Investigation of Gesture Control for Articulatory Speech Synthesis with a Bio-mechanical Mapping Layer

by

Johnty Yizhong Wang

BASc, University of British Columbia, 2008

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Applied Science

in

THE FACULTY OF GRADUATE STUDIES
(Electrical and Computer Engineering)

The University Of British Columbia
(Vancouver)

September 2012

Abstract

In the process of working with a real-time, gesture controlled speech and singing synthesizer used for musical performance, we have documented performer related issues and provided some suggestions that will serve to improve future work in the field from an engineering and technician's perspective. One particular, significant detrimental factor in the existing system is the sound quality caused by the limitations of the one-to-one kinematic mapping between the gesture input and output. In order to solve this a force activated bio-mechanical mapping layer was implemented to drive an articulatory synthesizer, and the results were and compared with the existing mapping system for the same task from both the performer and listener perspective. The results show that adding the complex, dynamic bio-mechanical mapping layer introduces more difficulty but allows a greater degree of expression to the performer that is consistent with existing work in the literature. However, to the novice listener, there is no significant difference in the intelligibility of the sound or the perceived quality. The results suggest that for browsing through a vowel space force and position input are comparable when considering output intelligibility alone but for expressivity a complex input may be more suitable.

Preface

This thesis is written from two perspectives. The first perspective is based on learnings and recommendations from the author's role as an engineer and studio technician with a musical background working in the building, deployment and technical support of a gesture controlled voice instrument (Chapter 3). The second perspective is based on the findings in the process of improving the synthesis system through exploring the use of an articulatory, force activated bio-mechanical mapping and the main findings are to be presented in the 12th International Conference on *New Interfaces for Musical Expression (NIME)* (Chapter 4).

Publications

The following publications were the result of direct and indirect work related to this thesis:

An extended abstract documenting the main proposed implementation [8], a poster demonstration [43] of a hybrid touch+force controller, a demonstration [7] of a mobile singing synthesis system and a performance[27] for "Interactivity" and a paper [42].

Ethics

Procedures and protocols employed in the work described by this thesis were approved by the UBC Behavioural Research Ethics Board, Certificate Number H11-02370.

Table of Contents

List of Tables

List of Figures

Glossary

DIVA	Digital Ventriloquized Actor
FSR	Force Sensitive Resistor
GUI	Graphical User Interface
HCI	Human Computer Interaction
IP	Internet Protocol
NIME	New Interface(s) for Musical Expression
MIDI	Musical Instrument Digital Interface
OSC	Open Sound Control
RBF	Radial Basis Function
UDP	User Datagram Protocol
USB	Universal Serial Bus
UI	User Interface

Acknowledgments

I would like to thank Dr. Sidney Fels, Dr. Nicolas d'Alessandro and Dr. Robert Pritchard for providing their supervision, support and guidance over the years. Thanks Mom, Dad, and all my friends for your love. I am also extremely grateful for the wonderful people at MAGIC, HCT and SUBCLASS that I've had the pleasure to work with.

Chapter 1

Introduction

The Voice

The human voice is one of the most interesting sonic instruments in existence. The voice is intimate, in the physical sense that it resides within the body, as well as its role as a communicative medium that forms the basis of social interaction. The voice is expressive, in the sense that a wide variety of sounds can be produced with subtle nuance to convey a vast spectrum of acoustic output. The voice is refined, in the sense that it takes many years to become a fluent speaker or singer. The voice is highly coordinated, in the sense that a huge number of muscles in the body must move in precisely timed trajectories to generate even a simple utterance. Finally, the voice is ubiquitous, in the sense that every (non-disabled) person is a trained listener and performer of the voice instrument at some level.

Therefore, the study of the voice is of interest to many including engineers, health practitioners, linguists, and musicians. Engineers work on models that synthesize or recognize voices in order build communication tools and user interfaces. Clinicians are concerned with the physical anatomy of the vocal apparatus for corrective procedures and rehabilitation. Linguists work with the voice since spoken word is an integral part of all languages. Musicians use the voice as an instrument, converting artistic intent through oral output. In each of these examples, working with the voice requires

knowledge in more than one domain and as such, the study of voice is an interdisciplinary field.

The DiVA Project

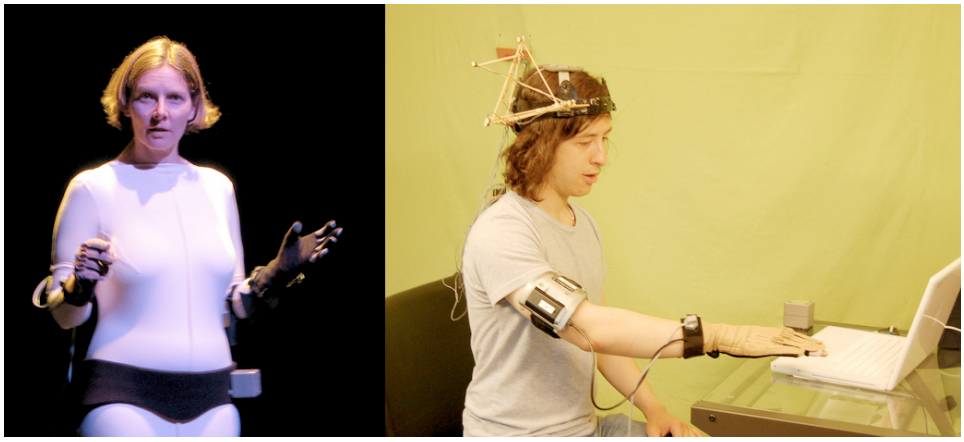


Figure 1.1: The DiVA system used during performance and speech research

The Digital Ventriloquized Actor (DIVA) project is an interdisciplinary research project centred around the voice. The DIVA is a gesture controlled speech and singing synthesizer used primarily for musical performance as a new instrument. This thesis describes the lessons learnt while working with the DIVA over the last 3 years from the perspective of an engineer building and maintaining the system, a technician supporting musicians and artists in performing with the system and a researcher in attempting to integrate the experience and explore methods to improve gesture to voice controllers and gain a better understanding of the human voice.

1.1 Research Contributions

The main research contributions of the work described in this thesis are as follows:

- By using an experimental system for performance, a number of insights have been collected and documented through building the system, setting it up for rehearsals and performance, and interaction with performers and the audience
- In the process of improving the system, mapping concepts from the existing literature for new instrument design were applied specifically in the context of gesture controlled speech synthesis

1.2 Thesis Structure

The structure of this thesis is as follows: in Chapter 2, relevant existing work in the literature will be presented. In Chapter 3, an analysis of the existing DIVA system will be provided, documenting the lessons learnt from using a gesture to speech synthesizer in a performance setting, and motivation for further exploration of mapping and synthesis methods. Then,the implementation of a new mapping and synthesis system is described in Chapter 4 and a comparison with the existing scheme presented in Chapter 5. Finally, concluding remarks and suggested future work are presented in Chapter 6.

Chapter 2

Background and Related Work

In order to understand the significance of the work presented in this thesis, it is useful to first obtain a basic level of understanding on the underlying mechanism of voice production, and well as voice synthesis. This chapter first briefly describes how speech is produced in the human body, and then some of the methods with which speech can be produced artificially (synthesized) along with some historical examples. Then, a summary of these systems is provided with a discussion and comparison of the challenges of gesture control of speech. Finally, the topic of using gestures to control speech production models is placed in context with relevant research in the new instrument field on input mapping for general music and sound synthesis.

2.1 Human Speech Production

A simplified diagram of the human vocal process is shown in Figure 2.1 and can be broken down into four major steps. For a more in-depth description, Titze's book [33] describes the subject in much greater detail.

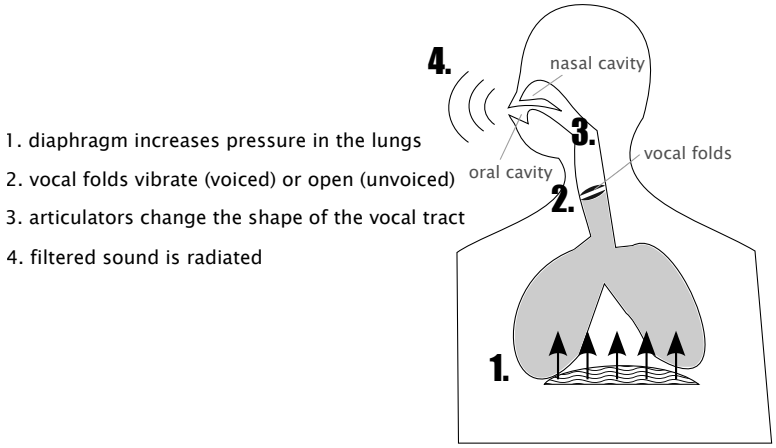


Figure 2.1: Simplified diagram of the human vocal process

1. The Lung and Diaphragm

The vocal process starts in the lungs, the vital organ of the human body that is responsible for breathing. The lungs are responsible for providing life sustaining oxygen to the bloodstream during inhalation, and the expiration of waste carbon dioxide during exhalation. The voice, with the exception of gasping sounds, occurs during the exhalation cycle of the breath. The inward and outward flow of air from the lungs is caused by contraction and expansion of lung volume that create differences between internal (lung) and external (atmospheric) pressure. The diaphragm, a sheet muscle under the ribs above the abdominal cavity, is responsible for this expansion and contraction of the lungs. Therefore, one could say that the diaphragm is the starting point of the vocal process. It is not surprising that singers and professional speakers spend a considerable amount of time practising their breathing (control of the diaphragm) as a part of their musical training.

2. The Vocal Folds

Once there is an excess of pressure in the lung (during the exhalation phase) air flows upwards past the vocal folds - a set of thin overlapping membranes that can open or close. The vocal folds are attached to the the arytenoid cartilage in the front and thyroid cartilage at the back. The area covered by the vocal folds and the space around it is called the glottis. As the structures move the posture and tension of the folds change, creating differences in physical response to the air flow. Under the right conditions, the vocal folds will undergo opening and closing at a regular interval due to the air flow pushing the folds open and the elastic forces pulling them back together - a forced oscillation. The oscillations results in an airflow profile over time that typically looks like Figure 2.2, and this is an example of a glottal source waveform. The steep downward slope of the closing phase (downward part of the waveform) contributes to a wide frequency range in the spectrum (in the same way an impulse function is infinitely wide). The exact shape of the waveform determines its spectral composition which affects the vocal quality, and is controlled by a large number of parameters including the lung pressure and positions of various structures that change the tension and rest position of the vocal folds.

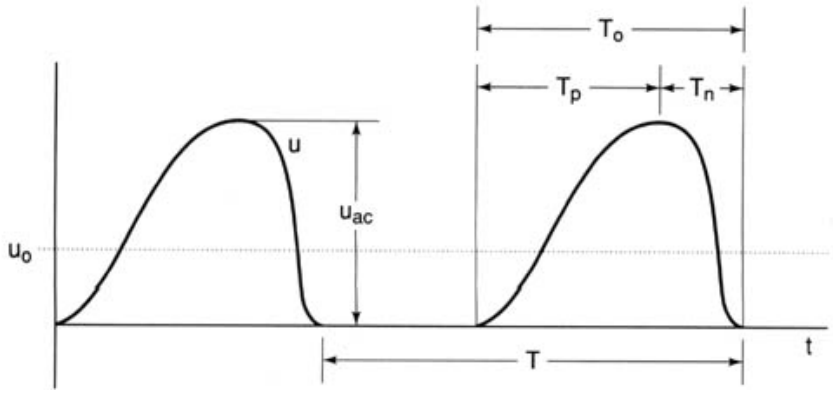


Figure 2.2: The glottal source waveform [15]

When the vocal folds are vibrating with a constant period, voiced sounds

with a discernible fundamental frequency (or pitch) are produced. Vowels and voiced consonants (such as "m" or "n") belong in this category of sounds. For unvoiced sounds such as consonants without discernible pitch, the vocal folds are completely open and air flows through the folds unobstructed. Since the configuration of the vocal folds is a continuous process, it is possible to have an "in between" glottal wave that is always partially open at its minimum (no complete closure) but with considerable oscillation behaviour as well, such as in the case of a "breathy" sound.

3. The Vocal Tract

The output from the glottis then goes through the pharyngeal and laryngeal cavities and then the oral and nasal cavities, collectively known as the vocal tract (which is also the upper airway). The varying cross sections along the vocal tract attenuate certain frequencies and amplify others. When a wide band signal (such as a glottal waveform) is passed through the tract, a particular vocal tract shape will create a number of resonant peaks in the spectrum. The amplitude, frequency, and bandwidth of the peaks are known as formants and their configuration define different sounds that are used to make up distinct phonemes, the basic building blocks of spoken vocabulary.

The way in which the vocal tract shape is manipulated is through the activation of various articulators (muscles) connected to the internal structures of the vocal tract. In the upper airway, the largest and most dominant structure for controlling the overall tract shape is the tongue. With a large number of muscles and a huge range of movement, the tongue is responsible for the most significant variations in the overall shape of the vocal tract.

As described in the previous section, for unvoiced sounds there is no vibration of the vocal folds and the vocal folds have minimum effect on the air flow, but constrictions at certain points in the tract create turbulence which generates high frequency noise responsible for making consonants sounds like "s" and "f". These sounds contain high frequencies caused by turbulent flow at the constriction points and are called fricatives. At the branching point between the oral and nasal cavity, the velum (soft palate) can function

as a switch which can block off airflow into the nasal passage to create oral sounds. When open, airflow through the nasal passage creates sounds such as "m" and "n" which are, aptly, called nasals.

4. Lip Radiation

In the final step of the vocal process, the filtered sound from the vocal tract is radiated through the teeth and lips into the atmosphere, creating acoustic pressure waves that are radiated outwards from the speaker. Another important function of the lip opening is during the creation of plosive sounds. Plosive sounds (such as "p" and "b") are created by first blocking off airflow completely by closing the mouth/lips followed by a sudden opening that creates an explosive burst of pressure that had been built up prior to the release.

Bringing it Together

Based on the above description, it should be plain to see that the act of speech and singing requires coordinated movements from a large number of structures related to the vocal process. Based on years of practice and refinement, when we speak this comes almost instinctively, but attempting to fully model the production and control of speech is highly challenging due to the complexity of the system consisting of these closely coupled components that have to work together with precise configuration and timing.

2.2 Speech Synthesis

The act of artificially producing (synthesizing) speech involves ultimately recreating the acoustic pressure waves that are radiated from the human body. This section describes some of the production models that have been employed to generate speech sounds. For each production model, there is also a directly coupled control process that affects how the user interacts with the system when attempting to speak with them, and these control systems will also be described. The production models are presented in order of increasing abstraction. First, physical mechanical models that attempt

to imitate the physical properties of the vocal tract are described, followed by articulatory synthesizers that model the various physical structures using software simulation. Then, a higher level of abstraction by representing the vocal process as a source-filter system is described. Finally, the pressure vs time model, based on the explicit specification of the output acoustic pressure waves produced is presented.

2.2.1 Mechanical Articulatory Models

Before the age of analogue and digital electronics, attempts have been made to recreate the human voice using physical mechanical models that imitated the features of the vocal tract. Mechanical systems, the earliest examples of speech synthesis, attempted to match the functionality of the physical anatomy by building a model vocal tract with an air pressure source to simulate the lung pressure, a vibrating membrane to simulate the vocal folds, and a deformable cavity that generated the resonance of the vocal tract made up of the laryngopharyngeal, oral and nasal structures.

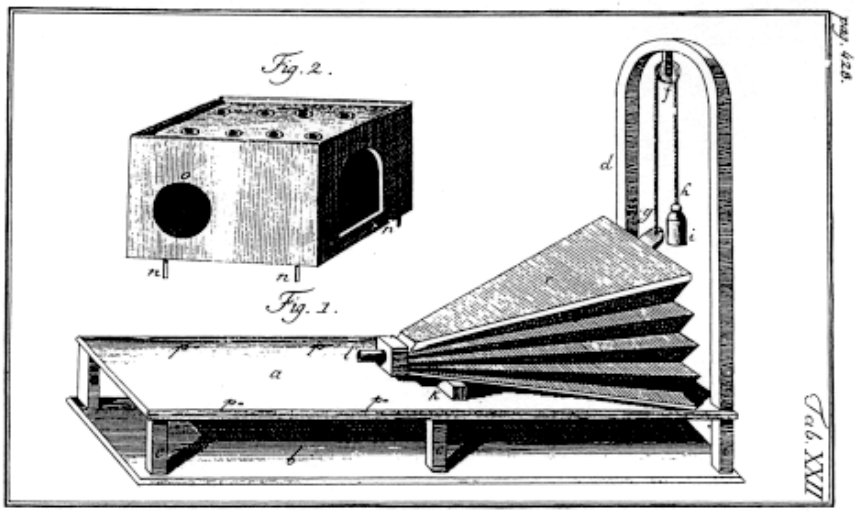
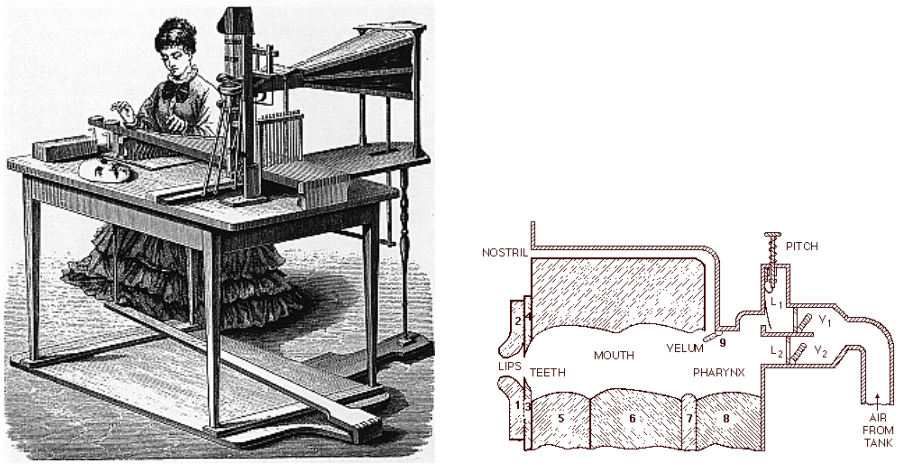


Figure 2.3: Bellow and box cavity components of the von Kempelen machine [34]

The von Kempelen machine [40], as shown in Figure 2.3, is an example

of an early speech synthesizer. Built in the 1700's, the device consists of a bellow that provided the lung pressure that caused a reed to vibrate, and a boxed cavity with switchable branches to simulate the the vocal tract. A deformable leather tube allows the effective tract area to change to model the changing resonance of the tract. Later, more refined systems were built by Faber (1846) that included a tongue model and Riesz (1937) [34] with a more natural vocal tract model (Figure 2.4).



(a) Faber's Euphonium **(b)** Riesz's Vocal Tract

Figure 2.4: Other mechanical synthesizers [34]

These devices provide the physical mechanisms that mimic the vocal tract in acoustic response, but of course the activation requires user interaction for generating the equivalent glottal source and tract movements. The user directly manipulated these devices using mechanical actuators: bellows, levers, switches, foot pedals and deformable tubes that generated, or affected the output sound. The physical nature of these systems provides a direct correlation between the user input and the sound output. Mechanical levers and coupling provide mapping between manual user input and the articulators. Because these are physical acoustic instruments, their output is generally quite constrained to the physical properties of the components. In more recent times, robotic systems such as the Waseda Talker [17] has been built. These systems use electronically actuated physical models which

means sequences of articulator commands can be generated from a computer, allowing for automated playback of a large number of control signals at a time. Determining a way for a user to directly manipulate these signals for real time control is still an open ended question.

2.2.2 Software Articulatory Models

An software articulatory synthesizer, as its name implies, models the articulators in the human vocal mechanism to synthesize speech. However, unlike a mechanical model, the simulation is done in software. [25] provides a more comprehensive overview of the developments in articulatory synthesis and a description of various technical challenges, and VocaltractLab [1], as shown in Figure 2.5, is an example of the state of art in articulatory synthesis.

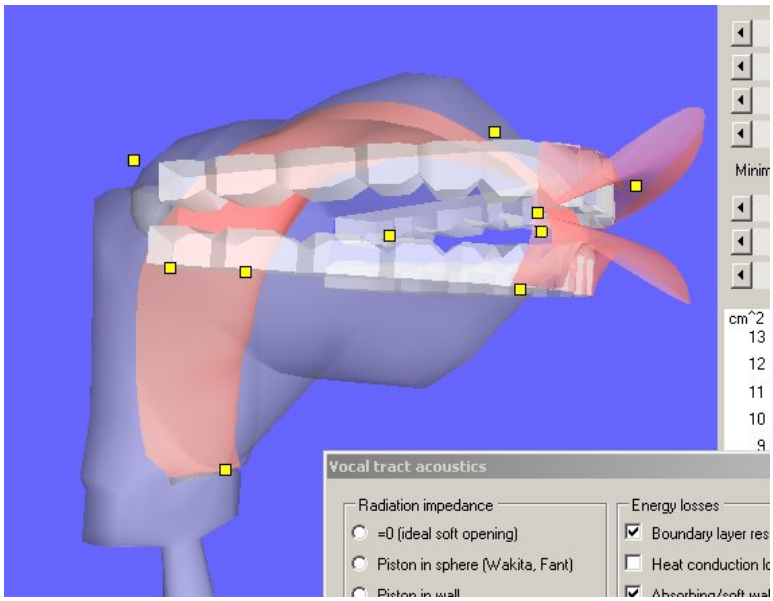


Figure 2.5: Screenshot of the VocaltractLab [1] tube model

Generally, an software articulatory synthesizer contains the following components: a source model, tube model, and synthesis model. The following sections describe the basic functioning of each component.

Source excitation

The source model simulates the effect of airflow through the glottis that generates the glottal source waveform (Figure 2.2). There are two main methods of generating this waveform: parametric and dynamic. In parametric models the glottal waveform is defined explicitly as a function over time, and the shape of the function is adjusted based on desired vocal qualities [29]. In dynamic representations such as the 2-mass model [23], the glottal waveform is derived from modelling the vocal folds as a mass spring system and solving the basic Newtonian equations. While it would be possible to tackle the source generation at an even more fundamental perspective (and indeed the entire acoustic simulation) from first-principles airflow modelling, the extensive computation times and incompleteness of such models make them unsuitable for synthesis at the speech level.

Tube model

The tube model simulates the relevant physical properties of the vocal tract in the form of a geometry that can be used to calculate the resonant parameters. Similar to the source excitation, the geometry can also be parametric or dynamic. In parametric representations, the explicit tube geometry is defined and varied over time based on pre-computed functions, while in dynamic bio-mechanical models the geometry is computed based on physical principles using numerical methods.

Synthesizer

Based on the source excitation waveform and the tube model configuration, compute the output audio samples that can be played back on a computers digital to analog converter (sound card). The synthesis can be offline, which generates a series of samples that can be stored for later playback, or in real-time where the samples are generated on the fly and sent directly to the sound card at the same rate as the real-time audio output. For the latter case it is critical that the system can produce the samples at a sufficient rate.

Controlling Articulatory Synthesis

Compared to a physical mechanical system, a software articulatory synthesizer allows easy adjustment of system parameters to create desired ranges of behaviour. Sequences of control for the articulators can be automated to reproduce natural speech trajectories [24]. Automation of control values can reduce the input bandwidth required, but a trade-off must be made in the degree of control. As an example, one extreme case is for a single button press to trigger an entire sequence of trajectories to control all the synthesis parameters, allowing pre-determined but inflexible input. On the other extreme, if each individual synthesis parameter is exposed to explicit control, a huge number of inputs would be required and likely to exceed the number of controls a user can effectively actuate at a time.

2.2.3 Filter Models

A higher level abstraction of the vocal process is to look at the source-filter nature of the vocal process. Instead of working at the physical tube model level, the source-filter production model deals directly with the resultant resonant property of the vocal tract and its effect on the frequency spectrum of the glottal source from the perspective of acoustic targets observed during actual speech. The source-filter model can be implemented in electronic circuits, or digitally on a computer.

Analog Electrical Circuits

An example of an analog circuit source-filter speech synthesizer is Dudley's Voder [10] as shown in Figure 2.6. The Voder consists of an oscillator to simulate the glottal source, a noise generator for unvoiced sounds and filters to represent the vocal tract shape. The system was controlled using a food pedal and keyboard. The foot pedal controlled the frequency of the oscillator that determined the pitch of the glottal source and the keyboard selected various resonant parameters. Each key's position determined the amplitude of a particular frequency band so the user had direct control over the frequency response. This meant that the formants had to be explicitly

defined by the user, and required a significant amount of practice. Typically, the Voder required the operator to train for a year before becoming a fluent speaker.

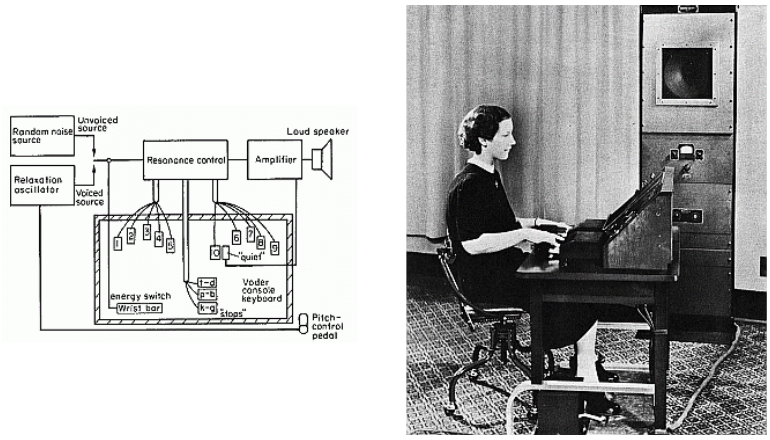


Figure 2.6: Dudley's Voder. [34]

Software Formant Synthesis

A similar kind of source-filter synthesis can be implemented in software. In a software formant synthesizer, a fixed or parametric glottal source waveform is bandpass filtered according to the spectral specification of the desired output using digital signal processing techniques. Just like the analog electronic version, descriptive parameters (and corresponding trajectories) are needed to produce the desired output. However, such digital systems are easy to control using pre-determined control trajectories. Given the correct set of inputs, these systems can produce good sounding speech. An example is the Holmes synthesizer [20] where pre-tuned sets of inputs can be used by the system to produce high quality and somewhat natural sounding output. However, when attempting to control the synthesis in real time, there is a similar challenge of mapping input with limited bandwidth into a large number of formant parameters.

2.2.4 Pressure vs Time Models

Since the ultimate result of the human vocal process is to create acoustic pressure waves that propagate through the air, a direct method of recreating speech is to generate these pressure waves through audio recording and playback, as shown in Figure 2.7.

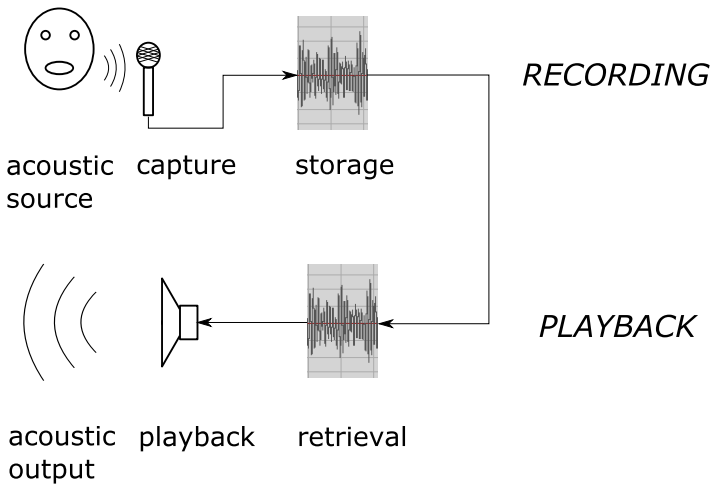


Figure 2.7: Overview of the audio recording and playback process

During the recording process, acoustic pressure signals are captured using a sensor (microphone), and then converted into a representation suitable for the storage medium. The storage medium can be analog (phonograph or magnetic tape), or digital (Compact Disc or digital memory), but the fundamental goal is the same: to provide an explicit representation of the acoustic pressure waves over time that can be retrieved and reproduced at a later time. In the playback process, the stored representation of the pressure waves are retrieved and then converted back using an actuator (speaker). Figure 2.8 shows a small segment of digitally recorded segment of "ah". The horizontal axis represents time and the vertical axis is the digital represen-

tation of the measured acoustic pressure.

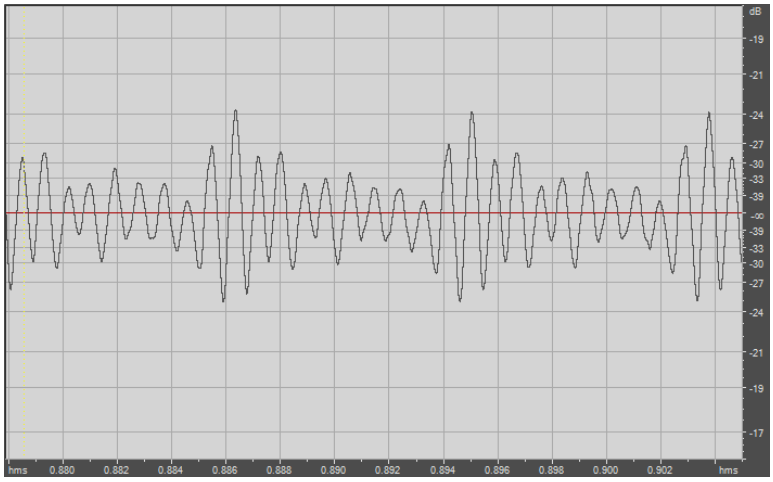


Figure 2.8: "Zoomed in" display of the "ah" sound

Of course, the playback of segments of audio does not allow for flexibility in terms of speech output produced. Nevertheless, systems based on the selective recording/playback of words, phrases and even entire sentences are employed in automated announcement and interactive user interfaces. One method of improving the degree of control is to segment recordings into short segments at the phoneme (or smaller) level and then recombining them based on the intended output.

By segmenting the recording at appropriate periods, manipulating the samples and blending with others, it would be possible to generate a wide combination of words that may not have been in the original recorded dataset. This method is known as concatenative synthesis. Using suitable rules to select and stitch samples together can yield natural sounding speech, but the process does not lend to direct real-time manipulation of the underlying vocal characteristics required for expressive speech and singing. However, systems such as CataRT [30] have used concatenative based synthesis for real-time performance - in this particular case, the performer interaction is built upon the selection and manipulation of utterances rather than control of actual vocal parameters. Another system, the RAMCESS

synthesizer [5], provides an example of using concatenative synthesis at the glottal source level but then using a formant-based filtering system for vocal tract manipulation - essentially a hybrid approach. The Handsketch controller based on a Wacom tablet [6], as shown in Figure 2.9 was used to control this synthesizer. The Handsketch mapped the X-Y position of the tablet stylus to formant specifications and pitch, while the pressure of the stylus determined the glottal source parameters attributed to vocal quality.

Figure 2.9: The Handsketch controller for singing synthesis

2.3 Summary of Gesture to Speech Systems

A general overview of the human speech production mechanism and a survey of gesture to speech systems have been provided in this chapter. From the description of existing systems, we can classify these systems based on their production and control methods, as shown in Table 2.1.

Table 2.1: Gesture to Speech systems

System	**Production**	**Control**
von Kempelen machine	Mechanical	physical manipulation
Faber's Euphonium	Mechanical	physical manipulation
Riesz	Mechanical	physical manipulation
Dudley's Voder	Source/Filter	buttons and switches
Dunn's Electric VT	Source/Filter	buttons and switches
d'Alessandro's Handsketch	Concatenative+ Source/Filter Hybrid	hand gesture digital mapping

While each production model has distinct features and requirements, a common challenge faced by speech synthesis is the method in which they are controlled. Generally, it is possible to either provide detailed input parameters to synthesizers that lead to well defined vocal trajectories, or to provide extensive control to individual synthesis parameters, but not both. As a result, there is generally a trade off between being able to output a flexible range of vocal sounds, or provide natural sounding output.

2.4 Gesture Mapping in the NIME Context

As one of the major challenges of gesture controlled speech is the mapping between input (gesture) and output (synthesis), the new instrument field provides some potential answers. New Interface(s) for Musical Expression (NIME) deals with the design, construction, practice, composition and performance of music on novel interfaces. Many NIMEs consist of new sensors (or traditional sensors used in new ways) connected to various synthesis techniques, and the behaviour of the instrument is determined largely by how the inputs are mapped to the synthesis parameters. An effective mapping

system allows the instrument to be intuitive and transparent to the listener while being controllable and expressive to the performer. However, because of the novel ways many NIMEs are constructed both on the sensing and synthesis side, it is often not the case. [21] proposed a mapping strategy shown in Figure 2.10 and suggests that the challenge of designing an effective NIME is to establish a set of *meaningful* parameters in the mapping layer.

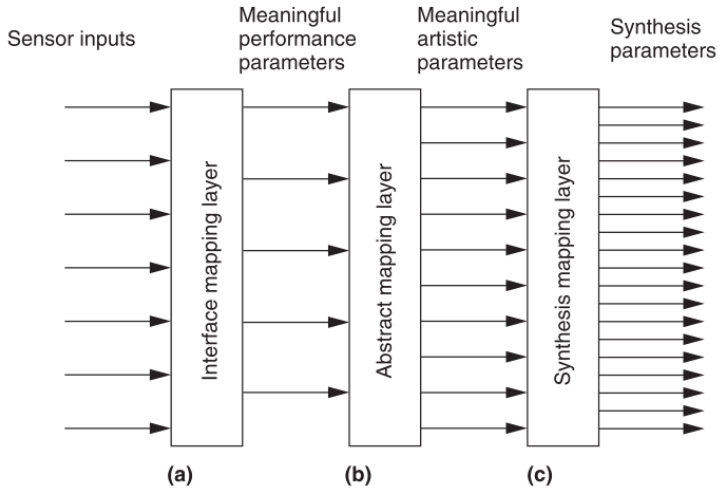


Figure 2.10: The "three layer mapping strategy" according to [21]

In terms of developing new musical interfaces, [26] provides a framework for selecting and evaluating musical input devices using tools from Human Computer Interaction (HCI). Using comparisons with existing HCI theories for movement and target acquisition a list of recommended musical tasks were selected. [18] develops the concept of "feature-based" synthesis, a formalized framework for mapping between acoustic and perceptual features onto synthesis. In [19] the same authors develop strategies for applying the concept in real time performance.

[38] categorizes musical functions into three basic groups: static (selection of range), relative dynamic (modulation of pitch, amplitude, or timbre) and absolute dynamic (selection of absolute pitch, amplitude, or timbre). Based on these categories, [41] provides physical descriptions of specific sen-

sor technologies, and their suitability for these tasks. Specifically, isometric force input (from sensors such as an Force Sensitive Resistor (FSR)) appears to be suitable for relative dynamic tasks while a position input (from a touch screen) appears to be better for absolute tasks. Regarding the mapping scheme of new instruments, [22] suggested that complex mappings are more expressive compared to a direct one-to-one mapping between gesture inputs and synthesis output.

While the findings in existing NIME literature inform the development of a voice instrument, when applying these concepts to speech there are many unanswered questions such as the categorization of speech/singing tasks within the same frameworks.

2.5 Chapter Summary

In this chapter, we introduced some basic theory of human speech production and provided examples of how speech is synthesized. From there we see that speech is a complex phenomenon requiring a large number of precisely coordinated movements. Existing speech synthesis research does not deal with the question of how to control synthesis parameters from real-time input gestures. In terms of gesture mapping for NIMEs, existing literature suggests that a complex, meaningful mapping layer is required for an effective instrument that is intuitive, transparent and expressive.

Chapter 3

Evaluation of the DiVA System

In this chapter, first an overview of the development of the DIVA system is presented. Then, the experience from working with the DIVA in performance is documented, describing various technical and musical issues encountered throughout the process. One of the issues discovered from working with the DIVA system is that the input mapping and synthesis system can be improved, and the final part of this chapter describes in further detail the motivation for exploring a new input mapping and synthesis approach.

3.1 History of the DiVA System

3.1.1 GloveTalk

GloveTalkI [13] was a gesture to word synthesizer that detected input hand gestures using an instrumented glove (VPL Dataglove) and mapped the gestures to target words for synthesis through a neural network. GloveTalkII [14] used a more advanced glove and mapped input gestures to synthesis parameters in real-time. Hand position and posture was detected by a Polhemus Tracker and Immersion Cyberglove, and the data was sent to a neural network that outputs formant parameters corresponding to specific

vowel and consonant target sounds to drive a formant synthesizer. Specific hand position and postures were defined as targets (sets of frequencies and amplitudes) that defined particular vowels and consonants and as the hand browsed through the input space the system blended between the formant parameters to create a continuous range of output. The neural network was trained by back-propagation using predefined pairs of hand position and posture combinations (input training sample) and its intended formant output (target).

3.1.2 GRASSP and DiVA

Based on GloveTalk, GRASSP [28], and its sucessor, DIVA1.0 [12], were used for artistic applications as a gesture controlled voice instrument. One of the key differences between DIVA1.0 and GloveTalkII was the way the neural network was trained, reducing the samples needed. DIVA1.0 was implemented using Max/MSP, a visual programming environment designed for music and multimedia. Due to the instability of input device drivers in Max/MSP, a decision was made to re-implement the system in a more robust package. DIVA2.0, the most recent stable version of the system used in the latest performances, reflects the porting and re-design of the collection of Max patches into a standalone C++ application.

Figure 3.1 shows a signal flow diagram of DIVA2.0. Various sensors for the two hands and foot function as input devices to drive a software formant synthesizer. The right hand sensors were the same as the ones used in GLoveTalkII (Polhemus Patriot tracker for hand position and CyberGlove for hand posture). The vertical position of the right hand controlled the fundamental frequency (pitch), while the horizontal position browsed between vowel targets in formant space when the hand is in an open posture. When the right hand is closed, specified closed hand postures mapped to consonants. The vowel and consonant targets were specified by formant parameters, and when the hand is in an in between position and posture, a continuous blending of the target parameters was performed based on a distance-based function. The left hand triggered the plosives by the means of

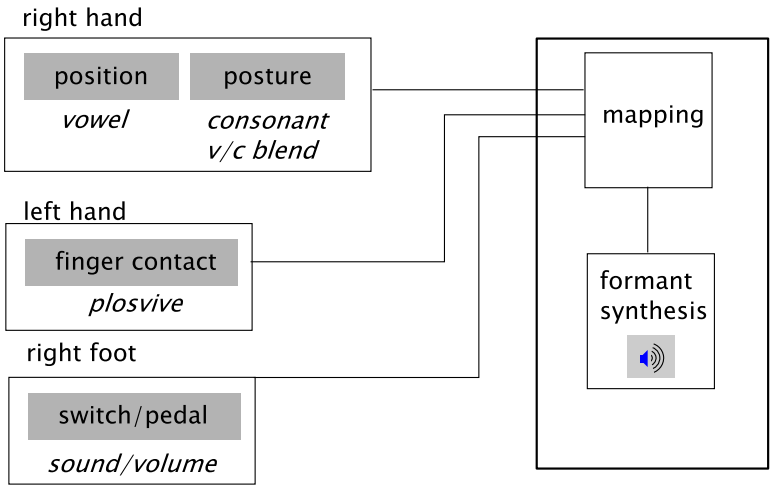


Figure 3.1: The DiVA 2.0 system

a custom made contact glove that functioned as 8 individual on/off switches that triggered stops and plosive sounds. A foot switch turned on the main sound for vowels and consonants.

While the overall functionality of the DIVA2.0 system was somewhat similar to GloveTalkII, there were distinct differences. First, the DIVA provided an user-adaptive training procedure using a radial basis function that drastically reduced the number of training samples required compared to back propagating a neural network. Secondly, considerable effort was spent during design based on the aesthetics and robustness of the system as required for use by musicians during rehearsal and performance. The following section of this chapter makes an attempt to document some of the performance and technical issues discovered through working with the DIVA systems as a technician supporting musicians in rehearsal and performance, as well as a developer working on various system components.

3.2 Lessons Learnt

One of the key features of the DIVA project is that, unlike many NIMEs where the builder and the performer is/are often the same technologically knowledgeable person, DIVA performers are skilled musicians with limited technical knowledge of the underlying system, and the instrument is used to perform a scored, composed piece. While using a new instrument in this setting creates certain challenges, it also provides a rich environment for exploring the musical implications through incorporating the creativity, experience and skill of trained musicians. In this section some general comments documenting the experience of working with DIVA musicians are presented.

3.2.1 Types of Musicians

Despite a small sample size of performers to date (3), it was clear from different observations that the type of training a musician received had an effect on their approach to the instrument. The DiVA group under discussion contained:

- a classically trained singer
- a trained singer - pianist with a great deal of experience in contemporary experimental music
- a classical guitarist, who was also an untrained vocalist

During the training process the guitarist became comfortable and adept with the DIVA much faster than the others. While there is an obvious distinction between singers and instrumentalists where the latter perform on physical, external instruments, there are also notable differences within instrumentalists that are based on the nature of interaction with the instrument. As an example, a pianist is never required to tune, regulate, restring or otherwise service their instrument (unless they have the additional vocation as a piano technician), whereas a guitarist is expected to tune the instrument (frequently) and restring (less frequently). On the other extreme,

many double reed players, such as oboists, make their own reeds, which requires a significant amount of invested effort in learning how to "build" a part of the instrument from scratch. It would be reasonable to suggest that when introducing a NIME to a classically trained musician, the perceived mutability of the instrument has an effect depending on the performer's musical background.

3.2.2 Semantics

When presenting a new instrument to a performer, it is important to delineate the boundaries between setting up the instrument, tuning and practising/performing. The description of each activity and expected outcome should be clearly communicated to the performer so the appropriate amount of effort can be applied to each task. The performer should have an idea of the limitations of the system so that a proper balance of time is spent tuning the system compared to rehearsing.

Musicians and engineers have very different vocabularies. The same words can mean different things so extra caution should be exercised when communicating. As an example, if the term "training" is used to describe the process of recording preset target positions, there is an embedded meaning that feeding the system with more samples would somehow improve the performance. As another example, many musicians make strong distinctions between the terms "practice" and "rehearse", distinctions that are not always apparent to engineers. Simply going by the definitions of words is insufficient: often there is a need to explicitly define certain processes to avoid ambiguity and misinterpretation.

3.2.3 Precision, Accuracy and Consistency

Repetition and reproducibility is a key aspect of musical practice. When a pianist approaches the keyboard and strikes a key it is expected that the same note and timbre is emitted from the instrument as on a previous occasion, assuming that an identical gesture was applied. The DIVA system has a high level of *precision*, and continuously blends various vocal sounds using

sensitive devices to measure input gestures with great resolution. However, due to slight variations in sensor mounting (caused by the physical nature of the wearable interfaces), the system did not always produce *consistent* results between practise sessions, and as the performer progressed to a certain level it was a source of frustration since it negated the effect of fine-detail practice and tuning. Therefore, methods of maintaining and checking for consistency should be implemented from both the technical perspective (in terms of tools and indicators) as well as practice routines (e.g. defining setup procedures, neutral positions for sensors, etc).

A even larger consistency issue lies in the dynamic nature of new instruments. Traditional instruments (with the exception of the saxophone and electric guitar) are the result of hundreds of years of slow, incremental development. In the NIME community, instruments are based on rapidly changing technologies and are often transformed in a very short time. When dealing with a device that is designed to serve multiple users with different, changing needs, the rapid development and iteration cycle, as mentioned by [35], is a positive attribute. However, when a musical instrument requires a great deal of practice to reach a certain level of virtuosity for performance, it is crucial that the physical dimensions and the responsiveness remain the same. Any "improvements" that are imposed upon the performer are potential setbacks.

3.2.4 Robustness and Stability

The issue of system robustness has crept up on numerous occasions with varying degrees of detriment. A musician typically spends a significant amount of time (on the order of months or even years) preparing for a performance that may last a few brief minutes. To quote Bill Buxton, an expert in interface and instrument design, on the reliability of instruments:

"... in the grand scheme of things, there are three levels of design: standard spec., military spec., and artist spec."[2]

More specifically we have found some of "Perry's Principle"s, such as avoiding battery powered and wireless devices [3] to be extremely relevant

here. Despite improvements in wireless technology (as mentioned in the updated article [4]) there is still a strong motivation to use wired connections for performance whenever and wherever possible. As an example, despite the ongoing developments in making Bluetooth communications robust, the instrumented gloves using this technology were the source of many problems throughout various iterations of the DIVA project, problems that included crashes and severely degraded performance during concerts. As one can imagine, the robustness and stability of the system is closely related to the confidence of the performer (see the previous section on consistency above).

3.2.5 Sound Quality

A major technical issue of the existing system is the sound quality. From both the performer and audience perspectives, the perceived synthesis quality of the existing DIVA system was poor in terms of expressiveness, naturalness and intelligibility. Some performers even admitted that the sound quality had a negative effect on their motivation for practising.

3.2.6 Summary of Learnings

Up to this point in the chapter we have described some of the lessons learnt through working with the DIVA system as a performance instrument, and they appear as follows:

- The type of musician may affect their attitude and approach towards new instruments, and consideration should be made (where appropriate) when selecting performers.
- The semantics is important especially when working in cross disciplinary teams.
- Musical instruments need to be consistent in function and behaviour.
- Musical instruments, especially those intended for stage use, must be reliable.
- The sound quality of the current system is an issue

With exception of the final item, the above points can be directly applied as guidelines when dealing with new instruments and performers. The sound quality issue requires in depth investigation, and will be discussed in the remainder of this chapter.

3.3 Motivation for Exploring New Mapping and Synthesis

Sound Quality: A Mapping Problem?

In Chapter 2, examples of speech synthesis production and control methods were presented. From there we identified the trade-off between being able to fully control synthesis parameters and well defined input trajectories that provided natural sounding output. The challenge faced by the current DIVA system can be attributed to this trade-off: compared to Dudley's Voder where the amplitude of each frequency band is individually controllable, the DIVA mapping is more constrained since it can only produce output existing within a pre-determined formant space. This formant space is chosen to resemble specific speech targets defined by a vocabulary. Although having these targets mean that the output is closer to what is expected in natural speech (and much more easier to control compared to individual manipulation of frequencies), when making gesture trajectories between these formant targets, the interpolations in frequency space do not necessarily correspond to the physical resonance behaviour of the vocal tract when moving between these articulatory targets. It may be possible to carefully craft these trajectories to match natural speech but the gestures required can be very complex. As an example, the offline-tuned preset trajectories developed by Holmes in his synthesizer, as mentioned previously, does create quite natural sounding speech but the gestures involved would be extremely difficult to make. The three cases, along with our ultimate goal, is presented relative to each other in Figure 3.2.

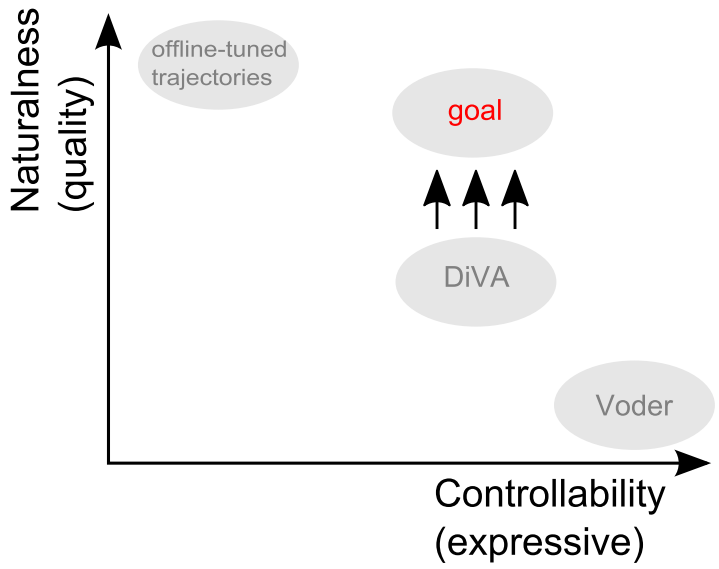


Figure 3.2: Controllability vs naturalness

3.3.1 Suggested Avenue of Exploration

To find a way of moving towards our goal of creating more natural sounding speech and preserve controllability, we need to look at the problem from both the synthesis and input mapping perspectives.

Since browsing through trajectories linearly interpolated in formant space do not necessarily translate to natural trajectories in the physical, vocal space, a reasonable approach is to look at synthesizing the sound within a space that has better connection with physical anatomy. Articulatory based synthesis methods appear to be an obvious choice.

As mentioned in Chapter 2, articulatory synthesis systems can employ a parametrically defined kinematic tube geometry or a dynamic one. Referring back to Figure 2.10 in Chapter 2.4 in the context of mapping in the NIME perspective, the bio-mechanical model serves as the middle layer providing a "meaningful" [21] translation between input and synthesis parameters since the model would constrain the vocal tract shape (and hence, resonant

response) to physically meaningful configurations, as opposed to arbitrary frequency values and trajectories. Additionally, from the findings of [22] on direct vs complex mapping, the extra layer imposed by the bio-mechanics adds a layer of complexity (compared to directly driving the formant values) should produce an interface that is more expressive.

In terms of the actual input device, the existing DIVA system used kinematic controllers where the spatial position is measured. There are also dynamic (force) controllers which, when dealing with a dynamic model with force activations, appear to make sense. It is not clear which input method is best for controlling speech.

Based on the above discussion, the proposed course of action is to investigate an articulatory based synthesis system and evaluate the effect of bio-mechanics and various input devices as the starting point for exploring ways of creating more natural sounding, but still controllable (expressive) speech.

Chapter 4

Implementation of Force and Position Input Controlling a Bio-mechanical Model for Articulatory Synthesis

This chapter describes the implementation of the bio-mechanically based mapping layer that drives an articulatory synthesizer and the process of setting it up to evaluate different input and mapping strategies.

First an overview of the system will be provided showing the relationship between various components and the signal flow from the starting point of user hand gestures to the final audio output. Then, a more detailed description of each individual module is presented, followed by the results of the integrated system. Finally, kinematic mapping is added to the system to allow comparison with the input mapping strategy of the existing DIVA system.

4.1 Overview

Figure 4.1 shows the overall system diagram in terms of signal flow. The gesture input from hands are detected, and sent to a mapping layer. From there the raw input values are scaled into muscle activations that are fed into a bio-mechanical model of the vocal tract, which calculates the geometry that then drives an articulatory synthesizer that finally outputs audio in real time.

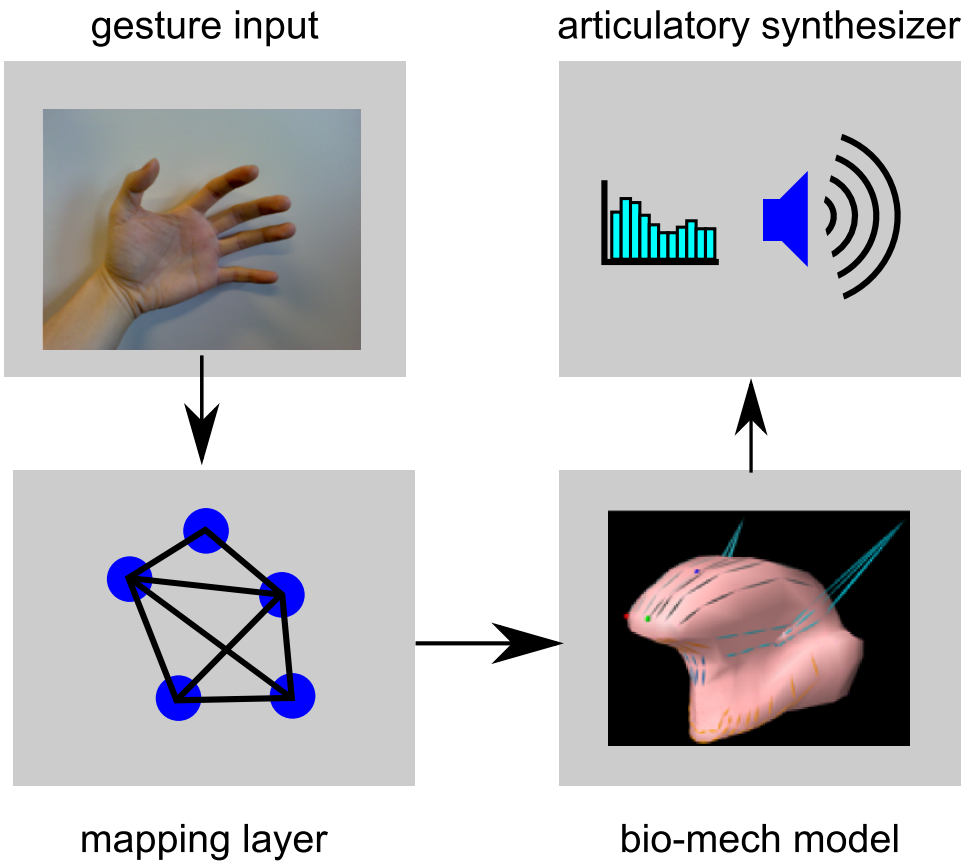


Figure 4.1: System diagram

Table 4.1: Development platforms

Component	Platform
Gesture Input	Arduino
Input Mapping	Max/MSP
Bio-mechanical Model	Java
Synthesizer	Java

4.2 Development Environment

Table 4.1 shows the various system components and the environment in which they were developed. The bio-mechnical model and synthesizer contained modified and extended parts of existing systems while the input system and mapping layers were implemented from scratch.

Max/MSP was chosen for the mapping layer due to its graphical representation and manipulation of signals. The graphical nature of the platform allowed rapid implementation, tuning and debugging of the mapping parameters that, while feasible for implementation in other languages such as Java or C++, would require a considerable increase in development overhead.

4.3 Inter-module Communication

The modular nature of the system was partially influenced by the experience from working on the various DIVA systems in the past, as well as the implementation of existing components and tools used in the system. The advantage of a modular system is that components can reside on multiple platforms which allow extra computational power or provide system features not available. One drawback is that the communication between the modules may introduce bandwidth limitations and latency. The next section describes the implemented communication system that takes into consideration the requirements of the system and available resources.

4.3.1 Open Sound Control

For the various modules to communicate, Open Sound Control (OSC) was used. OSC is a network-based messaging protocol designed to address some of

the limitations of the Musical Instrument Digital Interface (MIDI) standard, especially when dealing with new instruments implemented on increasingly powerful systems that require more flexible and descriptive control parameters. With a simple specification that sits on top of the User Datagram Protocol (UDP), there are software libraries for OSC on most existing platforms. The plain text nature of the protocol allows message specification in a human readable format. However, since OSC requires a network connection, there is the added requirement for hardware (either Ethernet or Wifi) and software which creates potential performance issues (especially on lower-powered devices). [16] provides a more detailed comparison of OSC and MIDI. One extremely attractive feature of OSC for the current implementation is the drastically reduced programming overhead without noticeable bandwidth and latency limitations. When running on a local machine the latency was less than 1ms for each connection and on a local wireless network, rarely above 10ms.

4.3.2 Sender and Receivers

Since OSC uses the UDP, the standard socket protocol setup applies. Communication is done between senders (clients) and receivers (servers). Each receiver listens on a specified port, and a sender requires a port and Internet Protocol (IP) address. If the modules run on the same machine, then the local loopback IP address (127.0.0.1) can be used.

4.4 Gesture Input

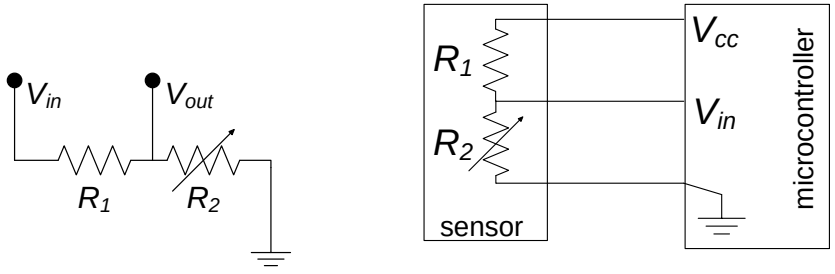
4.4.1 Input Hardware

The force input hardware consist of a number of FSRs attached to a microcontroller. Made of a semi-conductive sandwich that decreases in electrical resistance as force is applied, the amount of force can be measured by placing the FSR in a voltage divider configuration, as shown in Figure 4.2a. The output voltage in such a configuration is given by Equation 4.1 :

$$V_{out} = V_{in} \times \frac{R_2}{R_2 + R_1} \tag{4.1}$$

where the FSR is connected as R_2 for each sensor. The nominal value of the FSR is around 100kΩ and drops to around 10kΩ when first depressed. The resistance is roughly inversely proportional to the force applied, and reaches around 250Ω when saturated. While it may be useful for future studies to identify the consequence of mapping different force profiles and know the exact force applied, for now the main interest is to sense a monotonically increasing force for the mapping. With a V_{in} of 5V, the output is 5V when there is no applied force and close to 0V when the sensor is fully saturated. Four identical sensors were used and each V_{out} is connected to an analog input pin of the microcontroller as shown in Figure 4.2b.

For the microcontroller, the Arduino was chosen due to its availability for rapid prototyping with little development overhead and easy to use IDE. While more constrained in terms of features compared to other microcontroller solutions, the Arduino provides sufficient capabilities and performance for the task at hand, and allows room for easy expansion if more inputs are required in the future. The open hardware and software platform, along with widespread retail availability, allows the system to be easily rebuilt elsewhere in the future.



(a) FSR in Voltage Divider **(b)** FSR Interface

Figure 4.2: Force sensor circuit

The FSR's are mounted using clear tape and a layer of closed cell foam

for comfort on top of a plastic enclosure that houses the microcontroller as shown in 4.3.

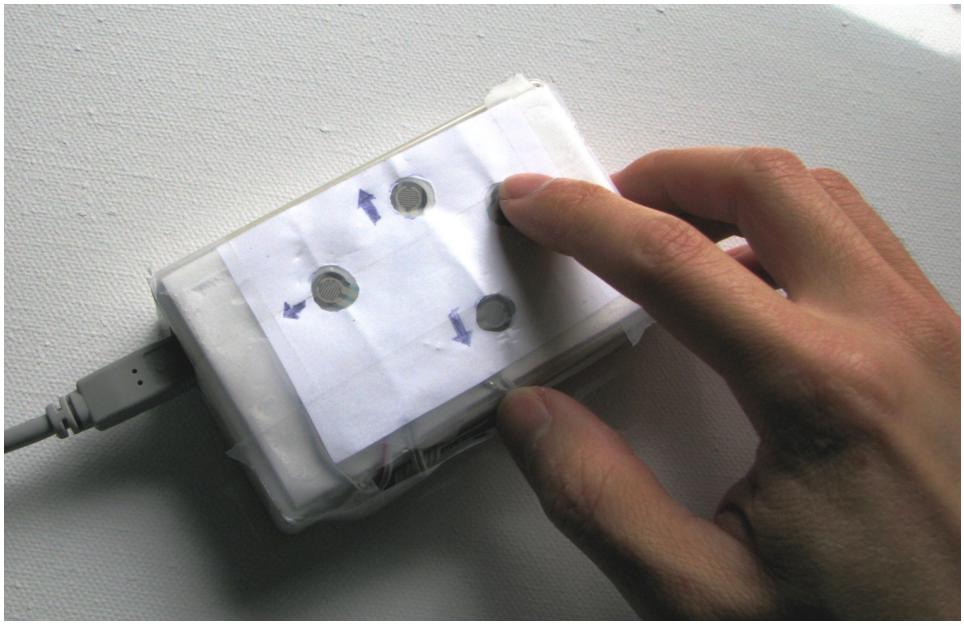


Figure 4.3: FSRs mounted on enclosure

4.4.2 Input Software

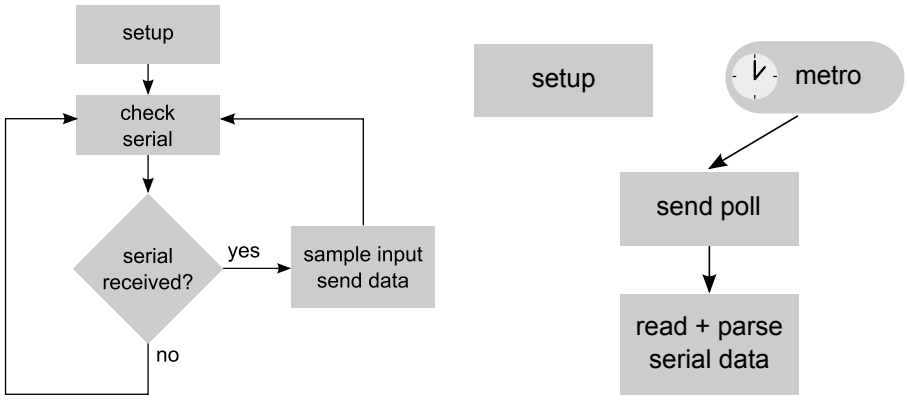
There are two pieces of software responsible for getting the force input values from the physical hardware to the computer software. First is the firmware that runs on the Arduino to control the hardware (the sender), and second is a Max/MSP patch (the receiver) to process the values. The system runs in a polled mode, and the rate is controlled by the receiver end.

Arduino Firmware

The Arduino firmware is responsible for setting up the hardware, reading the force inputs and sending the data to the computer. After a simple setup function to initialize the serial and input ports, the program enters a loop that constantly checks for a read command through serial input. When a

read command is received, the program enters a routine that samples the input pins and sends the result back to the computer.

The measured voltage reading per port is a 10 bit value (based on the resolution of the hardware) and the values are assembled into a plain text ASCII string, separated by the space ' ' character and terminated with a newline. The string is sent to the computer via the Arduino's serial port, which is connected to an on-board Universal Serial Bus (USB)-serial adapter. The serial port operates at 115200 baud. Figure 4.4a shows the flow diagram of the microcontroller firmware.



(a) Sensor End **(b)** Receiver End

Figure 4.4: Input system flow diagrams

Receiver Patch

Figure 4.4b shows the corresponding polling receiver flow diagram of the Max/MSP Patch that sends the poll commands and parses the input values. Here the setup function initializes the serial port by choosing the correct interface and baud rate, and starts a "metronome" object which issues events at a pre-set intervals of 15 ms. This sampling rate corresponds to existing hardware used in the DIVA system, as well as the touch pad system used in the experiment (explained in later sections). Each time the metro event triggers, a poll command is issued to the hardware and the response is parsed. A scaling function also inverts and maps the received value between

0.0 and 1.0 (from 1023 to 0).

4.5 Force to Muscle Mapping

The four force inputs represent directions, and are used to control muscles based on their effect on the tongue body: front, back, up, and down. This input system allows opposing muscles to be activated at the same time. The mapping between the input sensor values and muscle activations was implemented as a Max/MSP patch shown in Figure 4.5.

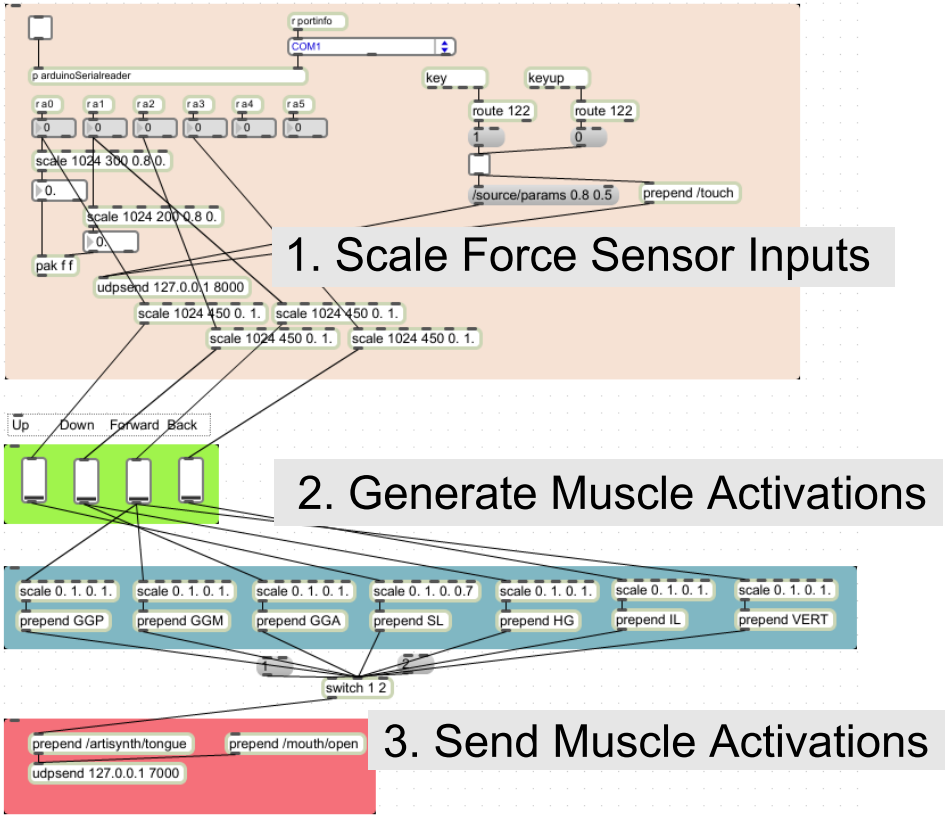


Figure 4.5: Force input and mapping

4.6 Bio-Mechanical Model

The model was implemented in the ArtiSynth modelling environment [11]. A series of beams were constructed around an existing tongue model [39] to represent sections of the vocal tract around the tongue, and 22 marker points were placed at set intervals along the tract surface. The distances between these marker points and the tongue surface are computed in real time which allows an effective cross sectional area function to be calculated. These area functions provide the main inputs required by the articulatory synthesizer in calculating the audio output.

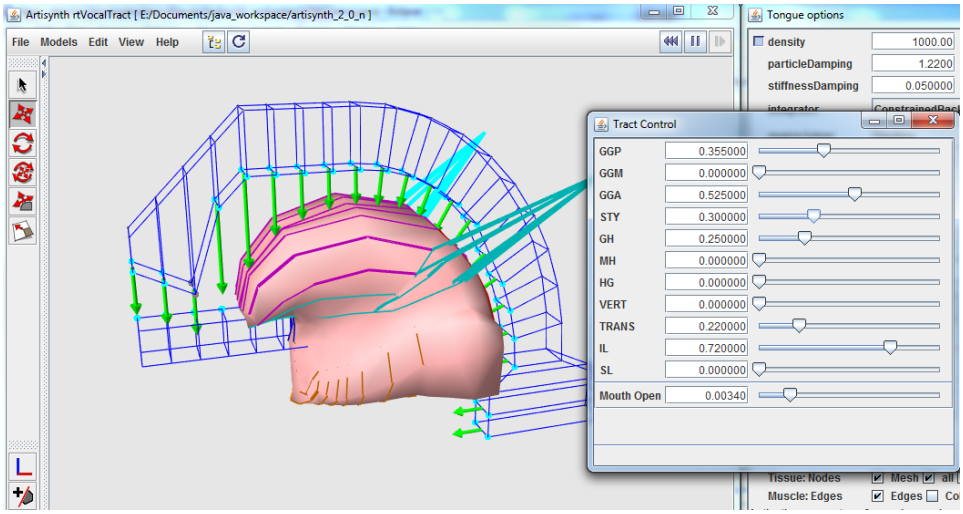


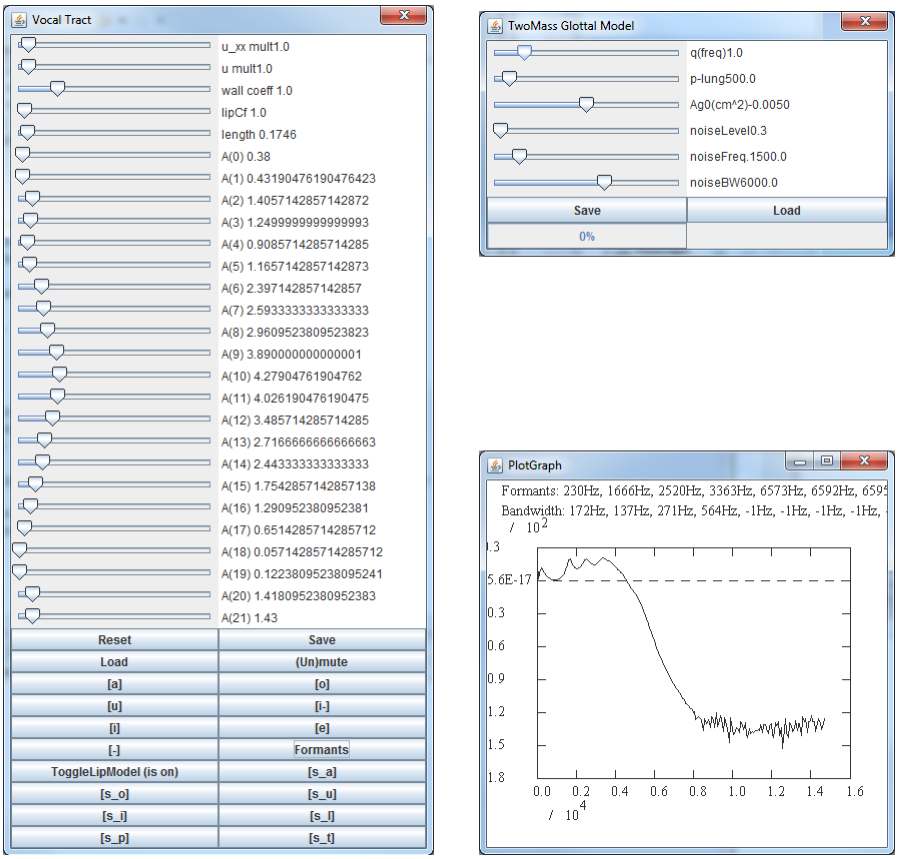
Figure 4.6: The Artisynth vocal tract model

Figure 4.6 shows the model and parts of its Graphical User Interface (GUI). The sliders in the "Tract Control" window allow interactive manipulation of the tract parameters (mostly muscle activations) for tuning, but in operation they are set via incoming OSC messages. As the simulation runs, the resultant cross sectional areas are then sent to the synthesis module.

4.7 Synthesizer

The synthesizer is from the jass library [37] and the implementation is described in [36]. The glottal source model used is based on [23]. The main addition made to the synthesizer was the implementation of an OSC listener that allows the tube shape and glottal source parameters to be controlled in real time. The number of tube sections is set to be the same as the output of the bio-mechanical model, although a linear interpolation function was also implemented which allows a different number of sections to be entered. The synthesizer is also able to output numerically, on request, the current tube parameters for debugging and comparison.

Figures 4.7a and 4.7b shows the interface for the synthesizer. In the original application, the sliders in the User Interface (UI) allow the tube parameters to be modified interactively. In the background the received OSC messages from the bio-mechanical model set the tube width parameters and change the output sound in realtime.



(a) Vocal Geometry Parameters **(b)** Two-mass model and Formants

Figure 4.7: Jass synthesizer

4.8 Integration and Output

Using OSC, the various system components were connected as shown in Figure 4.8. Of note is the extra connection on port 12002 between the force mapping Max patch and synthesizer that send initialization parameters for the synth. In theory they could be implemented in the bio-mechanical model, but for the sake of development it was much more convenient for the tuning process to send these messages from Max/MSP.

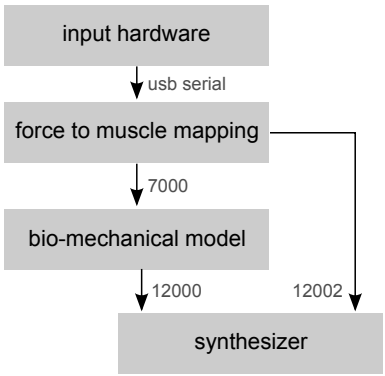
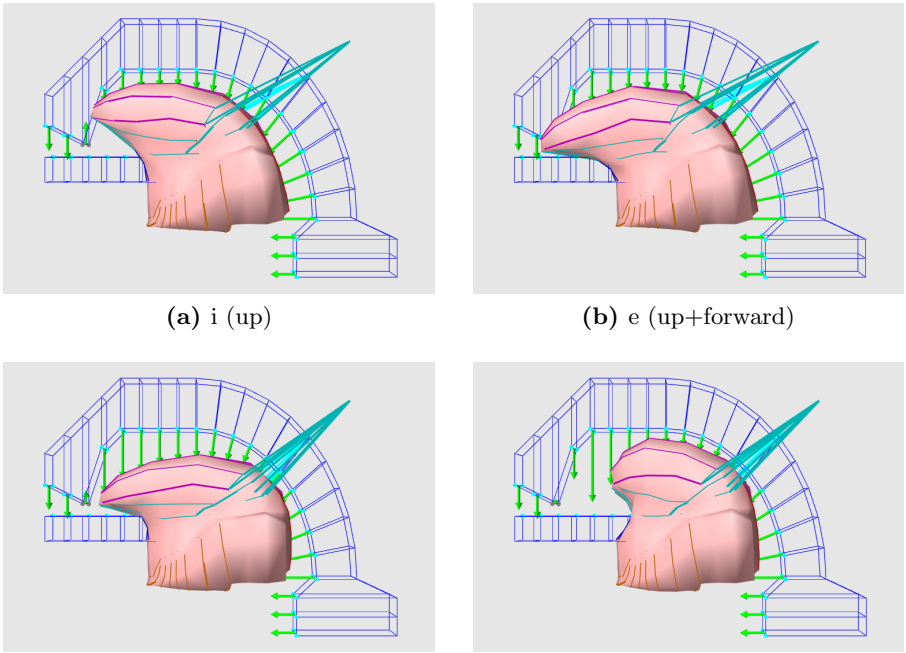


Figure 4.8: UDP ports for OSC messages

Once the components were connected and the OSC messaging system tested, the physical geometry of the bio-mechnical model was manually tuned to produce 4 positions that resembled target vowels when driven by 4 sets of saturated force inputs, as shown by Figure 4.9.

When the calculated cross sections are received by the synthesizer, the following spectral output was observed (Figure 4.10). Since the actual geometry differs, it is not possible to compare the actual frequencies with others in the literature [31]. However, the relative positions of the formants and trajectories when transitioning between the targets appear to be important for vowel identification as suggested by studies described in [32]. Indeed, the actual targets for vowels are often not reached for successful vowel trajectory production and identification [9].



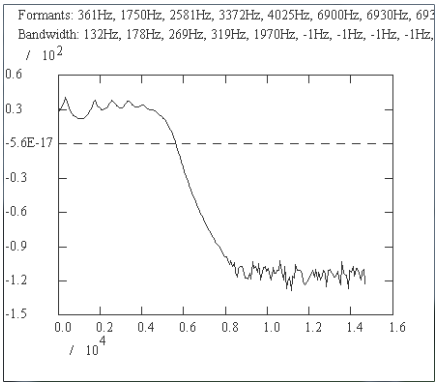
(a) i (up)

(b) e (up+forward)

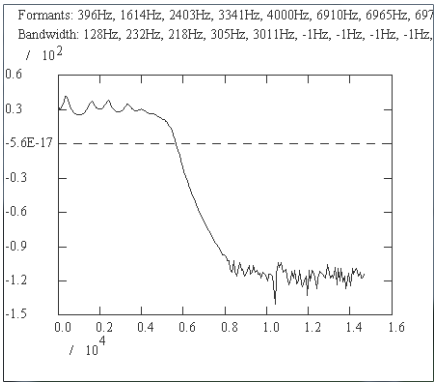
(c) a (down)

(d) u (up+back)

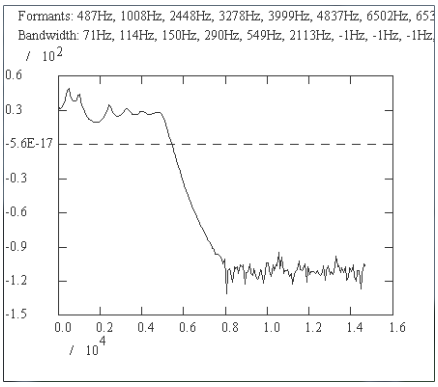
Figure 4.9: Tongue in target vowel positions



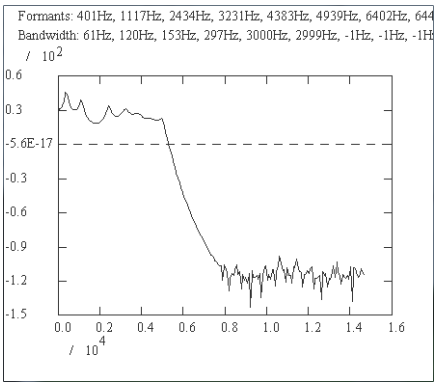
(a) i (up)



(b) e (up+forward)



(c) a (down)



(d) u (up+back)

Figure 4.10: Output spectrum for various tongue positions

4.9 (Re)Implementation of Position/Kinematic Input

In order to provide a balanced comparison of just the input mapping, we cannot simply use the existing system due to the different sounds of the synthesizers. In fact, a preliminary study was done using the existing synthesizer for the position input case and there were significant differences in the perceived audio quality which affected the validity of results. Therefore, the kinematic input for browsing the vowel space was reimplemented in the new system with the articulatory synthesizer.

Although the original DIVA system made use of a 3D position tracker for the control of vowel sounds (X-Y for vowel browsing and Z for pitch), we decided to use only 2D (X-Y) since we are only comparing the browsing of the vowel space. The position input was implemented on a touch screen and the same gaussian Radial Basis Function (RBF) based interpolation was applied to tube geometries (instead of formants). Four targets was laid out on the 2D surface, each corresponding to the target tube shapes represented geometrically in Figure 4.9 which produced the spectral output in Figure 4.10. The goal of the position/kinematic input system is to allow browsing between these tube shapes.

For a given input coordinate (x, y) on the touch screen the pixel distance between the touch point and the ith target position at (x_i, y_i) is calculated using the distance (Equation 4.2). The standard Gaussian RBF (Equation 4.3) was applied for each target.

$$r_i = \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{4.2}$$

$$\phi_i(r_i) = e^{-(\epsilon r_i)^2} \tag{4.3}$$

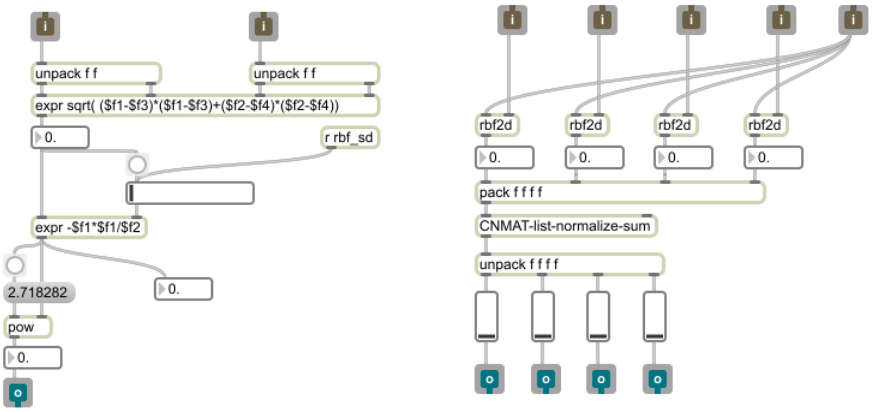
$$\bar{\phi_i} = \frac{\phi_i}{\sum_{j=0}^{N} \phi_j} \tag{4.4}$$

Then, the ϕ_i values were normalized to 1 (Equation 4.4), and the nor-

malized values function as weights for calculating the tube geometry for the tube T that is composed of a linear interpolation of N target tube shapes (in this case, N=4 for the target vowels) (Equation 4.5).

$$T(k) = \sum_{i=0}^{N} T(k)\bar{\phi}_i \tag{4.5}$$

The kinematic mapping layer was implemented in Max/MSP and, in a twist of certain irony, employed similar patching structures as sections of the original DIVA1.0 system. Figure 4.11a shows the sub-patch, rbf2d, that takes in two point coordinates (one for the input position, the other for the target)and calculates the RBF distance. Figure 4.11b the rbf2d subpatches being used together and their outputs normalized.



(a) RBF Calculation **(b)** Normalized Tube Parameters

Figure 4.11: Components of the RBF mapping patch

The physical input system was implemented as simple openFrameworks iPad application that detected and sent a single touch location to the mapping system. Figure 4.12 shows the force and position input devices side by side.

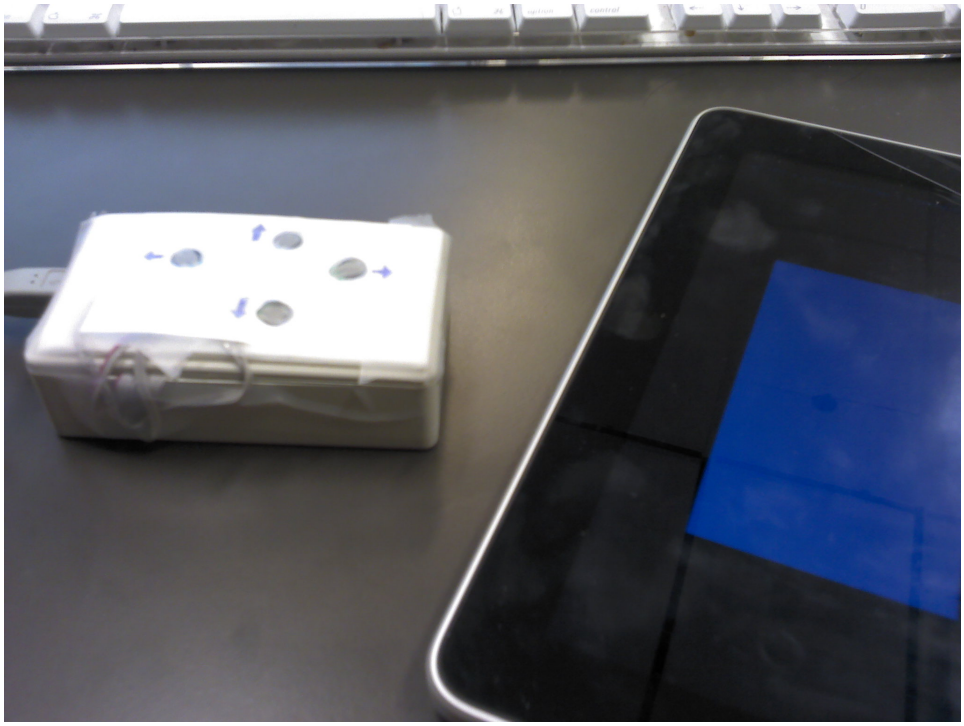


Figure 4.12: Force and kinematic input interfaces

Chapter 5

Evaluation

This chapter describes the evaluation and comparison of the input and mapping for the new DIVA system compared with the existing one. The browsing of the vowel space set up in Chapter 4 is compared from both the performer and listener perspectives. A discussion follows the presentation of results of the experiment.

5.1 Overview

The main goals of the evaluation are to compare the force and position input and mapping systems from the following perspectives:

- Performer: differences in usage (qualitative)
- Performer/Listener: intelligibility (quantitative)
- Listener: other characteristics of sound (qualitative)

5.2 Pilot Study

A pilot study was conducted to guide implementation of the experiment. three performers and four listeners were recruited for the pilot study. The experiment contained two separate parts. In the first part, each performer was introduced to the system and after some practice, asked to generate a

number of vowel sequences that were recorded and used for the second part, the listener evaluation. During the listener evaluation, the subjects were asked to identify audio samples produced from the first phase and provide qualitative comparisons between samples produced using different interfaces.

Through the pilot experiment various issues were discovered that motivated modification and refinement of the evaluation procedure. In the pilot one of the key issues was the use of the existing system's synthesizer, and a significant difference in output sound quality was noticed by listeners. There was a significant difference in the intelligibility results but the effect may have been due to the sound difference.

Another issue raised by the pilot study was dealing with the training of the performers. There was no strict metric for when the performer was ready to perform other than their personal response, and so for the final implementation a small pass-fail test was employed before the subject could proceed onto the recording portion of the experiment.

5.3 Experiment

The final evaluation consisted of two phases: the performer phase and listener phase. Six performers and eight listeners were recruited. In the first phase, the performers used the systems to produce a number of "words" composed of the four vowels (Section 4.9) which are recorded for later playback. In the second phase the listeners are provided with audio recordings of the samples and asked to identify them as well as provide qualitative comparisons. The following sections provide more detail about each phase of the evaluation. The same headphones and computer were used for all the experiments at the same volume to eliminate differences in sound quality due to audio hardware.

Table 5.1: Number of sample words

Syllables		# of words
1	4	4
2	4C2	12
3	4 × 3× 3	36
Total		**52**

5.3.1 Performer Experiment

Preparation

The performers were introduced to the system and given time to practice. The interfaces were presented in random order to eliminate potential learning effects. After the subject had some time to familiarize themselves with the interface, they were asked to produce a few sample words to ensure that they were able to reach within 10% of the entire input range for each syllable in the word. The experiment only proceeded if the subject was able to fulfil this criteria.

Main Test

To provide a reasonable number of samples, words of up to three syllables were used, yielding 52 samples in total (Table 5.1). The 2 and 3 syllable words were chosen such that consecutive syllables were different, since we are more interested in the transitions between the vowel sounds. During the recording portion of the experiment, the order of the word list was randomized for each subject.

As mentioned in the results from the pilot study, both the force input/biomechanical system and the existing position/kinematic one were connected to the same synthesizer, and the additional modifications implemented for the latter system are described in Chapter 4.9. After the entire list of words was recorded for the one interface, the second interface was introduced and the process repeated.

After the samples were recorded for both interfaces, the subject was

asked a number of preference questions based on the following:

- Ease of use
- Musical/Expressiveness
- Naturalness
- Fun/Enjoyment

5.3.2 Listener Experiment

The main part of the listener experiment was an identification test that played back samples recorded in the first phase of the experiment. 80 samples were randomly selected from the pool of samples recorded in the performer experiment. The two interfaces were represented equally (40 samples each). The listeners heard each sample only once and was asked to identify the word in the sample.

In the second part of the listener experiment 16 pairs of the same word (from both interfaces by a subject) were randomly selected from the entire pool of recordings. The order within each pair was randomized, and then played back to the listener for comparison. The metrics "sharp", "exciting", "natural", "speech-like" and "intelligible" were used.

5.3.3 Performer Evaluation Results

All the subjects were able to reach the within-10% input value for each target on both interfaces. After all the recordings were complete, the performers were asked a series of preference questions and rationale for their choices. The following results were obtained:

Ease of Use

5 out of the 6 subjects thought the position input was easier to use. The 6th subject initially stated that the force input was more "intuitive" and hence "easier", but then retracted his statement and admitted that in terms of simplicity, the position controller was preferable.

Musical/Expressiveness

In this category, the results were evenly distributed between no preference, the position and force inputs. The two users who preferred the position input provided similar reasons as to why they were able to more easily achieve their desired target. The two users who chose the force input both used the ability to vary the input and the many-to-many mapping as the main reason. The other two users who did not state the difference were not entirely sure why they chose them.

Naturalness

3 subjects preferred the force input, 2 the position input and 1 was neutral. The subjects choosing the force input provided the reason that they were able to hear the difference in trajectories between the two input/mapping systems, and the force/dynamic system provided closer resemblance to natural speech. The subjects preferring the position/kinematic system felt that the input movements required were more natural.

Fun

All but one subject felt the force input was more fun due to its higher level of difficulty. The subject choosing the position input realized non-linear movements in the position may potentially provide more interesting results. It appears all the subjects seemed to have attributed "fun" with attempting to do something more difficult.

Table 5.2: Identification accuracy

User	Position	Force
1	65%	63%
2	73%	68%
3	70%	68%
4	33%	35%
5	48%	48%
6	60%	45%
7	70%	58%
8	80%	78%
average	62%	58%

5.3.4 Listener Evaluation Results

The identification task showed much closer accuracy rates compared to the pilot study (where the difference was 15% in favor of the force input system), and suggest that the output sounds are very similar (at least for novice listeners).

For the qualitative descriptors (Table 5.3), there was no longer a significant perceived difference in the sound quality produced by the two input and mapping systems. This result suggest that the potential bias caused by the different synthesis has been removed.

	Position	Force
Sharp	48%	52%
Exciting	53%	47%
Natural	54%	46%
Speech-like	55%	45%
Intelligible	49%	51%

(a) Final Experiment

	Position	Force
sharp	98%	2 %
exciting	77%	23 %
natural	20%	80 %
speech-like	27%	73 %
intelligible	56%	44 %

(b) Pilot Experiment

Table 5.3: Qualitative comparisons of final and pilot experiments.

5.4 Summary of Results

From the above results, it appears that for vowel browsing, at least involving inexperienced listeners, the force and position mappings do not seem to sound that different. The intelligibility of the sound output is also not significantly different. However, there were strong responses from the performers regarding the two input interfaces. The effect of the differences in the underlying gesture to synthesis mapping was also apparent to some subjects. Most of the performers agreed that the force input was more difficult to use, but provided a greater level of input variability and through certain perspectives, expressivity. The response regarding the musical and expressive aspects were not as uniform and this may be attributed to personal definitions of the terms.

Chapter 6

Conclusion

6.1 Contribution Summary

In this thesis we have documented the experience of working with a gesture controlled vocal synthesizer used for music performance. We identified and documented a number of issues and provided recommendations and guidelines for future work of a similar nature. One of the issues we discovered that prompted an in depth look at the gesture mapping and synthesis was the sound quality - more specifically, a trade-off between controllability and voice quality. Based on that, we explored using a new synthesis system based on a more representative bio-mechanical model of the vocal tract. During the first-pass implementation and integration of the new system components, two input and mapping strategies were evaluated.

The evaluation results suggest that for the browsing of a vowel space involving inexperienced performers and listeners there is no significant difference in the intelligibility of the sound produced between the two input and mapping strategies. However, from the performer's perspective the force input system was generally identified as being more capable of producing more expressive output due to its ability to provide a more complex mapping and some performers noticed the stronger coupling between the input forces and natural speech trajectories.

The implemented system and evaluation mentioned in this thesis provide

an initiatory but significant starting point for the future development of gesture controlled vocal synthesis systems. The following section describes potential next steps based on the results obtained thus far.

6.2 Suggested Future Work

6.2.1 System Additions

While the synthesizer itself is capable of producing a wide variety of sounds, there is currently no mapping implemented for many of the synthesizer parameters (pitch and voicing, for example). While some exploration has been done in controlling the two mass model in the synthesizer using a hybrid interface [43] to provide rudimentary control of pitch and volume, the relationships between the parameters in the glottal source model and output pitch and volume are not linear.

The vowel space implemented so far is quite limiting - not only does it not cover the entire vowel space in the English vocabulary, the actual target configurations of the vowels are also not optimal. This is a result in the limitation of the current vocal tract model, which does not account for the actual physical tube shape. The implementation of a fully dynamic, anatomically correct bio-mechanical model is already in existence but at the same time, computational times are an issue. The currently implemented system represents a compromise between an accurate model and fast computation time.

6.2.2 Input Mapping Strategies

Given the choices of force or position on the input side, and dynamic and kinematic representations in the mapping layers, there are 4 possible combinations of mapping strategies:

1. Force-Dynamic: force input / dynamic model mapping layer
2. Force-Kinematic: force input / kinematic model mapping layer
3. Position-Dynamic: position input / dynamic model mapping layer

4. Position-Kinematic: position input / kinematic model mapping layer

The evaluation described in this thesis contain the first and third out of the four possible schemes. These were the most straight-forward choices and were explored first, but it would be useful to evaluate the second and forth methods as well to cover the full spectrum.

It should also be mentioned that the above discussion only deals with the browsing of vowel space. When consonants and plosives are implemented, a larger variety of input parameters will be required. It is not clear at this point what the optimal input interface and mapping scheme will look like.

Ultimately, the addition of vocal features result in an increase in the amount of input bandwidth required. From working with musicians in the existing DIVA project has shown that there are definite limits in user input bandwidth and there is a trade-off between the level of expressive control and usability.

6.2.3 Musical Evaluation

While an attempt was made to evaluate the expressive nature of the system, the scope of the musical evaluation was quite narrow due to the system limitations as mentioned in section (6.2.1). As more components of the system relevant in the context of musical expression are implemented (such as the input and mapping responsible for controlling pitch, volume, vocal effort, extended vocabulary etc.) appropriate testing procedures should be generated to evaluate their effectiveness. [26] suggests a list of common "musical tasks" that can be used. Additionally, a major challenge in evaluating musical interfaces in general is the amount of training required to reach proficiency and the availability of resources to do so.

6.3 Final Thoughts

The synthesis and control of the human voice provides a rich platform for NIMEs, and in turn, the requirements of such an instrument offer considerable challenges in terms of technical demands and refinement of knowledge across many fields. While there are still questions left unanswered at this point, one

of the fundamental goals of the system is quite clear: to build an expressive voice instrument - an instrument that will project the emotive intentions of a musician through a set of skills developed through learning and experience to the audience. In the end, the proof of the pudding is in the performance.

Bibliography

[1] P. Birkholz, D. Jackel, and K. Kroger. Construction and control of a three-dimensional vocal tract model. In *2006 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2006. → pages viii, 11

[2] B. Buxton. Artists and the art of the luthier. *SIGGRAPH Comput. Graph.*, 31(1):10–11, Feb. 1997. → pages 26

[3] P. Cook. Principles for designing computer music controllers. In *Proceedings of the 2001 conference on New interfaces for Musical Expression (NIME2001)*, pages 3–6, 2001. → pages 26

[4] P. Cook. Re-Designing Principles for Computer Music Controllers: A Case Study of SqueezeVox Maggie. In *Proceedings of the International Conference on New*, volume 11, pages 218–221, 2009. → pages 27

[5] N. d'Alessandro and T. Dutoit. HandSketch bi-manual controller: investigation on expressive control issues of an augmented tablet. In *Proceedings of the 2007 Conference on New Interfaces for Musical Expression (NIME07)*, pages 78–81, New York, New York, USA, 2007. → pages 17

[6] N. d'Alessandro and T. Dutoit. RAMCESS / HandSketch : A Multi-Representation Framework for Realtime and Expressive Singing Synthesis. In *Eighth Annual Conference of the International Speech Communication Association*, 2007. → pages 17

[7] N. d'Alessandro, B. Pritchard, J. Wang, and S. Fels. Ubiquitous Voice Synthesis : Interactive Manipulation of Speech and Singing on Mobile Distributed Platforms. In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, pages 335–340, 2011. → pages iii

[8] N. d'Alessandro, J. Wang, B. Pritchard, and S. Fels. Bringing Bio-Mechanical Modelling of the OPAL Complex as a Mapping Layer for Performative Voice Synthesis. In *International Seminar on Speech Production (ISSP2011)*, 2011. → pages iii

[9] P. Divenyi. Perception of complete and incomplete formant transitions in vowels. *The Journal of the Acoustical Society of America*, 126(3):1427–39, Sept. 2009. → pages 42

[10] H. Dudley and R. Riesz. A synthetic speaker. *Journal of the Franklin Institute*, 2(2), 1939. → pages 13

[11] S. Fels, J. Lloyd, K. Van Den Doel, F. Vogt, I. Stavness, and E. Vatikiotis-Bateson. Developing Physically-Based, Dynamic Vocal Tract Models using ArtiSynth. In *International Seminar on Speech Production*, volume 6, pages 419–426, Ubatuba, Brazil, 2006. Citeseer. → pages 39

[12] S. Fels, R. Pritchard, and A. Lenters. Fortouch: A wearable digital ventriloquized actor. In *New Interfaces for Musical Expression (NIME2009)*, pages 274–275, 2009. → pages 22

[13] S. S. Fels and G. E. Hinton. Glove-Talk: a neural network interface between a data-glove and a speech synthesizer. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 4(1):2–8, Jan. 1993. → pages 21

[14] S. S. Fels and G. E. Hinton. Glove-TalkII-a neural-network interface which maps gestures to parallel formant speech synthesizer controls. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 9(1):205–12, 1998. → pages 21

[15] N. C. for Voice and Speech. Voice production tutorials, Aug. 2011. http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/graphing.html. → pages viii, 6

[16] A. Fraietta. Open Sound Control: Constraints and Limitations. *Proceedings of the Conference on New Interfaces for Musical Expression (NIME2008)*, 2008. → pages 34

[17] K. Fukui, K. Nishikawa, and T. Kuwae. Development of a new human-like talking robot for human vocal mimicry. In *2005 IEEE International Conference on Robotics and Automation (ICRA 2005)*, number April, pages 1437–1442, 2005. → pages 10

[18] M. Hoffman. Feature-based synthesis: mapping acoustic and perceptual features onto synthesis parameters. In *Proceedings of the International Computer Music Conference*, 2006. → pages 19

[19] M. Hoffman and P. R. Cook. Real-time feature-based synthesis for live musical performance. *Proceedings of the 7th international conference on New interfaces for musical expression - NIME '07*, page 309, 2007. → pages 19

[20] J. Holmes, I. Mattingly, and J. Shearme. speech synthesis by rule. *Language and Speech*, 7(3):127, 1964. → pages 14

[21] A. Hunt and R. Kirk. Multiple Media Interfaces for Music Therapy. *Multimedia, IEEE*, pages 50–58, 2004. → pages viii, 19, 29

[22] A. Hunt, M. Wanderley, and R. Kirk. Towards a model for instrumental mapping in expert musical interaction. In *International Computer Music Conference*, 2000. → pages 20, 30

[23] K. Ishizaka and J. Flanagan. Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell Systems Technical Journal*, 51(6), 1972. → pages 12, 40

[24] B. Kröger and P. Birkholz. A gesture-based concept for speech movement control in articulatory speech synthesis. *Verbal and Nonverbal Communication Behaviours*, pages 174–189, 2007. → pages 13

[25] B. Kröger and P. Birkholz. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 306–319, 2009. → pages 11

[26] N. Orio and N. Schnell. Input devices for musical expression: borrowing tools from HCI. *interfaces for musical expression*, 2001. → pages 19, 57

[27] B. Pritchard. Performance What Does A Body Know ? In *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, pages 2403–2407, 2011. → pages iii

[28] B. Pritchard and S. Fels. GRASSP : Gesturally-Realized Audio , Speech and Song Performance. In *Proceedings of the 2006 conference on New Interfaces for Musical Expression (NIME2006)*, pages 272–276, Paris, France, 2006. → pages 22

[29] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2):Suppl 2:583+, Feb. 1971. → pages 12

[30] D. Schwarz, G. Beller, and B. Verbrugghe. Real-time corpus-based concatenative synthesis with catart. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, number September, pages 279–282, 2006. → pages 16

[31] B. H. Story, I. R. Titze, and E. a. Hoffman. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 100(1):537–54, July 1996. → pages 42

[32] W. Strange. Evolving theories of vowel perception. *The Journal of the Acoustical Society of America*, (May 2012), 1989. → pages 42

[33] I. Titze. *Principles of voice production*. Prentice Hall, 1994. → pages 5

[34] H. Traunmller. History of speech synthesis, 1770 - 1970, Sept. 2000. http://www2.ling.su.se/staff/hartmut/kemplne.htm. → pages viii, 9, 10, 14

[35] O. Vallis, J. Hochenbaum, and A. Kapur. A Shift Towards Iterative and Open-Source Design for Musical Interfaces. In *Proceedings of the 2010 conference on New Interfaces (NIME2010)*, number Nime, 2010. → pages 26

[36] K. van den Doel and U. Ascher. Real-Time Numerical Solution of Webster's Equation on A Nonuniform Grid. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(6):1163–1172, Aug. 2008. → pages 40

[37] K. van Den Doel and D. K. Pai. Jass: A java audio synthesis system for programmers. In *Proceedings of the 2001 International conference on Auditory Display*, 2001. → pages 40

[38] R. Vertegaal and T. Ungvary. Towards a musician's cockpit: Transducers, feedback and musical function. In *International Computer Music Conference (ICMC1996)*, number criterion 2, 1996. → pages 19

[39] F. Vogt, J. Lloyd, S. Buchaillard, P. Perrier, M. Chabanas, Y. Payan, and S. Fels. Efficient 3d finite element modeling of a muscle-activated tongue. *Biomedical Simulation*, (Figure 1):19–28, 2006. → pages 39

[40] W. R. von Kempelen. *Mechanismus der menschlichen Sprache nebst Beschreibungeiner sprechenden Maschine. Mit einer Einleitung von Herbert E. Brekle und Wolfgang Wild*. Stuttgart-Bad Cannstatt F. Frommann, 1970. → pages 9

[41] M. Wanderley, J. Viollet, F. Isart, and X. Rodet. On the choice of transducer technologies for specific musical functions. In *Proceedings of the 2000 International Computer Music Conference (NIME2000)*, pages 244–247, 2000. → pages 19

[42] J. Wang, N. d'Alessandro, S. Fels, and R. Pritchard. Investigation of Gesture Controlled Articulatory Vocal Synthesizer using a Bio-Mechanical Mapping Layer. In *Proceedings of the 2012 conference on New interfaces for Musical Expression (NIME2012)*, 2012. → pages iii

[43] J. Wang, N. d'Alessandro, B. Pritchard, and S. Fels. SQUEEZY: Extending a Multi-touch Screen with Force Sensing Objects for Controlling Articulatory Synthesis. In *Proceedings of the 2011 International Computer Music Conference (NIME2011)*. → pages iii, 56

Appendix A

Consent Forms and Sample Questionnaires

THE UNIVERSITY OF BRITISH COLUMBIA

Media and Graphics Interdisciplinary Centre
FSC 3640 – 2424 Main Mall
Vancouver, BC V6T 1Z4

20 September, 2011

Consent Form

Gesture-based Articulatory Speech Synthesizer

Principal Investigator

Dr. Sidney Fels, Associate Professor, Department of Electrical and Computer Engineering, University of British Columbia, 604-822-5338

Co-Investigators

Johnty Wang, Master's Candidate, Department of Electrical and Computer Engineering, University of British Columbia, 604-822-9248

This research is to be used as material in a thesis, which is a publically available document. Your identity will remain confidential and the information collection during the study will be used in an anonymous way.

Purpose

You are being invited to take part in this research study that involves a gesture controlled speech synthesizer that allows you to speak using hand gestures. You will either be asked to use the system to make sounds, or

Study Procedures

This study will take between 30 minutes and 1 hour. You will be either a "Performer" or a "Listener"

As a "Performer":

Get introduced to the speech synthesis system and try it out
Practice various words until you can reach a certain level of accuracy
Perform a series of words
Share your thoughts and provide feedback through a short questionnaire

As a "Listener":

Listen to recorded audio samples
Write down what you hear
Share your thoughts and provide feedback through a short questionnaire

Confidentiality

Your identity will be kept strictly confidential and will not be known to anybody except the interviewer. In order to assure this confidentiality, any information that may identify you as an individual will not be written on any data collection sheets. Instead, your consent form will be linked to your data collection sheets using an arbitrary number identifier. Furthermore, consent forms and data collection sheets will be stored in two different locked cabinets. Any computerized files will be stored on password protected internal servers at the Media and Graphics Interdisciplinary Center that is not accessible over the Internet.

Remuneration/Compensation

Each participant will be receive an honorarium in the amount of $10.

Contact Information About the Study

If you have any questions or require further information about the project you may contact Johnty Wang

Contact for Information About the Rights of Research Subjects

If you have any concerns about your treatment or rights as a research subject, you may contact the Research Subject Information Line in the UBC Office of Research Services at 604-822-8598 or if long distance, email to RSIL@ors.ubc.ca.

Consent

We intend for your participation in this project to be pleasant and stress-free. Your participation is entirely voluntary and you may refuse to participate or withdraw from the study at any time.

Your signature below indicates that you have received a copy of this consent form for your own records.

Your signature indicates that you consent to participate in this study.

_____ _____

Participant's Signature Date

_____

Participant's Printed Name

Sample Listener Questions

Note: This document contains sample segments of the questionnaires to be used in the experiment. The actual forms will be longer/shorter depending on number of trails, but the following document provides all the possible material that may appear in the questionnaires/experiment forms.

In order:

1.) **Listener Identification Form**

The played back sample may contain up to syllables, and the listener is asked to circle the ones they hear.

2.) **Listener Qualitative Comparison Form**

Two samples are played back to back and the listener is asked to compare the two based on the provided quality/metric.

1.) **Questionnaire: Listening/Comparative:**

1.
i i i i
e e e e
a a a a
u u u u

2.
i i i i
e e e e
a a a a
u u u u

3.
i i i i
e e e e
a a a a
u u u u

Questionnaire: Listening/Comparative:

Which sample sounded SHARPER

1. First Second
2. First Second
3. First Second
4. First Second
5. First Second
6. First Second
7. First Second
8. First Second

Which sample sounded more SPEECHLIKE?

1. First Second
2. First Second
3. First Second
4. First Second
5. First Second
6. First Second
7. First Second
8. First Second

Which sample sounded more EXCITING?

1. First Second
2. First Second
3. First Second
4. First Second
5. First Second
6. First Second
7. First Second
8. First Second

Which sample sounded more INTELLIGIBLE?

1. First Second
2. First Second
3. First Second
4. First Second
5. First Second
6. First Second
7. First Second
8. First Second

Which sample sounded more NATURAL?

1. First Second
2. First Second
3. First Second
4. First Second
5. First Second
6. First Second
7. First Second
8. First Second