su et al., 2003; Wu, 2013). Facial feature tracking techniques are generally distinguished on the basis of whether they employ a shape model or not.

For facial feature tracking techniques that do not employ a shape model, they track every single feature on the face independently without recognizing the whole face. Consequently, any changes such as pose change or obstruction do not impact the performance of this technique.

A subsequent section will address facial feature tracking techniques which use shape models.

2.5.8 *Detection of Facial Features with a Shape Model*

Facial feature tracking using a shape model is a strong method for identifying and for locating objects in the presence of occlusions. Dependencies among the facial features can be captured with this method. Consequently, any shape of the face can be determined and tracked by this method.

Examples of tracking using a shape model include Active Shape Model (ASM) (Cootes, Taylor et al., 1995), and Active Appearances Model (AAM) (Antonakos, Alabort-i-Medina et al., 2014). In both cases, a deformable shape model is employed and is fit to different shapes from an image or video stream. Both of these models (ASM and AAM) fit a generative model in order to detect the shape or appearance variations of each face. Another variant, using a combination of Support Vector Regression and Markov Random Fields, first prioritizes the location of facial features corresponding to stable points on the face (nose, and eye corners), as opposed to the other parts which may vary with facial expression (Valstar, Martinez et al., 2010). After identification of the stable facial features, detection of the different facial feature then takes place. In some situations, if the stable points cannot be detected, the facial feature detection might fail (for instance, if a person's eyes are covered by sunglasses).

2.5.9 *Facial feature detection under adverse conditions*

This research project makes certain assumptions: about the quality of the images (specifically, medium quality), about the pose of the subject face in the camera's field (specifically that there is a consistent frontal view) and about the direct view of the face (specifically that it can be assumed to be free of occlusions). However, the issues of image quality, face pose and occlusion are briefly covered here.

Facial feature detection using low-quality images

There is a body of research that focuses on facial feature detection specifically in low quality images. Although there are numerous techniques available which perform real-time facial feature detection on medium and high quality of images (Belhumeur, Jacobs et al., 2013), relatively few are tailored to work well on low quality images. Dantone, Gall et al. (2012) developed a real-time technique to operate on low quality images which employed a technique using conditional regression forests (Dantone, Gall et al., 2012). The technique locates facial points on the face by using support vector regressions and random forests (Breiman, 2001). Random forests have been used in different field such as regression (Fanelli, Weise et al., 2011) and classification (Shotton, Sharp et al., 2013). This technique works by using a set of 2D images in order to learn the relations between the locations of several facial feature points. To detect the facial features, the random regression forest is trained and tested. In order to achieve the head pose, the forest regression was trained to create five different subsets, corresponding to the head pose types of left, right, front, right profile, and left profile. Figure 4 shows a set of facial feature points on 2D images. The super-imposed aqua-outlined shapes predict the head pose.

Figure 4. Facial feature points on 2D images (Dantone, Gall et al., 2012)

Feature Detection in the Presence of Occlusions

The Annotated Facial Landmarks in the Wild (AFLW) database was created, containing 25,993 faces with a variety of face appearances such as pose and expressions (Köstinger, Wohlhart et al., 2011). In the AFLW database, there are many differences in pose and expressions and in the accessories that people are wearing (such as sunglasses, hats, so on). Thus, many of these faces are partially occluded. Figure 5 shows some of these images in the database.

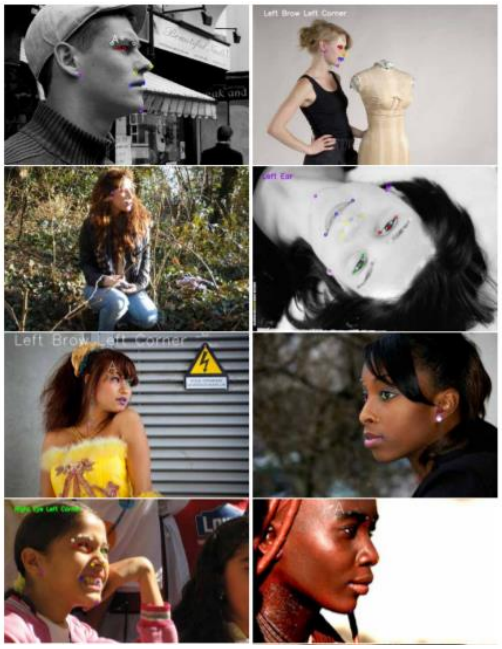

Figure 5. pictures in AFLW database (Köstinger, Wohlhart et al., 2011)

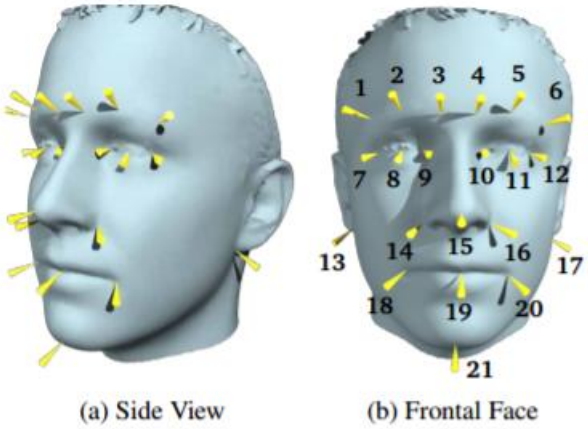

Figure 6. Landmark Locations from Different View (Köstinger, Wohlhart et al., 2011)

Making use of the AFLW database, a subsequent investigation aimed to locate 21 facial landmarks on the face (such as eyebrows, eyes, lips, and chin) across a variety of poses:

most of face poses in the AFLW database are in the frontal view; however, many are not (and instead are in three-quarter or side views). The output of the technique, as shown in Figure 6, is demonstrated on a 3D face model, shown in 2 different views (the side view and the frontal view).

Dollár, Welinder et al. (2010) proposed *Cascaded pose regression* (CPR) which is an iterative technique to identify the object pose parameters. In each iteration of this technique, not only is a new regressor learned, but also the image features are assessed from the last predictions. Cao, Wei et al. (2014) have shown that CPR is particularly good for estimating face landmarks. Following on this work, Burgos-Artizzu, Perona et al. (2013) developed a technique to detect facial features in the presence of occlusions, using a model called *Robust Cascaded Pose Regression (RCPR)*. RCPR as an extension of CPR treats occlusions in a more principled way than CPR does. RCPR is able to predict many occlusions on the face by using regression. In this method, the face image is categorized into 3×3 rectilinear uniform blocks. To identify the location of the landmarks within all the blocks, RCPR uses a block without any occlusion. Figure 7 shows the estimated landmark locations by using RCPR method, where red dots show the occluded parts on each face and green dots shows the un-occluded parts. The evaluation shows very good results in regards to occlusions, the source code is also available online as a library.

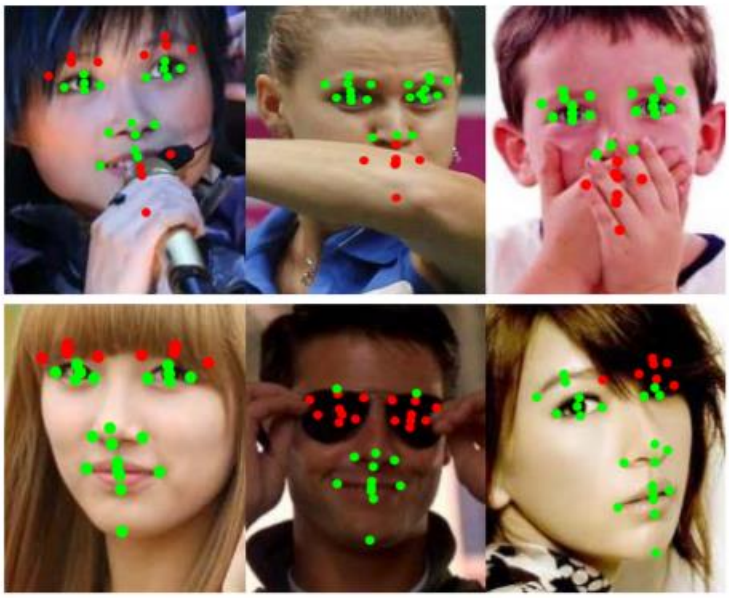

Figure 7. Landmark locations by RCPR method (Burgos-Artizzu, Perona et al., 2013)

2.5.10 *Face Detection using Facial Landmarks*

Computational techniques to detect facial landmarks (i.e., specific points located on the face, such as outer corner of the eyes) are useful for many purposes such as estimation of the head pose (Murphy-Chutorian & Trivedi, 2009), alignment of faces to one another (Kumar, Berg et al., 2009), and tracking of the speech articulators. Detection of facial landmarks is usually accompanied by a bounded box surrounding the detected faces (Yang, Kriegman et al., 2002).

Eight Point Tracker

Uřičář, Franc et al. (2012) have developed a technique to track a set of eight landmarks on the face: the corner of the eyes, the center of the nose, the center of the face, and the

corner of lips (Uřičář, Franc et al., 2012). The facial features allow the extraction of complete contours of some face locations such as eyes, mouth, and nose. The face detector technique processes each frame separately, which means that the temporal continuity of landmark positions is not exploited. Figure 8 shows the bounding box (red rectangle) returned by the face detector; the blue rectangle represents the bounding box used to construct the input to the detector (the normalized image frame).

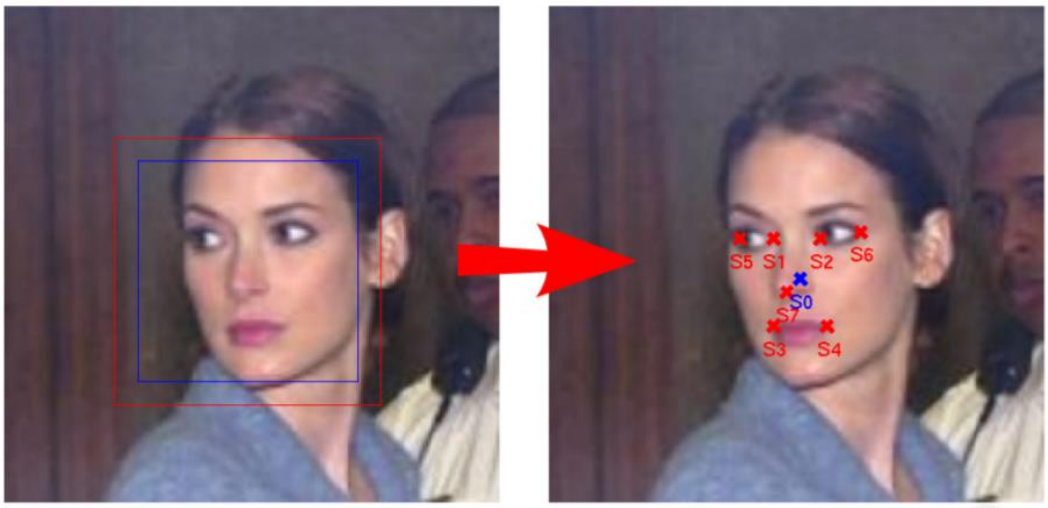

Figure 8. Landmark Detection (Uřičář, Franc et al., 2012)

The face detection algorithm uses a structured output classifier based on the Deformable Part Models (DPM) as follows:

For given image I and M landmark components:

$S = (S_0 \times \cdots \times S_{M-1})$ (1) (Uřičář, Franc et al., 2012)

The variables S and M correspond to the search space among all the small boxes on the face, and landmark location respectively. The location of each landmark and the search space for each landmark is shown in Figure 9 .

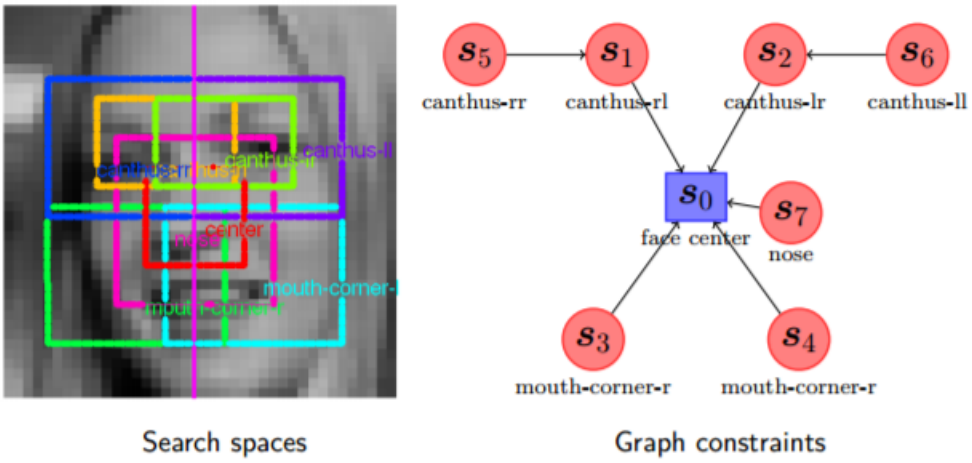

Figure 9. Search Spaces and Graph location of Landmarks (Uřičář, Franc et al., 2012)

The exact formulation of the landmarks derived from the graph constraints is as follows:

$$f(I,S) = \sum i = 0M - 1qi(I,si) + \sum i = 1M - 3gi(s,si) + g5(s1,s5) + g6(s2,s6) + g7(s0,s7)$$ (2) (Uřičář, Franc et al., 2012)

In equation (2) above, the functions qi (I,si) and gi (si,sj) (appearance fit and deformation cost) are parameterized. These parameters are learned from annotated examples using the structured output SVM algorithm.

Canny Edge Detection

Edge detection is an image processing technique intended to recognize the margins of objects within the image. There are numerous types of edge detection such as the Sobel method, the fuzzy logic method, and the canny edge detector method. The Canny edge detector is an algorithm derived by Canny (Canny, 1986), which was devised to come up with optimal edge detection.

The main criteria for this detector are:

- Small error rate, meaning a decent identification of just-existent edges
- Great localization, implying that the distance between the edge pixels recognized
- Insignificant responses, meaning that for each edge, one detector response

Haar Cascade

Face detection has become an active research area for computer vision researchers. This detection has been enhanced regarding speed with the use of Haar cascades and with the impact of some object detection frameworks such as OpenCV (Padilla, Costa Filho et al., 2012). Wilson and Fernandez (2006) have introduced a method for face detection which automatically detects the facial features. As an example, the eye, nose, and the lips are detected based on their position on the face.

Deformable Face Alignment and JFT

Active shape modeling is based on the contours that gets the landmarks and put little patches on different points along the face. Each landmark is detected using the patches among the face on the image (Lucey, Cohn et al., 2010).

Landmarks positions using JFT

FaceTracker[1] is a library for deformable face tracking, originally developed by Jason Saragih, now maintained by Kyle McDonald, and described in several publications (Bashier, Abusham et al., 2013; J. Saragih & Goecke, 2007; J. M. Saragih, Lucey et al.,

[1] https://github.com/kylemcdonald/FaceTracker

2011). For brevity, we use the name Jason Face Tracker (JFT) from this point on to refer to the software library. The JFT software uses a 3D model of the face which locates 66 landmarks on the face, positioning on the eyebrows, eyes, among the face edge, across the nose. The model has different triangles and a stretchable mask to be placed on the face. There are some deformable models and parameters on the face, so not only can arbitrary deformations be detected on the face, but also particular types of deformations, such as the mouth moving. The landmarks from the face model correspond to the same point on the object of the input image. The objects might have some additional decorations on their face such as beards or glasses, and so there may be some difficulties in detecting each patch.

Face detection using JFT

The JFT software uses Haar face detection, previously described, to decompose the image into light regions (cheeks, forehead) and dark regions (eyes, chins, eyebrows, lips). It then looks for a comparison between the light and dark regions. There may be problems tracking faces with darker skin tones. In order to get good results over all skin tones, pre-processing on the image may be needed in the form of histogram equalization. A known limitation with the Haar detector is requirement of a frontal view of the face; so if the face moves, the Haar detector cannot detect it. The JFT software first runs the Haar cascade detector to figure out generally whether something that looks like a face is in the image. From the general idea of where the face could be, the JFT software throws down some initial landmarks and makes some guesses based on the Haar cascade detection.

Given an initial guess for landmarks, the JFT software then identifies relevant features, normalizing and inverting each for the consistency between different skin colors. Once the features of each landmark have been detected, the mean shift algorithm is used to slowly refine the positioning of the landmarks. The positions of the landmarks do not move completely independently of one another. For example, if one eye is closed, it becomes difficult for the JFT software to detect the other eye if open. Another example is the mouth: once the mouth is open, both side of the mouth will be detected either as closed mouth or open mouth. The opening of the mouth exists in one of the states that are either the mouth is open in a certain amount or the mouth is being wider in certain amount. Since this research project focused on visible features of jaw opening, detection of the mouth via JFT will be used intensively in the subsequent sections.

2.6 **Conclusion**

This chapter covered the three main components. First, I provided an overview of PROMPT therapy, a clinical intervention to Childhood Apraxia of Speech which has a series of seven phases of intervention and both in-clinic and at-home components. Among these phases, I identified two phases — phonatory control (stage 2) and mandibular control (stage 3) — as the particular focus for a potential computer based speech therapy (CBST) system.

Second, an analysis of stakeholder needs and requirements was presented. This analysis was contextualized with respect to an overarching design methodology: User

Centered Design (UCD), which is the methodology which frames the actions in this research project.

Third, and finally, a review and synthesis is presented of the prior research that is relevant to the design proposition: the development of an effective and engaging computer-based system to support delivery of PROMPT therapy for Childhood Apraxia of Speech in the home setting. This review included the topics of modes of computer-based speech and language rehabilitation and the potential utility of serious games in computer-based delivery of interventions for speech and language disorders. I presented a summary of techniques for tracking the speech articulators, including techniques based on camera and other types of sensor input. I covered some of the main challenges for camera-based tracking techniques, including occlusions, and low quality images. I also identified and discussed a number of facial feature tracking techniques such as Haar Cascade, Canny Edge Detector, Eight Point Tracker, and Deformable Face Alignment (and the JFT software library). Among all these algorithms, the JFT algorithm had been chosen for further analysis in the next chapter.

Chapter 3
Study 1: Visible Speech Feature Identification from Video

In this chapter, I present the first of two studies conducted for this research project. This chapter concerns the phase concerned with the identification of visible features of speech from a video corpus, the development of relevant computational techniques, and the evaluation of the most promising technique. First, I start by describing the preparation of the video corpus and how I collected the data (section 3.3). Next, I describe the development of several different computational techniques (section 3.4), and last, I describe the evaluation of the most promising computational technique using a fidelity study (section 3.6).

3.1 **Objectives**

The first objective is to define an inventory of visible facial features of speech articulation that are relevant to the PROMPT program of therapy — those that have the potential to have a high degree of correlation with jaw opening and/or jaw sliding. The second objective is to identify a slate of ready-made, camera-based, face tracking techniques and libraries that could potentially be the basis for the automatic identification of the facial features in the PROMPT inventory and, from among these, to identify the most promising technique. The third objective is to perform fidelity study of the selected technique.

3.2 **Methodology**

To address the first objective, which is to identify an inventory of relevant visible facial features, I will first prepare a video corpus which records participants articulating relevant speech stimuli items. Next, I will use that corpus to develop the inventory of visible facial features of interest, using the PROMPT therapy as the basis.

To address the second objective, which is to develop a set of computational techniques, I will develop a set of pilot computational techniques, using the video corpus from the previous step to test and to evaluate them. I will make use of existing libraries for facial feature tracking, and I will use an exploratory approach to develop several techniques iteratively, drawing on a variety of sources. From these, I will choose the most promising technique to be evaluated in more detail in a fidelity study.

To address the third objective, which is to perform a feasibility study, I will set up an observational study to compare the features extracted via two methods: a ground truth technique and the most promising technique from the previous step.

3.3 **Preparation of Video Corpus**

In order to study facial features during speech and to have a set of testing materials, I developed a corpus of video segments.

3.3.1 *Data collection*

This data was collected under HPRC Certificate "Computer Games Using Automatic Speech Recognition, Video and Interactive Toys" #2010 - 090 (Approval period 08/02/13-08/02/14).

3.3.2 *Subjects*

The subjects participated (one male, aged 9 y.o., and one female, aged 7 y.o.), both healthy children with normal speech and hearing. The participants are representative of the target age group of children who typically receive PROMPT treatment (between the ages of 4-9), but did not have diagnoses of Childhood Apraxia of Speech. The reason for this choice of participants was the need to develop the baseline for which other children with CAS strive to achieve. No financial incentives were given for participation in the experiments. Participants were de-identified and assigned identifiers 001 and 002.

3.3.3 *Stimuli*

In consultation with a speech language pathologist who is also a PROMPT practitioner, we created a stimuli set of consonant-vowel-consonant (CVC) segments to provide coverage across 5 different vowel classes:

High-Front Vowel /i/ CVC segments (Level 1): Bead, Beep, Peep, Pin, Bin

Mid-Front Vowel /e/ CVC segments (Level 2): Bed, Head, Ten, Pen, Hen

Low-Front Vowel CVC /ae/ segments (Level 3): Bat, Dad, Hat, Tap, Mat

High Back Vowel /u/ CVC segments (Level 4): Boom, Boot, Food, Hoop, Moon

Low Back Vowel /a/ CVC segments (Level 5&6): Mom, Bob, Hop, Pot, Top

Each word list corresponds to a vowel class that has a class-specific jaw position, as identified via the PROMPT therapy guidelines. Thus, this set of words will generate the full range of jaw positions that will be targeted in treatment. The particular words in each list were developed by the collaborating speech-language pathologist.

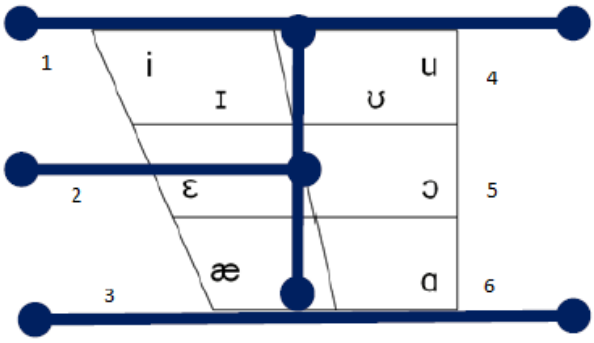

Figure 10. Mapping of vowel place of articulation to vowel classes

A vowel sound is created when air passes through the vocal tract. One of the features that make the vowel sounds different from one other is the position of the tongue in the mouth. Figure 10 presents a schematic illustrating the scope of vowel articulation and the respective location of the back of the tongue being raised relative to the soft palate, which is driven, in part, by the relative jaw opening (Low, Mid, High positions).

Additionally, the back of the tongue can adopt a frontal or back position, creating classes of vowels that are frontal and back.

The schematic shows placement of the tongue for the different vowels. At the left of the chart are the frontal vowels — those pronounced with the back of the tongue towards the front of the mouth. The vowels along the right of the chart are pronounced with the back of the tongue pulled back in the mouth. The vertical dimension of the chart refers to how open or closed the opening between the back of the tongue and the pharynx is during articulation of a vowel. This opening is driven, to a large extent, by the degree of jaw opening. Therefore, the vowels at the top of the chart are pronounced with the jaw in a more closed position, whereas the vowels towards pronounced with the jaw in a more open position.

The Low, Mid, and High levels are distributed over the frontal and back vowels and each of them represents the relative distance between a lower and upper jaw. By and large, the Low level corresponds to when talkers slightly open their mouth. The Mid-level corresponds to when talkers open their mouth a bit more than the low level. Lastly, the High level corresponds to when talkers widely open their mouth.

3.3.4 *Elicitation software*

In order to present the stimuli to the study participants, elicitation software was developed for this research project. The software presents a GUI which displays the stimuli words, one by one, and drawn from the defined stimuli set.

The software draws the words randomly, as the subject should not be able to anticipate the next word and to avoid effects arising from grouping all the stimuli words from a given vowel class together. A screenshot of the software in operation can be seen in Figure 11.

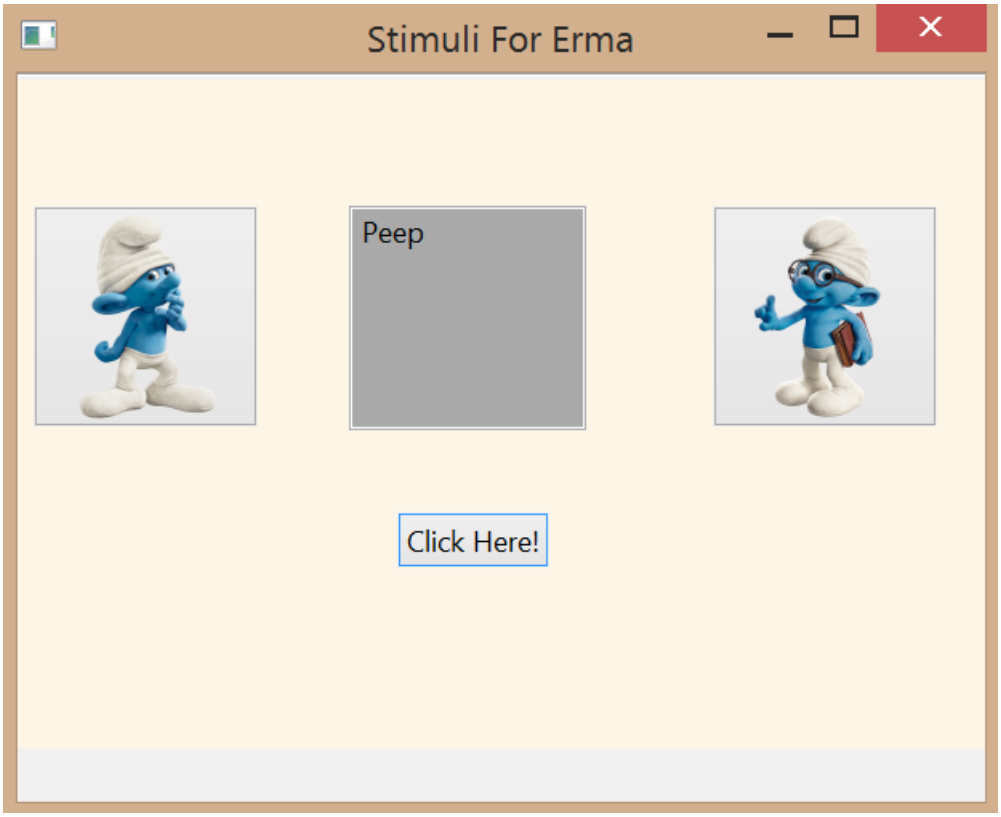

Figure 11. Sample screenshot of the program that was developed to display stimuli words to the study participants.

3.3.5 *Procedure*

For each data collection trial, the participant was seated comfortably and the procedure was explained. The elicitation software was demonstrated and the participant performed a few practise runs. Participants were recorded using an iPhone6 camera. The elicitation

software then presented each of the 125 stimulus words (25 words with 5 repetitions each) in a randomized sequence.

3.3.6 *Video Post-Processing*

The video file from each data collection trial was subsequently post-processed. I first viewed each file and identified the onset time for each of the individual CVC segments. I prepared a shell script that takes the onset timestamp file and the video file as inputs and then segments the video into smaller, CVC-specific individual files.

We employed a specialized naming scheme for each video segment, based on the method below:

<Subject-ID>-<Date>-<Level>-<Type>-<Stimulus Word>-<Repetition number>.avi

For example, the file 001-20151610-1-HF-Bead-1.avi corresponds to a video segment of subject 001 articulating the stimulus CVC segment “Bead” on 16th of October 2015, where the CVC segment belongs to the level 1 vowel class.

For example, for the dataset for subject 001, each segment is approximately 5 seconds in duration, for a total of 76 segments. We expected 125 segments, but 49 segments were missing due to the fact that the subject got bored and distracted during the data collection.

3.4 **Inventory of Relevant Visible Facial Features**

The main tasks in identifying an inventory of visible facial features that are relevant to PROMPT is to be able to characterize, on the basis of a frontal view of the face, the degree of jaw opening and the degree of jaw 'sliding' (lateral movements from left to right).

One of the clinical targets of PROMPT is getting the child to calibrate his/her degree of jaw opening correctly. Jaw opening is controlled by a hinge-type joint, so the degree of opening could, in principle, be measured directly, if the joint could be instrumented directly. In place of this, we instead look to features of jaw opening that can be ascertained from a frontal view of the face, specifically using relatively low-quality video from a commodity mobile-device camera. Another clinical target of PROMPT is to gain appropriate control of lateral jaw movements, specifically to eliminate jaw sliding. Jaw opening and jaw sliding are discussed in section 3.4.1 and 3.4.2, respectively.

3.4.1 ***Jaw Opening***

Jaw opening can be directly derived by calculating the angle of the temporomandibular joint, but the camera does not have access to this view. Instead, we need to derive measures that correlate with jaw opening and that can be obtained from a frontal view of the face. Five measures were identified as potential useful indicators of jaw opening.

These measures are based on different facial features specifically located on the articulators (jaw, lip, tongue, and mouth), as described in section 2.5.1 of this thesis. The

objective measurements could be derived on a frame-by-frame basis, and these measurements subsequently will be used as features in different types of analyses.

Figure 12. Jaw opening features: (a) d_0, Nose-Chin Distance, with points p_1 and p_2 shown; (b) d_1, Outer lip Distance, with points p_3 and p_4 shown; (c) d_2, Inner lip Distance; (d) d_3, Lip Corner Distance; and (e) d_4, Speaker's "mouth roundness"

(a) Nose-Chin Distance (d_0)

The distance between the tip of the nose and the chin was identified as a potential measure of jaw opening. To measure this feature d_0, two anchor points, p_1 and p_2, need to be located on the middle of the nose and the chin, respectively. We expect this distance to be highly correlated with jaw opening: as the jaw opens, the d_0 value should get bigger.

(b) Outer lip Distance (d_1)

Another feature for measuring the degree of jaw opening was identified as the vertical distance between the top of the upper lip margin and the bottom of the lower lip margin. In order to measure this distance d_1, two anchor points, p_3 and p_4, need to be located: one in the center and on top of the vermilion border of the upper lip, and the other in the center and on the vermilion border of lower lip. As the jaw opens, the mouth opens more and thus d_1 gets bigger; d_1 correlates positively with jaw opening.

(c) Inner lip Distance (d_2)

The vertical distance between the bottom of the upper lip and the top of the lower lip is another feature that was identified for calculating the jaw opening. As the jaw opens, d_2 is expected to get bigger. Therefore, d_2 correlates systematically with the degree of jaw opening.

(d) Lip Corner Distance (d_3)

The distance between the innermost corner of the left side of the mouth to the innermost corner of the right side of the mouth (the oral commissures) is another potential feature

for jaw opening. To calculate d_3, two anchor points need to be located; one in the right corner and another one in the left corner of the lip. As the jaw opens, the lips corners often move inward. Therefore, this measure d_3 is expected to be negatively correlated with degree of jaw opening.

(e) Speaker's "mouth roundness" (d_4)

A parameterized characterization of the speaker's "mouth roundness" was developed as another measure of degree of jaw opening. When the jaw is closed, the mouth opening is flattened and wide, whereas when the jaw opens, the mouth gets rounder. We expect "mouth roundness" to be positively correlated with degree of jaw opening.

To characterize mouth roundness, the points that correspond to the inner outline of the lips should first be located. Then an ellipse is fitted to these outline points.

To fit an ellipse on the points, the following equation (3) was used:

$$\frac{\left((x-h)\cos(A)+(y-k)\sin(A)\right)^2}{(a^2)} + \frac{((x-h)\sin(A)-(y-k)\cos(A))^2}{(b^2)} = 1 \quad (3)$$

Values for the variables are chosen such to minimize RMSE to the outline points. The values (h, k) correspond to the center of the ellipse, and a, b are the major and minor axis length of the ellipse respectively. The value A captures the rotation of the ellipse. A rotated boundary box is then overlaid for display purposes. This boundary box is shown in Figure 12 (e).

To derive d_4, which is a measure of the "roundness" of the ellipse, the ratio between the minor and major axis was computed. A preliminary analysis revealed a mean ratio of approximately 2:1 overall the frames, with a higher ratio when a vowel is being pronounced and a lower ratio when the mouth is closed. Consequently, the ratio expresses that the rounder the mouth is, the larger the jaw opening.

For the first iteration of this project, we looked at a large number of features. Since this work features would be undertaken iteratively, we expected that, in the subsequent stages, we might add or remove features from this initial set.

3.4.2 *Jaw Sliding*

We developed a measure to characterize the degree of jaw sliding. It is based on calculation of angles between the following two lines:

- u = a straight line starting from the bridge of the nose to the tip of the nose (nose-bridge line)
- v = an oblique line that passes through two points: mid of bottom of chin, and tip of the nose

We calculate the angle between those two lines on the participant face using equation (4):

$$\cos a = \frac{u_1.v_1 + u_2.v_2}{\sqrt{{u_1}^2+{u_2}^2} \cdot \sqrt{{v_1}^2+{v_2}^2}} \quad (4)$$

We convert the result from radians to degrees.

This measure is demonstrated in Figure 13, which shows a frame of a video of a young girl speaking and demonstrating jaw sliding.

Figure 13. A frame taken from a video of a young girl with CAS (UrbanKowboy, 2010). In this frame, her jaw is "sliding" during articulation. Blue lines have been added to illustrate the degree of the jaw asymmetry.

3.4.3 *Measures that are Invariant to Facial Expression*

All of these measurements are derived using 2D images. The values depend on the extent to which the subject's face fills the frame and the degree to which the subject faces the camera directly. Since subjects often shift from frame to frame, we sought to normalize these measures. For this, we sought to identify pairs of points on the face whose distance would be invariant to facial expression. We did not deal with the case when the subject rotates their face relative to the camera.

We identified four pairs of points: Figure 14 shows the invariant measurments.

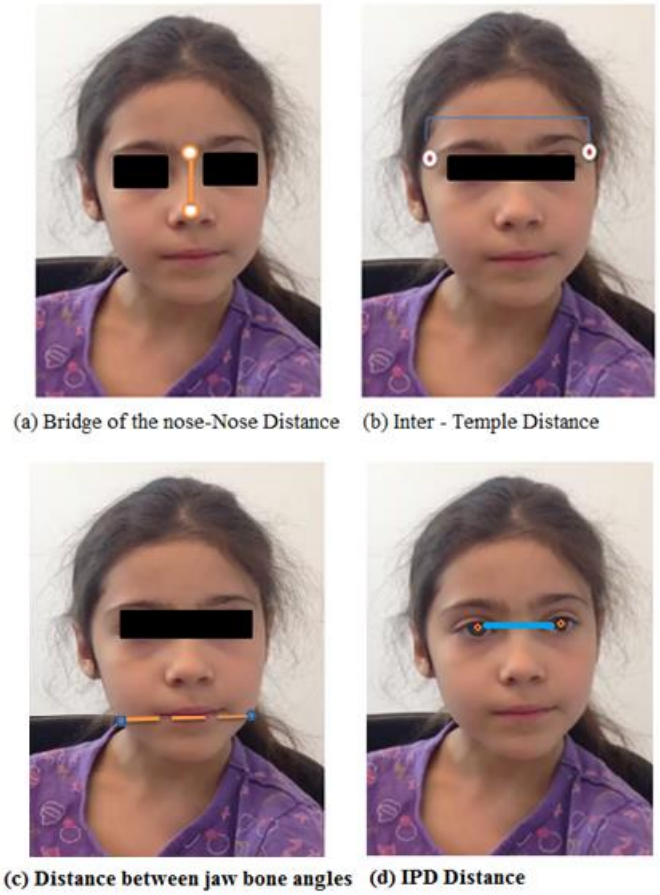

Figure 14. Invariant measurements

(a) Bridge of the nose-Nose Distance

A pair of points located on the bridge of the nose is expected to stay at the same fixed distance during changes in facial expression, with perhaps some small changes with extreme face "scrunching". Thus, this was used as an invariant feature.

(b) Inter-Temple Distance

The distance between two points, one located on each temple, was identified as invariant. The temples indeed move as a function of head poses; however, the distance between the temples is expected to be invariant to facial expression.

(C) Distance between Points Located on the Jaw Bone Angles

Another distance measurement that is expected to be invariant to facial expression is the distance across the jaw. To measure this, the distance is derived between two points located on each of the jaw angles (right and left side).

(d) Interpupillary Distance (IPD)

A distance that is expected to be speech invariant is interpupillary distance (IPD). IPD in a participant can be considered a fixed measure, since a person's eyes stay relatively fixed in their head, and eyes movement is yoked for fixed focal distances.

3.5 Computational Technique Development

3.5.1 *Techniques Implemented*

In this section, four different techniques are described: Haar Cascade, Canny Edge Detector, Eight Point Tracker, and the Jason Face Tracker. I instantiated and performed a preliminary evaluation of each technique.

The face tracker techniques, in general, face challenges such as:

- Obstructions of the face (by hands, etc.)
- Changing distance from camera to the face (i.e., moving back and forth)
- Head swiveling
- Subject moving out of frame

3.5.1.1 Requirements

As per the requirements analysis, which was described in chapter 2, the computational techniques should be based on input from a single, consumer-grade camera that is built into a tablet device and should have the ability to extract features of interest. The technique should function reasonably well under adverse conditions, such as situation when the user is moving around within the frame. As well, the computational technique should reasonably run on CPU/GPU of the tablet device.

3.5.1.2 Haar Cascade

I employed an app that made use of the Haar Cascade algorithm. I used a Haar Feature-based Cascade classifier which has been deployed within the OpenCV platform. The app takes the entire set of frames from a video segment and output the same set of frames, but with the eye and face detected. Figure 15 shows a sample result from face and eye detection on one of the segments from the video corpus.

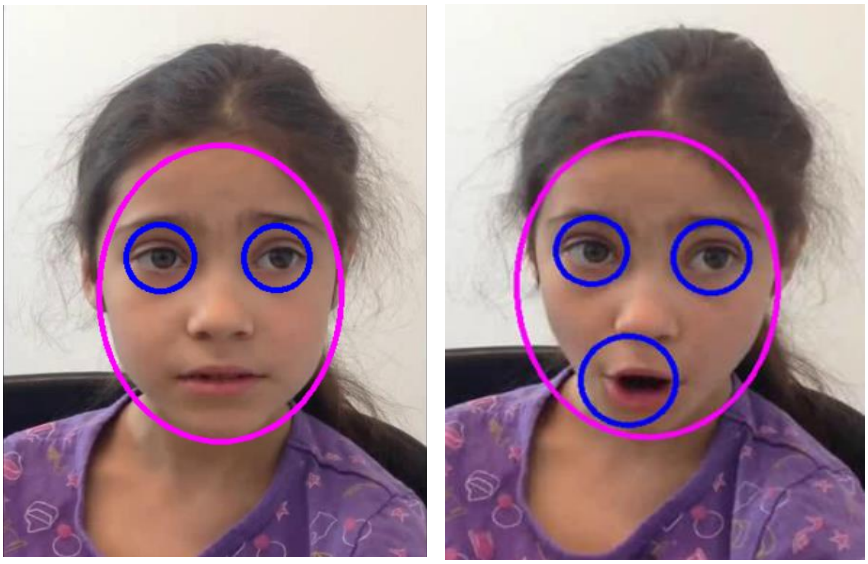

Figure 15. Sample use of Haar feature-based facial cascade classifier to determine and to extract the face and eyes from two images

The Haar Cascade algorithm performed very well for eye and face detection. However, locating the target points on the chin presented a challenge that would require subsequent work. Before undertaking this task, I instead explored the other techniques.

3.5.1.3 Canny Edge Detector

I implemented an app that made use of the Canny Edge Detection algorithm; an open source library (OpenCV) was used and modified for this exploration. Figure 16 shows a sample output (a frame of video is taken from one of the video segments from the corpus).

This technique provided strong results in terms of detecting the edges; further development would be needed to derive the particular features from the edges. Before undertaking this task, I decided to pursue other techniques.

Figure 16. Detection with the canny edge detector algorithm

3.5.1.4 Eight-Point Tracker via Landmarks

I implemented an app that detects eight points on the face: two points for the corner of each eye (4 eye corner points in total), two points for detecting the corner of the lips, one point to detect the middle of the face, and finally the last point to detect the middle of the nose. The app made use of the OpenCV libraries for the detection process.

To augment the OpenCV library I developed additional functionality to extract the distance, roundness, and angular displacement features described earlier in this chapter. A sample output is shown in Figure 17, which shows the positions of the landmarks on the face. The red rectangle is bounding box returned by the face detector, the blue rectangle represents the bounding box used to construct the input to detector (the normalized image frame).

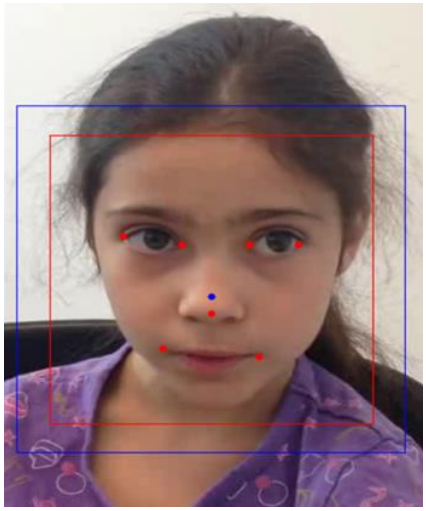

Figure 17. Sample of Landmarks which were chosen on the face

With the help of the landmarks, I augmented the software to then find the relative position of the chin using an arced line segment and then the middle of the chin. The identification of the mid-chin point is based on the two inner points of the eyes and the bridge of nose location. With the help of the distance between the center of the face and the bridge of the nose, the same measurements were allocated to the face starting from the center of the nose to the mid-chin location. Thus, this defines a “nose-to-chin line” for each of the participants. Figure 18 shows these results.

Figure 18. The defined chin and chin line based on the different landmarks which were set

3.5.1.5 JFT (Jason Face Tracker)

I created an app to use the Jason Face Tracker (JFT) algorithm, which is available as an open source library. The JFT makes use of active shape modeling as described in section 2.5.6. Figure 19 shows the JFT algorithm output, given an input video frame from the video corpus. Based on an ad hoc examination of the video segments, I determined that this technique performed the best in regards to chin detection, and I continued developing different feature extractions using this technique.

Figure 19. JFT algorithm on participants

3.5.1.6 Methodological Challenges

For any facial feature tracking technique, one must assess the possibility of error. Possible source of error includes:

i. Error in the facial feature point identification (deviation from the ground truth)

ii. Error in the invariant distance calculation. The distance measurements are based on the assumption that the points co-occurred on a Cartesian plane that is parallel to the image plane. However, the facial feature points are, in actuality, moving in a 3D-space and are projected from 3D onto the 2D plane of the image field. A participant's head movement may move in the sagittal plane, or their head pose may swivel (so that the paired points do not remain at the same distance from the camera). This projection is subject to distortion and thus can be considered a source of error. Resultant distances are derived in pixel units.

iii. Error in the distance assumption. The distance between the camera and the person may change if the person moves around in their seat. This also impacts the distance measures. If the true distance between the camera and the person is known, on a frame-by-frame basis, then the resultant distances in pixel units can be transformed into mm.

Error of type (ii) must be accepted as a limitation of this technique in general. However, error of type (iii), if significant, can be mitigated through normalization.

In order to assess the degree of noise due to error (iii), we examined the degree to which the invariant measures (see section 3.4.3) changed over the duration of the video segment.

A histogram of a large-set of IPD values are provided below to illustrate the examination. These values were derived on a frame-by-frame basis. Any changes in IPD over the duration of a video segment can be assumed to be due to error of type (i) or (ii). The histogram, shown in Figure 20, demonstrates the distribution of values for the IPD values over all of the frames.

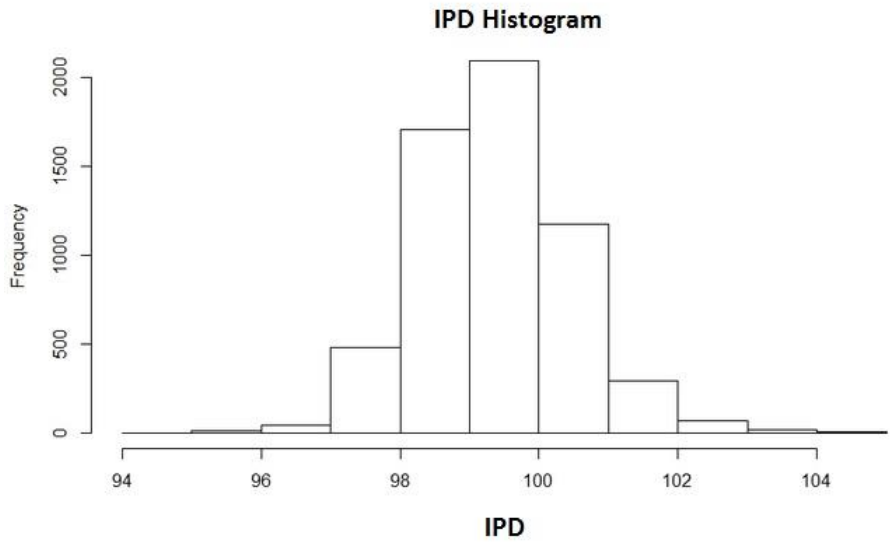

Figure 20. IPD Histogram

We expect this variation to be due both to error in pupil tracking and to changes in the subject's distance to the camera. In order to mitigate the factor of subject motion, normalization can be performed. To do so, we derive a frame-specific scaling factor using the following equation:

$$\text{scaling for frame } f_i = \frac{actual\ IPD\ (mm) \times 100}{\text{observed IPD (pixel) for f}_i}$$

Using this ratio, we can convert each distance that is computed relative to the image frame (in pixels) into millimeter units.

3.5.1.7 Discussion

Each of the four computational techniques demonstrates strengths and weaknesses. Although the Canny Edge Detector provided strong results in terms of detecting edges, it did not directly locate the particular features of interest and does not represent a solution

(not, at least, without further work). The Haar Cascade algorithm performed very well for eye and face detection. However, locating the chin points presents challenges for this technique. Given the particular importance of the chin for PROMPT therapy, the Haar Cascade algorithm was found to be lacking.

Both the JFT and Eight Point Tracker performed well in regards to detecting the eyes and nose, however, the JFT performed better than the Eight Point Tracker with regard to locating the chin points. Figure 21shows a typical illustration of the differences between the two algorithm (a) Eight point tracker, and (b) Jason Face tracker algorithm.

Figure 21. (a) Eight Point Tracker (b) Jason Face Tracker

3.6 **Fidelity Study**

3.6.1 ***Objectives***

We decided to focus on the JFT technique and to evaluate it more specifically.

Among all the features described in section 3.8, we focus specifically on degree of jaw opening and will focus on the following measures specifically: (1) d_0, the distance between the philtrum to the chin and (2) d_1, the vertical distance between the top of the upper lip margin and the bottom of the lower lip margin. The objective of the fidelity study is to answer the following research questions:

How does the measurement of the d_0 and d_1 distances via the JFT differ from the ground truth measurement of d_0 and d_1?

Which of d_0 and d_1 is measured most accurately using the JFT technique?

To what extent is JFT tracking accuracy related to the different vowel classes?

3.6.2 ***Methodology***

To address the first objective, I will compare, using a representative sample of video, the measurement of the d_0 and d_1 distances via JFT to a "ground truth" measurement of these distances. This will entail two analyses (1) a scaled, frame-by-frame comparison of the JFT vs "ground truth" measurement, (2) a segment-by-segment comparison of the JFT vs "ground truth" measurement (to be based on the particularly crucial frames of each segment). To address the second objective, I will compare the accuracy of the d_0 measurement to the d_1 measurement over all of the frames. To address the third objective,

I will perform a statistical analysis of the dependent variable of error using a one way ANOVA, to examine the impact, if any, of vowel class on tracking inaccuracy.

3.6.2.1 Ground Truth: Wave System

To establish the "ground truth" of the facial features, we employed the WAVE system (Northern Digital Inc., Canada), which is an electromagnetic articulography system that tracks the kinematics movements and provides three dimensional (3D) tracking of 5 or 6 degree-of-freedom (DOF) sensors. This system consists of a hardware and software components, including an electromagnetic field generator, 2mm-sized sensors, a microphone, and a standard PC for data collection.

This particular Electromagnetic Articulograph (EMA) system operates by generating a calibrated electromagnetic field within a 50cmx50cmx50cm volume. Participants are then fitted with sensors (attached to the face and to the speech articulators) and seated with the sensors within the field. The EMA system then decodes positional sensor movements by interpreting perturbations in the EMA field. Through this technique, movements of the face and speech articulators can be tracked. The Wave system samples at maximum 400Hz within 0.5 mm of accuracy (Berry, 2011).

The Wave system requires that sensors be attached on the points of the participant's face where the data is to be collected. The data in the Wave system is collected through a desktop computer (PC). The PC in our data collection study had the following features:

- Microsoft Windows 7 Professional
- Intel i3-2100 CPU @ 3100MHzi3-2100 CPU @ 3100MHz
- 4 Processors
- 4.00 GB of RAM

The Wave system also includes the WaveFront software, which performs head correction and ensures all the data points are timestamped and logged.

3.6.2.2 Sensor Placement

For derivation of the ground truth, a set of 7 sensors were positioned as follows:

- The first sensor (S_1), which is a six degree-of-freedom reference sensor, was attached on what is meant to be a head-static position --- a headband which is positioned on the participant's forehead. This sensor is subsequently used as the basis for head correction.
- Two sensors (S_2, S_3) were attached on the participant's temples. These and the sensors described below were attached using medical-grade adhesive tape, applied to the surface of the participants' skin.
- Jaw movements were captured by attaching two sensors (S_4, S_5) to either side of the participant's jaw bone.
- Finally, the last two sensors (S_6, S_7) were attached under the nose (philtrum) and on a participant's chin respectively. The locations did not correspond precisely to the points (p_1, p_2, p_3, p_4) as identified in section 3.4.1.

The data collected via the sensors S_2-S_6 was head-corrected relative to sensor S_1. Placement of the 7 sensors on the participant's face can be seen in Figure 22. It is not practical to position sensors on the lips due to interference effects with speech.

Figure 22. Sensor attachment on the participant's face

3.6.2.3 Stimuli

The stimuli set of 25 CVC segments described in section 3.3.3 was used, which provides coverage across 5 different vowel classes. Five repetitions were elicited for each stimulus, for a total of 125 stimuli items.

3.6.2.4 Data Collection

WAVE data collection using the WaveFront software was undertaken for 1 participant in 1 session of 30 minutes. The participant was drawn from the Toronto Rehabilitation Institute volunteer pool and the data was collected under a certificate issued by the University Health Network Research Ethics Board (Certificate: 13-6235-DE Visual

Feedback Systems in Speech Rehabilitation). After the sensor placement, the participant was asked to articulate each stimulus item, each of which was shown using the elicitation software, at their comfortable zone and usual loudness.

Acoustic data was collected via a laboratory-grade microphone (Crown head-worn microphone CM311), attached on the participant's scarf about 20cm from the mouth, and connected to and calibrated with the WaveFront software to ensure kinematic-acoustic alignment. In addition, a camera rig was set up and audio-video data was collected for the same duration.

The camera rig was selected to be representative of the design specification, as per the requirements identified in section 2.3. An iPhone6 camera with the following features was used:

- 8-megapixel iSight camera with 1.5µ pixels, collecting video with frame width and height of 1280×720 pixels, and frame rate of 30 fps, 1080p HD video recording, and audio sampling via AAC (8 to 320 Kbps).

We deliberately employed a low-quality video recording device (mobile phone), since this best represents the type of input in the intended clinical application.

The start time of data collection via the WAVE system and the start time of the iPhone video recording were not synchronized and thus alignment was performed in the data preparation stage.

Last, the participant's interpupillary distance (IPD), the distance between the centers of the pupils, was measured using a ruler and an established IPD measurement procedure (MacLachlan & Howland, 2002).

Pilot Study

In order to ensure our data collection procedure functioned as intended, we ran a pilot study. Through the pilot study, we learned that sometimes the adhesive tape comes loose and results in data loss. From this, we revised our sensor attachment process to use more adhesive tape to ensure a more secure site of fastening.

3.6.3 ***Data Preparation***

Once data collection was complete, we prepared the data from each of the two sources: the WAVE system and the iPhone camera.

We used the NDI WaveFront export function to produce two files: the kinematic and the acoustic data files:

[Original filename].wav

[Participant ID Number] _data_log.csv

The first file contains the audio of the participant's speech, time-aligned to the kinematic data file. The second file is a csv-formatted file contained the kinematic data, containing time-stamped positional and orientation data for each of the 7 sensors. As per common practise in speech laboratories, movement with respect to the least significant axis (the WAVE system's x axis, the sagittal axis) was discarded for dimensionality reduction (Roweis & Alwan, 1997). The coordinates from the transverse and vertical axes were

retained (corresponding to the transverse plane, which is composed by the y and z coordinates under the WAVE system). A total 74612 samples of kinematic was collected (corresponding to 3 minutes 13 seconds, collected at 100Hz)

From the iPhone, we collected 5810 frames of video (corresponding to 3 minutes and 3 seconds of video, which was collected at 30 frames/second). There was an approximately 10-second delay between the start of the WAVE data collection and the start of the video data collection.

3.6.3.1 Alignment

As described earlier, the start time of the WAVE and iPhone data collection was not synchronized and there approximately 10-second delay between their start times. Thus an alignment step was performed.

For this, we examined the audio track from the iPhone video collection and the audio file derived via the WAVE system. The offset was derived using (i) the transient detection tool in ProTools (Avid Inc, Burlington, Massachusetts) to detect the speech onset in each of the two audio samples and then (ii) using the selection tool to derive the exact number of samples between the two onset times. The corresponding prefix of samples was removed from the kinematic data file in order to align it with the video data file.

A small sample of the time-aligned WAVE and JFT distances is shown in Figure 23.

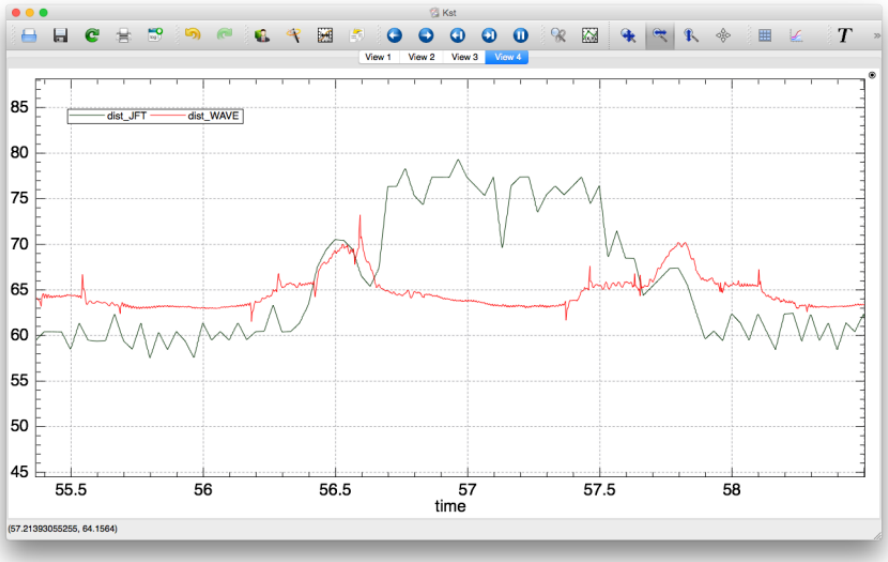

Figure 23. Distance_JFT VS Distance_WAVE

3.6.3.2 Tracking-Based Distances

In order to identify the facial features, I ran the entire video file through the JFT algorithm. The (x, y) values of each tracked feature were derived on a per-frame basis and captured in a tsv text file. Each frame was annotated with frame count and the feature locations. The video's audio track was lost in this process. The resultant set of output frames (a set of 5810 image files in total) was restitched into a video file. For this, the ffmpeg command line utility for transcoding video and audio files was used (Tomar, 2006). By using ffmpeg, all the image files were combined into a video file with the same frame rate and duration as the original video corpus frame rate.

The next step was to extract the audio track from the original video corpus and then to add it back into the processed video.

The feature-annotated video was then segmented on a word-by-word basis. To calculate the pixel-based distances from the facial features, for each measurement, the distance between the relevant feature points was calculated.

A second phase of derivation was performed. The maximum values on a per-segment basis were derived.

3.6.3.3 Normalization

Frame-specific scaling factors were derived as described in section 3.4.3. The subject's IPD was derived in pixels for each frame, and then a scaling factor was derived by computing the ratio of pixel IPD to the known IPD (66mm). This scaling factor was applied to all the derived distances.

3.6.3.4 Ground-Truth Distances

The ground truth of the d_0 is defined as follows:

d_0_WAVE: ground truth of the philtrum to chin distance

The d_0_WAVE value was calculated for each kinematic sample.

It was not possible to position sensors on the lips and so we do not have a d_1_WAVE ground truth.

3.6.3.5 Difference between Tracked and Ground Truth Distances

Since the sampling rates of the JFT and WAVE-derived data were not the same, a direct sample-by-sample comparison could not be performed immediately and resampling was needed.

An R script was used to perform down-sampling of the WAVE data, producing an interpolated kinematic sample for each video frame. Another R script was used to perform up-sampling of the JFT tracked-point data file, producing an interpolated tracked sample for each kinematic sample. I decided to use the downsampled WAVE data as the basis for the sample-by-sample differences.

For this aim, the difference between the Wave and the JFT data was derived on a frame-by-frame basis for both of the features of interest: philtrum to chin distance (1) and outer lip distance (2). The comparison was obtained from the following equations:

$$\text{Diff\_}d_0 = (d_0\_Wave\_mm) - (d_0\_JFT\_mm) \quad (1)$$

$$\text{Diff\_}d_1 = (d_0\_Wave\_mm) - (d_1\_JFT\_mm) \quad (2)$$

3.6.3.6 Per-Segment Features

A segmented version of the kinematic data file was prepared (each segment containing an individual CVC segment). The audio file was used as the basis for manually labelling the onset and offset timestamps of each segment. The onset and offset timestamps were recorded in a tab-delimited text file. Then I wrote a MATLAB script that would accept the tab-delimited file and use it as the basis for segmenting the kinematic data file into the

set of smaller, segment-specific segment kinematic files. As a result, 125 CVC segments were obtained from the data.

For each segment, a peak_ d_0 value and a peak_d_1 value were obtained.

In terms of fidelity measures, two measurements were calculated as follows:

Diff_max_d_0 and Diff_max_d_1: locate the sample demonstrating the maximum distance value within a given WAVE segment, and then find the deviation between WAVE and JFT distance (for both d_0 and d_1) for that sample.

Diff_mean_d_0 and Diff_mean_d_1: determine the deviation between WAVE and JFT distance (for both d_0 and d_1) for each sample and derive the mean over those values.

3.6.4 *Results and Data Analysis*

3.6.4.1 Deviation from ground truth

Since the sensors were not placed exactly in the same position as the points tracked via the JFT technique, a direct comparison could not be made. Deviation from the ground truth was calculated in two ways.

Technique (1) for deriving the deviation from ground truth entails the application of a scaling factor to the JFT values, so the range of the JFT values get rescaled into the interval that matches the ground truth.

Using this approach, we define Diff_d_0_S to be the frame-specific difference between the d_0 distance, as derived via the JFT technique, and the scaled d_0_WAVE

distance, as derived via the WAVE system data. The mean value, over all the frames, of Diff_d_0_S was calculated as 4.69 mm.

We define Diff_d_1_S to be the frame-specific difference between the d_1 distance, as derived via the JFT technique, and the scaled d_0_WAVE distance, as derived via the WAVE system data. The mean value, over all the frames, of Diff_d_1_S was calculated as 2.49 mm.

Technique (2) for deriving the deviation from ground truth entails the use of the raw values. Using this approach, sample-by-sample differences of Diff_d_0 and Diff_d_1 were calculated. Diff_d_0 is calculated per frame by calculating the difference between the d_0 distance, as derived via the JFT technique and the d_0_WAVE distance, as derived via the WAVE system. The mean difference for d_0 calculated as -21.26mm.

Diff_d_1 is calculated per frame by calculating the difference between the d_1 distance, as derived via the JFT technique and the d_0_WAVE distance, as derived via the WAVE system. The mean difference for d_1 calculated as 42.0335 mm.

As well, mean_d_0, max_d_0, mean_d_1, and max_d_1 values were derived on a per-segment basis. The vales of each were determined to be:

mean_d_0 = 84.52mm

max_d_0 = 99.43mm,

mean_d_1 = 21.23mm, and

max_d_1 = 36.19mm.

The difference between the two techniques for deriving difference indicates that scaling should be used when deriving fidelity measures. To investigate further, I examined the histograms of the values derived using the raw values.

3.6.4.2 Performance comparison

Differences between features of interest in JFT vs WAVE

In order to determine which of d_0 and d_1 is measured most accurately using the JFT technique, I compared the histogram of the Diff_d_0 and the Diff_d_1 values, as seen in Figure 24 and Figure 25.

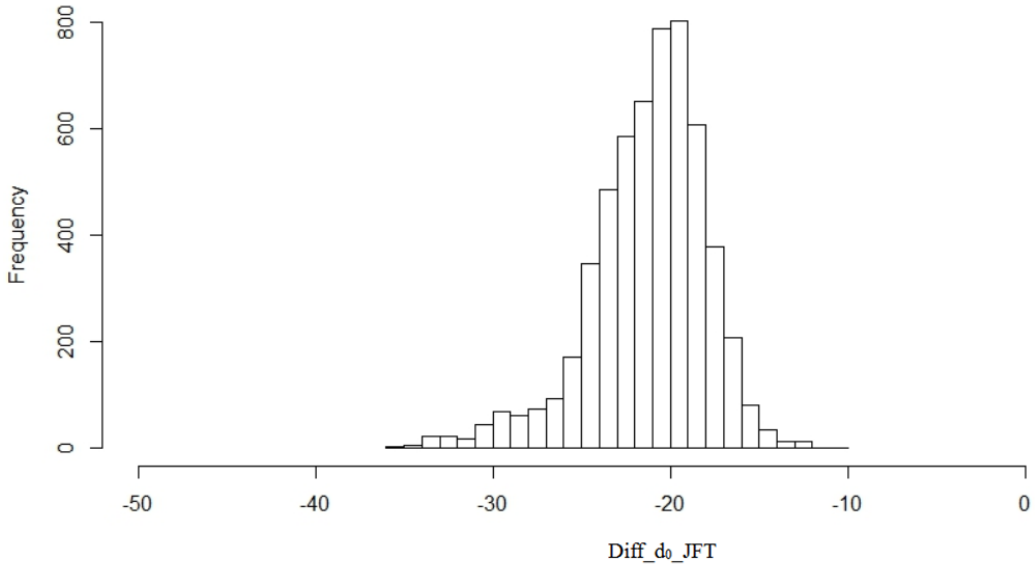

Figure 24. Histogram of the difference of Philtrum to chin distance (Diff_d_0)

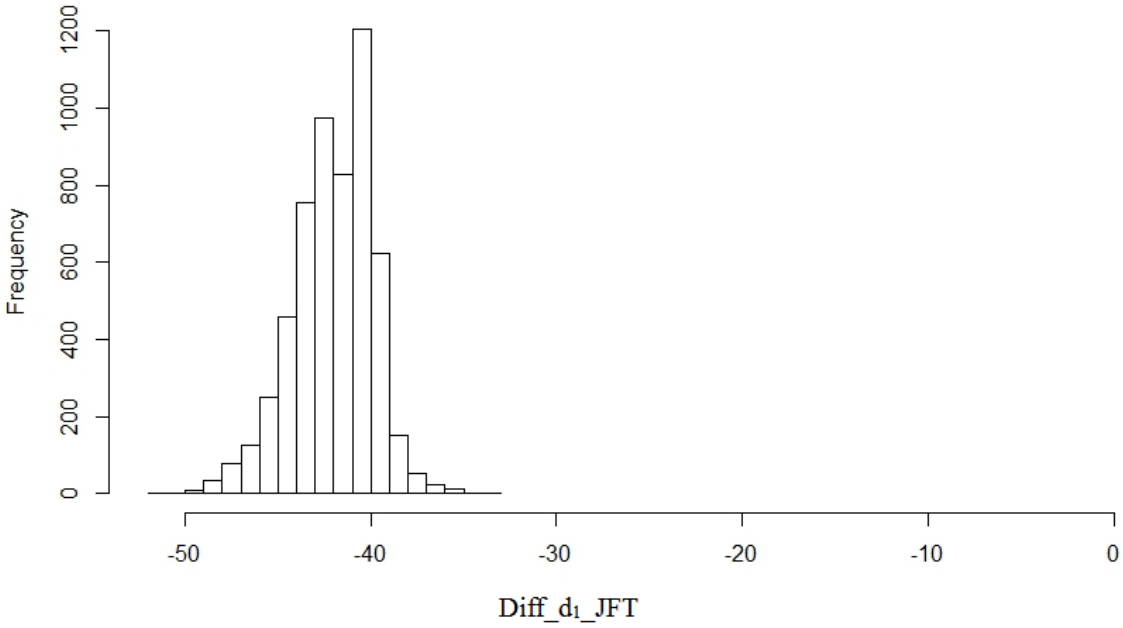

Figure 25. Histogram of the differences of vertical displacement (Diff_d_1)

The two histograms demonstrate that the Diff_d_1 values have a smaller range, from -33mm to -50mm, compared to the range of -10mm to -36mm. This leads to the conclusion that d_1 is more accurate than d_0. This is affirmed by the mean error rate derived in the preceding section using rescaling: Diff_d_1_S = 2.49mm whereas Diff_d_0_S = 4.69mm.

3.6.4.3 Statistical analysis

In order to identify whether type of vowel class has a systematic impact on the fidelity of the tracked points, a one-way analysis of variance was performed. It is possible that certain vowel classes entail movements that are more ballistic than other vowel classes.

Since highly ballistic movements may present additional challenges to tracking software, it is possible that tracking performance may depend on the type of speech sounds being articulated.

The one-way analysis of variance is a statistical test that is used to determine whether there are any significant differences between the means or max of more than two independent groups (Tabachnick, Fidell et al., 2001).

I examined 4 different dependent variables: Diff_mean_d_0, Diff_max_d_0, Diff_mean_d_1, and Diff_max_d_1 for the 5 independent groups (each of the five vowel classes).

A one-way ANOVA was conducted to compare the effect of (IV) vowel class on (DV) Diff_mean_d_0 in five articulation conditions (High, Mid, Low, Front, and Back vowels). The results demonstrated that there was not a significant effect of IV vowel class on DV Diff_ mean _d_0 at the $p<.05$ level for the three conditions [$F(4, 122) = 2.28$, $p = 0.064$].

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	-569.827249	-529.833497	-549.12155	-514.723897	-508.096643	-2671.60283
Mean	-21.104713	-21.19334	-21.964862	-20.588956	-20.323866	-21.036243
ΣX^2	12121.17678	11332.21225	12167.48907	10725.70201	10428.40253	56774.98266
Variance	3.659087	4.302952	4.421251	5.336434	4.246441	4.559498
Std.Dev.	1.912874	2.074356	2.102677	2.310072	2.060689	2.135298
Std.Err.	0.368133	0.414871	0.420535	0.462014	0.412138	0.189477

standard weighted-means analysis

ANOVA Summary

Source	SS	df	MS	F	P
Treatment [between groups]	39.990575	4	9.997644	2.28	0.064522
Error	534.506158	122	4.381198		
Ss/Bl					
Total	574.496732	126			

Figure 26. One-way ANOVA on Diff_mean_d_0

A one-way ANOVA was conducted to compare the effect of (IV) vowel class on (DV) Diff_max_d_0 in five articulation conditions (High, Mid, Low Front, and Back vowels). There was not a significant effect of IV vowel class on DV Diff_max_d_0 at the $p<.05$ level for the three conditions [$F(4, 122) = 1.14$, $p = 0.229$].

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	-515.729703	-505.511806	-515.462448	-395.088756	-492.406527	-2424.19924
Mean	-19.1011	-20.220472	-20.618498	-15.80355	-19.696261	-19.088183
ΣX^2	9958.14076	10485.7110	10831.0356	13364.5878	9816.56018	54456.0355
Variance	4.120618	11.000986	8.457259	296.699285	4.916361	64.940292
Std.Dev.	2.029931	3.316773	2.908137	17.224961	2.217287	8.058554
Std.Err.	0.39066	0.663355	0.581627	3.444992	0.443457	0.715081

standard weighted-means analysis

ANOVA Summary

Source	SS	df	MS	F	P
Treatment [between groups]	369.567357	4	92.391839	1.44	0.224859
Error	7812.909449	122	64.040241		
Ss/Bl					
Total	8182.476806	126			

Figure 27. One-way ANOVA on Diff_max_d_0

A one-way ANOVA was conducted to compare the effect of (IV) vowel class on (DV) Diff_mean_d_1 in five articulation conditions (High, Mid, Low Front, and Back vowels).

There was not a significant effect of IV vowel class on DV Diff_ mean_d_1 at the p<.05 level for the three conditions [F (4, 122) = 1.23, p = 0.302].

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	-1142.49179	-972.570391	-1041.59269	-1050.76399	-1082.26007	-5289.67893
Mean	-42.314511	-38.902816	-41.663708	-42.03056	-43.290403	-41.651015
ΣX^2	48398.4867	44306.7145	43460.1399	44222.8923	46922.9158	227311.149
Variance	2.09637	269.624495	2.646935	2.445576	2.976719	55.48136
Std.Dev.	1.447885	16.420247	1.62694	1.563834	1.725317	7.448581
Std.Err.	0.278646	3.284049	0.325388	0.312767	0.345063	0.660955

standard weighted-means analysis

ANOVA Summary

Source	SS	df	MS	F	P
Treatment [between groups]	271.496307	4	67.874077	1.23	0.301685
Error	6719.155038	122	55.075041		
Ss/Bl					
Total	6990.651345	126			

Figure 28. One way ANOVA on Mean_d_1_JFT Error (vertical displacement)

A one-way ANOVA was conducted to compare the effect of (IV) vowel class on (DV) Diff_max_d_1 in five articulation conditions (High, Mid, Low Front, and Back vowels). There was a significant effect of IV vowel class on DV Diff_ max_d_1 at the p<.05 level for the three conditions [F (4, 122) = 4.1, p = 0. 004].

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	-1207.9847	-1114.56883	-1077.14833	-1074.13861	-1136.27949	-5610.11996
Mean	-44.740174	-44.582753	-44.88118	-42.965545	-45.45118	-44.524762
ΣX^2	54167.5512	49863.1435	48424.5471	46285.2511	51774.4795	250514.972
Variance	4.696368	7.19152	3.515591	5.595861	5.384836	5.805747
Std.Dev.	2.167111	2.681701	1.874991	2.365557	2.320525	2.409512
Std.Err.	0.417061	0.53634	0.382731	0.473111	0.464105	0.214656

standard weighted-means analysis

ANOVA Summary

Source	SS	df	MS	F	P
Treatment [between groups]	86.620978	4	21.655244	4.1	0.003761
Error	639.097383	121	5.281797		
Ss/Bl					
Total	725.71836	126			

Figure 29. One way ANOVA on Max_d_1_JFT Error

In conclusion, the difference between max_d1 and the ground truth (Diff_max_d1) is significantly correlated to vowel class and is the only measure with a significant effect.

The results mean that we have determined that there the difference between the means of these independent groups is significant for IV vowel class on DV Diff_ max_d1 at the $p<.05$ level. The ANOVA test tells us whether the differences among the means are

significant for this particular subject. It is yet to be determined the degree to which the results for this particular subject can be generalized to other adults or to children.

3.7 **Conclusion**

In this chapter, I presented three different phases of development toward the goals of this research project: the identification of visible features from a video corpus, the development of a set of different computational techniques, and the evaluation of the most promising technique among those developed with a formal fidelity study. The study revealed that the best-performing feature, in terms of accuracy to ground truth, is the vertical distance between the top of the upper lip margin and the bottom of the lower lip margin (d_1), which has the fidelity of 2.49mm when compared to ground truth. The relationship between fidelity and vowel class was evaluated through a one way ANOVA analysis, which revealed that the fidelity of the distance d_1 is indeed significantly impacted by the type of vowel being articulated, at least for the maximal extent of the ballistic movement of mouth opening. It remains to be seen whether the degree of fidelity offered by the measure d_1 provides a sufficient basis for performing the classification of CVC productions using visible features of speech, which is the topic of the next chapter.

Chapter 4
Study 2: Classification of CVC Productions using Visible Features of Speech

4.1 Introduction

In this chapter, I present an investigation of the degree to which CVC productions can be classified using visible features of speech. In the previous chapter, we sought to characterize the degree of fidelity of our particular camera-based facial feature tracking technique. Now, we employ this technique to derive features that concern degree of jaw opening, and use these features as the basis for classification of CVC productions.

4.2 Objective

The objective of this investigation is to answer the following research questions:

1. To what extent does maximal mouth opening correlate with the five vowel levels that are targeted in PROMPT therapy? These include: High Front [HF], Mid Front [MF], Low Front [LF], High Back [HB], Low Back [LB], corresponding to levels 1-5.
2. With what accuracy can we distinguish between the high front [HF] and the low front [LF] vowel levels on the basis of visible features of speech? These two levels, at least in principle, should be the most visually distinct, since the jaw will be at the extrema of the vowel space (low vs high).

3. With what accuracy can we distinguish among the different frontal vowel segments, specifically the low, mid, and high frontal vowels: [LF], [MF], and [HF]? These correspond to levels 1-3.
4. With what accuracy can we tell the differences between the high front [HF] and high back [HB] vowels? The two levels are expected to be the most visually similar.
5. Can the classifier reliably distinguish a correct production of a vowel segment category from the competitor segments? (all five classes together)

4.3 **Methodology**

Each of the questions is answered via its own study.

- Study 1: To examine correlation, I will examine the data using various data visualizations and then perform a statistical analysis using a one-way ANOVA.
- Study #2: To assess the accuracy with which we can distinguish between high front [HF] and low front [LF] segments, I will perform a ML experiment using the HF+LF subset of the dataset.
- Study#3: To evaluate the accuracy with which we can distinguish among the three frontal vowel segments high front [HF], Mid front [MF], and low front [HF], I will apply a ML experiment using LF+MF+HF subset of the dataset.
- Study #4: To determine the accuracy that we can differentiate between high front [HF] and high back [HB], I will conduct a ML experiment using the HF+HB subset of the dataset.

- Study #5: To determine the classifier performance among all the vowel segments, I will conduct a ML experiment using the entire dataset: High Front [HF], Mid Front [MF], Low Front [LF], High Back [HB], and Low Back [LB].

4.3.1 *Classification*

For the machine learning experiments, I will use two techniques. One technique will employ support vector machines (SVM), which is a popular supervised learning approach. For this technique, I make use of WEKA (Weikato, 2015). The other technique will employ the k-nearest neighbors algorithm (k-NN), which is an unsupervised machine learning approach. k-NN classification is considered as a lazy learning algorithm and it classifies the data set based on its similarity with its neighbors. In k-NN classification, to determine the k number as a number of nearest neighbour, a range of numbers are chosen, starting from one and ending to the root of the size of the training data set and, as a result, the maximum accuracy among all the range numbers is selected (Hassanat, Abbadi et al., 2014). For k-NN, I used a script developed in MATLAB (MATLAB 8.0, The MathWorks, Inc., Natick, MA, US).

4.4 **Data Collection**

All five of the planned studies will make use of the dataset #1, which was collected for the fidelity study (as described in section 3.4). The dataset consists of 127 video segments, where each segment consists of subject 003 (an adult female) uttering a particular CVC segment, drawn from among the stimulus set of 25 words covering 5 difference vowel classes. The corpus contains up to 5 repetitions of each CVC segment.

As described previously, for each video segment, several distance measures were derived on a frame by frame basis. These measures were as follows:

- Philtrum to chin distance (d_0)
- Outer lip Distance (d_1)
- Inner lip Distance (d_2)
- Lip Corner Distance (d_3)
- Speaker's "mouth roundness" (d_4)
- Theta features for characterizing jaw sliding

On a per-segment basis, the two measures were derived:

- Peak_d_0_JFT
- Peak_d_1_JFT

A second dataset, dataset #2, was also employed, the one collected at an early stage of this research and used in Chapter 2 for the feature inventory phase of this research. This dataset corresponds to the articulation of the same stimulus set of 25 words, with up to 5 repetitions each, by a child (subject 001). For each frame in each video segment in dataset #2, we derived the following measures:

i. the distance from the nose tip to the bottom of the chin (d_0)
ii. the total distance from the bridge of the nose (midpoint between the two inner eye points) to the bottom of the chin (D).

The maximum values for these two distance measures were derived on a per-segment basis.

4.5 Evaluation

4.5.1 *Study #1*

In order to examine the correlation between degree of jaw opening and vowel class, I first created a visualization of the data for feature d_0 for dataset #1, which is the vertical displacement between the philtrum and the chin, derived on a frame-by-frame basis. These visualizations are shown in Figure 30.

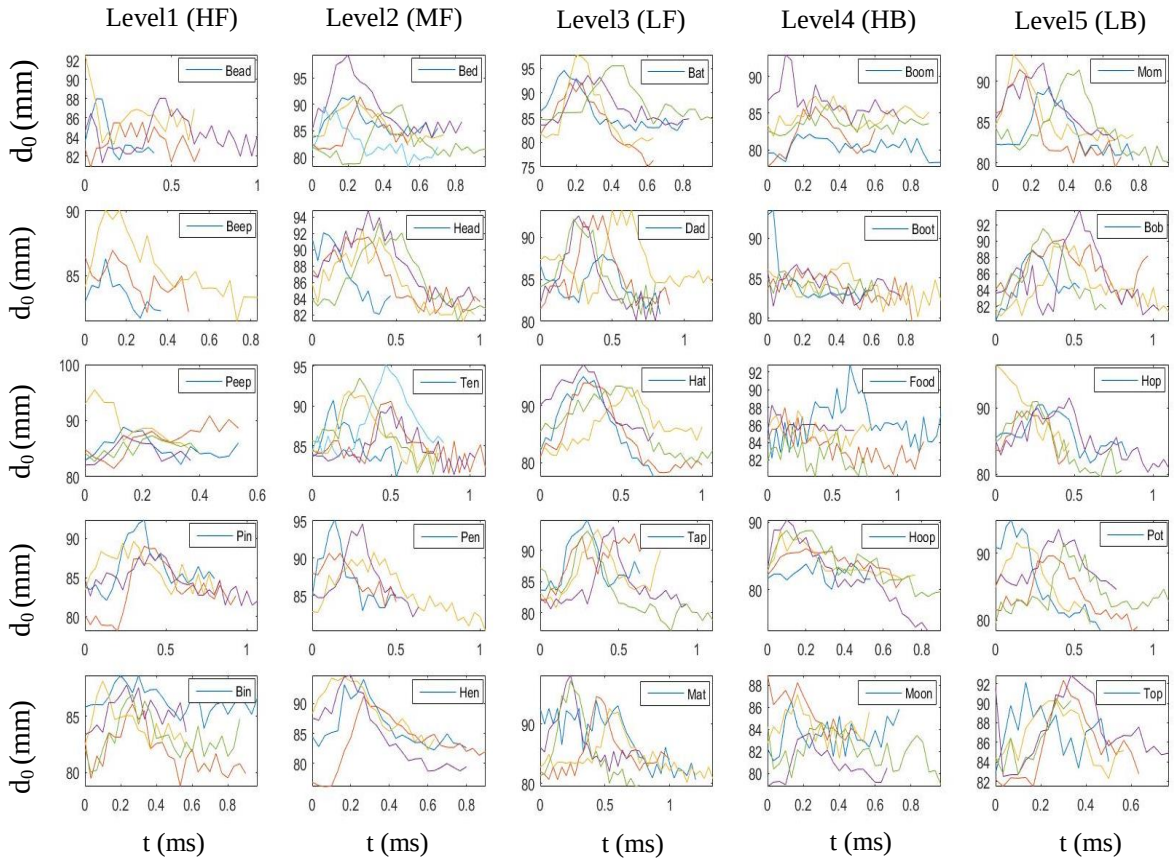

Figure 30. Visualization of frame-by-frame d_0 values. Each column of line charts represents the CVC segments of a particular level, with one row for each of the five different words for the CVC for that level. The particular CVC is indicated in the chart legends.

A visualization of the data for feature d_1, which is the vertical distance between the upper and lower lips, was also derived and is shown in Figure 31.

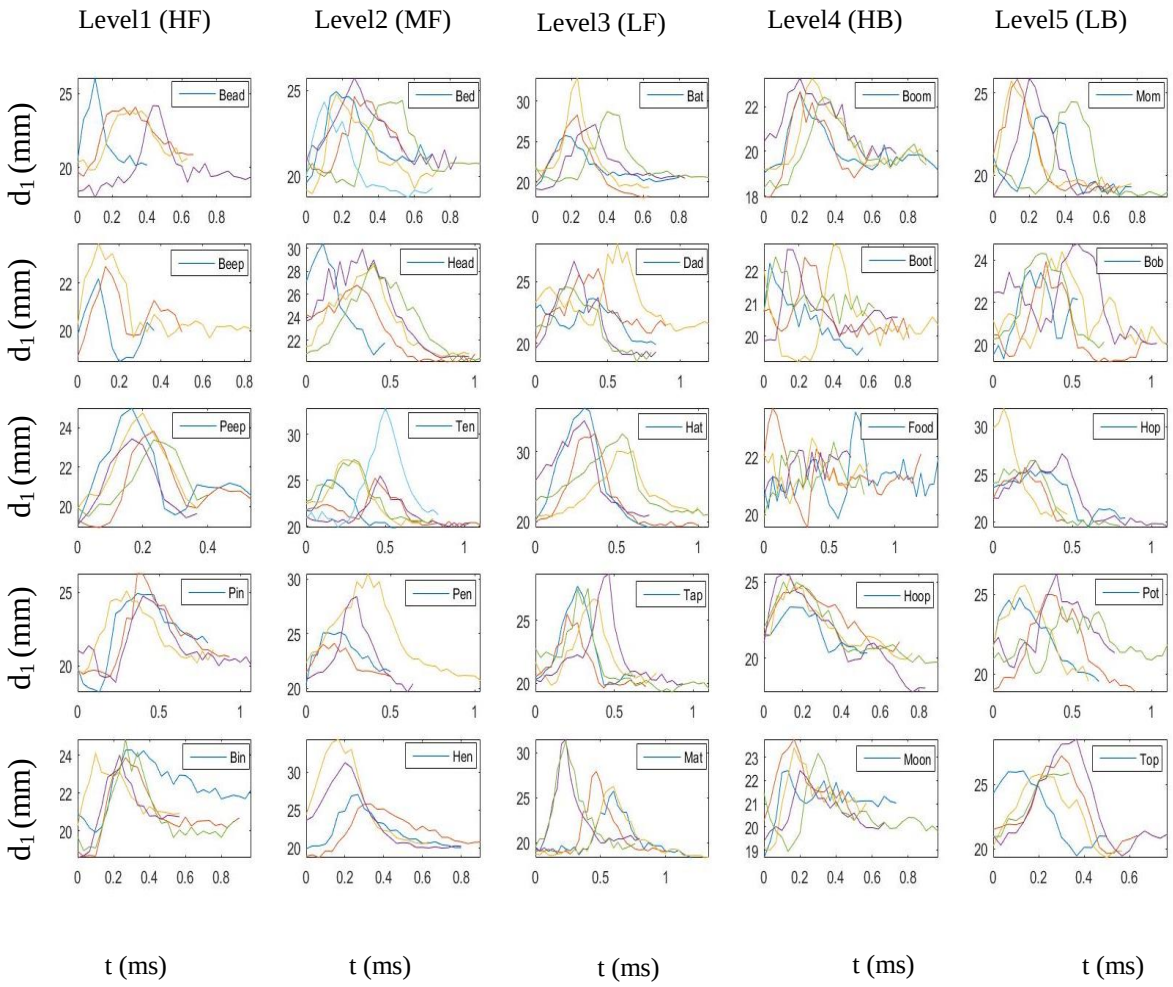

Figure 31. Visualization of frame-by-frame d_1 values: Each column of line charts represents the CVC segments of a particular level, with one row for each of the different CVCs for that level. The particular CVC is indicated in the chart legends. Each line graph illustrates the 4-6 repetition per CVC. For the sake of conciseness, the repetition labels are omitted.

For each CVC segment, the peak displacement of the respective features was identified and labeled peak_d_0 and peak_d_1, respectively. Scatterplots of the peak_d_0 and peak_d_1 values among the 5 different levels are shown in Figure 32 and Figure 33.

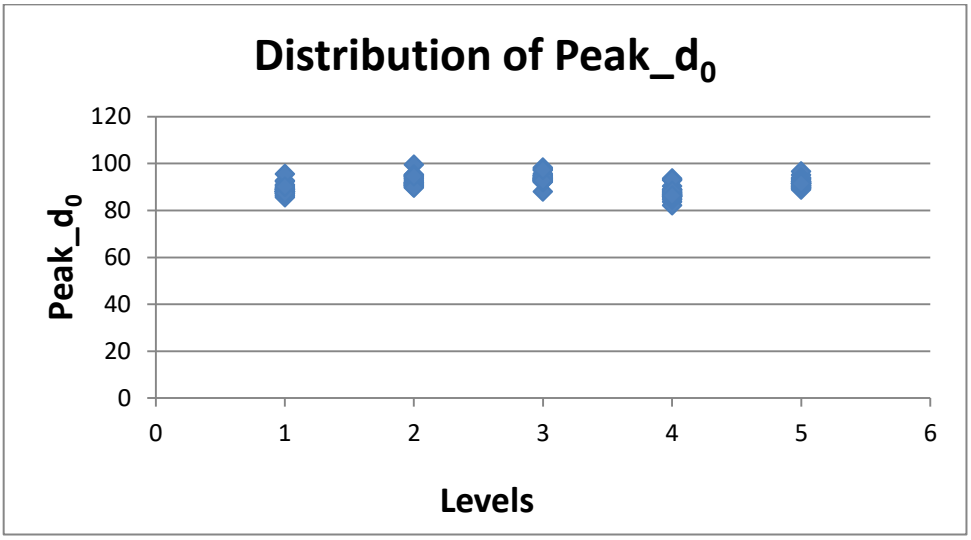

Figure 32. Distribution of data on peak_d_0

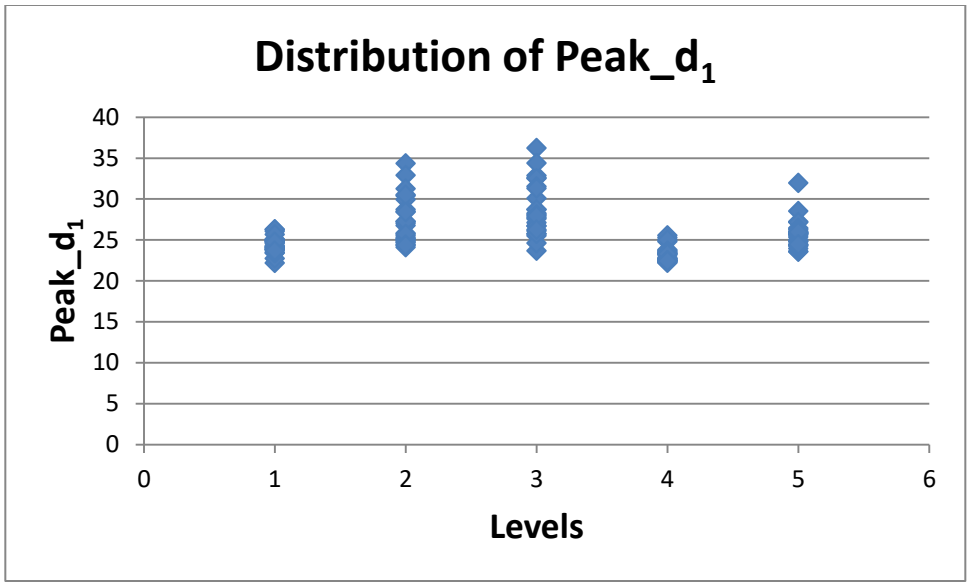

Figure 33. Distribution of data on peak_d_1

In order to determine the degree of correlation between the segment classes and the degree of jaw opening, a one-way ANOVA was performed for each of peak_d_0, and peak_d_1.

There was a significant effect of CVC class on peak_d_0 at the $p<.01$ level for the five conditions [$F(4,122) = 36.07$, $p < 0.0001$], as shown in Figure 34.

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	2397.5607	2315.4481	2348.9825	2185.929	2290.2502	11538.1704
Mean	88.7985	92.6179	93.9593	87.4372	91.61	90.8517
ΣX^2	213021.915	214576.670	220820.340	191314.605	209893.280	1049626.81
Variance	4.693	5.1951	4.6495	7.6327	3.4767	10.8254
Std.Dev.	2.1663	2.2793	2.1563	2.7627	1.8646	3.2902
Std.Err.	0.4169	0.4559	0.4313	0.5525	0.3729	0.292

standard weighted-means analysis

ANOVA Summary

Source	SS	df	MS	F	P
Treatment [between groups]	739.0882	4	184.772	36.07	<.0001
Error	624.9131	122	5.1222		
Ss/Bl					
Total	1364.0013	126			

Figure 34. One-way ANOVA of peak_d_0

There was a significant effect of CVC class on peak_d_1 at the $p<.01$ level for the five conditions [$F(4,122) = 25.69$, $p < 0.0001$], as shown in Figure 35.

Data Summary

	Samples					
	1	2	3	4	5	Total
N	27	25	25	25	25	127
ΣX	659.4903	682.8205	716.4853	580.6327	642.5273	3281.9562
Mean	24.4256	27.3128	28.6594	23.2253	25.7011	25.8422
ΣX^2	16133.5945	18846.295	20783.1001	13508.312	16587.9354	85859.2369
Variance	0.9681	8.1893	10.3771	0.9557	3.095	8.3044
Std.Dev.	0.9839	2.8617	3.2214	0.9776	1.7593	2.8817
Std.Err.	0.1894	0.5723	0.6443	0.1955	0.3519	0.2557

standard weighted-means analysis

ANOVA Summary 1

Source	SS	df	MS	F	P
Treatment [between groups]	478.3708	4	119.5927	25.69	<.0001
Error	567.9811	122	4.6556		
Ss/Bl					
Total	1046.3519	126			

Figure 35. One-way ANOVA on peak_d_1

Therefore, these two studies demonstrate that maximal mouth opening for a CVC is strongly correlated with the class of the vowel within the word.

4.5.2 ***Study #2***

In order to assess the accuracy with which we can distinguish between High Front [HF] and Low Front [LF] vowel segments, I designed and perform a ML experiment using the HF+LF subset of the data.

4.5.2.1 The HF+LF Dataset

The HF+LF dataset #1 consists of philtrum to chin distance (d_0), as produced by subject 003, obtained via the JFT, for the articulation of the HF and LF vowels. Articulation of vowels from these two classes, in principle, should be the most visually distinct. The subset of the dataset corresponding to the HF vowel class is made up of the 5 words "Bead", "Beep", "Peep", "Pin", "Bin", with 5 or more repetitions per word, with 27 segments in total. Moreover, the other subset of the dataset corresponding to the LF vowel classes is made up of the 5 words "Bat", "Dad", "Hat", "Tap", "Mat" with 5 repetitions for each word, with 25 segments in total. Therefore, the dataset consisted of, in total, 52 segments.

There is also an HF+LF dataset #2, which consists of the features, as produced by subject 001, as derived using the Eight Point Tracker. This dataset was derived at an early stage of the project, prior to the completion of the fidelity study described in section 3.4.1. This dataset contains articulation features for the LF and HF vowel classes. The HF vowel classes includes 5 words "Bead", "Beep", "Peep", "Pin", "Bin", with 4 repetitions each, for 20 segments. The LF vowel class consists of the 5 words "Bat", "Dad", "Hat", "Tap", "Mat", between 2-3 repetitions each, for 12 segments in total. Therefore, in total, the dataset consisted of the features taken from 30 segments.

4.5.2.2 Procedure

I employed SVM (Boser, Guyon et al., 1992; Cortes & Vapnik, 1995) as a supervised machine learning approach. For this technique, I used 85% of dataset #1 for training and

15% for testing (randomly-chosen). I also ran the k-NN classifier on both dataset #1 and dataset #2. For dataset #1, I used 85%/15% training/testing split. For dataset #2, I also used 85%/15% training/testing randomly-chosen split and I ran the split 10 times and calculated the mean value among them.

4.5.2.3 Results

Using the SVM classifier, the mean classification accuracy for the High Front (level1) and Low Front (level3) categories of vowel classes based on the feature max_d_0 was 92.3% (maximum accuracy).

The classification accuracy over the two categories using the k-NN classifier was a mean of 89.8% (mean is taken over all true positive accuracy values, over all values of k=1,…,10) and a maximum of 92.3% for dataset #1. For dataset#2 the mean accuracy was 57.9% (mean is taken over all true positive accuracy values). The k-NN results for dataset #1 and dataset #2 are shown in Figure 36 and in **Table 1**, respectively.

k	Results
1	86.5
2	84.6
3	88.4
4	84.6
5	92.3
6	92.3
7	92.3
8	92.3
9	92.3
10	92.3
Mean	89.8

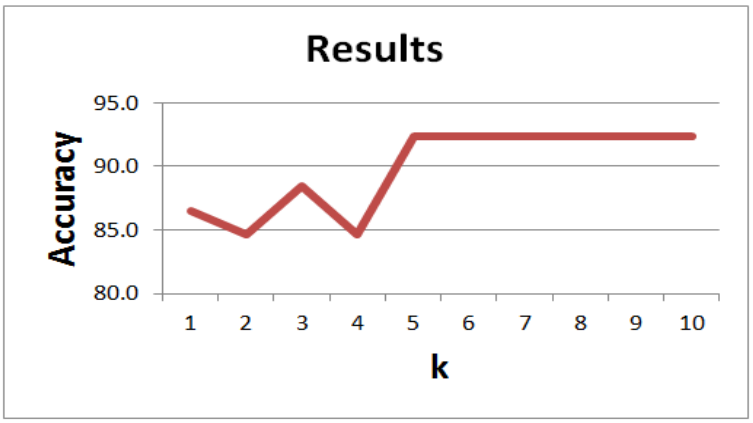

Figure 36. Classification between level 1 and level 3 for JFT

Using the k-NN classifier with dataset #2, which contains the features derived using the Eight Point Tracker, the mean classification accuracy was 62.5%, The results are shown in **Table 1**.

Confusion Matrix		Detected	
		Level1 (%)	Level3 (%)
Actual	Level1	50	50
	Level3	25	75

Table 1. k-NN results, confusion matrix for dataset #2 between Level 1 and level 3 (features derived using the Eight Point Tracker)

The best accuracy is observed from the first dataset using features derived from the JFT technique, with the accuracy of 92.3% using SVM classifier.

4.5.3 *Study 3*

In order to evaluate the accuracy with which we can distinguish among the three frontal vowel segments, High Front [HF], Mid Front [MF], and Low Front [HF], I performed a ML experiment using LF+MF+HF subset of the data.

4.5.3.1 The HF+MF+LF Dataset

The HF+MF+LF dataset #1 consists of the features, as produced by subject 003, derived using the JFT for the HF, MF, and LF vowels. This classification task introduces a degree

of difficulty, but does not address the factor of distinguishing between front and back vowels for a given degree of jaw opening. The subset of the dataset equivalent to the HF vowel class is made up of the 5 words “Bead”, “Beep”, “Peep”, “Pin”, “Bin”, with 5 or more repetitions per word, with 27 segments in total. The MF vowel class consists of 5 words “Bed”, “Head”, “Ten”, “Pen”, “Hen”, and 5 repetitions for each word, with the total of 25 segments. Moreover, the LF vowel class consists of 5 words “Bat”, “Dad”, “Hat”, “Tap”, “Mat”, with 5 repetitions for each word, and 25 segments in total is used in this dataset. Therefore, the dataset consists of, in total, 77 segments.

The HF+MF+LF dataset #2 consists of the features, as produced by subject 001, derived using the Eight Point Tracker algorithm. In this dataset, HF includes 5 words “Bead”, “Beep”, “Peep”, “Pin”, “Bin”, with 3 to 4 repetitions for each word, and 16 segments in total. The MF dataset consists of 5 words “Bead”, “Beep”, “Peep”, “Pin”, “Bin” with up to 4 repetitions for each word, for 20 segments in total. Finally, the LF dataset includes 5 words “Bat”, “Dad”, “Hat”, “Tap”, “Mat”, with 2 to 3 repetitions for each word, and 14 segments in total. Therefore, the dataset consists of the features taken from 50 segments in total.

4.5.3.2 Procedure

In this study, for the first dataset, SVM and k-NN classifiers were used (as described in section 4.3.1). In both SVM and k-NN techniques, I used 85% of dataset #1 for training and 15% for testing (randomly chosen). For dataset #2, the k-NN classifier was used (see

section 4.3.1). For this dataset, I used 85%/15% training/testing randomly-chosen split and I ran the split 10 times and calculated the mean value among them. In this dataset, out of 50 segments, 43 segments were used for training and 7 segments were used for testing.

4.5.3.3 Results

Using dataset #1, the mean classification accuracy over the High Front (level1), Mid Front (level2), and Low Front (level3) vowel classes was 59.7% using the SVM classifier.

The classification accuracy over the three categories using the k-NN classifier had a mean of 64.5% (mean is taken over all true positive accuracy values, over all values of k=1,…,10) and a maximum of 68.8% for dataset #1. For dataset#2 the mean accuracy was 57.9% (mean is taken over all true positive accuracy values). The k-NN results for dataset #1 and dataset #1 are shown in Figure 37 and in Table 2, respectively.

k	Results
1	58.4
2	51.9
3	61.0
4	63.6
5	68.8
6	67.5
7	67.5
8	68.8
9	68.8
10	68.8
Mean	64.5

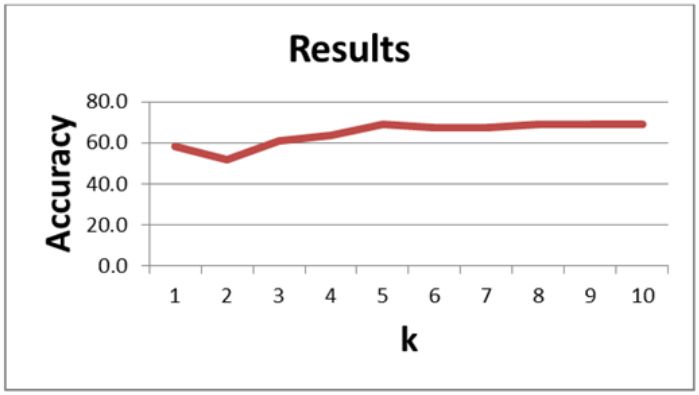

Figure 37. k-NN results between level 1, 2, and 3 for dataset #1 (features derived using JFT)

Confusion Matrix		Detected		
		Level1 (%)	Level2 (%)	Level3 (%)
Actual	Level1	54.54	27.27	18.18
	Level2	23.07	69.23	7.69
	Level3	50	0	50

Table 2. k-NN results, confusion matrix for dataset #2 (features derived using the Eight Point Tracker)

The best accuracy is observed for the first dataset using features derived using the JFT technique, with the accuracy of 64.5% using k-NN classifier.

4.5.4 *Study 4*

In order to determine the accuracy with which we can differentiate between High Front [HF] and High Back [HB] vowel segments, I conducted a ML experiment using the HF+HB subset of the data only.

4.5.4.1 HF+HB Dataset

The HF+HB dataset #1 includes the features, as produced by subject 003, derived from JFT algorithm. This dataset contains of the two vowel classes (level1 and level4) only. These two classes represent the High Front (HF) and High Back (HB) vowels, respectively. In principle, these two classes are less visually distinct and should be challenging for a vision-based algorithm to distinguish. The subsection of the data relating to the HF vowel class includes 5 words “Bead”, “Beep”, “Peep”, “Pin”, “Bin”, with more than 5 repetitions for two of these words, for 27 segments in total. The HB vowel class corresponds to the 5 words “Boom”, “Boot”, “Food”, “Hoop”, “Moon”, with 5 repetitions for each word, and 25 segments in total. Therefore, the dataset consists of the features from 52 segments.

The HF+HB dataset #2 contains of the features, as produced by subject 001, derived using the Eight Point Tracker algorithm for articulation of the High Front [HF] and the High Back [HB] vowels. The [HF] vowels are found in these 5 words “Bead”, “Beep”, “Peep”, “Pin”, “Bin”, with 3-4 repetitions for each word, for 16 segments in total. The [HB] vowel class comprises of 5 words “Boom”, “Boot”, “Food”, “Hoop”,

"Moon", with 3 repetitions for each word for 15 segments in total. Thus, dataset #2 consists of a total of 31 segments.

4.5.4.2 Procedure

In this study, for HF+HB dataset #1, the SVM and k-NN classifiers were used (as described in section 4.3.1)., For the HF+HB dataset #2, the k-NN classifier was used (see section 4.3.1).

In both SVM and k-NN techniques, 85% and 15% of the dataset was used for training and testing respectively (randomly chosen). Moreover, for the second dataset, k-NN classifier as supervised machine learning was used (see section 4.3.1). For this dataset, I used a 85%/15% training/testing random-chosen split and I ran the split 10 times and calculated the mean value among them. In this dataset, out of 31 segments, 26 segments were used for training and 5 segments were used for testing.

4.5.4.3 Results

The mean classification between the High Front (level1), and High Back (level4) vowel classes, computed across individual speakers, was 57.6% using the SVM classifier and 62.1% using the k-NN classifier with dataset #1 (mean is taken over all true positive accuracy values, over all values of k=1,…,10) and a maximum of 67.3% for dataset #1. The k-NN results can be seen in Figure 38.

k	Results
1	61.5
2	55.7
3	55.7
4	57.6
5	55.7
6	69.2
7	61.5
8	69.2
9	67.3
10	67.3

Mean	62.1

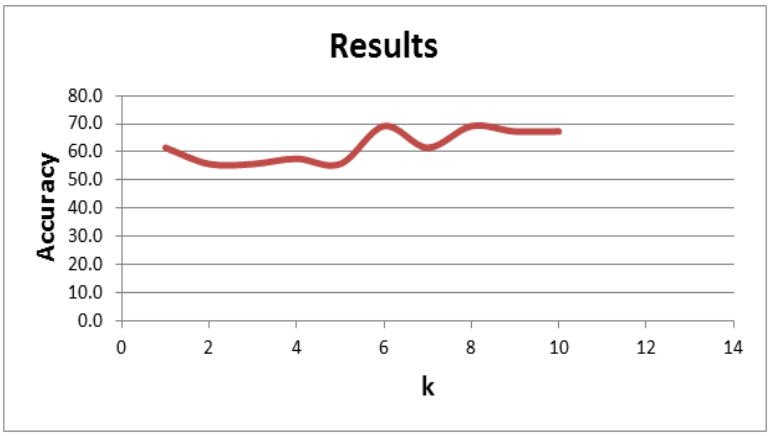

Figure 38. k-NN Results between level1 and 4 in JFT

Using the k-NN classifier with dataset #2, which contains the features, derived using the Eight Point Tracker, the mean classification accuracy was 31.2% (mean is taken over all true positive accuracy values). A breakdown of the classifier results are shown in Table 3.

Confusion Matrix		Detected	
		Level1 (%)	Level4 (%)
Actual	Level1	12.50	87.50
	Level4	50	50

Table 3. Confusion Matrix with Percentage

The best accuracy is observed from the first dataset using features derived using the JFT technique, with the accuracy of 67.3% using k-NN classifier.

4.5.5 ***Study 5***

To determine the classifier performance among all the vowel segments, I conducted a ML experiment using the entire dataset: High Front [HF], Mid Front [MF], Low Front [LF], High Back [HB], and Low Back [LB].

4.5.5.1 HF+MF+LF+ HB+LB Dataset

The HF+MF+LF+HB+LB dataset #1 includes the features, as produced by subject 003, derived using JFT for all five vowel 5 classes: level 1, level 2, level 3, level4, and level 5&6 vowels. To begin with, level1 [HF] consists of 27 segments in total, with 5 words "Bead", "Beep", "Peep", "Pin", "Bin" and more than 5 repetitions for each word. For level 2 [MF], there were 25 segments in total: 5 words, 5 repetitions each of "Bed", "Head", "Ten", "Pen", "Hen". For level 3 [LF], there were 25 segments in total, 5 words, 5 repetitions each of "Bat", "Dad", "Hat", "Tap", "Mat". Level 4 [HB] adds up to 25 segments in total, 5 words, 5 repetitions for each word "Boom", "Boot", "Food", "Hoop", "Moon", and, lastly, level 5 [LB] is made up to 25 segments in total, 5 words, 5 repetitions each of "Mom", "Bob", "Hop", "Pot", "Top". Therefore, the dataset consists of features drawn from a total of 127 segments.

The HF+MF+LF+HB+LB dataset #2 includes the features, as produced by subject 001, derived using the Eight Point Tracker for all five level classes [HF+MF+LF+ HB+LB]. Level 1 [HF] is represented by 16 segments in total, 5 words, with 3-4 repetitions for each word "Bead", "Beep", "Peep", "Pin", "Bin". Level 2 [MF] is represented by 20 segments in total, 5 words, 4 repetitions of each word "Bed", "Head",

"Ten", "Pen", "Hen". Level 3 [LF] is represented by 14 segments in total, 5 words, with 2-3 repetitions of each word "Bat", "Dad", "Hat", "Tap", "Mat". Level 4 [HB] is represented by 16 segments in total, 5 words, with 3-4 repetitions of each word "Boom", "Boot", "Food", "Hoop", "Moon", and, finally, level 5 [LB] is represented by 10 segments in total, 5 words, 2 repetitions each of "Mom", "Bob", "Hop", "Pot", "Top".

4.5.5.2 Procedure

In this study, for the first dataset, SVM and k-NN classifiers were used (as described in section 4.3.1). In both SVM and k-NN techniques, 85% and 15% of the dataset was used for training and testing respectively (randomly chosen). Moreover, for the second dataset, the k-NN classifier was used (see section 4.3.1). For this dataset, I used a 85%/15% randomly-chosen training/testing split, and I ran the split 10 times and calculated the mean value among them. In this dataset, out of 76 segments in total, 65 segments were used for training and 11 segments were used for testing.

4.5.5.3 Results

The mean classification accuracy among the five vowel classes for dataset #1, computed across individual speakers, was 67.7% using SVM classifier, and was 58.6% (mean is taken over all true positive accuracy values, over all values of k=1,…,10) and a maximum of 71.6% using the k-NN classifier. The k-NN results can be seen in Figure 39.

k	Results
1	62.2
2	71.6
3	60.6
4	61.4
5	54.3
6	55.9
7	54.3
8	54.3
9	54.3
10	57.4

Mean	58.6

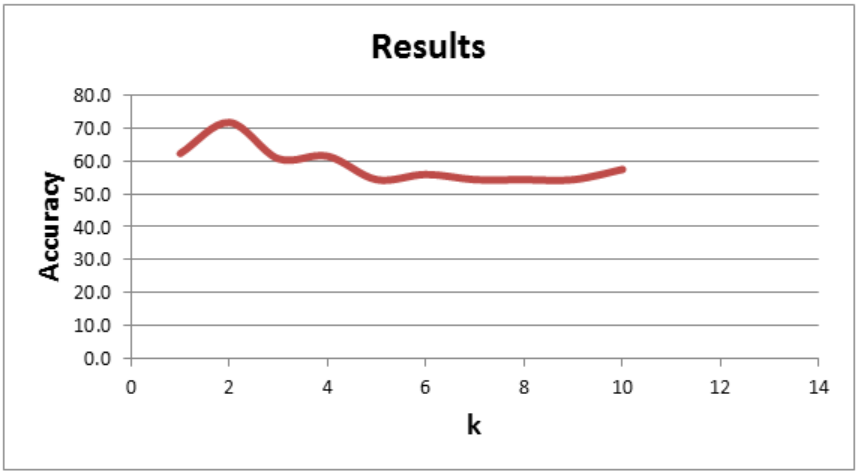

Figure 39. k-NN Results between all levels

Using the k-NN classifier with dataset #2, which contains the features, derived using the Eight Point Tracker, the mean classification accuracy was 52.8% (mean is taken over all true positive accuracy values). The results are shown in Table 4.

Confusion Matrix		Detected				
		Level1 (%)	Level2 (%)	Level3 (%)	Level4 (%)	Level5&6 (%)
Actual	Level1	37.50	0	25	37.50	0
	Level2	0	77.77	11.11	11.11	0
	Level3	37.50	50	12.50	0	0
	Level4	0	0	10	70	20
	Level5&6	0	22.22	0	11.11	66.66

Table 4. Confusion matrix between Level 1, 2,3,4,5, and 6

The best accuracy is observed from the first dataset using features derived via the JFT technique, with the accuracy of 67.7% using SVM classifier.

4.5.6 *Summary of Accuracy Results*

The mean accuracy rates for all the studies can be seen in Table 5.

		Mean Accuracy Results (%)		
		Features from JFT		***Features from Eight Point Tracker***
Study	**Classes**	SVM (Subject_Id:003)	K-NN (Subject_Id:003)	K-NN (Subject_Id:001)
2	Levels 1, 3	92.3	89.8	62.5
3	Levels 1, 2, 3	59.7	64.5	57.9
4	Levels 1 ,4	57.6	62.1	31.2
5	Levels 1, 2, 3, 4, 5&6	67.7	58.6	52.8

Table 5. Mean Accuracy Results

The findings from this investigation demonstrate the impact of classification technique. For instance, if we consider only the features derived via the JFT, we see only minor differences in terms of classification technique: for instance, in study 2, for k-NN, the mean accuracy was 89.8% whereas for SVM the mean accuracy was 92.3%. The difference between the two classification techniques was more striking for studies 2-5. The k-NN technique produced better classification results than SVM only for studies 3 and 4.

The findings from this investigation also demonstrate that relative quality of the features extracted via the JFT tracker as opposed to the Eight Point Tracker. We can compare the k-NN classification results between dataset #1 (JFT) and dataset #2 (Eight Point Tracker). The results demonstrate that poorer quality of the features extracted using the Eight Point Tracker algorithm than the features extracted using the Jason Face Tracker. For instance, for study 2, the mean classification accuracy was 89.8% for JFT-based features and 62.5% for the analogous features extracted using the Eight Point

Tracker. However, since the subjects are not the same across the two datasets, it is possible that other factors may have played a role (for instance, facial feature tracking maybe be more challenging for a child-sized face than an adult-face).

These results demonstrate that camera-based tracking is likely accurate enough as the basis for a providing feedback within a game that supports at-home practice in distinguishing between level 1 and level 3 vowel classes. Classification accuracy rates below 90% are likely not high enough, and would create too many false positives (positive feedback would be given for incorrect articulations) and of false negatives (the system would fail to recognize a correctly-produced articulation). These results demonstrate that these techniques require further improvement for training scenarios among the other vowel classes. Study 1 demonstrates that the features for each of the five vowel classes, in principle, are distinct enough, so we can conclude that poor results are attributable to problems with the quality of the features employed by the classifiers and/or the classification techniques themselves.

4.6 **Conclusion**

This chapter described a series of five studies to investigate the degree to which features derived from the JFT and the Eight Point Tracker could serve as the basis for distinguishing among five different vowel classes. Both the SVM and the k-NN classifiers were used. The study 2 results demonstrate that, for the first dataset, the best results are observed when attempting to distinguish between level 1 and 3 vowels, with a mean accuracy of 92.3% and 89.8%, by using SVM and k-NN classifier, respectively.

Moreover, a side-by-side comparison of the features derived using the JFT vs the Eight Point Tracker can be made, using the same k-NN classification approach for features derived from each of these techniques, at least for the specific data used in these analyses. The results demonstrate that the JFT produced better visible features of speech than the Eight Point Tracker. The worst result is observed when endeavoring to distinguish between level 1 and level 4 vowels: both are high vowels, distinguished only by front vs back placement. Neither classification technique nor feature set produced good results. As it was expected, distinguishing between level 1 and level 4 is very challenging for a camera-based algorithm due to the subtly of the visual features.

Chapter 5
Conclusion

This research project set out to undertake the first iteration of a design process for a computer-based speech therapy (CBST) system to support a PROMPT-based intervention to Childhood Apraxia of Speech (CAS). It reports on an investigation of camera-based tracking of visible features of speech. To structure this project, a framework of long term, mid-term, and short term goals was established.

A long term goal of this research is to develop and to evaluate a user-friendly game that supports a PROMPT-based approach to speech therapy and that makes use of camera-based facial feature tracking, such as the techniques developed in this research study. A middle term goal of this research is to design a game-like scenario which delivers support for one or two particular stages of PROMPT therapy. To support these mid- and long-term goals, a series of short-term goals was developed: (1) to identify an inventory of visible facial features that would be suitable for camera-based tracking and that are relevant for PROMPT; (2) to identify suitable, ready-made facial feature tracking software libraries and to appropriately extend them for the present purposes, in order to develop a PROMPT-relevant tracking software module; and (3) to determine the fidelity and the accuracy of the PROMPT-relevant facial feature tracking module that was developed.

In support of the first short-term goal, the following questions were posed: what is PROMPT protocol and how can this protocol be supported by a CBST system to treat

people with different types of speech disorder, especially Childhood Apraxia of Speech (CAS)? What methodology should be employed for the design of a system that supports home-based computer-supported therapy?

In support of the second short-term goal, the following questions were posed: what are the relevant technologies in facial feature tracking? What characterizes the performance of these techniques? What are the most promising options for the identification of visible speech features from video?

In support of the third short-term goal, the following questions were posed: What is the fidelity of camera-based tracking of PROMPT-relevant facial features of speech? What are the results of off-line classification of speech productions by vowel segment, on the basis of those tracked, visible features of speech?

5.1 **Findings**

5.1.1 ***How can the PROMPT protocol be supported by a CBST system to treat CAS?***

First, a review of research literature provided the background on different ways CBSTs provide therapeutic benefits. It was demonstrated that children with Childhood Apraxia of Speech (CAS) could benefit greatly from Computer-based speech therapy. In PROMPT therapy, the clinician provides tactile ‘prompts’ to provide children with feedback about their speech motor control during speech articulation. To further develop the idea of camera-based tracking of the facial features that correspond to PROMPT-relevant speech motor control characteristics, I developed a video corpus of speech segments for a set of relevant speech stimuli items (described in Chapter 3). Using this

video corpus, I identified key facial features of speech which are relevant to the delivery of Stage 2 and Stage 3 PROMPT therapy and developed formal feature-based measures for the jaw, encompassing the following:

- Nose-Chin distance
- Outer lip distance
- Inner lip distance
- Corner lip distance
- Speaker's "mouth roundness”
- Theta" feature for characterizing jaw sliding

5.1.2 *What methodology should be employed for the design of a system that supports home-based computer-supported therapy?*

In chapter 2, I described how a User Centered design (UCD) methodology should be employed. This methodology is an iterative approach to interactive systems. The first iteration, undertaken in the project described here, involved a SLP who is a practitioner of PROMPT, who represented this key stakeholder group, and who provides an initial set of key requirements for the system.

5.1.3 *What are the relevant technologies in facial feature tracking? What characterizes the performance of these techniques?*

In chapter 2, I presented an overview of the following techniques: Feature tracking (Haar Cascade), Canny Edge Detector, Eight Point Tracking, and the Jason Face Tracker (JFT). I have performed an analysis of relative merits on each technique (e.g., accuracy, processing requirements, flexibility, functioning). I identified, for each, how the

PROMPT-relevant features could be identified and tracked, and I developed prototype versions using each, which operated upon the corpus of video segments.

5.1.4 *What are the most promising options for the identification of visible speech features from video?*

Among all the techniques surveyed, the Jason Face Tracker (JFT), a technique that employs active shape modeling, was selected for the next phases of the study. This tracking technique represented a good match to requirements, in terms of being able to track the PROMPT-relevant features and being free and open source. I created a software module that extended and augmented the JFT technique and produced the specific, PROMPT-relevant different features of interest on a frame-by-frame basis for a given input video segment. This work made use of the video corpus developed at an earlier phase.

5.1.5 *What is the fidelity of the camera-based tracking of PROMPT-relevant facial features of speech?*

To determine fidelity, a comparison study was performed to evaluate JFT relative to the output from a 3D kinematic system (the WAVE electromagnetic articulograph system), which served as a ground truth.

Among all the PROMPT-relevant features, I focused specifically on degree of jaw opening and the following measures specifically: (1) d_0, the distance between the philtrum to the chin and (2) d_1, the vertical distance between the top of the upper lip margin and the bottom of the lower lip margin. The study revealed that the best-

performing feature, in terms of accuracy to ground truth, is the vertical distance between the top of the upper lip margin and the bottom of the lower lip margin (d_1), which has fidelity of 2.49mm when compared to ground truth.

5.1.6 *What are the results of off-line classification of speech productions by vowel segment, on the basis of visible features of speech?*

I described a series of five studies with supervised machine learning (SVM) and unsupervised machine learning (k-NN) techniques, to investigate the degree to which facial features of speech, derived via the JFT and the Eight Point Tracker techniques, could serve as a basis for distinguishing among the five vowel classes.

The best classification results were observed using the feature of philtrum to chin distance (d_0), as produced by subject 003, obtained via the JFT to distinguish between level 1 (high front, such as the /i/ in "pin") and level 3 (low front, such as the /ae/ in "bat") vowels, with accuracies of 92.3% and 89.8%, by using the SVM and k-NN classifiers, respectively. These results demonstrate that it is possible to obtain decent accuracy in classifying jaw movements during speech, even in the presence of factors, such as low-resolution camera input. The worst results were observed when attempting to distinguish the most difficult scenario between level 1 (high front, such as the /i/ in "pin") and 4 (high back, such as the /u/ in "moon") with mean accuracies of 57.6% and 62.1% for SVM and k-NN classifiers, respectively.

5.2 Future work

This section describes a number of directions for future work.

Video Corpus

Different phases of this work each depend on corpora of relevant video segments. For instance, one corpus was used in this research project at an early stage for investigating facial features of visible speech and another corpus was developed additionally for the fidelity and classification studies. These corpora each contain approximately 127 segments (approximately 5 repetitions of each of 25 different stimuli items, with 5 stimuli items per vowel class). For future work, a more comprehensive corpus of video segments should be developed for training and testing purposes. This will be needed for the development of more refined classifiers, including on-line and off-line classifiers. To create a more comprehensive corpus, the articulations made by a wider variety of subjects should be included, because the shapes and the kinematics of human facial features for speech articulations can vary significantly through individuals, sex, or even due to the facial expression. In particular, it is very important to develop a corpus of video segments of children with CAS articulating key stimulus items.

The video corpus plays a key role in developing the classifier for the vowel classes on the basis of facial features of speech. By using only a small number of different speakers for training, it may be the case that the trained classifier works on one speaker more effectively than for other speakers. Thus, this also motivates the need for a larger corpus of video segments in future work.

Computational techniques for facial feature tracking

Numerous modifications and extensions could be made, in the future, to the software module that was used to track the facial features of speech articulation. In this research project, the JFT library was used. Although this approach shows the good performance when compared to the ground truth, the fidelity should be improved as a future step by refining the camera-based technique. Further improvement could be targeted to obtain a fidelity that is better than 2.49mm. Moreover, occlusion was not addressed in this work, but future work could consider and provide a robust handling of occlusion that may occur in the eventual use scenario.

Augmented Feature Set

In future work, the features used in machine learning experiments should be more closely coordinated with the outcome of the fidelity study. In this work, I employed the d_0 measures in the machine learning experiments, yet the d_1 measure was shown to have the higher fidelity (lower mean difference and smaller variability). Moreover, peak_d_1 is more variable in relation to the CVC classes. The classification experiment should be repeated with d_1.

In this study, different facial features — such as the distance between the philtrum to chin, vertical displacement (the distance between the outer lips), and corner lip distances — were calculated. In future work, additional features to detect degree of jaw opening and sliding should be identified and investigated. I was not able to distinguish between the front and back vowels very effectively, so the additional features may be

able to serve as the basis for distinguishing between these subtly-different two vowel classes.

Classification Techniques

The use of different classification techniques should be considered as another component for future work. By increasing the number of features, different classifiers, such as neural network, decision tree, or naïve Bayes, could be employed and could be considered in the future iterations of this work. Another recommended next step of this research would be conducting a user study to determine the impact of false positive and false negatives in terms of user experience and clinical efficacy. These two types of errors can be possibly minimized, but are likely inevitable in this application domain.

Game Development

A next step of this research should be to develop and to derive the clinical efficacy of a game scenario that supports PROMPT therapy. This game would have the objective of prompting the user for certain, relevant speech stimuli, tracking their speech articulators (such as chin movements), and providing constructive and helpful corrective feedback. This would support children with CAS in practicing speech therapy exercises under the supervision of a Speech-Language Pathologist. The development of the game should continue to follow the user centered design (UCD) methodology. And, as mentioned in chapter 2, since the focus of this thesis is on children, a low-cost, easily-accessible tablet-based platform should continue to be used.

Bibliography

Abras, C., Maloney-Krichmar, D., & Preece, J. (2004). User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications, 37*(4), 445-456.

Adolphs, R., Sears, L., & Piven, J. (2001). Abnormal processing of social information from faces in autism. *Journal of cognitive neuroscience, 13*(2), 232-240.

Al-Nafjan, A., Al-Wabil, A., & Al-Ohali, Y. (2015). *Augmenting Speech-Language Rehabilitation with Brain Computer Interfaces: An Exploratory Study Using Non-invasive Electroencephalographic Monitoring*.

Amberg, B., & Vetter, T. (2011). *Optimal landmark detection using shape models and branch and bound*. Paper presented at the Computer Vision (ICCV), 2011 IEEE International Conference on. (pp. 455-462)

Antonakos, E., Alabort-i-Medina, J., Tzimiropoulos, G., & Zafeiriou, S. (2014). *Hog active appearance models*.

ASHA. (1993). Definitions of communication disorders and variations. Retrieved from http://www.asha.org/policy/RP1993-00208/

ASHA. (2016). Childhood Apraxia of Speech. Retrieved from http://www.asha.org/public/speech/disorders/ChildhoodApraxia/

Baek, E.-O., Cagiltay, K., Boling, E., & Frick, T. (2008). User-centered design and development. *Handbook of research on educational communications and technology*(1), 660-668.

Baker, F., & Uhlig, S. (2011). *Voicework in music therapy: research and practice*: Jessica Kingsley Publishers.

Bakker, S., Vorstenbosch, D., van den Hoven, E., Hollemans, G., & Bergman, T. (2007). *Tangible interaction in tabletop games: studying iconic and symbolic play pieces*.

Ballard, K. J., Maas, E., & Robin, D. A. (2007). Treating control of voicing in apraxia of speech with variable practice. *Aphasiology*, *21*(12), 1195-1217.

Bälter, O., Engwall, O., Öster, A.-M., & Kjellström, H. (2005). *Wizard-of-Oz test of ARTUR: a computer-based speech training system with articulation correction*.

Bandini, A., Ouni, S., Cosi, P., Orlandi, S., & Manfredi, C. (2015). *Accuracy of a markerless acquisition technique for studying speech articulators*.

Baranek, G. T. (2002). Efficacy of sensory and motor interventions for children with autism. *Journal of autism and developmental disorders*, *32*(5), 397-422.

Bartle-Meyer, C. J., Goozée, J. V., Murdoch, B. E., & Green, J. R. (2009). Kinematic analysis of articulatory coupling in acquired apraxia of speech post-stroke. *Brain injury*, *23*(2), 133-145.

Bashier, H. K., Abusham, E. E., Abdullah, M. F. A., Liew, T. H., Yusof, I., & Lau, S. H. (2013). Real time face tracker based on local graph structure threshold (LGS-TH). *Australian Journal of Basic and Applied Sciences, 7*(2), 632-638.

Behnia, B., & Duclos, É. (2003). Children with disabilities and their families Retrieved from http://www.statcan.gc.ca/pub/89-585-x/89-585-x2003001-eng.pdf

Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., & Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence, 35*(12), 2930-2940.

Bellini, S., & Akullian, J. (2007). A meta-analysis of video modeling and video self-modeling interventions for children and adolescents with autism spectrum disorders. *Exceptional children, 73*(3), 264-287.

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

Beukelman, D., & Mirenda, P. (2005). Augmentative and alternative communication: Supporting children and adults with complex communication needs.

Bose, A., Square, P. A., Schlosser, R., & van Lieshout, P. (2001). Effects of PROMPT therapy on speech motor function in a person with aphasia and apraxia of speech. *Aphasiology, 15*(8), 767-785.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). *A training algorithm for optimal margin classifiers*.

Boyle, E. A., Hainey, T., Connolly, T. M., Gray, G., Earp, J., Ott, M., Lim, T., Ninaus, M., Ribeiro, C., & Pereira, J. (2016). An update to the systematic literature review of empirical evidence of the impacts and outcomes of computer games and serious games. *Computers & Education, 94*, 178-192.

Bradski, G. R. (1998). Computer vision face tracking for use in a perceptual user interface.

Brederode, B., Markopoulos, P., Gielen, M., Vermeeren, A., & De Ridder, H. (2005). *pOwerball: the design of a novel mixed-reality game for children with mixed abilities*.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Bunnell, H. T., Yarrington, D., & Polikoff, J. B. (2000). *STAR: articulation training for young children*.

Burgos-Artizzu, X. P., Perona, P., & Dollár, P. (2013). *Robust face landmark estimation under occlusion*.

Canny, J. (1986). A computational approach to edge detection. *IEEE transactions on pattern analysis and machine intelligence*(6), 679-698.

Cao, X., Wei, Y., Wen, F., & Sun, J. (2014). Face alignment by explicit shape regression. *International Journal of Computer Vision, 107*(2), 177-190.

Case-Smith, J., & Bryan, T. (1999). The effects of occupational therapy with sensory integration emphasis on preschool-age children with autism. *American Journal of Occupational Therapy, 53*(5), 489-497.

Chumpelik, D. (1984). *The PROMPT system of therapy: Theoretical framework and applications for developmental apraxia of speech*.

Cockburn, J., Bartlett, M., Tanaka, J., Movellan, J., Pierce, M., & Schultz, R. (2008). *Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder*.

Cootes, T. F., Taylor, C. J., Cooper, D. H., & Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding, 61*(1), 38-59.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

Craiger, J. P., Goodman, D. F., Weiss, R. J., & Butler, A. (1996). Modeling organizational behavior with fuzzy cognitive maps. *International Journal of Computational Intelligence and Organizations, 1*(3), 120-123.

Cumley, G., & Swanson, S. (1999). Augmentative and alternative communication options for children with developmental apraxia of speech: Three case studies. *Augmentative and Alternative Communication, 15*(2), 110-125.

Dabbs, A. D. V., Myers, B. A., Mc Curry, K. R., Dunbar-Jacob, J., Hawkins, R. P., Begey, A., & Dew, M. A. (2009). User-centered design and interactive health technologies for patients. *Computers, informatics, nursing: CIN, 27*(3), 175.

Dantone, M., Gall, J., Fanelli, G., & Van Gool, L. (2012). *Real-time facial feature detection using conditional regression forests*.

Dollár, P., Welinder, P., & Perona, P. (2010). *Cascaded pose regression*.

DomíNguez, A., Saenz-De-Navarrete, J., De-Marcos, L., FernáNdez-Sanz, L., PagéS, C., & MartíNez-HerráIz, J.-J. (2013). Gamifying learning experiences: Practical implications and outcomes. *Computers & Education, 63*, 380-392.

Earnest, M. M., & Max, L. (2003). En route to the three-dimensional registration and analysis of speech movements: instrumental techniques for the study of articulatory kinematics. *Contemporary Issues in Communication Science and Disorders, 30*, 5-25.

Fanelli, G., Weise, T., Gall, J., & Van Gool, L. (2011). *Real time head pose estimation from consumer depth cameras*.

Fell, H., MacAuslan, J., Gong, J., Cress, C., & Salvo, T. (2006). *visiBabble for pre-speech feedback*.

Ferster, C. B. (1964). Positive reinforcement and behavioral deficits of autistic children *Conditioning Techniques in Clinical Practice and Research* (pp. 255-274): Springer.

Flores, E., Tobon, G., Cavallaro, E., Cavallaro, F. I., Perry, J. C., & Keller, T. (2008). *Improving patient motivation in game development for motor deficit rehabilitation*.

Forrest, K. (2003). Diagnostic criteria of developmental apraxia of speech used by clinical speech-language pathologists. *American Journal of Speech-Language Pathology, 12*(3), 376-380.

Frauenberger, C., Good, J., & Alcorn, A. (2012). *Challenges, opportunities and future perspectives in including children with disabilities in the design of interactive technology*.

Frauenberger, C., Good, J., & Keay-Bright, W. (2011). Designing technology for children with special needs: bridging perspectives through participatory design. *CoDesign, 7*(1), 1-28.

Freund, Y., & Schapire, R. E. (1995). *A desicion-theoretic generalization of on-line learning and an application to boosting*.

Gabbard, J. L., Hix, D., & Swan, J. E. (1999). User-centered design and evaluation of virtual environments. *IEEE computer Graphics and Applications, 19*(6), 51-59.

Geng, L. (2012). ORAL MOTOR DYSFUNCTION; EXERCISES AND THERAPY FOR AUTISM AND APRAXIA. Retrieved from http://pursuitofresearch.org/2012/07/22/oral-motor-dysfunction-exercises-and-therapy-for-autism-and-apraxia/

Georgopoulos, V. C., Malandraki, G. A., & Stylios, C. D. (2003). A fuzzy cognitive map approach to differential diagnosis of specific language impairment. *Artificial intelligence in Medicine, 29*(3), 261-278.

Gros, B. (2003). The impact of digital games in education. v. 8, n. 7: jul.

Gu, H., Su, G., & Du, C. (2003). Feature points extraction from faces. *Image and Vision Computing NZ*, 154-158.

Hamidi, F. (2016). A LIVING MEDIA SYSTEM FOR MOTIVATING TARGET APPLICATION USE FOR CHILDREN. Retrieved from https://yorkspace.library.yorku.ca/xmlui/handle/10315/32261?show=full

Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., & Alhasanat, A. A. (2014). Solving the problem of the k parameter in the KNN classifier using an ensemble learning approach. *arXiv preprint arXiv:1409.0919*.

Hayden, D., & Stockman, I. (2004). PROMPT: A tactually grounded treatment approach to speech production disorders. *Movement and action in learning and development: Clinical implications for pervasive developmental disorders*, 255-297.

Hayden, D. A., & Square, P. A. (1994). Motor speech treatment hierarchy: A systems approach. *Clinics in Communication Disorders*, *4*(3), 162-174.

Heimann, M., Nelson, K. E., Tjus, T., & Gillberg, C. (1995). Increasing reading and communication skills in children with autism through an interactive multimedia computer program. *Journal of autism and developmental disorders*, *25*(5), 459-480.

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

Hodge, M. M. (1998). Developmental coordination disorder: A diagnosis with theoretical and clinical implications for developmental apraxia of speech. *SIG 1 Perspectives on Language Learning and Education, 5*(2), 8-12.

Hoque, M. E., Lane, J. K., El Kaliouby, R., Goodwin, M., & Picard, R. W. (2009). Exploring speech therapy games with children on the autism spectrum.

Hummels, C., Van der Helm, A., Hengeveld, B., Luxen, R., Voort, R., Van Balkom, H., & De Moor, J. (2006). Explorascope: an interactive, adaptive educational toy to stimulate the language and communicative skills of multiple-handicapped children. *Proceedings ArtAbilitation*, 6-24.

Johnson, W. L., Vilhjálmsson, H. H., & Marsella, S. (2005). *Serious games for language learning: How much game, how much AI?*

Koegel, R. L., O'dell, M. C., & Koegel, L. K. (1987). A natural language teaching paradigm for nonverbal autistic children. *Journal of autism and developmental disorders, 17*(2), 187-200.

Kortemeyer, G., Fish, J., Hacker, J., Kienle, J., Kobylarek, A., Sigler, M., Wierenga, B., Cheu, R., Kim, E., & Sherin, Z. (2013). Seeing and experiencing relativity—A new tool for teaching? *The Physics Teacher, 51*(8), 460-461.

Köstinger, M., Wohlhart, P., Roth, P. M., & Bischof, H. (2011). *Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization*.

Kumar, N., Berg, A. C., Belhumeur, P. N., & Nayar, S. K. (2009). *Attribute and simile classifiers for face verification*.

Lan, T., Aryal, S., Ahmed, B., Ballard, K., & Gutierrez-Osuna, R. (2014). *Flappy voice: an interactive game for childhood apraxia of speech therapy*.

Lohse, K., Shirzad, N., Verster, A., Hodges, N., & Van der Loos, H. M. (2013). Video games and rehabilitation: using design principles to enhance engagement in physical therapy. *Journal of Neurologic Physical Therapy, 37*(4), 166-175.

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). *The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression*.

Maas, E., Robin, D. A., Hula, S. N. A., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17*(3), 277-298.

Maassen, B., & van Lieshout, P. (2010). *Speech motor control: New developments in basic and applied research*: Oxford University Press.

MacLachlan, C., & Howland, H. C. (2002). Normal values and standard deviations for pupil diameter and interpupillary distance in subjects aged 1 month to 19 years. *Ophthalmic and Physiological Optics, 22*(3), 175-182.

Magerkurth, C., Stenzel, R., & Prante, T. (2003). STARS-a ubiquitous computing platform for computer augmented tabletop games. *Proceedings of Video Track of Ubiquitous Computing (UBICOMP'03)*.

Moore, M., & Calvert, S. (2000). Brief report: Vocabulary acquisition for children with autism: Teacher or computer instruction. *Journal of autism and developmental disorders, 30*(4), 359-362.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence, 31*(4), 607-626.

Murphy, N. A., & Carbone, P. S. (2008). Promoting the participation of children with disabilities in sports, recreation, and physical activities. *Pediatrics, 121*(5), 1057-1061.

Murray, T. G., & Parker, V. (2004). Integration of Computer-Based Technology into Speech-Language Therapy. *Educational Technology, 31*, 53-59.

Öster, A.-M., House, D., Hatzis, A., & Green, P. (2003). Testing a new method for training fricatives using visual maps in the Ortho-Logo-Paedia project (OLP). *Proc of Fonetik 2003, Umeå University, Dept of Philosophy and Linguistics PHONUM, 9*.

Padilla, R., Costa Filho, C., & Costa, M. (2012). Evaluation of haar cascade classifiers designed for face detection. *World Academy of Science, Engineering and Technology, 64*.

Papageorgiou, E., Spyridonos, P., Glotsos, D. T., Stylios, C. D., Ravazoula, P., Nikiforidis, G., & Groumpos, P. P. (2008). Brain tumor characterization using the soft computing technique of fuzzy cognitive maps. *Applied Soft Computing, 8*(1), 820-828.

Papageorgiou, E. I., Papadimitriou, C., & Karkanis, S. (2009). *Management of uncomplicated urinary tract infections using fuzzy cognitive maps*.

Papageorgiou, E. I., Stylios, C. D., & Groumpos, P. P. (2003). An integrated two-level hierarchical system for decision making in radiation therapy based on fuzzy cognitive maps. *IEEE transactions on Biomedical Engineering, 50*(12), 1326-1339.

Parnandi, A., Karappa, V., Son, Y., Shahin, M., McKechnie, J., Ballard, K., Ahmed, B., & Gutierrez-Osuna, R. (2013). *Architecture of an automated therapy tool for childhood apraxia of speech*.

Pinnell, C. (2015). Computer games for learning: An evidence-based approach. *Educational Technology & Society, 18*(4), 523-524.

Piper, A. M., O'Brien, E., Morris, M. R., & Winograd, T. (2006). *SIDES: a cooperative tabletop computer game for social skills development*.

Prensky, M., & Prensky, M. (2007). *Digital game-based learning* (Vol. 1): Paragon house St. Paul, MN.

Rao, R. P. (2013). *Brain-computer interfacing: an introduction*: Cambridge University Press.

Rapp, V., Senechal, T., Bailly, K., & Prevost, L. (2011). *Multiple kernel learning svm and statistical validation for facial landmark detection*.

Roweis, S. T., & Alwan, A. (1997). *Towards articulatory speech recognition: learning smooth maps to recover articulator information*. Paper presented at the Eurospeech, Rhodes, Greece. (pp. 1227-1230)

Rupp, R., Kleih, S. C., Leeb, R., Millan, J. d. R., Kübler, A., & Müller-Putz, G. R. (2014). Brain–computer interfaces and assistive technology *Brain-Computer-Interfaces in their ethical, social and cultural contexts* (pp. 7-38): Springer.

Saragih, J., & Goecke, R. (2007). *A nonlinear discriminative approach to AAM fitting*.

Saragih, J. M., Lucey, S., & Cohn, J. F. (2011). Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision, 91*(2), 200-215.

Shahin, M., Ahmed, B., Parnandi, A., Karappa, V., McKechnie, J., Ballard, K. J., & Gutierrez-Osuna, R. (2015). Tabby Talks: An automated tool for the assessment of childhood apraxia of speech. *Speech communication, 70*, 49-64.

Shibata, T., Mitsui, T., Wada, K., Touda, A., Kumasaka, T., Tagami, K., & Tanie, K. (2001). *Mental commit robot and its application to therapy of children*.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., & Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM, 56*(1), 116-124.

Square, P. A., Chumpelik, D. A., Morningstar, D., & Adams, S. (1986). Efficacy of the PROMPT system of therapy for the treatment of acquired apraxia of speech: A follow-up investigation. *Clinical aphasiology, 16*, 221-226.

Squire, K., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight, 3*(1), 5-33.

Strand, E. A. (1995). *Treatment of motor speech disorders in children*.

Styblinski, M., & Meyer, B. (1991). Signal flow graphs vs fuzzy cognitive maps in application to qualitative circuit analysis. *International Journal of Man-Machine Studies, 35*(2), 175-186.

Stylios, C. D., Georgopoulos, V. C., Malandraki, G. A., & Chouliara, S. (2008). Fuzzy cognitive map architectures for medical decision support systems. *Applied Soft Computing, 8*(3), 1243-1251.

Svirsky, M. A., Robbins, A. M., Kirk, K. I., Pisoni, D. B., & Miyamoto, R. T. (2000). Language development in profoundly deaf children with cochlear implants. *Psychological science, 11*(2), 153-158.

Tabachnick, B. G., Fidell, L. S., & Osterlind, S. J. (2001). Using multivariate statistics.

Takada, K., Miyawaki, S., & Tatsuta, M. (1994). The effects of food consistency on jaw movement and posterior temporalis and inferior orbicularis oris muscle activities during chewing in children. *Archives of oral biology*, *39*(9), 793-805.

Tomar, S. (2006). Converting video formats with FFmpeg. *Linux Journal*, *2006*(146), 10.

Uppal, S., Kohen, D., & Khan, S. (2008). Educational services and the disabled child. *Health Analysis and Measurement Group. Ottawa: Statistics Canada*.

UrbanKowboy. (2010). Childhood Apraxia of Speech 3 year old girl. Retrieved from https://www.youtube.com/watch?v=szjfC9K190U&t=45s

Uřičář, M., Franc, V., & Hlaváč, V. (2012). Detector of facial landmarks learned by the structured output SVM. *VIsAPP*, *12*, 547-556.

Valstar, M., Martinez, B., Binefa, X., & Pantic, M. (2010). *Facial point detection using boosted regression and graph models*.

Van Velsen, L., Van Der Geest, T., Klaassen, R., & Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *The knowledge engineering review*, *23*(03), 261-281.

Vaughan, T. M., Heetderks, W., Trejo, L., Rymer, W., Weinrich, M., Moore, M., Kübler, A., Dobkin, B., Birbaumer, N., & Donchin, E. (2003). Brain-computer interface technology: a review of the Second International Meeting.

Vicsi, K., Roach, P., Öster, A., Kacic, Z., Barczikay, P., Tantos, A., Csatári, F., Bakcsi, Z., & Sfakianaki, A. (2000). A multimedia, multilingual teaching and training system for children with speech disorders. *International Journal of speech technology, 3*(3-4), 289-300.

Viola, P., & Jones, M. (2001). *Rapid object detection using a boosted cascade of simple features*.

Ward, R., Leitão, S., & Strauss, G. (2014). An evaluation of the effectiveness of PROMPT therapy in improving speech production accuracy in six children with cerebral palsy. *International journal of speech-language pathology, 16*(4), 355-371.

Whalen, C., & Schreibman, L. (2003). Joint attention training for children with autism using behavior modification procedures. *Journal of Child Psychology and Psychiatry, 44*(3), 456-468.

Wiepert, S. L., & Mercer, V. S. (2002). Effects of an increased number of practice trials on Peabody Developmental Gross Motor Scale scores in children of preschool age with typical development. *Pediatric Physical Therapy, 14*(1), 22-28.

Williams, A. L., McLeod, S., & McCauley, R. J. (2010). *Interventions for Speech Sound Disorders in Children*: ERIC.

Wilson, P. I., & Fernandez, J. (2006). Facial feature detection using Haar classifiers. *Journal of Computing Sciences in Colleges, 21*(4), 127-133.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., & Vaughan, T. M. (2002). Brain–computer interfaces for communication and control. *Clinical neurophysiology*, *113*(6), 767-791.

Wu, C. (2013). *Towards linear-time incremental structure from motion*.

Yang, M.-H., Kriegman, D. J., & Ahuja, N. (2002). Detecting faces in images: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *24*(1), 34-58.

Zakari, H. M., Ma, M., & Simmons, D. (2014). *A review of serious games for children with autism spectrum disorders (ASD)*.

Ziegler, W. (2008). Apraxia of speech. *Handbook of clinical neurology*, *88*, 269-285.

Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). *Research through design as a method for interaction design research in HCI*.

Zyda, M. (2005). From visual simulation to virtual reality to games. *Computer*, *38*(9), 25-32.