

Exploring the Usability of Non-verbal Vocal Interaction (NVVI) and a Pitch Based Implementation

Samuel Williams

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Application

Denis Gračanin, Chair

R. Benjamin Knapp

Jeffrey T. Ogle

December 15, 2023

Blacksburg, Virginia

Keywords: Natural User Interfaces, NVVI, HCI, Usability, Controller Scheme, Pitch

Copyright 2024, Samuel Williams

Exploring the Usability of Non-verbal Vocal Interaction (NVVI) and a Pitch Based Implementation

Samuel Williams

(ABSTRACT)

Natural user interfaces, including verbal vocal interactions like speech processing, are ubiquitous and commonly used in both industry and academic settings. However, this field is limited by the speech and language components. Non-verbal vocal interaction (NVVI) provides further opportunities for people to use their vocals as an input modality. Despite the many possibilities of NVVI input modalities, such as whistling, humming, and tongue clicking, the field is niche and literature is few and far between. This work attempts to address these gaps, as well as the small sample sizes of performed studies of prior work. The problem definition is defined as to perform a large-scale study exploring a pitch-based NVVI modality that uses a relative pitch interaction technique to offer a continuous mode of one-dimensional interaction. A user study is outlined and performed via an ecosystem comprising of Amazon Mechanical Turk for recruitment and study access, a modularized study website, and a secure server that stores the study results, tasks users with controlling a slider with the NVVI technique by humming and whistling, in addition to using the computer mouse to perform these tasks as a baseline. In total, 72 participants' results are considered for analysis. Results show that the pitch based NVVI technique used in this study does not follow Fitts' Law, is not as performant as the computer mouse, humming is a more performant modality with the NVVI technique than whistling, and that participants experienced a significantly higher task workload using the NVVI technique than the computer mouse. Using the results of this

study and from reviewed literature, an NVVI framework is developed and implemented as a contribution of this work.

Exploring the Usability of Non-verbal Vocal Interaction (NVVI) and a Pitch Based Implementation

Samuel Williams

(GENERAL AUDIENCE ABSTRACT)

Speech processing allows people to use language to interact with systems. However, there is much more to the human voice than just language. Non-verbal vocal interaction (NVVI) allows people to use non-speech as a method of interaction. Even though there are many forms of NVVI including humming, whistling, and clicking of the tongue, the amount of existing research is limited in number and high in variety. This work attempts to address these research gaps and to provide a large-scale study. The study conducted in this work explores the usability of a pitch based NVVI system using a relative pitch approach. This NVVI system allows users to control a slider with their voice by whistling and humming. The study is outlined and performed using Amazon Mechanical Turk for recruiting participants, a study website where NVVI is introduced and participants control sliders with the NVVI technique, and a secure server that stores the study results. The study tasks users with controlling a slider with the NVVI technique by humming and whistling, and to use the computer mouse to control the slider as a baseline. In total, 72 participants' results are analyzed. Findings show that the pitch based NVVI technique used in this study has different patterns compared to mouse use. The users perform better with the mouse than the tool, users perform better by humming with the tool than by whistling, and participants felt that the NVVI technique required more effort and energy to use than a computer mouse. Combining the results of this study with the reviewed works, an NVVI software framework

design and coded implementation is described in detail as an additional contribution of this work.

Dedication

I dedicate this work to the pioneers of technologies we take for granted and the people who worked to make them a reality.

Acknowledgments

I'd like to thank my wonderful committee and the Distributed Virtual Environments lab at Virginia Tech.

Contents

List of Figures	xiv
-----------------	-----

List of Tables	xvii
----------------	------

1	Introduction	1
1.1	Background	1
1.2	Motivation	2
1.3	Research Questions and Challenges	3
1.4	Approach	3
1.5	Contributions	4
1.5.1	A Functional NVVI System	5
1.6	Thesis Structure	6
2	Review of Literature	8
2.1	Introduction	8
2.2	Narrative Review	9
2.3	Systematic Literature Review	12
2.3.1	Methods	13
2.3.2	Research Questions	13

2.3.3	Inclusion Criteria	14
2.3.4	Search Methodology	15
2.3.5	Key Words	16
2.4	Results	17
2.4.1	Screening Results	17
2.4.2	Characteristics of Human Acoustics	19
2.4.3	Existing Acoustic Detection and Analysis Algorithms	22
2.4.4	Existing NVVI Systems and Evaluations	27
2.5	Discussion	28
2.5.1	Limitations	29
3	Problem Definition	31
3.1	Research Questions	32
3.2	Research Challenges	32
3.2.1	Recruiting	32
3.2.2	Interaction Design	33
3.2.3	Peoples' Conceptual Understanding of Pitch Based NVVI	33
3.2.4	Human Vocal Characteristics	33
3.2.5	Musical Background	34
3.2.6	Microphone Device Variance and Noise	34

3.2.7	Generalizability	34
3.2.8	Novelty	35
4	Approach	36
4.1	Relative Pitch Interaction Technique	36
4.1.1	Overview	37
4.1.2	Considerations	37
4.1.3	Trade Offs	39
4.2	Pitch Based NVVI technique	40
4.2.1	Sampled Audio Data Input	41
4.2.2	Usage Overview	41
4.2.3	Processing Function	41
4.2.4	Selection	43
4.2.5	Pitch Based NVVI Prototype	45
5	Study Design	49
5.1	User Study	50
5.1.1	Addressing Research Questions	51
5.1.2	Addressing Challenges	52
5.1.3	Study Limitations	56
5.1.4	Study Ecosystem	58

5.1.5	Study Procedure	59
5.2	Statistical Methods	63
5.2.1	Research Questions	66
5.3	Hypotheses	67
6	Results	69
6.1	Pilot Study	69
6.2	Study Results	70
6.2.1	Demographics	71
6.2.2	User Performance	73
6.2.3	Task Workload	86
6.2.4	System Usability	87
7	Discussion	89
7.1	User Experience	89
7.1.1	Performance	89
7.1.2	Feedback	91
7.2	Research Questions	98
7.2.1	RQ1: Does pitch-based NVVI follow Fitts' Law?	98
7.2.2	RQ2: How does a pitch-based NVVI technique compare with a com- puter mouse?	99

7.2.3	RQ3: With which mode of input for the tool do users perform better, whistling or humming?	100
7.2.4	RQ4: Which modality (whistling or humming) do users experience less task workload?	101
7.3	Limitations	102
7.3.1	Methodology	102
7.3.2	MTurk	103
7.3.3	Study Ecosystem	104
7.4	Future Work	105
8	NVVI Framework	107
8.1	NVVI Framework	107
8.1.1	Discussion	108
8.1.2	Key NVVI Implementation Insights	110
8.1.3	Structure and Interaction Flow	112
8.2	Framework Implementation	113
8.2.1	Specifications	114
8.2.2	Examples and Demonstrations	117
9	Discussion	119
10	Conclusion	127

Bibliography	129
Appendices	143
Appendix A IRB Approval Letter	144

List of Figures

4.1	A high level overview of the relative pitch based NVVI system.	40
4.2	A high level overview of transient detection used in this pitch based NVVI system.	44
4.3	The result of the (continuous) ACF algorithm, where the first peak present is located one period from the start.	45
4.4	The RMS algorithm, where x is the audio frame (array containing audio data) and n is the length of the audio frame.	47
5.1	A high level overview of tasks for participants to complete. Participants should use the pitch based NVVI technique to move the bottom slider's knob position to horizontally align with the top slider's knob position.	50
5.2	A high level overview of the study design and flow using MTurk.	53
5.3	The structure of a phase and the special case of Phase 1.	60
5.4	The structure of Phase Tasks and Rounds.	61
5.5	The chronological procedure of the user study.	63
5.6	Fitts' Law, which states the time taken (T) to perform a movement is proportional to the logarithm of the distance travelled (D) divided by the width of the target (W .) [23]	66
6.1	The age distribution of participants of the user study.	71

6.2	The microphone device distribution of participants of the user study.	71
6.3	The classifications of task performances.	74
6.4	Classifications of phase tasks.	76
6.5	A graph representing the average path taken during a phase task, where each phase was normalized on the time axis by the total time taken and on the distance axis by the distance between the starting and ending positions. . . .	77
6.6	Graphs containing all phase task paths in categories 5 and 6 (attempted and successful,) where each phase was normalized on the time axis by the total time taken and on the distance axis by the distance between the starting and ending positions. Additionally, the distance axis is scaled by a factor of 16. .	78
6.7	Graphs plotting the relationship between the time taken to complete tasks and the distance between starting and ending positions.	79
6.8	Common phenomena occurring during tasks.	80
6.9	Box plots comparing NVVI performance with the computer mouse.	81
6.10	A box plot comparing the performance of using humming and whistling with the NVVI technique.	82
6.11	Box plots depicting changes of performance between Phase 1 and Phase 2. .	83
6.12	Box plots comparing NVVI performance using whistling and humming. . . .	85
6.13	Box plots comparing NVVI performance with the computer mouse.	85
6.14	A box plot of the System Usability Scale (SUS) questionnaire scores for the pitch based NVVI prototype used in the study.	87

7.1	Bar charts depicting participants' comfort using the NVVI technique. Answered according to the Likert statement "I felt very comfortable using the NVVI technique with ____."	93
7.2	Bar charts depicting participants' preferred modality to use with the NVVI technique and whether they regretted using the modality they used in phase 2."	94
7.3	Bar charts depicting participants' preferred modality to complete phase tasks (the pitch based NVVI technique or computer mouse.)	96
8.1	The flow and structure of three interfaces customizable by clients implementing an NVVI system using an NVVI framework.	112

List of Tables

6.1	A table containing the NASA TLX questions.	84
6.2	A table containing the average scores for each NASA TLX question for both whistling and humming with the NVVI technique.	85
6.3	A table containing the average scores for each NASA TLX question for using the NVVI technique and the computer mouse.	85

List of Abbreviations

CSS Cascading Style Sheets

DOM Document Object Model

HCI Human-Computer Interaction

HTML Hypertext Markup Language

NUI Natural User Interface

NVVI Non-verbal Vocal Interaction

NVVI is a field of human-computer interaction exploring the use of non-speech vocal acoustics as a form of input. Some examples of non-speech vocal acoustics include blowing, clicking the tongue, and humming.

Hypertext Markup Language is a programming syntax used to render web page components.

The Document Object Model is a live and stateful representation of web page components on a website. DOM elements can be manipulated via interaction with the web page or through JavaScript Code.

Cascading Style Sheets are documents that specify the styling of DOM elements as they are rendered and displayed on a web page.

Human-Computer Interaction is a field of computer science exploring the ways humans interact with machines, computers and programs.

Natural User Interfaces aim to provide intuitive and natural feeling interaction techniques for humans.

Chapter 1

Introduction

1.1 Background

Vocal interaction allows people to use their voice to interact with a system. As technologies have developed over the years, an increase of computational power and cutting-edge research has paved the way for innovative vocal interaction techniques. In the 1950's, one of the first speech recognition systems developed by BELL Labs could identify the digits 0-9 with a 97% accuracy rating [17]. In the mid 1970's, a leap in speech processing capabilities was made by using prediction with Markov models [10, 77]. Fast forward to today, powerful algorithms that power Automatic Speech Recognition (ASR), has led to major developments in speech processing by harnessing artificial intelligence and natural language processing [4]. Speech processing is now incorporated in many consumer systems, such as in Siri and Alexa, two notable user-friendly speech processing systems[12].

These systems share the common trait of using speech as a form of interaction. However, there are more capabilities provided by the human voice than just speech. The field of non-verbal vocal interaction (NVVI) explores the possibilities of leveraging non-speech vocal interactions.

1.2 Motivation

NVVI offers a wide variety of possible interaction schemes. Examples of NVVI include using human acoustics, such as whistling, humming, and clicking, as interaction techniques. Certain characteristics of these types of these acoustics, such as volume, pitch, formant information (harmonic resonances produced during speech [34]), and spectral information from vowel sounds, can be detected, analyzed, and converted to numeric input. The bounds of NVVI extend to the range and variety of possible sounds humans can produce, sans speech.

Some fields NVVI can benefit include accessibility, as they can serve as inputs for physically/motor impaired individuals. For example, a person could control the dimness of their lighting with the pitch of their voice instead of physically adjusting a slider. Language barriers present in speech processing also limit the audience and functionality of speech processing tools.

The variety of possible NVVI types can provide continuous user-generated control inputs, like pitch, not possible with speech processing, in addition to discrete inputs/actions. Some example inputs include humming, whistling, clicking of the tongue, and blowing, among others. NVVI can also serve as an effective alternative/orthogonal mode of interaction, working on top of other modes of interaction.

For example, NVVI could be used in eXtended Reality (XR) environments to allow extended capabilities instead of just the buttons and axes provided by controllers. Compared with other controller schemes in mixed reality, such as any controller scheme that relies on point-to-point interactions involving distance like raycasting, a human acoustic controller scheme does not involve distance and therefore the precision or accuracy of any parameter modulation should theoretically not be affected by distance.

1.3 Research Questions and Challenges

Currently, there is appears to be a lack of frameworks and design guidelines. Moreover, existing studies have small sample studies and are few and far between in terms of variety. The papers considered in this research are only written by a handful of authors.

Overall, there is a limited supply of NVVI literature available. This limits the development of more powerful and effective NVVI techniques and general knowledge of NVVI in academics and industry, which may, in turn, limit the development and inclusion of NVVI in industry.

From this, I aim to address the following research questions with my work.

1. What are existing NVVI systems and how effective are they?
2. How effective and usable is a pitch-based NVVI prototype?
3. What is an effective design framework for an NVVI prototype?

1.4 Approach

This work focuses on exploring and evaluating the usability of NVVI to answer the above research questions. To accomplish this goal, a narrative and systematic literature review using PRISMA [55] guidelines was conducted to identify research gaps and existing NVVI works and evaluations.

Results from the literature review were synthesized into a high level NVVI framework. From this framework, the idea for a user study was conceived by designing a pitch based NVVI technique for users to control one dimensional parameters like a slider. In the study, the usability of the pitch based NVVI technique could be assessed by people using the tool to

control an HTML slider on a web page. Quantitative data, like the time taken and a sampled path of the slider and task workload data, as well as qualitative data, such as user feedback, was recorded and used to evaluate the tool.

The user study was then evaluated to not only assess the usability of the pitch based NVVI technique, but to provide generalizeable design guidelines and user data regarding the user experience with NVVI.

1.5 Contributions

A number of contributions are made by this work. We address the research gap of lack of NVVI user data by conducting a user study with 72 participants. The pitch-based NVVI technique used for the study, as well as the website and tools used to conduct and capture study data, are provided to Virginia Tech. From the results of the literature review and user study, a general NVVI framework is implemented in JavaScript and provided to Virginia Tech along with a website with some demonstrations using the tool. In summary, this work provides the following contributions:

1. A discussion of the current state of NVVI resulting from a systematic literature review conducted following PRISMA guidelines.
2. A detailed explanation and implementation of a pitch based NVVI technique coded in JavaScript.
3. A modular framework for conducting user studies and questionnaires, coded in JavaScript, HTML and CSS.
4. An evaluation of a pitch based NVVI technique with the presentation of both qualita-

tive and quantitative results from the conducted user study.

5. A discussion of relevant usability phenomena observed using the study and design guidelines considering these phenomena and results from the user study evaluation.
6. An implementation of a general NVVI framework coded in JavaScript, as well as a variety of implemented demonstrations using the framework.

1.5.1 A Functional NVVI System

A functional NVVI system will first need to capture a person's audio data. This data should ultimately be processed and served as an input to an application. Before that, the audio needs to be processed to translate the audio data into a meaningful input value, such as a numeric value, to use as input for an application. This result should be sent as an input to an application expecting the type of input. For example, a slider on a web page will probably expect to receive a particular distance to shift by. This would require the NVVI system to translate the audio data into some kind of number that can be considered as the distance to move the slider. This number may be the result of a pitch identification algorithm, for example.

However, the processing will require background knowledge of a variety of factors. Since there are many possible types of inputs for an NVVI system to consider, simply choosing the type(s) of input to be used can help narrow down necessary knowledge to build an NVVI system. However, before building an NVVI system, a designer should have some understanding of 1. characteristics of the human voice (e.g. whistling, humming,) and 2. algorithms to detect characteristics (e.g. pitch detection, noise detection.) Both of these areas may be affected by characteristics of the speaker like biological sex, age, and other factors. For example, people may have nasally voices that influence the presence of

harmonics in the voice.

1.6 Thesis Structure

In this chapter, the motivation, research questions, and approach for this work were discussed. The remainder of this work is organized in the following way.

- In Chapter 2, a narrative and systematic literature review are conducted and the results are discussed. The narrative review was conducted to direct a follow-up systematic review. The systematic review was conducted to explore research gaps and the current state of NVVI in literature.
- In Chapter 3, results from the systematic literature reviews are discussed and developed into the research questions considered for the remainder of the work. Challenges regarding the processes of answering these questions are also discussed.
- In Chapter 4, the proposed approach to answering the research questions is discussed in detail. The rationale is provided for conducting a study using a pitch based NVVI technique to help answer the research questions, as well as its high-level design.
- In Chapter 5, the user study design for exploring the effectiveness of the pitch based NVVI technique is detailed, including the study environment, statistical methods and hypotheses to be conducted.
- In Chapter 6, results from the user study and explored, including qualitative and quantitative data collected. Results from a pilot and actual study conducted with the approval of Virginia Tech's IRB are both considered.

- In Chapter 7, the results are discussed in answering the research questions posed by this work.
- In Chapter 8, a NVVI framework designed considering the results of the user study and systematic literature review is detailed along with an implementation coded in JavaScript with several demonstrations of how the tool can be used.
- Conclusions to be made from this work are explored in Chapter 9 and the work is summarized in Chapter 10.

Chapter 2

Review of Literature

2.1 Introduction

Most people that use computers are familiar with the windows, icons, menus and pointer (WIMP) paradigm, a standard user interface and set of interaction techniques for many common computer operating systems like Windows, MacOS, and Linux. There are also non-traditional user interfaces that aim to improve upon or redesign user interfaces for specific contexts. Natural user interfaces (NUIs) aim to provide interaction techniques that feel natural and intuitive to people and aim to reduce cognitive load.

For example, many mobile devices and some laptops today use touch screens. One such non-traditional interaction technique using touch screens is the finger walking in place (FWIP) technique, where users interact in a way similar to walking with their fingers as an intuitive metaphor leveraging body-based senses [\[36\]](#).

In three-dimensional environments like virtual reality, non-traditional three-dimensional user interfaces and interaction techniques allow users immersed in three-dimensional environments to have the ability to provide more types of interaction. For example, Sardana et. al found that three-dimensional coordinated-multiple views interaction techniques is more efficient than two-dimensional interfaces [\[62\]](#).

Multi-modal techniques allow users to leverage multiple input modalities as a controller,

which can be highly useful for abstract applications. For example, Handosa et. al suggest that multi-modal interactions are preferable to uni-modal interactions for controlling Smart Homes [27].

In this work, voiced-based natural user interfaces are explored, referred to as non-verbal vocal interaction (NVVI.) Two literature reviews were performed to explore NVVI in literature, a narrative and a systematic review. The narrative review was conducted as an exploratory analysis of existing NVVI systems and to give a more clear direction for a follow-up systematic review.

The following three key goals were created for both literature reviews.

1. What are the characteristics of human acoustics?
2. What are existing pitch detection and analysis algorithms?
3. What are existing human acoustic controller schemes?

2.2 Narrative Review

The narrative review was performed to examine the state of existing human acoustic control methods, or methods of controlling parameters involving non-speech user produced audio.

Characteristics of Human Acoustics

According to Stephanie Watson, women generally speak at a pitch that is about an octave higher than mens'. [76] She also claims an adult woman's average range is from 165 to 255 Hz, while a man's is 85 to 155 Hz. This is supported by Sulter et al. who found men's average vocal range to range from 120.2 Hz to 258.7 Hz and women's average vocal range to be from

203.6 Hz to 261.1 Hz [74]. Sulter et al. also devise a phonetogram, a two-dimensional graph with the axes of frequency and intensity (volume), and provide a normal examples [75].

Yoon et al. create and evaluate a speech production model to replicate whistling [80]. In the paper, they discuss and analyze the time-domain and frequency-domain spectra of human whistling. Specifically, they identify 3 formant frequencies, similar to human speech.

Clendinning's dissertation examines and goes in depth into human speech spectra (both time-domain and frequency-domain components) and the human "singing" spectra [15].

Pitch Detection and Analysis Techniques and Algorithms

Cuadra et al. delve into four popular pitch detection algorithms - Harmonic Product Spectrum (HPS), Maximum Likelihood (ML), Cepstrum-Biased HPS, and Weighted Autocorrelation Function (ACF) [18]. These algorithms vary, as some are more efficient while some are better at detecting multiple frequencies. They also discuss limitations with detecting pitch in real time.

In addition to an overview of the human voice for speech and singing, Clendinning's dissertation describes pitch detection algorithms for both the time domain (zero-crossing) and frequency domain analysis (spectrum analysis) [15]. Clendinning also proposes a novel form of pitch detection by using autocorrelation and zero-crossing with the sample's second derivative.

Muludi et al. evaluate a pitch detection technique called Pitch Class Profile (PCP) suitable for identifying multiple notes in a chord [49]. They identify struggling points with the algorithm when tested with guitar chords and particular FFT frame sizes.

Dubnowski et al. create and describe an efficient real-time hardware pitch detector using clipping and simplified autocorrelation analysis [20].

Existing Human Acoustic Controller Schemes

Sporka et al. describe a human acoustic controller scheme for controlling a mouse pointer with human whistling [71]. Their technique analyses the frequency and volume of input audio. They propose and evaluate two control modes. The orthogonal mode moves the mouse only horizontally or vertically, set by preference. It is based on the deviation from an initial frequency, where the speed is dependent on the input volume. The melodic mode wraps an octave into a two-dimensional circle such that moving from a C3-C4, for example, will rotate the mouse in a clockwise motion, while descending the same scale would rotate the mouse in a counter-clockwise motion. Both modes determine if the length of the produced audio is less than t , and if so to perform a mouse click. Results from a small study indicated that orthogonal mode was easier to use than melodic mode. Users also indicated that humming and singing was less tiring than whistling.

Harada et al. conducted and evaluated a study where users learned to use a voice-based user control interface called the Vocal Joystick [28]. The Vocal Joystick control scheme works by creating a two-dimensional vowel space to act as a continuous input. The four cardinal dimensions in this vowel space are the “a” in cat corresponding to north, the “a” in law corresponding to east, the “oo” in boot corresponding to south, and the “ee” in feet corresponding to west, with a continuous interpolated vowel classification in between. Pitch and volume are recorded as additional one-dimensional parameters. Results from a study found that the system was possible to learn and use, as all novices were able to reach to at least 70% of the skill level of “experts” by the end of 10 1 hour sessions.

Limitations and Need for Future Work

There have been many pitch detection algorithms described in existing literature for a long time [15, 18, 20, 49]. These resources, among others, provide plenty of information on different pitch detection algorithms and their strengths and weaknesses. However, Dubnowski et al.'s hardware pitch detector may be of little practical use in today's primarily software-based algorithm ecosystem, though its insights are beneficial.

As indicated by Sporka et al. and Harada et al., there is limited work and existing solutions for human acoustic controlling schemes [28, 71]. Clearly, there is a need for more future work in this niche area. Additionally, Sporka et al.'s study was performed with only 4 participants, so statistical results may not be representative of a larger population.

Discussion

Results from the narrative review indicated that little research and few experiments have been performed that harness human acoustics as a controller scheme. However, the tools to create such a mechanism are readily available and plentiful, such as information regarding the characteristics of the human voice and whistling and many different pitch detection algorithms and implementations. Future work should use this information to develop such mechanisms and experiment with them to contribute to the limited existing literature.

2.3 Systematic Literature Review

The purpose of the systematic review was to examine the current state of literature exploring NVVI and perform a meta-analysis on existing NVVI systems. As discovered in the narrative review, the quantity of literature appears to be limited. Despite this, the aim of this review

was to present the background knowledge necessary to understand the structure of NVVI systems, the overall effectiveness of NVVI systems at this point in time, and a framework for constructing an NVVI system.

More precisely, the goals of the systematic review were as follows:

1. Provide background information on NVVI systems.
 - (a) What are the characteristics of human acoustics?
 - (b) What are existing acoustic detection and analysis algorithms?
 - (c) How can these be used to create a controller scheme (NVVI)?
2. Qualify the effectiveness of current NVVI systems.
 - (a) What are current existing NVVI systems?
 - (b) What have studies shown about their effectiveness?

2.3.1 Methods

This systematic review aimed to solve the aforementioned research questions by following the PRISMA framework [55]. The Search Methodology, Key Words and Inclusion Criteria were performed on a per-RQ basis.

2.3.2 Research Questions

The three main research questions considered in this systematic literature review are:

1. **RQ1** What are the characteristics of human acoustics? (1(a))

2. **RQ2** What are existing acoustic detection and analysis algorithms? (1(b))
3. **RQ3** What are current existing NVVI systems and what have studies shown about their effectiveness? (2(a,b))

2.3.3 Inclusion Criteria

The following criteria were considered for each research question.

RQ1: What are the characteristics of human acoustics?

The inclusion criteria were the following: the paper was in a journal, conference proceedings, or a book, and had at least 15 citations. The exclusion criteria were the following: there were no exclusion criteria.

Publications from any year were considered. The paper was considered if it was written primarily in English or if it was already translated into English.

The publication was included if it explored any characteristics of human acoustics or vocals as previously described.

RQ2: What are existing acoustic detection and analysis algorithms?

The inclusion criteria were the following: the paper was in a journal, conference proceedings, or a book, and had at least 15 citations or was a part of a dissertation. The exclusion criteria were the following: there were no exclusion criteria.

Publications from any year were considered. The paper was considered if it was written primarily in English or if it was already translated into English.

The publication was included if it explained acoustic detection algorithms in depth.

RQ3: What are current existing NVVI systems and what have studies shown about their effectiveness?

The inclusion criteria were limited because not much literature was found on this topic, so as many papers as possible were included. The exclusion criteria were the following: there were no exclusion criteria.

Publications from any year were considered. The paper was considered if it was written primarily in English or if it was already translated into English.

The publication was included if it included an implementation or an analysis of NVVI systems using some continuous acoustical parameter such as pitch or formants (i.e. not speech processing), and was preferred if it contained a study assessing the systems. Publications that performed a systematic or narrative review of NVVI were also included. Publications on speech or verbal interaction schemes, such as voice commands, were not included in this compilation process.

2.3.4 Search Methodology

RQ1: What are the characteristics of human acoustics?

Research for this research question was conducted using the following databases: Science Direct.

RQ2: What are existing acoustic detection and analysis algorithms?

Research for this research question was conducted using the following databases: Science Direct, IEEE Xplore.

RQ3: What are current existing NVVI systems and what have studies shown about their effectiveness?

Research for this research question was conducted using the following databases: Science Direct, IEEE Xplore, ACM Digital Library, Springer Link.

Review Phases

The review phase had 2 parts. First, publications were evaluated based on their title, abstract, introduction and conclusion. Next, some credibility requirements were enforced. For publications on the characteristics of human acoustics and on acoustic detection algorithms, we ensured the number of citations was at least 5. For publications on existing NVVI systems, citations were not considered as the field is sparse and may have excluded a significant proportion of literature. Finally, the publications were thoroughly read and scrutinized based on the eligibility criteria.

2.3.5 Key Words**RQ1**

The primary search terms were: human, acoustics, characteristics. The secondary search terms were: vocals, pitch, whistle/whistling, hum/humming.

RQ2

The primary search terms were: acoustic, pitch, timbre, formant, detection, algorithms, sound. The secondary search terms were: human, voice.

RQ3

The primary search terms were: nonverbal vocal interaction controller, NVVI, acoustics, human. The secondary search terms were: sound, pitch, formant, timbre, scheme.

2.4 Results

2.4.1 Screening Results

Initially, 13 publications were added regarding human vocal characteristics, 28 publications were added regarding acoustic detection algorithms and 14 publications were added regarding existing NVVI systems. Hence, the initial number of publications considered for this paper was 56.

Then, the number of citations of selected publications were considered. 4 publications regarding the characteristics of human acoustics were found with less than 5 citations and therefore removed from consideration in this review. 1 of these papers was a duplicate and was also removed. Another 1 of these papers was inaccessible due to Virginia Tech subscription so it was removed.

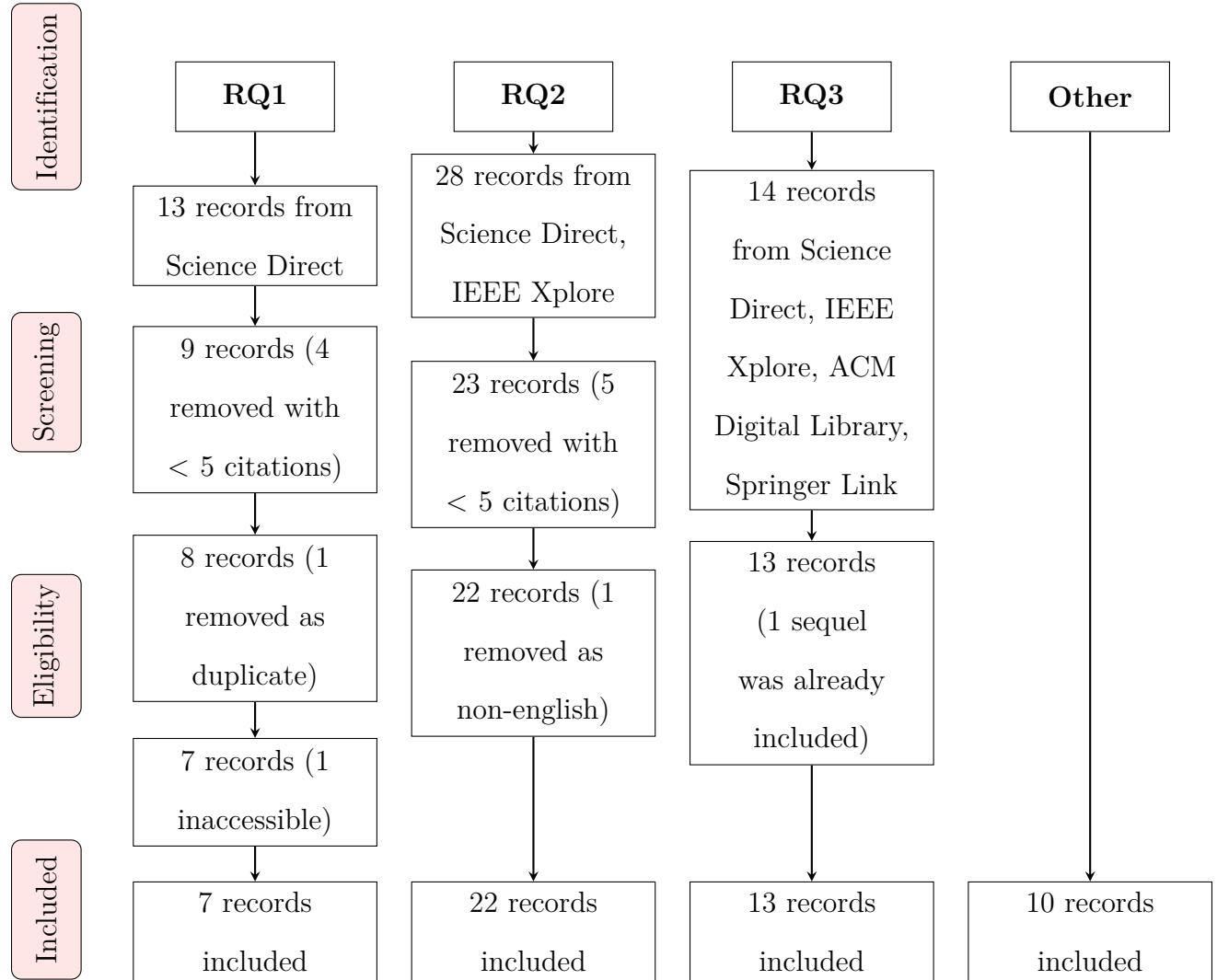
5 publications regarding existing acoustic detection algorithms were found with less than 5 citations and therefore removed from consideration in this review. 1 of these papers was found not to be written or translated into English and was removed.

As mentioned in Section 2.2, the number of citations for publications regarding NVVI were not considered, as the number of available publications was found to be limited. However, 1 was removed as a continuation of that paper was already included and studied instead.

Finally, remaining papers were considered with a thorough screening ensuring their relevance to this research topic. Zero papers were deemed irrelevant in any of the considered categories.

From the initial 54 publications regarding all considered research questions, only 42 remained after the appraisal process, including 7 regarding human vocal characteristics, 22 regarding acoustic detection, and 13 regarding existing NVVI systems. Common finds and trends are explored regarding the first two research questions, while common themes, differences, and limitations of existing NVVI systems are explored.

10 additional publications were added during the review that were deemed important background and credibility publications, for a grand total of 52 publications examined in this review.



2.4.2 Characteristics of Human Acoustics

To answer this question, the characteristics of human acoustics and voice were aggregated. The context of human acoustics and characteristics focused primarily on spectral information, as this identifiable information can be used to identify timbre, pitch, and volume. In turn, this quantifiable information can be used in NVVI systems via acoustic detectors.

Theory

A widely-accepted theory of vocal production is Gunnar Fant's *Acoustic Theory of Speech Production* [34, 47]. In short, the theory proposes that the characteristics of a signal are comprised of the characteristics of the individual components of the system, and that any time variance (delays or advancements) input to a system will result in similar variances as an output.

It is well established that sound is a phenomenon of pressure variance over time. Fant's source-filter theory further explains human speech production as a two-part process [34, 47]. First, sound energy is expelled from the lungs to the vocal folds in the larynx, creating an oscillation (the source). This oscillation describes the fundamental frequency of the speaker, which can vary depending on the length of the vocal folds. Then, the sound energy passes through the vocal tract, where filtering occurs from the nasal, oral, and pharyngeal cavities selectively allowing certain frequencies to pass and blocking others (effectively simulating multiple band-pass filters.) The filtering process will vary depending on factors such as the length and shape of the vocal tract.

Vocal Classifications

Though there are various terminologies reported in the considered papers, sounds produced by humans can be classified into 3 main groups. Vowel sounds are such sounds that are sustained and form syllables [34]. Sibilants, or fricative consonant sounds, are sounds like shushing and hissing [34, 47], and transient sounds, which describe the release of quick stop consonants [34].

More generally, human produced audio can be categorized into harmonic and inharmonic components [63], the former being vowel sounds and the latter describing sibilants and tran-

sient sounds. In addition to the fundamental frequency produced by the source, harmonics created by resonances in the vocal tract are found in the human voice, though they vary from speaker to speaker [34]. The relatively larger or maxima of these harmonics are often referred to as formants [34, 63]. The vocal tract length ultimately determines the spacing between formants [34].

Vowel sounds are unique formant patterns that are created by humans. They can be created by different physical configurations of the vocal tract, such as opening/closing the jaw/throat or placement of the tongue [34].

These are more generally described as laryngeal mechanisms, and are described as the physical adjustments and configurations of the vocal tract [41]. The concepts of registers in singing are a subset of laryngeal mechanisms, and are often described in terms of the vocal fry, head, falsetto, and whistle registers [41].

Inharmonic sounds, specifically sibilants, can also be created by different arrangements of the vocal tract, though they have are uniquely described by airflow channeled with precise arrangements of the tip of the tongue and upper-front teeth (upper incisors), known as the sibilant groove [81]. Yoshinaga et al. create a multimodal model of a simplified vocal tract that can simulate an \s\ sibilant, and validate its results with a airflow supply [81]. They model the \s\ sibilant with 3D geometries (waveguides and rectangular cross sections) and different positions of the airflow supply.

The overall characteristics of an individuals voice are often referred to as timbre, consisting of the harmonic and inharmonic components of ones voice [63]. Thus, due to the many individual components of the source-filter vocal production process, such as vocal fold length, the size of the nasal, oral, pharyngeal cavities, etc., each individual's timbre can be expected to be different.

Influences of Age and Sex

Studies have shown that there is little difference in the vocal tract length between men and women until around 15 years of age, in which men's lengths increase disproportionately, leading to a lower fundamental frequency [47]. But people also have different natural formant frequencies. For an example regarding different types of singers, basses have lower formant frequencies than tenors [34].

External Influences

The surrounding environment can play a role in how people annunciate. With mixed results, ambient sounds such as speech, white noise, and office noise can effect task performance [60]. Additionally, the auditory cortex has been found to become activated during both speech production and detection [21]. Eliades and Wang report vocalization-based suppression of neural activity in the audio cortex during vocal production, in addition to heightened sensitivity of alterations in auditory feedback from the vocalizations.

2.4.3 Existing Acoustic Detection and Analysis Algorithms

To answer this question, acoustics were defined as any speech or non-speech vocals produced by humans. Working off the information described in section 3.1, acoustic detection algorithms were considered for vowel sounds, sibilant sounds, and transient sounds. Speech-specific algorithms were not considered, i.e., any form of speech processing.

It is important to note that the perceived pitch humans here is not always in line with what pitch is detected by algorithms. For example, often times the predominant pitch heard in a signal corresponds to the greatest common divisor of harmonic tones f_0 , even if f_0 is not

actually detected as present in the signal [65].

Spectral Analysis

Most of what we know about vocal and general acoustic characteristics today comes from spectral analysis [21, 34, 41, 47, 60, 63, 81]. Spectral analysis generally involves performing a transformation on an audio signal to analyze and characterize its frequency components. A single “window” of analysis of a signal at a given point in time is often referred to as a spectral envelope [34, 63].

The **Fourier Transform** can be used to transform an audio signal from its time-domain (pressure over time) format to its corresponding frequency-domain format, in the form of energy density (power/intensity over frequency) format, where the energy density is taken as the square of the resulting magnitude of the Fourier Transform for each sampled frequency [1]. This is usually performed as a Short Time Fourier Transform (STFT), generally implemented as a Fast Fourier Transform (FFT), which has a lower time complexity ($n \log n$) than the Discrete Fourier Transform (DFT) (n^2) [16].

The energy density is used to describe the amount of “presence” of each frequency in an audio signal, and hence is useful in quantifying the presences of particular frequencies present in an audio signal. For example, spectral analysis may be useful in identifying a dominant frequency in a signal, though an algorithm must be implemented to analyze the frequency spectrum to make this determination.

Linear Predictive Coding (LPC) is another form of spectral analysis that works by estimating the series of band-pass filters applied to a voice based on the previously described source-filter model [7]. Though the algorithm is often used for speech re-synthesis (and contains more steps than described here), it can be used just for analysis after its filter

identification process. Like the FFT power spectrum, the LPC's estimated filter can be used to estimate the spectral envelope of an audio signal [34].

Spectral Analysis Algorithms

Spectral analysis algorithms use some form of spectral analysis to analyze a spectral envelope in the frequency domain. Some of the most well-known spectral analysis algorithms include the Harmonic Power Spectrum (HPS) [64] and Cepstrum analysis [53].

HPS considers a spectral envelope and multiplies the envelope with downsampled versions of itself, with the logic being that harmonics will align during these multiplications and give a clear indication of present frequencies.

Cepstrum analysis considers the spectral envelope of the signal as a mathematical convolution between the source and filter aspects of the human voice (recall the source-filter model). The logarithm of this spectrum is taken, following the logic that a convolution in the time domain is equivalent to a multiplication in the frequency domain, and a property of multiplications inside of a logarithm is that they may be represented by an addition of two logarithms. Therefore, performing the logarithm effectively separates the source information from the filter information. When an Inverse Fourier Transform (IFT) is taken, the resulting envelope is referred to as the *cepstrum*, and, in effect, the source and filter information is separated, which is helpful for examining fundamental pitch and harmonic/formant data independently from each other.

Modern implementations of frequency detection algorithms use variations and optimizations of these fundamental algorithms or spectral analysis in general [3, 8, 22, 40, 66].

Time Domain Analysis

Time domain analysis is another mode of audio analysis, where the audio sample data is the basis of analysis. Unlike spectral analysis, time domain algorithms are not directly converted to the frequency domain.

Time Domain Analysis Algorithms

Some well-known and frequently used time domain analysis include the auto-correlation function (ACF) [68] and the average magnitude difference function (AMDF) [61]. The **ACF** correlates an audio signal with itself, as a periodic wave will correlate with a phase shifted version of itself at every period. The **AMDF**, a popular variation of the ACF, creates a difference signal based on the magnitudes of an input signal and a delayed version of itself, and when the phases align, this difference should be zero. Another classic but primitive algorithm is the zero-crossing rate analysis, which estimates the period of a signal based on the number of times the signal crosses the zero (hypothetically, the center of the wave).

There are many variations of the ACF function and the AMDF function [5, 22, 26, 32, 35, 38, 48], many of which contain optimizations and improvements that have been built upon these algorithms over the years. Some even allow for polyphonic (multiple pitched) frequency detection [9, 38].

Other Algorithms

Algorithms with a combination of frequency and time domain analysis have been explored [30, 31]. **Wavelet** algorithms have been explored to detect frequencies in audio signals [54] that work by performing a wavelet transform, an alternative to a Fourier Transform where a unique wavelet, or wave shape, can be convolved with a signal and can provide more localized

information than a Fourier Transform [78].

Some frequency detecting algorithms explore more statistical and probabilistic approaches of spectral frequency detection, including maximization and minimization algorithms [9] and even artificial neural network algorithms [67].

Some algorithms have been developed specifically for individuals with pathological, or atypical in terms of harmonics, voices [37, 43]. Note that some of these algorithms are also used for noise detection and removal, including room, machine noise and audible artifacts like jitter and shimmer [19, 35, 67]. Voice detection algorithms have also been developed specifically to determine if a voice is present [79].

Algorithm Evaluations

There are no clear best picks when it comes to the most accurate frequency detector algorithms [33], as there are many different conditions and metrics to be considered. However, there are some general takeaways to be considered by the evaluation and comparisons of different algorithms.

Overall, pitch detection algorithms' accuracy as the level of noise increases [33, 45]. Different algorithms react to changes in noise levels and types of noise differently [33]. Additionally, babble noise, or background noise created by other speakers, is more harmful than other types of noise. The distance of a speaker to the microphone effects the accuracy of pitch detection algorithms [33]. Pitch detection algorithms have been found to not perform as effectively on dysphonic voices [43].

Kolhatar et al. found a cepstrum based algorithm to have the highest accuracy of pitch identification for pathological voices, though the ACF was identified to have the fastest mean time for classifying a voice as pathological [37]. Similarly, Luengo et al. found an

implementation of cepstrum analysis to perform with highest accuracy compared to other algorithms, and also found the ACF to perform the best at classifying voiced or unvoiced signals [45].

2.4.4 Existing NVVI Systems and Evaluations

NVVI literature is limited, and thus the quantity and variety of existing prototypes is as well. However, NVVI systems may be useful during hands-busy situations and a useful design for individuals with physical disabilities [29, 39], and thus further research may prove NVVI to be a valuable mode of interaction across disciplines. The types of interactions described in the papers analyzed in this review were found to fit into 2 categories: trigger and continuous input. Additionally, some of the examined work explore vocal gestures, which are short melodic or rhythmic patterns to trigger commands [58].

Trigger Input

One study used blowing as an input to trigger clicking [82]. User input was captured and processed to determine if they were blowing, and if they were, a mouse click was triggered. The Vocal Joystick NVVI system processes a discrete k sound to trigger a mouse click [46].

Continuous Input

Pitch is a common mode of continuous NVVI input [29, 30, 57, 58, 69, 70, 73]. Many of the systems use pitch as the basis for a vocal gesture-based interaction system, explained in the next section. However, properties of pitch alone can be used as an interaction technique [29, 30]. Formant information can also be used as input to an NVVI system, as demonstrated by Harada et al. [11]. Their technique analyzed the user-produced formants pertaining to a

set of vowels and mapped them to a two-dimensional grid to be used like a joystick. Whistling and humming appear to be the most used forms of pitch input.

Vocal Gestures

Vocal gestures are short melodic or rhythmic patterns to trigger commands [58]. For example, if a user starts by humming a pitch at frequency f_0 and slides it up to a new pitch f_1 , and the NVVI system has a mapping from this gesture to a command, the corresponding command will execute (if it is recognized properly). Gestures were used in several of the examined publications [57, 58, 59, 70, 73].

2.5 Discussion

In this systematic literature review, the niche field of non-verbal vocal input (NVVI) is explored. The methodology of acquiring materials used in the review is described following PRISMA guidelines [55]. Results of the review explore human vocal characteristics, frequency and other acoustic detection algorithms, and a limited set of NVVI systems.

Results from this review provide valuable insights into how to design an NVVI system. Human vocal characteristics were described, specifically the source-filter theory and formants/harmonics. Frequency and noise detection analysis and algorithms are explored. A limited set of existing NVVI systems are described and informally evaluated.

User performance with NVVI compared with traditional techniques varied from study to study, and there does not appear to be any conclusive quantitative assessment of the effectiveness of NVVI in general compared to other interaction techniques. This is largely due to the variety and limited quantity of available work.

Users may prefer alternatives to NVVI systems [59], but some users still think favorably of NVVI in terms of interest and enjoyment [29, 30]. A major limitation of pitch and gesture-based NVVI systems appears to be skill and physical ability to produce the required frequencies [11, 58]. For example, fatigue is a common response to the use of NVVI systems. However, NVVI appears to be a strong alternative of interaction for the physically impaired [57, 73].

Poláček et al. found that users preferred non-complicated gestures, such as flat tones [58]. They also found users preferred absolute pitch gestures to relative pitch gestures. NVVI control may be a more favorable control style than speech commands [70].

Design Considerations

Using volume as an input is a considerable interaction technique, but user embarrassment in vocalizing NVVI input may play a factor in their comfort using this technique. [29]. Poláček and Míkovec note that pitch gestures are a cultural and language independent mode of vocal interaction [59].

The ACF appears to be a common implementation for pitch detection and control as a continuous input [29, 58, 70]. Zielasko et al. used machine learning techniques to more accurately determine if a user was blowing [82]. Multiple algorithms may be used for better resolution depending on the frequency range [30].

2.5.1 Limitations

In the reviewed studies, the main limitation was the limited number of accessible publications found on existing NVVI systems. However, another two major limitations were the number of authors and the size of studies. Regarding the former, for example, Sporka was included

as an author to 8 of the 14 included studies of NVVI systems [39, 46, 57, 58, 69, 70, 72, 73], while Poláček was included as an author in 4 [57, 58, 59, 73]. This supports the claim that the area of NVVI is a niche field of research. Only 3 studies were conducted with more than 20 participants [58, 59, 82], while 4 studies had 10 or less participants [11, 46, 70, 73]. Additionally, only 3 studies were conducted with disabled or impaired participants [29, 57, 73]. Some studies did declare to perform studies with disabled persons in future work [11, 46, 59].

A smaller limitation was the variety of projects and prototypes used these studies, including video games [29, 30], keyboard alternatives [57, 73], to controlling radio cars [70], among others. In conjunction with the limited quantity of literature and relatively small sample sizes, it is difficult to accurately synthesize corroborating information from the results of the variety of studies performed.

With the limited quantity of accessible literature specifically about NVVI systems, and in addition to the relatively small study counts over a wide variety of task contexts, it is difficult to find many obvious trends among NVVI systems. Additionally, most of the analyzed work was designed and performed by a small circle of authors, meaning their biased design and implementation choices may limit the variety of NVVI systems regarding both design and implementation. Furthermore, only 3 studies were performed on users with impairments and disabilities who would specifically benefit from NVVI. Overall, conclusions are heavily reliant on corroborating results from user studies including qualitative feedback (user experience) and task measurements.

Chapter 3

Problem Definition

Findings from the literature reviews indicate a lack of quantity of literature and studies necessary to understand broad trends and general design principles for NVVI systems. Additionally, the existing studies were few and far between in variety and were not performed with enough participants to make generalizeable claims.

To conduct further investigation into NVVI, one of many routes was pursued - an exploration of a pitch based NVVI system. A pitch based NVVI analysis was chosen because of the unique advantage it offers over other forms of NVVI - continuous input. Though there are many types of NVVI modalities, pitch is a well known and researched human vocal characteristic that has been studied extensively, explored in the review of literature. Other forms of continuous NVVI input, including formants, could have been considered as well. Humming and whistling are two modalities present in prior NVVI systems studies that are considered for pitch input for studies performed in this work, as they are unique acoustics that have not been adequately compared in prior work.

Thus, this work aims to address some of the limitations of prior work by addressing the following research questions.

3.1 Research Questions

1. Does pitch-based NVVI follow Fitts' Law?
2. How does a pitch-based NVVI technique compare with a computer mouse?
3. With which mode of input for the tool do users perform better, whistling or humming?
4. Which modality (whistling or humming) do users experience less task workload?

3.2 Research Challenges

Pitch based NVVI systems are not novel themselves and have been implemented in prior work explored in the review of literature. However, prior work is limited in quantity and the studies they performed were limited in the number of participants, which affected generalizability of the authors' claims, as well as the ability to identify consistent patterns in usability. This work aims to address that by performing a user study with more participants, but some challenges are expected.

3.2.1 Recruiting

Since prior studies did not include many participants, a particular challenge of this work will be to recruit many users to build more well supported claims and be more statistically sound. To address this challenge, a different type of recruiting approach from different studies will be necessary.

3.2.2 Interaction Design

Though pitch has been used in prior work, it has been applied variously in terms of exact translation to an input. For example, Sporka, Adam J. and Slavík used pitch gestures to control a remote control car[70], Hedesly et al. use pitch to control a video game[29], and Mahmud et al. explore using pitch to control a computer mouse [46]. Figuring out a robust but effective approach, as well as one that is learnable in the span of a single user study, may prove to be challenging.

3.2.3 Peoples' Conceptual Understanding of Pitch Based NVVI

Though it is expected most participants will be aware of the concept of pitch, it is foreseeable that understanding how it can be used as a mode of interaction will be difficult for some to understand, and therefore difficult to complete tasks. To address this, the concept of NVVI and an overview of any pitch based NVVI should be explained thoroughly and provide an example usage for any study conducted as a part of this work.

3.2.4 Human Vocal Characteristics

Since pitch is the modality considered for this design, it is important to recognize the domain of input and recall that pitch is an extracted characteristic of the human voice. As described in section LIT REVIEW, the human voice is comprised of a fundamental pitch as well as formants (harmonics.) It will be important to consider particularly divergent voices, including not only the vocally impaired (cite,) but even people with very formant-heavy vocal spectrums (e.g. nasally voices.)

3.2.5 Musical Background

Since this study is exploratory, it is unknown how much a person's ability to control the pitch of their voice will effect the usability of a pitch-based NVVI system. Thus, a person's musical background may affect the performance of successfully using the tool, and a high musical background requirement may alienate much of the tools' intended audience.

3.2.6 Microphone Device Variance and Noise

Anecdotal experience and experimentation while designing the NVVI prototype and tools indicated a variance of performance depending on the type of device and microphone used. The type of microphone device that a user of the pitch based NVVI technique will use may have a considerable impact on the performance abilities for users to complete tasks. A robust pitch detection algorithm that works well with a variety of audio input qualities and characteristics, including background noise and low input volume, should be used for implementation with this tool.

3.2.7 Generalizability

Pitch is one of many modalities for an NVVI system. This work aims perform analysis of a larger study than prior work to develop generalizeable claims for a pitch based NVVI system. However, the challenges of choosing one of many pitch interaction techniques (as described in 3.2.2) for analysis, as well as the additional of many other forms of NVVI (e.g. blowing, tongue clicking, formants) may make it hard to claim any findings are generalizeable in the broad scope of NVVI as a whole.

3.2.8 Novelty

Though existing work on particular NVVI applications have been performed, there is a lack of existing design frameworks and evaluation systems for the area. A partial explanation for this may be related to 3.2.4, i.e. there are many different forms of NVVI systems, and each may have their own. Even for a pitch based NVVI system, results and conclusions presented in this work will be novel in their own right, but will not cover enough territory to necessarily make generalizeable claims about NVVI systems as a whole. Ultimately, more studies examining different facets of NVVI will need to be conducted in the future with many participants, and then a broad outline of general NVVI guidelines may be devised.

Chapter 4

Approach

In Chapter 1, the motivation for this work was presented, to explore NVVI systems. Literature reviews were performed and the resulting research gaps, as well as limitations of prior work, were presented in Chapter 2. In Chapter 3, the resulting research questions regarding an analysis of a pitch based NVVI system were introduced, as well as foreseeable challenges in conducting the research and applying analysis while answering them. In this chapter, a pitch based NVVI design is detailed to help address the aforementioned research questions.

In Section 4.1, the particular pitch based interaction technique is detailed. This section includes a high level description and rationale behind the particular technique, and includes trade offs with other pitch based techniques.

In Section 4.2, the particular pitch based NVVI system is detailed. This section explores the high level design, flow, and intended use of the system.

4.1 Relative Pitch Interaction Technique

Though there are numerous pitch interaction techniques to choose from as described in the review of literature, the one chosen for this work is a relative pitch system.

4.1.1 Overview

The idea behind this approach is that any change of the user's pitch will result in a positive or negative numerical value to be used as the output of the system, depending on if the pitch increases or decreases. If the pitch is not estimated to have changed, then the tool should output 0. The numerical output can then be used move some other application's input in a continuous, one-dimensional direction (e.g. left or right, up or down.)

4.1.2 Considerations

This pitch interaction technique is robust but requires several considerations before implementing in a pitch based NVVI system.

Identification of the User's Initial Pitch

This pitch interaction technique relies on a relative change of pitch, but what exactly is the user's initial pitch? Most pitch detection algorithms will be able to estimate a fundamental pitch given any input audio data, but identifying the user's *intended* pitch is a separate matter.

Imagine a user has enabled a pitched based NVVI technique using this pitch interaction technique but has not yet provided input. The input audio data into the NVVI system will likely be only room noise, and the pitch detected by the algorithm will be inaccurate. It may deduce a particular humming sound present in the room as the pitch to use, or the pitch may fluctuate wildly depending on the noises present.

Ultimately, the tool should be able to discern the user's intended pitch from any background noise. Though there are many approaches that could be taken to tackle this problem,

including noise detection, some *confidence* of the user's intended pitch should be present in an NVVI system. That way, once a user starts producing an intended pitch, it should only then be considered by the NVVI technique, and likewise when the user stops producing an intended pitch, the NVVI technique should recognize this and stop considering an pitch input until the user starts again.

Real Time Analysis

This approach will require the fundamental pitch of the user's voice to be estimated in real time. There are different approaches to perform pitch analysis in real time, explored in the review of literature. The algorithm used to estimate the pitch should not significantly affect the performance of this tool. However, all pitch extraction algorithms described are performed in frames, or segments of audio data.

A general rule of thumb is that the more data considered during frequency analysis, the higher resolution and more accurate the pitch algorithm is. That said, there are not only computational tradeoffs to larger frame sizes, but also potentially less precise frequency measurements over a period of time. As an extreme example, if a pitch detection algorithm was highly accurate but was only performed analysis once per second, the tool would be difficult to use for fine adjustments.

A countermeasure for this drawback is to perform overlapped analyses (cite) at a fast rate. This means that larger frames can be considered by also including some previous audio data more frequently, so that the dramatized example above does not occur even with large frame sizes. However, it will incur additional computational overhead.

Frequency's Exponential Scale

Pitch and frequency are semantically different terms. Frequency usually refers to an absolute measurement taken in the units of Hertz (Hz) referring to the number of cycles or oscillations occurring per second. Pitch is more of a musical notation expressing the relationships between multiple frequencies. Pitch and frequency are related by an exponential/logarithmic scale due to the physical properties of sound [2].

What this means for the design of this tool is that changing a pitch from one octave to another is not a linear shift in frequency. Expected behavior for this NVVI system is that a change of pitch in one direction should be equivalent to the same difference in pitch in the other (e.g. going up an octave should result in the same amount of change as going down an octave.) Therefore, a pitch based NVVI system will likely want to consider this mathematical relationship so that a translation up or down an octave in pitch translates into some shift up or down equal in magnitude but opposite in direction. This can be done by using a mathematical conversion from frequency to pitch.

4.1.3 Trade Offs

The primary benefit of this approach is that it does not restrict the range of input frequencies. This particular property allows for a configuration-less tool that users can use regardless of the average pitch or range of their speaking voice.

One negative aspect of this approach is that it is not as detailed as some other approaches may be in considering and matching a particular user's vocal characteristics. For example, an alternative scheme that was based on recording the user's voice and extracting their unique vocal characteristics may prove to be more accurate and consistent, with the trade off of using a more expensive algorithm and requiring some initial configuration. Additionally,

an approach that requires even a basic configuration or calibration of a specified range of frequencies (e.g. one octave comfortable for the user) may allow for easier usage.

That said, some hybrid approaches could be created that leverage some of the advantages described in this design with others. Future work may choose to explore some of these options.

4.2 Pitch Based NVVI technique

The NVVI system design described in this work is a relative pitch based NVVI technique that converts user audio data into two components: a *pitch* and *confidence score*. These two pieces of information can be used together for an application using this system. The tool also provides a method of *selection* to determine when the tool should be active (listening), controlled via *transient detection*. Figure A. depicts a high-level visual explanation of the NVVI prototype. The tool is also designed to be generalizeable, and points are made where applicable for implementers to consider design decisions.

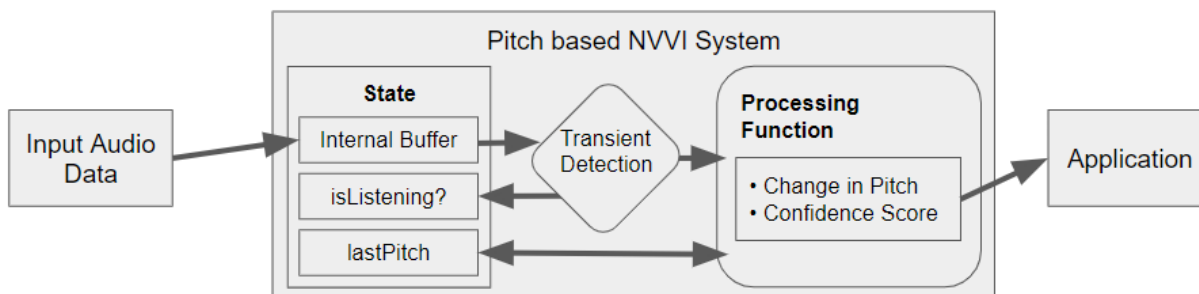


Figure 4.1: A high level overview of the relative pitch based NVVI system.

4.2.1 Sampled Audio Data Input

The NVVI prototype is fed sampled audio data in real time to provide a pitch and confidence score. Since the NVVI prototype is meant to work in real-time, the audio input should be segmented, or *framed*, into buffers as they are sent in to the tool. The tool will maintain an *internal audio buffer* where it will store a user-defined-sized buffer. Any processing will be performed with audio data from the *internal buffer*. The client's sole responsibility is to keep passing in frames of audio data in real time.

4.2.2 Usage Overview

The NVVI prototype will initially be inactive (not listening.) It will continue to update the internal buffer with data, but processing will not occur unless an *activation* is performed, and likewise will continue until a *deactivation* is performed. The activation and deactivation will be synonymous operations defined in our tool as the process of *selection*, or indicating when the tool should be in a *listening state*.

In a listening state, the tool will perform a processing operation every *frameSize* samples, as defined by the user. This will occur once every $\frac{frameSize}{sampleRate}$ seconds. Pitch detection and confidence scoring will be the result of this processing function.

4.2.3 Processing Function

The pitch will be detected using a specified pitch detection algorithm. Every pitch detection algorithm should have a corresponding confidence scoring function. Thus, our tool couples the pitch detection and confidence scoring in one function. Additionally, we provide an interface for defining custom pitch detection and confidence scoring functions, which here

on out will be referred to as *processing functions*.

Specifically, any *processing function* should return an absolute frequency that is in the units of Hertz, along with a corresponding confidence function. It may make sense to have unique or universal confidence functions for whichever pitch detection algorithm is chosen, considering things like efficiency and approach (e.g. a Fourier Transform's confidence function could include some analysis of the frequency spectrum as a whole, whereas this may not be an efficient approach for a time-domain pitch detection algorithm.)

A processing function should load data from the *internal buffer* into a separate *frame buffer*. This frame buffer should be used as the input for the pitch detection algorithm, and possibly used for the confidence scoring.

The *processing function* is responsible for returning a coupled pitch value (in the units of Hertz) and confidence score (in the range $[0, 1]$.) The function should only access a provided *frame buffer* to consider as the audio input data to be processed. The function may also use necessary variables, such as the *frameSize* and *sampleRate*, to derive an accurate pitch and confidence score.

Pitch Detection

Our tool provides the ability to use a Harmonic Power Spectrum (HPS) or Auto-correlation Function (ACF) pitch detection algorithm. Anecdotally, the ACF has higher performance and better resolution. Both algorithms, however, suffer from low resolution for lower frequencies that correspond to typical biologically male vocal ranges. Some hybrid or unique approaches, such as low-frequency DFTs or other pitch detection algorithms, can be implemented to account for this, so long as they abide by the requirements described in the Section 4.2.3.

Confidence Scoring

Confidence scoring is necessary for a pitch-based NVVI technique to allow the user to decide when or when not to trust a detected pitch value. Some reasons not to trust a value 100% are highly present harmonics, background noise, multiple voices, and low input volume. The confidence score should be a decimal number in the range $[0, 1]$, where 0 indicates no confidence and 1 indicates pure confidence.

It is up for the programmer of the processing function to decide how to derive this value. It makes sense to include the mentioned reasons in the decision-making process when developing a confidence score. Unique pitch detection algorithm may provide programmers an opportunity to capture information they may not with other algorithms without sacrificing efficiency.

The confidence should represent a confidence value for the pitch detected of the **current frame being analyzed**. We suggest not implementing multi-framed confidence at this stage to try and encapsulate the low-level *instantaneous* pitch detected and confidence scored. Experimentation and future work may suggest otherwise.

4.2.4 Selection

In order to specify when the tool should or shouldn't be active (in a listening state,) we decided to design the ability to set the listening state of the NVVI prototype via audio input. Specifically, we design this process to consider a *transient*, a short and sudden peak or impulse of audio.

Our tool allows for an arbitrary number of transients to be considered as the activation cue. However, there are two main concerns with detecting a transient. First, the transient must

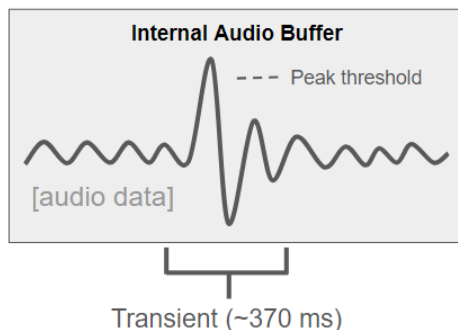


Figure 4.2: A high level overview of transient detection used in this pitch based NVVI system.

be able to fit inside of the *internal buffer*, that is, its length must not exceed the length of the internal buffer. Additionally, we hypothesize that at 1 transient could incidentally be detected but not be intentionally created by the user, so we suggest considering at least 2 consecutive transients to trigger activation.

Figure 4.2 provides a visual explanation of the general process of transient detection.

Transient Detection Algorithm

We define a threshold for detecting if a transient is present or not in the signal. We consider the audio data in the *internal buffer* to detect a transient. We experimentally defined a 370 millisecond window to be considered as a transient for our NVVI prototype.

A variable inside of the NVVI system stores the current number of transients detected inside the internal audio buffer. Just before the processing function runs, the transient detection algorithm checks for a change in the number of transients currently present in the buffer. If the number of transients increases, this will cause the system to check if the system should activate or deactivate (based on the configuration specifying the number of transients required to trigger activation.) If the number of transients decreases, nothing happens except

that this variable is decremented by 1.

4.2.5 Pitch Based NVVI Prototype

The NVVI prototype designed and implemented in this study follows the design outlined in Section 4.2. It is implemented in JavaScript and leverages the WebAudioAPI to access user audio data. It is detailed in this section. The code used for implementation is provided along with this document as a contribution.

Overview

The pitch based NVVI technique leverages the relative pitch interaction technique along with selection, allowing the user to specify when the tool should be listening for input.

Pitch Estimation

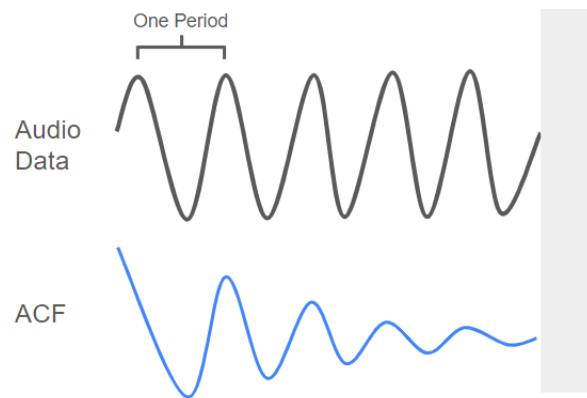


Figure 4.3: The result of the (continuous) ACF algorithm, where the first peak present is located one period from the start.

The pitch estimation algorithm implemented for this tool is a variant of the auto-correlation function. The auto-correlation function described is similar to the algorithm described in

the review of literature [68], which is depicted at a high level in Figure 4.3. In short, the pitch estimation algorithm used in the pitch based NVVI prototype works by performing the following procedure.

1. Normalize the audio frame so that the maximum value has a magnitude of 1.
2. Perform the auto-correlation function and store the results in an array. The array is indexed by the offset used when comparing the audio data with itself. This is the key insight to the ACF algorithm, as the value indexed by the offset corresponding to one period will be the maximum value.
3. The location of the first peak indicates the presence of the lowest detected frequency. This is important for voices with loud formants/harmonics.
4. Parabolic interpolation is performed to get a more precise estimation of the peak.
5. The frequency is calculated by dividing the sample rate of the audio data by the location of the peak.
6. This frequency is compared to the last frequency and the difference in pitch (not frequency) is reported.

Confidence Scoring

The pitch based NVVI prototype performs confidence scoring by considering the root-mean square (RMS) value of the current audio frame. In short, this translates to the average perceived loudness level of the sound for that audio frame. It is performed by taking the square root of the summation of the squares of all the values present in the audio frame, and dividing it by the size of the audio frame.

$$RMS = \sqrt{\frac{\sum_{i=0}^{n-1} x[i]^2}{n}}$$

Figure 4.4: The RMS algorithm, where x is the audio frame (array containing audio data) and n is the length of the audio frame.

A low RMS value indicates that a speaker is not producing a pitch, and that any input is likely just background noise. A high RMS value indicates a high likelihood that the user is producing a pitch. This approach is logical but perhaps not suitable for quiet inputs or inputs with lots of background noise.

Aggregation

In our pitch-based NVVI prototype design, we propose methods to aggregate resulting pitches and confidence scores to report back to users. This implies an additional layer of the NVVI prototype design that is responsible for providing an aggregate pitch and confidence score. Some metrics such as *smoothness*, *precision*, and others may be considered.

Smoothness is a term developed for our prototype describing the number of pitch and confidence scores considered during aggregation. Higher values may result in a smoother pitches and confidence values upon input frequency and volume modulation. Lower values may appear to lead to more jerky results.

Precision is a term developed for our prototype describing some form of rounding of resulting pitch values. For example, instead of considering Hertz accurate to the ones position, we may round a pitch so that it increments less frequently so that similar pitch values will not change when reported back to the users. Importantly, the rounding, if done in units of Hz, needs to consider the logarithmic distribution of frequencies so that low frequencies are not

affected more than higher ones.

Inputs and Outputs

Input. The input to the pitch based NVVI technique was audio data collected by microphone access granted via the WebAudioAPI. If the user does not grant the NVVI technique access to the microphone, the tool cannot start. The pitch based NVVI technique internally encapsulates an `AudioWorkletNode`, a WebAudioAPI component that processes audio in its own thread. Most of the state, including the internal audio buffer, frame buffer, and processing functions are stored inside the `AudioWorkletNode`.

Output. The output of the pitch based NVVI prototype is a parameter sent to a specified callback function. This parameter contains a *session* object. The session object contains information about the state of the NVVI technique, including its listening status, as well as the result of the processing function.

However, as indicated in the design of the tool, the result should be a relative change of pitch, not an absolute frequency as specified in Section 4.2.3. In this implementation, a layer outside of the `AudioWorkletNode` holds the state of the last frequencies detected. Then, as mentioned in the previous section, aggregation may be applied and the relative change in pitch may be propagated to the callback function.

For this particular implementation, a *session* object is passed as the parameter to the callback function, but contains not only just the (aggregated) difference in pitch, but the initial frequency, current frequency, and current deviation from the initial frequency. The initial frequency was the first detected frequency that exceeded the specified confidence threshold to start the session. This provides more information to applications in case they wish to use the data differently.

Chapter 5

Study Design

In previous chapters, one of the primary research gaps considered was the limited number and participant count of prior studies. In Section 1.5, one of the contributions this work aimed to achieve was to conduct a large study with a large number of participants, as well as to present the qualitative and quantitative results recorded during the study. In Section 3.2, a research challenge of recruiting a large number of participants was discussed. This study aims to address these goals and challenges by conducting an extensive large-scale study with detailed data points and demographic information, as well as an extensive analysis of results.

The entire study and NVVI prototype implementations are available as a part of this work's contributions.

In this chapter, the study design will be laid out and detailed. In Section 5.1, the user study will be outlined, including how research questions will be addressed in 5.1.1, how research challenges will be addressed in 5.1.2, limitations of the study in 5.1.4, the implementation of the pitch based prototype in 5.1.3, the ecosystem of the user study environment in 5.1.4 and the procedure in section 5.1.5. In Section 5.2, the particular statistical methods used for analysis will be detailed. In Section 5.3, the hypotheses that will be tested for each research question will be outlined.

5.1 User Study

A study was performed to examine user feedback of an implementation of a pitch-based NVVI prototype that was used to perform the task of controlling a continuous, one dimensional input. The parameter decided to be controlled by the pitch based NVVI technique was an HTML ranged input element (a slider.) Additionally, we wanted to compare this feedback with a baseline traditional method, so we chose to compare task performance using the pitch based NVVI technique with a computer mouse (where participants would click and drag the slider instead of using the pitch based NVVI technique.) Figure 5.1 provides an example usage of the tool for this study. The study was conducted via Amazon Mechanical Turk[6], and hosted on a Github Pages website [25]. Participants were rewarded \$1 US Dollar. The Institutional Review Board of Virginia Tech approved the study.

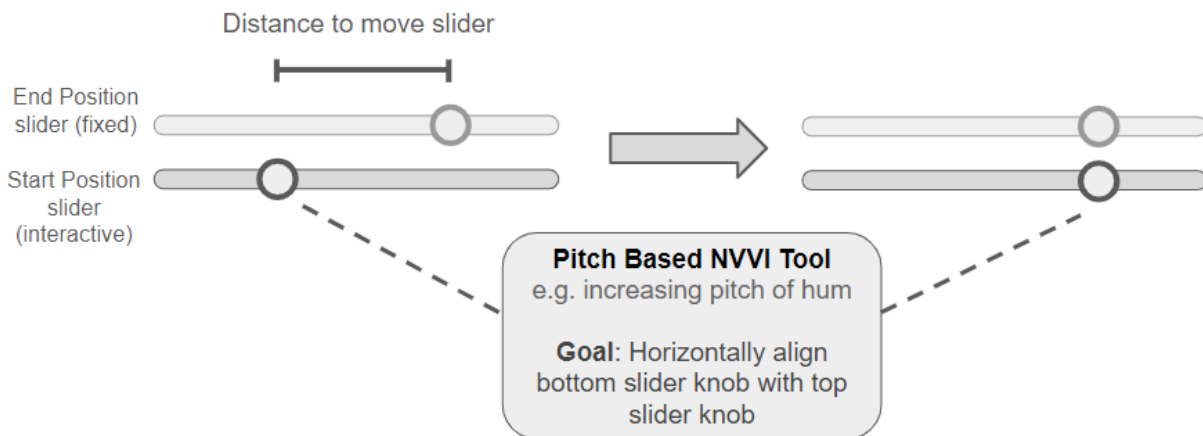


Figure 5.1: A high level overview of tasks for participants to complete. Participants should use the pitch based NVVI technique to move the bottom slider’s knob position to horizontally align with the top slider’s knob position.

5.1.1 Addressing Research Questions

In summary, the research questions considered in this study are answered in the following ways.

RQ1: Does pitch-based NVVI follow Fitts' Law? Data should be collected representing the users' interaction with the pitch based NVVI technique. Effectively, the position of the slider should be recorded for the duration of the task so that the path and time taken to complete tasks can be analyzed. Additionally, the resulting data should be aggregated to find interaction patterns and trends, and ultimately to see if this form of pitch based NVVI follows Fitts' Law.

RQ2: How does a pitch-based NVVI technique compare with a computer mouse?

A baseline test should be conducted to compare the paths and times taken using the pitch based NVVI technique and the computer mouse.

RQ3: With which mode of input for the tool do users perform better, whistling or humming?

Participants should have the option to decide whether to hum or whistle using the tool, but have the opportunity to try both modalities. When assessed, the user should indicate which modality they are using so that the resulting data may be correctly compared.

RQ4: Which modality (whistling or humming) do users experience less task workload?

A NASA TLX[14] (a brief task workload questionnaire) should be conducted just after the participant has been assessed to gauge the common types of fatigue experienced by participants while using this tool.

5.1.2 Addressing Challenges

In Section 3.2, foreseen challenges with conducting research were identified. Though it is not expected that all challenges are adequately addressed, the work done in conducting the user study attempts to alleviate their affect on the validity of results in the following ways.

Challenge 1: Recruiting

Since a particular goal of this work is to conduct a large scale study with many participants, a shortcoming of prior work, a different approach is necessary. In this work, Amazon Mechanical Turk [6] (MTurk) was chosen as the method to conduct the user study. MTurk allows for the conduction of an asynchronous study, provides advertising for the study, and allows participants to be monetarily rewarded after completing the study. A high level explanation of MTurk is as follows.

A *Requester* can develop a High Intensity Task (HIT) *task* for *Workers* to complete. This *task* is separate concept from tasks related to this particular study. Workers are advertised these tasks, which offer different incentives (usually direct payment.) A Requester can decide the amount to reward workers for completing a particular task.

A task will usually include two elements, a **website link** to an external survey for workers to complete, and a **text input** for workers to enter a *completion code*. The completion code is usually provided at the end of the survey and serves as proof that the user completed the task successfully. Sometimes, completion codes are given out upon error encounters. The worker submits the completion code as the final step of the task.

Requesters review all task submissions before awarding workers. If workers enter the correct completion code, a requester can choose to reward the worker. The requester can also choose

to reject the participant’s submission and not reward a worker if the requester believes the task was not successfully completed, e.g. an incorrect completion code was submitted, or if the time taken to complete the task is deemed too short.

For this study, the MTurk task links to an external website that the study is completed on. The ecosystem of this study is detailed in section 5.1.4. In short, participants will be MTurk workers who access the task from MTurk, complete the study on the study’s website, and receive the correct completion code upon completion of the study. Workers who complete the tasks successfully were rewarded \$1 US Dollar. The task flow of the study is presented in Figure 5.2.

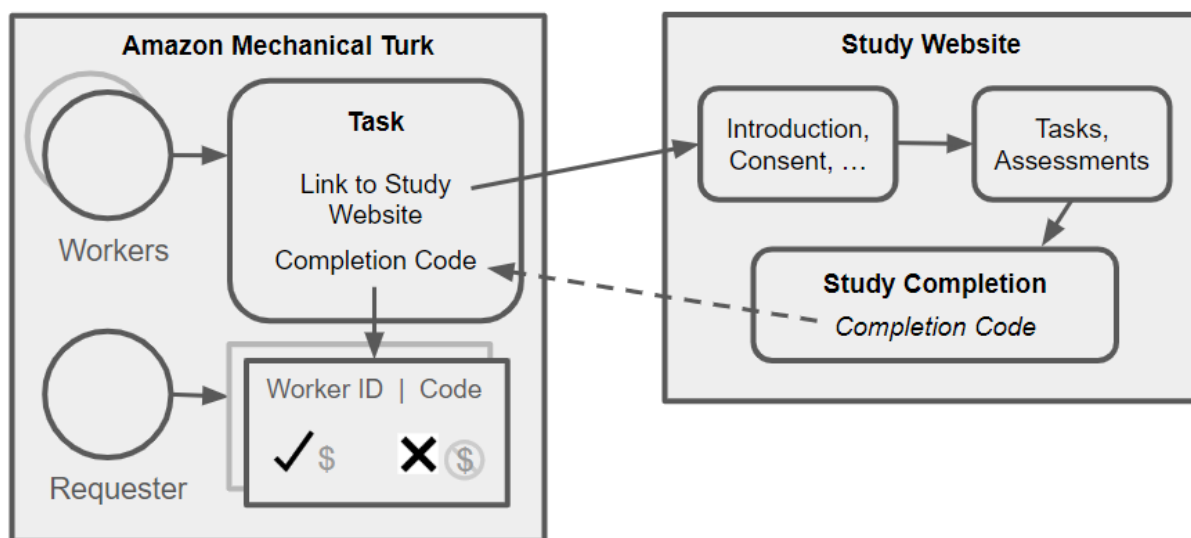


Figure 5.2: A high level overview of the study design and flow using MTurk.

Challenges 2 and 3: An Understandable Pitch Interaction Design

Though multiple ideas for pitch interaction design were considered, a relative pitch interaction technique was chosen to use for the pitch based NVVI technique, as detailed in Section 4.1. The primary motivation for this approach was lack of a necessary setup and configura-

tion for participants, such as presented in prior work[58].

To help explain this approach to participants so they could effectively use it to complete tasks, the study included an education portion to begin with. In this portion of the studies, users were given a brief but concise explanations of the field of NVVI, the relative pitch interaction technique, and an overview of the pitch based NVVI technique. Examples of using the tool with both whistling and humming were provided, as well as an example of using selection to activate and deactivate the tool.

Challenge 4: Human Vocal Characteristics

Given the broad and unknown domain of potential MTurk workers who were to complete the study, as well as the asynchronous nature of conducting this study, it was unknown how the vocally impaired or people with pathological voices would use the tool. This particular metric was not gauged in the study. However, a demographic question asked users to indicate their biological sex and their subjective ranking of their speaking voice (relatively low, medium, or high pitched.)

Additionally, a robust pitch estimation algorithm had to be chosen for implementation for the NVVI prototype. Though there are numerous pitch detection algorithms, many described in the review of literature section of this work, an implementation of the auto-correlation function was chosen, as it was found to be robust [45] and therefore good for a variety of voice types that may have a noticeable amount of other formant frequencies present [34]. It also experimentally seemed to be a more noise resilient algorithm than others.

Challenge 5: Musical Background

It was unknown at the time of the study how a person's musical background would affect their performance with the pitch based NVVI technique. However, this work was aimed at exploring using the NVVI technique with a more general audience, which had low expectations for musical background and skill. However, a few demographic questions were included in the study, including questions on musical background, skills producing a constant humming or whistle pitch, and a question asking the user about their pitch identification skills.

Challenge 6: Microphone Device Variance

As mentioned while addressing challenges 2 and 3, a robust pitch detection algorithm, the auto-correlation function, was chosen to implement the pitch based NVVI prototype. This algorithm was anecdotally found to be accurate with a variety of microphone devices, background noises, and speaker voice types, also identified in the pilot study. However, this algorithm does not completely eliminate this research challenge, and the device type was still considered for performance. A demographic survey question asked users about the microphone device type they used for the study.

Challenges 7 and 8: Novelty and Generalizability

This work strives to make generalizeable claims about user interaction with a pitch based NVVI technique being used to control a continuous, one dimensional parameter (the slider.) The study aims to provide a large quantity of data to help support these claims, not present in the extent of prior studies. However, this work provides one implementation of a pitch based NVVI prototype that uses a relative pitch interaction technique. This is a very specific

instance of NVVI, and it is unknown how the findings of this study will translate up the latter into broader interaction patterns of general NVVI techniques. With regards to answering RQ1, exploring Fitts' Law with a pitch based NVVI prototype, it is unknown whether the findings will translate to other continuous NVVI input modalities, such as blowing or formants. This work will not claim that findings from the studies done in this work are translatable to these areas, even though there may be similarities.

5.1.3 Study Limitations

Though the study aims to offer a large quantity of data with a diverse group of participants exploring the usability of the pitch based NVVI technique described, there are foreseen limitations with some of the approaches outlined in the this chapter.

The MTurk Ecosystem

Though MTurk can provide many services for user studies, including access to a large pool of participants, advertising the study, and providing a compensation mechanism for awarding participants, there are some notable downsides. One is that money may be primary motivating factor for MTurk workers to complete tasks [24], which may affect the quality of results and findings from studies, especially those that are conducted asynchronously without direct participant oversight.

Web Browsers and Devices

The Web Audio API[50] provides a guideline for how web browsers should provide functionality regarding audio. For example, at the time that this research was performed, the guidelines state web browsers must ask permission before the microphone is accessed, and

the user must explicitly grant a website permission to access the microphone before any audio data is collected.

Most commonly used web browsers, including Microsoft Edge, Google Chrome, Mozilla Fox, and Safari implement the vast majority of specifications described in the WebAudioAPI. However, it is possible that participants use a web browser that does not entirely support the WebAudioAPI, which may severely impact the quality of the study for that particular experience. A “whitelist” check for web browsers was not performed, so though unlikely, some participants could experience issues with the study.

It is unknown if all web browsers, especially on mobile devices, will provide the performance, features, and functionalities necessary for the study. For example, if an outdated web browser does not support the most recent version of JavaScript used for most of the study, it is possible errors or glitches may occur. Additionally, if a participant is performing the study on a very low performance device, it is possible the user experience and intent will not be ideal for the study.

Computer Mouse Ambiguity

Though the study explains that performance using the NVVI technique will be compared with the baseline of a computer mouse, there are no possible checks to be performed that a participant is using a computer mouse. They may use a track pad on their computer, for example. It is expected that most participants will use a computer mouse, but it is possible participants may use alternative inputs that are supported the same way as a computer mouse by web browsers.

Choice of Algorithms

The choice of algorithms used for pitch estimation and confidence detection, as well as the parameters used for these algorithms, were chosen at a balance of robustness and simplicity. There are likely much better algorithms at the cost of computational complexity and overhead, and many others providing lower cost but less accurate. The algorithms chosen to be implemented in the pitch based NVVI prototype were deemed worthy to include in the system based on design choices but may affect the quality of data collected.

5.1.4 Study Ecosystem

The study website was deployed as a static web page on GitHub pages. The web page was developed as a sequence of document object model (DOM) elements presented as a series of modules. The study website was developed using Hyper-text Markup Language (HTML,) JavaScript (a programming language) and cascading style sheets (CSS.)

Since the study was conducted asynchronously and without state via MTurk, user data had to be stored automatically and stored somewhere afterward. GitHub pages does not offer access to a database, so a secure Virginia Tech server was setup to store the results of the study.

NVVI Interaction

Since the pitch based NVVI prototype was also developed in JavaScript, the integration was designed so that any module could create an instance of the tool to use for a task or demonstration.

As mentioned in the previous section, the input sent to the NVVI prototype is collected from

the user microphone accessed via the WebAudioAPI. If the user rejects microphone access, the study cannot complete. Therefore, a module early on in the study performs a check to grant microphone access. If the user grants access, the web browser does not have to prompt the user for access again. If the user rejects access, the web page tells the participant that they must allow access to the microphone, otherwise they may not complete the study. The module refers participants to enable access via web browser settings.

The NVVI technique was configured at the start of each round, which instantiated a set of 10 or 20 randomized tests. The output of the NVVI technique was configured to send the session data through a callback function that adjusts the slider knob by a particular amount based on a configuration variable that was hard-coded for the study.

5.1.5 Study Procedure

The study is conducted in a modular format. Participants complete a series of readings (e.g. what is NVVI, how does the pitch based NVVI technique work,) *phases*, and *surveys*.

Phases and Phase Tasks

In this study, a *phase* contains an introduction of what is assessed in the phase, the phase tasks that assess the user using the NVVI technique, a NASA TLX task workload questionnaire, and a follow-up qualitative survey. The structure of a phase is depicted in Figure 5.3.

Phase tasks are comprised of a number of *rounds* of *tests* that assess a user's performance with the pitch based NVVI technique. A round is a collection of these tests that works with one customizable instance of a pitch based NVVI technique. The structure of phase tasks is depicted in Figure 5.4.

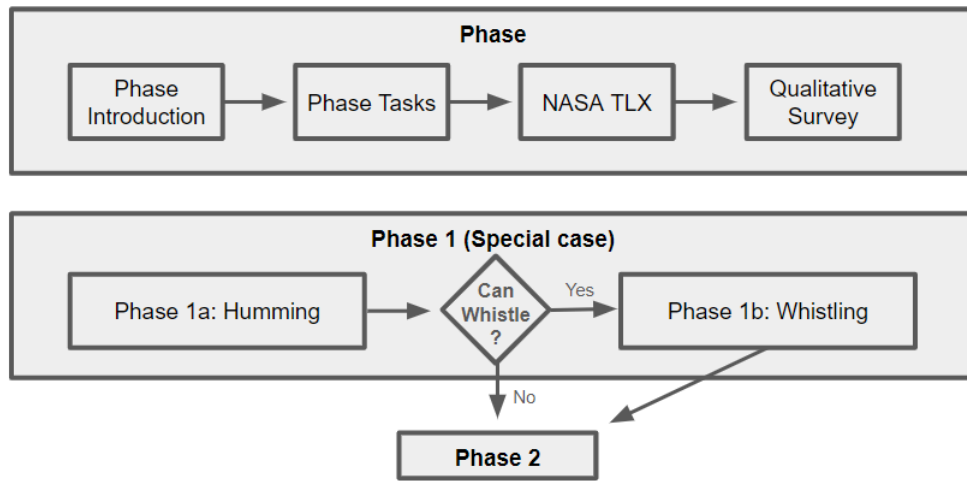


Figure 5.3: The structure of a phase and the special case of Phase 1.

In each test, a participant is tasked with moving and aligning the bottom of two HTML slider knobs with the top as precisely as they feel appropriate. The test is essentially completing the task as described and depicted in Figure 5.1. In order to begin the test, participants have to enable the NVVI technique by creating two transients before they can use the pitch of their humming or whistling to adjust the slider knob's position. To proceed to the next test, participants also must create two transients. This should occur when the user is comfortable with the bottom slider knob's position. Additionally, the starting and ending positions of the slider knobs are randomized for each test, allowing for a variety of distances to be assessed with in either direction. Phase tasks are depicted in Figure 5.1.

Three phases of the study are performed to assess participants' performance with the pitch based NVVI technique.

The first phase consists of two assessments of task completion using the pitch based NVVI technique, one for humming and one for whistling. Since not all participants are expected to be able to whistle, the second phase is optionally skipped, as described in the next section. In each part of this phase (humming and whistling, if applicable,) 10 tests are performed.

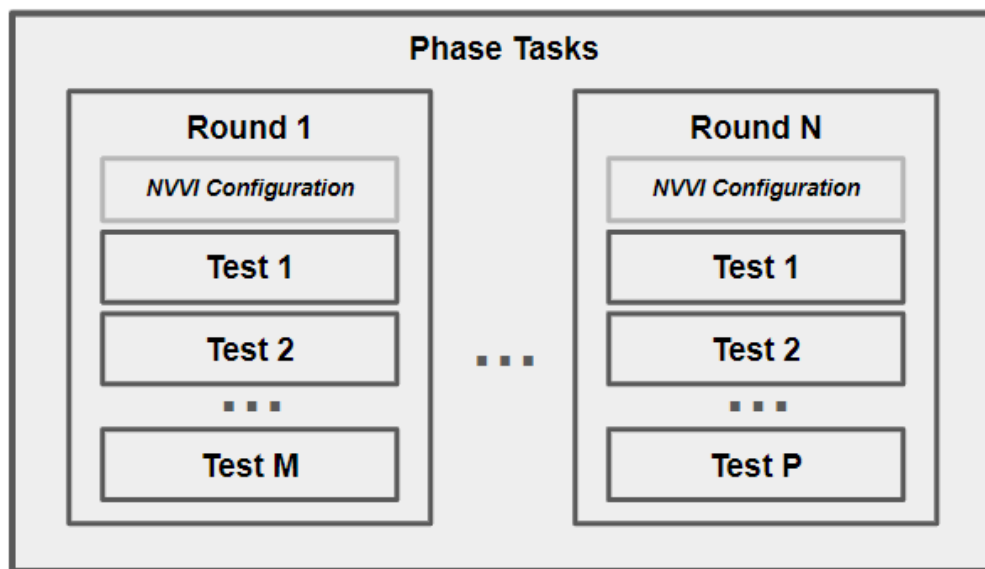


Figure 5.4: The structure of Phase Tasks and Rounds.

The second phase is a more comprehensive assessment of user performance with the NVVI technique. In this phase, the user chooses the modality, humming or whistling, to be assessed on. In this phase, 20 tests are performed.

In the third phase, 10 computer mouse baseline tests are performed. The tests are similar to the tests used for NVVI testing of the tool, but instead of using the pitch based NVVI prototype to adjust the slider knob, the computer mouse is used to drag the slider to its destination.

Surveys

Several questionnaires are administered throughout the study. The first is a demographic questionnaire, asking participants for basic information (appendix.) There are also follow-up questions to every phase that allow participants to provide more context and expression of their experience using the tool.

In the demographic questionnaire, participants are asked if they can whistle (yes/no.) They are also asked to rank their whistling skills according to a Likert [44] scale question in the form “I have strong whistling skills (with accurate, stable pitch).” If they choose the option “Disagree” or “Strongly Disagree,” or choose that they cannot whistle, the study deems them ineligible for the second part of Phase 1, which is comprised of the introductory tests to gain familiarity with whistling with the NVVI technique.

NASA TLX. After each phase assessment, a NASA TLX is administered to gauge the task workload for using the tool during the phase. In total, three or four TLXs are administered, depending on if the user can whistle and completes the second part of Phase 1 or not.

System Usability Scale (SUS) Questionnaire. At the end of the study, a SUS questionnaire is administered. The goal of the SUS survey is to collect a “quick and dirty” understanding of the usability of a tool[13]. In this case, the goal of the SUS is to gain an understanding of the usability of the NVVI technique developed and used throughout this study.

Data Collection

Data collection occurs by storing the results of phases and surveys inside of a JSON object. As the user completes more phases and surveys, this JSON object gradually fills up with this data. At the end of the study, the data is sent to the server to be recorded. If the data is successfully uploaded (indicated by the server response,) the completion code is given to the participant. If an error occurred, the completion code is still given to the participant for the work they put into completing the study. The result files hold data collected from *phases* and *surveys*. The connection established between the study website and secure server is made via a secure WebSocket (WSS) connection [51].

Phase Task Data Phase data is a collection of round data, which is itself a collection (array) of test data. Each test datum contains two pieces of quantitative data, the time taken to complete the test, as well as an array of tuples consisting of the time and position of the slider knob recorded every 100 milliseconds. Metadata about the test, including the start and end values (positions) of the slider knob are also included.

Chronological Procedure

The chronological procedure for the official study was conducted in the order depicted in Figure 5.5. The order was the same for the pilot study but MTurk was not involved in its conduction.

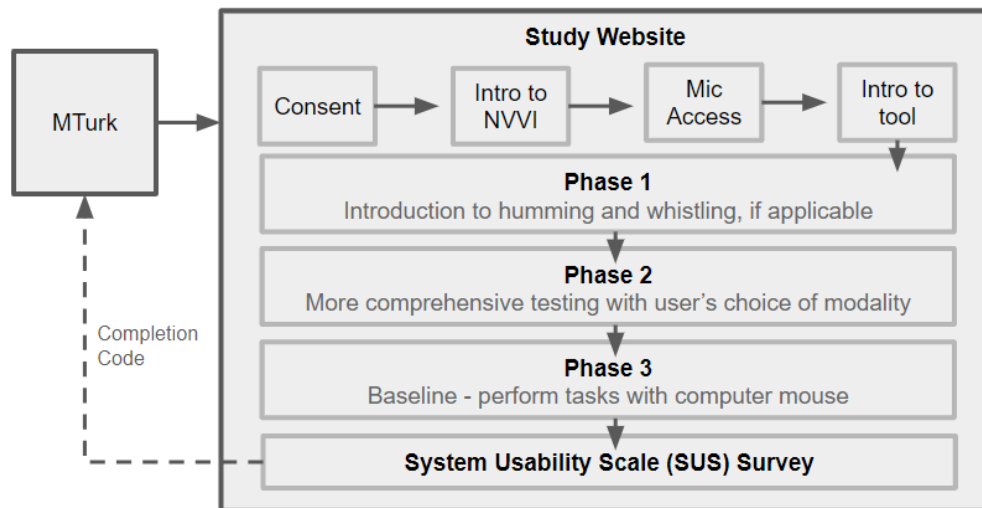


Figure 5.5: The chronological procedure of the user study.

5.2 Statistical Methods

The statistical methods used for this study vary by research question, with some exceptions. Statistical methods used in multiple research questions are described first, and then each

research question's particular statistical methods used are detailed.

Average times

For some analyses performed on the results of this study, the average time taken to complete a task is used as the data point. The way this is calculated is best explained by example. Some analyses consider tasks that fall within a particular category that were performed in a particular phase. To calculate the average time using this method, the results are filtered to only include the phases brushed by the client. Then, individual tasks are removed if they do not fall under the brushed category. Finally, the average times are calculated by taking the average of all remaining task times, and this number is used per participant.

Comparative Analysis

Multiple research questions can be partially or fully answered with comparative analyses. One factor Analysis of Variance (ANOVA) tests will be conducted for comparative analysis to help validate hypotheses provided in Section 5.3. For example, an ANOVA test will be administered to help answer research question 3, comparing user performance with whistling and humming. An alpha value of 0.05 will be used for the ANOVA tests, as it is a typically used significance value that translates to a 5% chance of rejecting the null hypothesis when it is actually true[\[42\]](#).

Linear Regression

To see if the task performance observes Fitts' Law, linear regression will be used to see how closely the plotted distance and time graph resembles a linear relationship. This will be performed twice, once to see if the task data observes Fitts' Law (time and the logarithm of

distance should have a linear relationship) and once with just a plot of distance and time.

NASA TLX Analysis

A NASA Task Load Index - TLX is administered after every set of phase tasks to gauge the task workload experienced while completing tasks using the pitch based NVVI technique. The NASA TLX contains seven response options for participants regarding their experience with the NVVI technique from “Very Low” to “Very High.” The TLX will be conducted in a 21-point manner. The results will be the average response value for each question.

SUS Analysis

At the end of the study, a SUS questionnaire is administered to collect a “quick and dirty” understanding of the usability of the pitch based NVVI system. 10 items are present in the questionnaire on a five-point scale from “Agree” to “Disagree.” During analysis, the scores will be converted to the numbers 0-4 summed together (for a maximum of 40,) and multiplied by 2.5 so that the maximum score would be 100. A score above 68 would be considered an “above average” usability score [13].

Likert Scale

The Likert Scale [44] is a scale used for questionnaires that gauges participants’ agreement with a statement. The Likert scale is used in this study to quantify participant agreement with particular experiences during this study. For example, participants are asked to indicate how much they agree with the statement “I felt very comfortable using the NVVI technique with humming.” In the five point Likert scale used in this study, the possible responses are “Strongly Disagree,” “Disagree,” “Neither Agree nor Disagree,” “Agree,” and “Strongly

Agree.”

5.2.1 Research Questions

RQ1: Does pitch-based NVVI follow Fitts’ Law? To answer this research question, Phase 3 results will be analyzed to see if tasks follow a Fitts’ Law. This may be shown by drawing a graph that strongly linearly correlates time with logarithm of distance, i.e.:

$$T = a + b * \log_2(2D/W)$$

Figure 5.6: Fitts’ Law, which states the time taken (T) to perform a movement is proportional to the logarithm of the distance travelled (D) divided by the width of the target (W .) [23]

This can be actualized by plotting the equation with T = the time taken to move the slider knob with the NVVI technique and D = the distance traveled from the start to end position of the slider knob. Since W is not changed, it can remain as a constant.

These data points can be substituted into the Fitts’ Law equation and plotted to create a graph. Linear regression can then be used to see how closely the data points resemble a line (have a linear, proportional relationship.) If they do, which is indicated by a high R^2 value as a part of the linear regression process, this would indicate the observation of Fitts’ Law.

RQ2: How does a pitch-based NVVI technique compare with a computer mouse?

To answer this research question, user performance will be measured between Phase 2 (which comprises of comprehensive testing with NVVI technique) and Phase 3 (the same tasks but performed with computer mouse.)

The difference in average time taken to complete tasks between the 2 modalities will be explored with ANOVA analysis as a “within subject test” (a test comparing two metrics

that all participants performed, as participants completed both Phases 2 and 3.)

RQ3: With which mode of input for the tool do users perform better, whistling or humming?

Similar to answering the last research question, to answer the third research question, user performance will be measured between Whistling and Humming in Phase 2.

Phase 2 was a comprehensive assessment of using the NVVI technique where the user chose the modality of whistling or humming to control the tool. The difference in the average time taken between the 2 categories will be explored.

The difference in average time taken to complete tasks between the 2 modalities will be explored with ANOVA analysis as a “between subject test” (a test comparing two metrics that were performed by two groups of participants, as participants only completed Phase 2 by whistling or humming.)

RQ4: Which modality (whistling or humming) do users experience less task workload?

The NASA TLX scores between participants who chose whistling and humming in Phase 2 will be compared. ANOVA analysis will be used as a “between subject test” for the same reason described in the previous research question.

5.3 Hypotheses

The hypotheses for each research question are ordered and listed below.

RQ1: Does pitch-based NVVI follow Fitts’ Law?

H_1 : Pitch based NVVI using the relative pitch interaction technique described in this study

will observe Fitts' Law.

RQ2: How does a pitch-based NVVI technique compare with a computer mouse?

H_2 : User performance with a computer mouse and the pitch-based NVVI technique will not significantly differ.

RQ3: With which mode of input for the tool do users perform better, whistling or humming?

H_3 : User performance with the pitch based NVVI technique will not significantly differ between the modalities of whistling and humming.

RQ4: Which modality (whistling or humming) do users experience less task workload?

H_4 : User perception of task workload experienced while using the pitch based NVVI technique will not significantly differ between the modalities of whistling and humming.

Chapter 6

Results

Two studies were performed. A pilot study was conducted to test the functionality of the study and receive feedback before the primary study. This feedback, along with other testing and refinements, was used in making slight adjustments to the study. Additionally, neither MTurk or the secure server were implemented at the time of the pilot study. Participants of the pilot study emailed results to the researchers instead.

6.1 Pilot Study

The pilot study was conducted informally with 6 participants. Results are not used to answer research questions, but to provide feedback with the study and pitch based NVVI technique, as well as improvements to be made for the primary study.

Participants mentioned that some of the instructions and functionality descriptions were unclear. One participant mentioned that they thought that this tool was particularly helpful for continuous controls. Most participants noted some difficulty learning the tool. Overall, every participant completed the study. Participants of the pilot study said that they enjoyed the study.

Based on this feedback and some study considerations, refinements made to the study after feedback from the pilot study included:

- More educational information was provided before the assessments, including a deeper explanation of how the tool works.
- More tests were added and the pitch based NVVI technique configuration was updated to be less smooth but more precise.
- Questionnaires were visually updated to be more clear when required questions weren't answered. (they were implemented from scratch for this study.)

6.2 Study Results

The user study was initially planned to be performed with 100 participants via MTurk. However, a few factors played into a total of 74 successful studies. The two most likely sources of error in conducting the study was a logical code error with the secure server's WebSocket server code that could reject a submission if the participant couldn't whistle.

Additionally, an outage occurred that may have resulted in a few missing submissions. It is unclear why the goal of 100 total participants (MTurk Workers) was unable to be met, but MTurk closed access to the study after 74 successful submissions were approved. Several participants received errors and emailed the researchers, who rewarded the subjects for participation.

Two more submissions were removed for erroneous result files. It is unclear exactly why, but two submissions missed some phase and survey data necessary for analysis.

Ultimately, 72 results were considered for analysis in this study. They were downloaded from the secure server to a secure lab computer at Virginia Tech, though the data was not considered to be sensitive by the IRB.

6.2.1 Demographics

Out of the 72 participants, 32 identified as biological males and 40 identified as biological females. The age distribution of participants is depicted in Figure 6.1, with most participants being between the ages of 24 and 50. 0 participants indicated that they suffered from a disability that impaired their use with a computer mouse, though it is unlikely they would be completing the study via MTurk if they had one. One participant noted a vocal impairment.

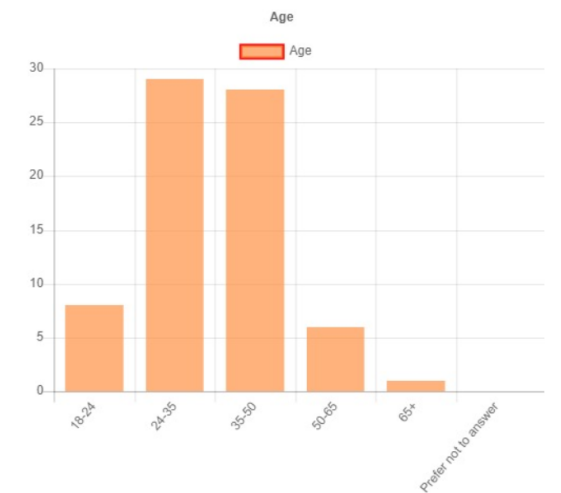


Figure 6.1: The age distribution of participants of the user study.

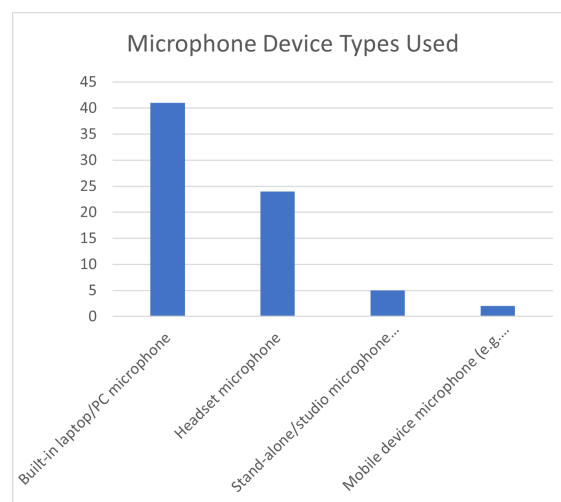


Figure 6.2: The microphone device distribution of participants of the user study.

Among biological women, 2 considered their voices to be “Low” in character, 31 said theirs were “Medium,” and 7 said theirs were “High.” Among biological men, 6 considered their voices to be “Low” in character, 23 said theirs were “Medium,” and 3 said theirs were “High.”

Participants were asked for the type of microphone device they were using for the study. Figure 6.2 shows the distribution of devices used, the most common two being a built-in laptop or PC microphone and a headset microphone.

40.85% of participants had moderate background knowledge of NVVI before completing the study (chose “Agree” or “Strongly Agree” with the Likert question “I have strong background knowledge in NVVI.”

Musical Background and Capabilities

54.17% of participants agreed or strongly agreed that they had a strong musical background. 44.44% of participants agreed or strongly agreed with the statement “I have strong pitch identification ability (i.e. perfect pitch).” 62.5% participants agreed or strongly agreed with the statement “I am proficient at humming/singing (with an accurate, stable pitch).”

19 participants claimed that they outright could not whistle. In a follow-up question for participants who claimed that they could whistle, 23 participants disagreed or strongly disagreed with the statement “I have strong whistling skills (with accurate, stable pitch).” In total, the study logic determined that based on the responses to these two questions, 27 participants were not able to whistle, and thus only 45 participants were deemed to be able to whistle.

6.2.2 User Performance

Participants were able to successfully complete tasks with the pitch based NVVI technique. Task completion was verified by the fact that users were able to progress and ultimately complete the study by performing the selection operation, as described in Section 4.2.4. Additionally, each task contained the times and positions of the slider knob intervally recorded during the task, as explained in Section 5.1.5. An example of this data visualized is shown in Figures 6.8(a-c).

However, upon closer examination of data, it was clear that some participants did not try to complete some tasks. There are a variety of reasons why a participant may not have tried to complete a task, but two primary cases are considered. The first is when the user really did not try to complete a task for some reason. The second is that the participant may have perceived the starting and ending position of the slider knob to be close enough to simply proceed to the task. As a reminder, the starting and ending positions of the slider knob are randomized, as described in Section 5.1.5.

Therefore, to get a better gauge of results, a classification system was devised to allow researchers to brush a variety of tasks. The classifications are explained in the next section.

Task Classifications

Tasks were assorted into 3 groups of 2 categories as depicted in Figure 6.3. The classifications appear to be complete (partitions,) as no task was given an undefined category. The groups are divided into the three categories of not attempted, attempted and failure, and attempted and success. The metrics used for categorization were *local variance*, *global variance*, and the *accuracy*.

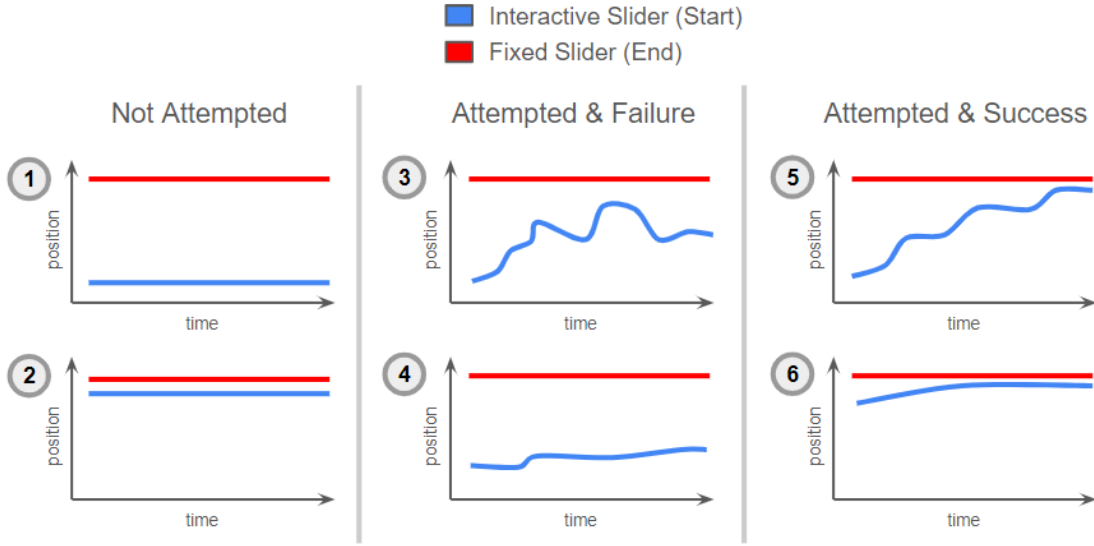


Figure 6.3: The classifications of task performances.

Metrics. The metrics used for classification are *local variance*, *global variance*, and the *accuracy*. The calculations occurred after normalizing the y axis (position) by the value of $|\text{end} - \text{start}|$.

Local variance was calculated by summing the square of the difference of the values at time t and $t - 1$, and taking the square root of this number. The threshold used for determining a high local variance was the value 0.1. In mathematical notation where x is the normalized array of data points and n is the size of the array:

$$\text{Local Variance} = \sqrt{\sum_{t=1}^{n-1} (x[t] - x[t - 1])^2}$$

$$\text{High Local Variance} = \begin{cases} \text{true}, & \text{if Local Variance} \geq 0.1 \\ \text{false}, & \text{if Local Variance} < 0.1 \end{cases}$$

Global variance was calculated as the standard deviation for the data points. This effectively ruled out any categories in the first group because they were straight lines (had no variance.) The threshold used for determining a high global variance (which is not necessarily high, just a little more than the user doing nothing) was the value 0.00005. In mathematical notation where x is the array of data points, n is the size of this array and μ is the average position value:

$$\text{Global Variance} = \sqrt{\frac{\sum_{t=1}^{n-1} (x[t] - \mu)^2}{n - 1}}$$

$$\text{High Global Variance} = \begin{cases} \text{true,} & \text{if Global Variance} \geq 0.00005 \\ \text{false,} & \text{if Global Variance} < 0.00005 \end{cases}$$

A task was determined to be accurate if the final location of the slider knob was within a $\pm 7\%$ range of the end value after normalization. In mathematical notation:

$$\text{Error} = |\text{endValue}_{\text{normalized}} - \text{startValue}_{\text{normalized}}|$$

$$\text{Accuracy} = \begin{cases} \text{accurate,} & \text{if Error} \leq 0.07 \\ \text{inaccurate,} & \text{if Error} > 0.07 \end{cases}$$

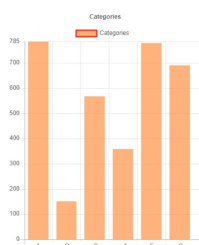
Categories. The first of the three groups contains categories 1 and 2, essentially the two previous forms of task performances. **Category 1** contains tasks that were not attempted and likely not because participants thought it was close enough. The metrics identifying these tasks are a low global variance and inaccuracy. **Category 2** contains tasks that were

not attempted likely because participants thought the slider knob was close enough to the end value. The metrics identifying these tasks are a low global variance and accuracy.

The second group contains tasks that were attempted but unsuccessful. Both categories in this group were inaccurate. **Category 3** contains tasks with a high local variance and high global variance, while **Category 4** contains tasks that had high local variance but low global variance.

The third group contains tasks that were attempted and successful, meaning both categories in this group were accurate. **Category 5** contains tasks with a high local variance and high global variance, while **Category 6** contains tasks that had high local variance but low global variance.

Figure 6.4: Classifications of phase tasks.



(a) Classifications of tasks for all tasks.



(b) Classifications of tasks for only phase 3 tasks.

Results.

Figure 6.4(a) depicts the distribution of category scores for all tasks completed. In total, 3,330 tasks were considered. Successful tasks were those in Categories 5 and 6. In total, 935 tasks were not attempted, 926 tasks were attempted but unsuccessful, and 1469 tasks were attempted and successful. Therefore, the total attempted success rate of all tasks was 61.34%.

Since Phase 1 was the participants' introduction to using the pitch based NVVI technique,

the tasks considered for analysis were only from Phase 2, the comprehensive assessment phase. In total, 1,440 tasks were considered. The distribution of task classifications is depicted in Figure 6.4(b). Successful tasks were those in Categories 5 and 6. In total, 475 tasks were not attempted, 378 tasks were attempted but unsuccessful, and 587 tasks were attempted and successful. Therefore, the total attempted success rate of all tasks was 61.34%.

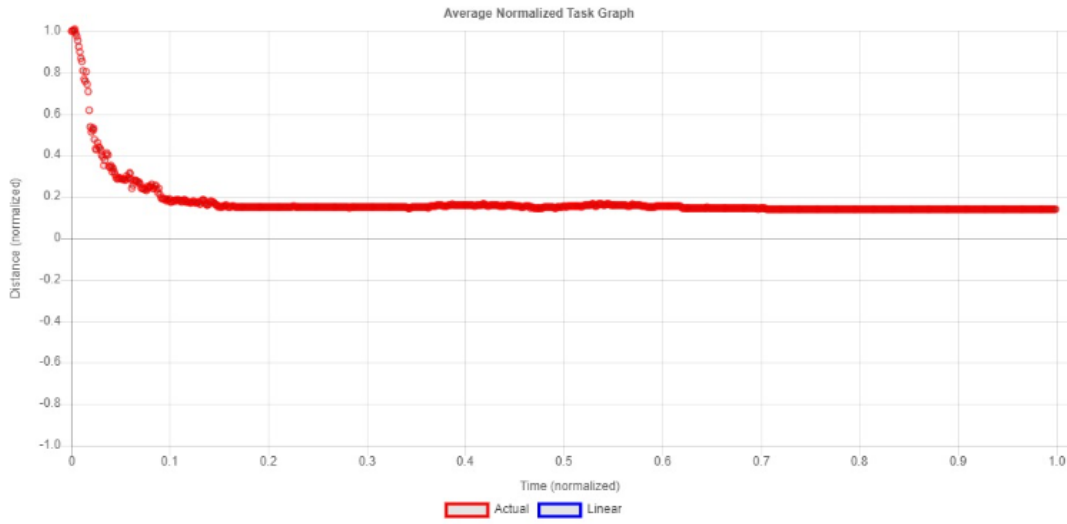


Figure 6.5: A graph representing the average path taken during a phase task, where each phase was normalized on the time axis by the total time taken and on the distance axis by the distance between the starting and ending positions.

Aggregate Task Data

Figure 6.5 provides a graph representing the average path taken in successful completing tasks. The domain of tasks used in the construction of this graph are successful tasks (from categories 5 and 6.) The graph was made by normalizing all task data from these categories in both the time and position axes, and then using linear interpolation to plot an average graph with 1000 sample points.

Figure 6.6: Graphs containing all phase task paths in categories 5 and 6 (attempted and successful,) where each phase was normalized on the time axis by the total time taken and on the distance axis by the distance between the starting and ending positions. Additionally, the distance axis is scaled by a factor of 16.

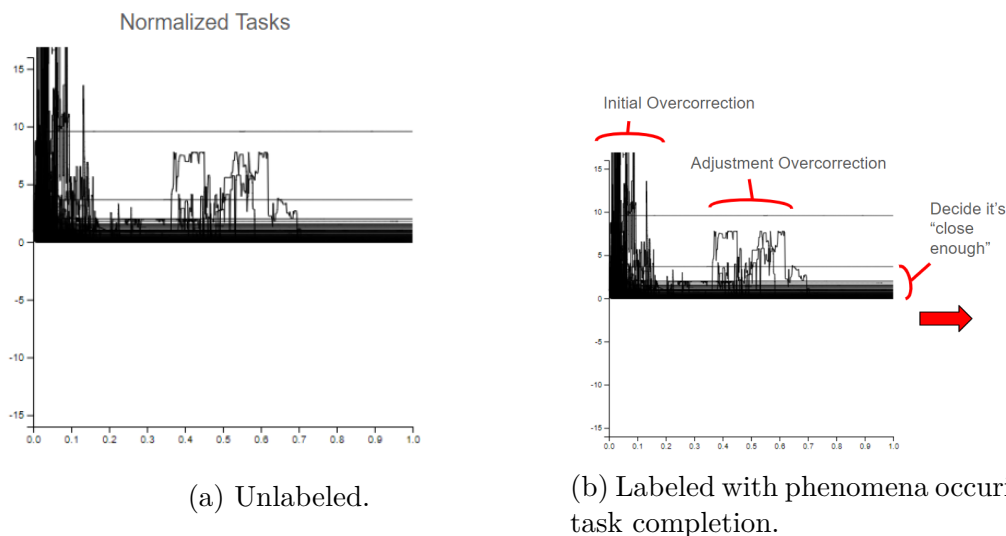
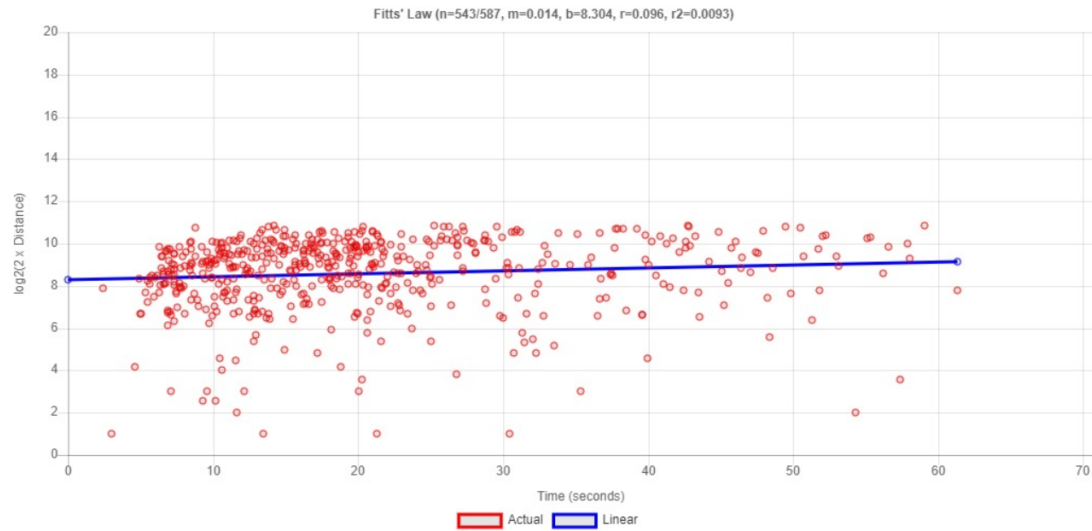


Figure 6.6(a) provides a graph with all normalized task data in both the time and position axes. Additionally, the absolute values of the positions are plotted. However, it is important to note that the position (y) axis is stretched by a multiple of 16. Some values are clipped from the graph as well. This is explained by a severe over-correction made that is multiple times the distance between the starting and ending positions. Since the graphs are normalized by the distance between these two points, if they are already close together and the user significantly changes the slider knob's position, the result will be a change that is multiple sizes of the distance used for normalization, which will be a value greater than one when normalized.

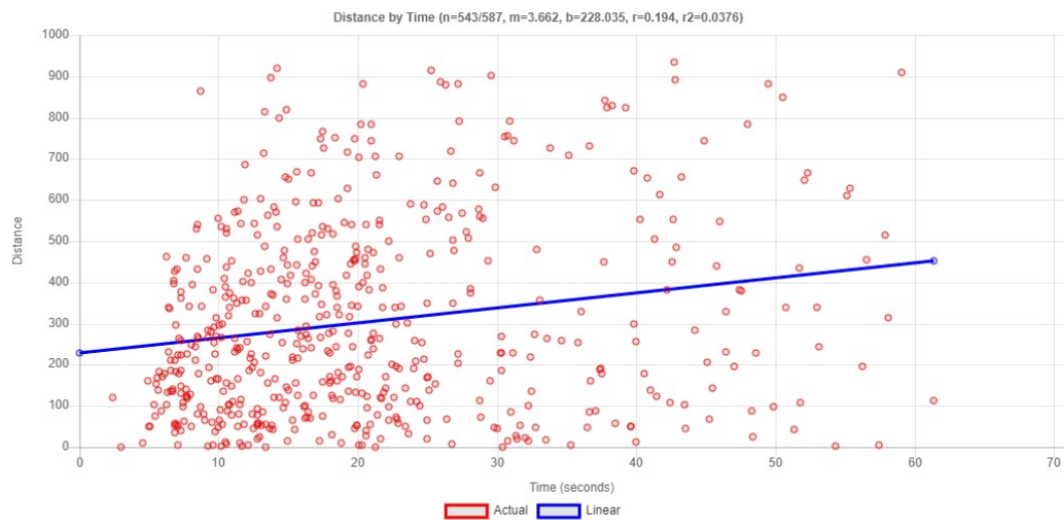
Fitts' Law

Figure 6.7(a) contains a plot depicting a relationship between the time taken to complete tasks and the logarithm of the distance from the start to end position of the slider knob. The

Figure 6.7: Graphs plotting the relationship between the time taken to complete tasks and the distance between starting and ending positions.



(a) Logarithmic (Fitts' Law.)



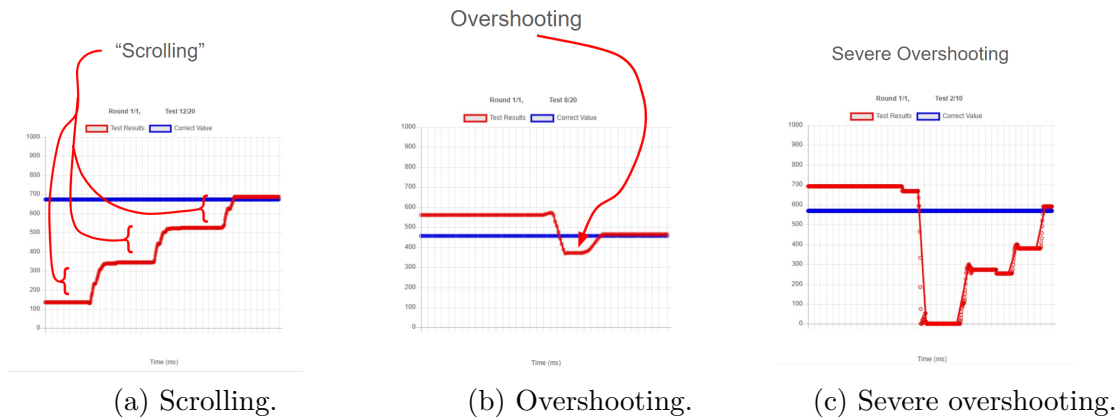
(b) Linear

top 7.5% of times were removed as a part of cleaning. A linear relationship would indicate an observance of Fitts' Law. However, the r^2 value of a linear regression model of this graph is 0.0093, which is not large enough to presume a linear relationship.

Figure 6.7(b) contains a plot depicting a relationship between the time taken to complete tasks and the distance from the start to end position of the slider knob. The top 7.5% of times were removed as a part of cleaning. A linear relationship would indicate a linear relationship between the time a task takes and the distance traveled by the slider. However, the r^2 value of a linear regression model of this graph is 0.0376, which is not large enough to presume a linear relationship.

r^2 is higher in this latter linear regression model than the former, which may indicate a stronger linear relationship between distance and time than a logarithmic one as defined in Fitts' Law.

Figure 6.8: Common phenomena occurring during tasks.



Representative/Exemplary Task Data

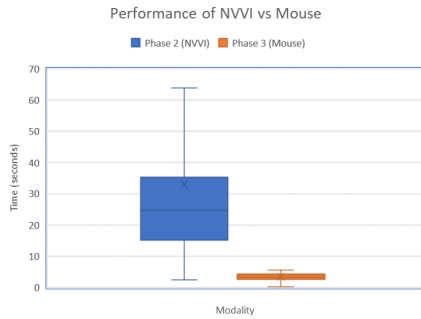
Many individual task data points were examined. In this section, a selection of individual task graphs are presented.

Figure 6.8(a) shows an instance of a phenomenon of what this work refers to as “scrolling.” As shown in the figure, the position of the slider in similarly sized increments until the desired position is reached. The exact number of tasks exhibiting this phenomenon was not recorded.

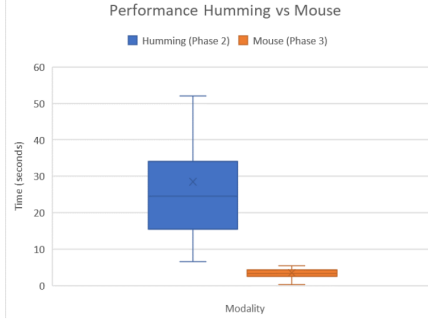
Figure 6.8(b) shows an instance of a participant overshooting using the tool. When trying to move the slider knob to the correct end position, they move it past this position to a distance almost equal to the initial distance from the start and end values.

Figure 6.8(c) shows an instance of a participant severely overshooting using the tool. When trying to move the slider knob to the correct end position, they move it past this position to a distance almost equal to six times the initial distance from the start and end values.

Figure 6.9: Box plots comparing NVVI performance with the computer mouse.



(a) NVVI vs. computer mouse.



(b) Humming with NVVI vs. computer mouse.

Pitch Based NVVI and the Mouse

In the study, participants were tasked with using the pitch based NVVI technique and a baseline using a computer mouse. In the second phase, participants chose the modality of their choice to use for a comprehensive assessment with the NVVI technique, and in the third phase participants were tasked with using the mouse. The average times taken to complete

the tasks that were in categories 5 or 6 were recorded per participant.

Figure 6.9(a) shows a box plot comparing the performance between the use of the pitch based NVVI technique and the computer mouse, as described above. The number of participants considered varied between the two categories, as some participants did not complete any tasks in categories 5 or 6. ANOVA analysis revealed a significant difference of performance between the NVVI technique and the computer mouse, in favor of the computer mouse.

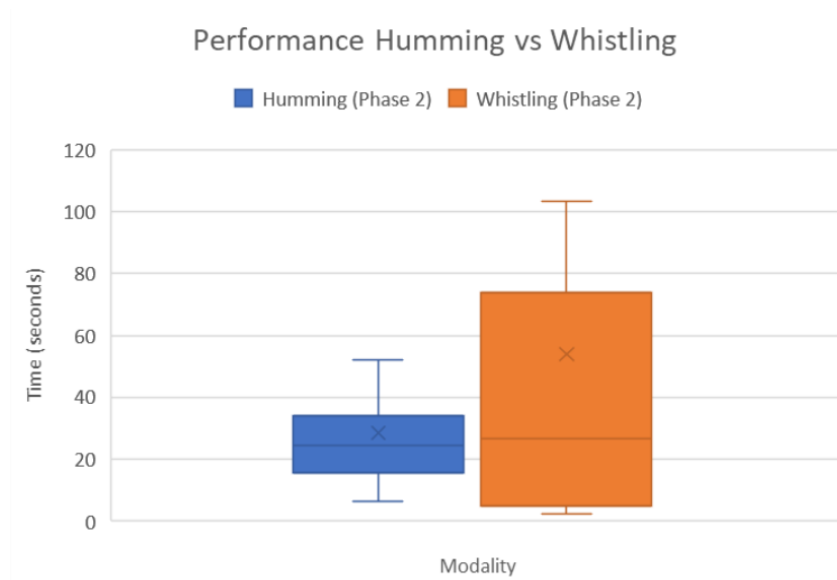


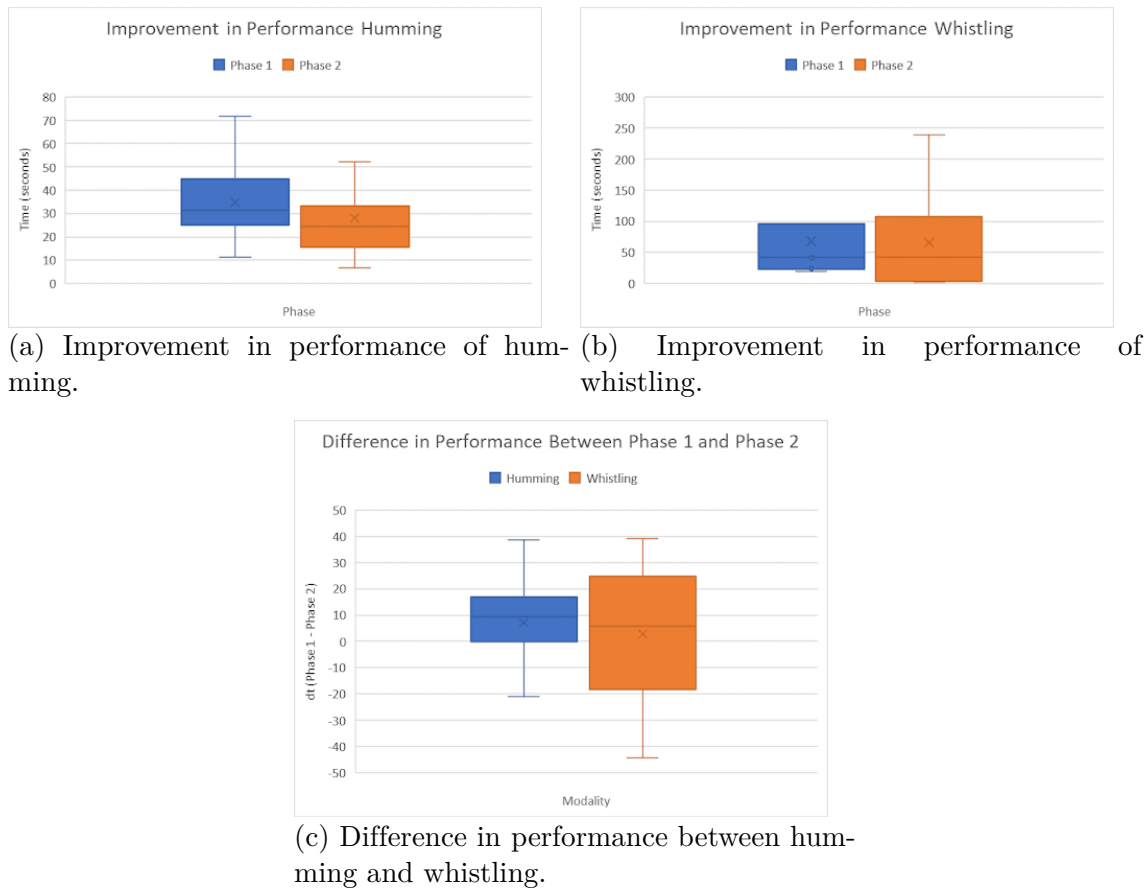
Figure 6.10: A box plot comparing the performance of using humming and whistling with the NVVI technique.

Whistling and Humming

In this study, participants chose a modality to be comprehensively assessed with in the second phase. 53 participants chose to hum and 19 chose to whistle. The average times taken to complete the tasks in the third phase that were in categories 5 or 6 were recorded per participant.

Figure 6.10 shows a box plot comparing the performance between the modalities of whistling

Figure 6.11: Box plots depicting changes of performance between Phase 1 and Phase 2.



and humming used in phase 2. Since some participants did not perform tasks in categories 5 or 6 for phase 2, the total number of participants considered was 47 for humming and 10 for whistling. ANOVA analysis revealed a significant difference of performance between humming and whistling for phase 2, in favor of humming.

Since humming was significantly better performance-wise, an ANOVA analysis similar to pitch based NVVI and the mouse as described in the previous section was performed, but with only humming and not whistling. A box plot comparison is depicted in Figure 6.9(b). ANOVA analysis revealed a significant difference between humming with the NVVI technique and using the computer mouse, in favor of the computer mouse.

Table 6.1: A table containing the NASA TLX questions.

Q1	How mentally demanding was the task?
Q2	How physically demanding was the task?
Q3	How hurried or rushed was the pace of the task?
Q4	How successful were you in accomplishing what you were asked to do?
Q5	How hard did you have to work to accomplish your level of performance?
Q6	How insecure, discouraged, irritated, stressed, and annoyed were you?

NVVI Performance Improvement

In phase one of this study, participants learned to use the humming modality with the pitch based NVVI technique, and if they were determined to be able to whistle by the criteria in Section 5.1.5, they also learned to use the whistling modality with the tool. In phase 2, they chose a modality to be comprehensively assessed with. The performance differences between phases 1 and 2 may be of relevance to understand the learnability of the tool for each modality.

Two analyses were performed to examine improvement of performance by comparing the differences of performance between phases 1 and 2.

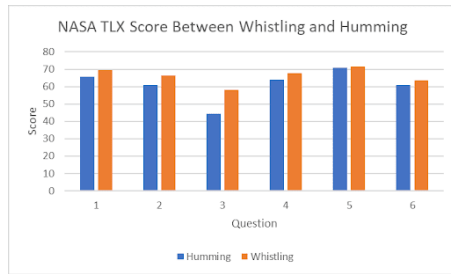
A between subjects analysis was performed comparing the average times taken to complete the tasks in the first and second phase that were in categories 5 or 6 were recorded per participant for each phase. The box plots for this comparison are presented in Figure 6.11(a) and 6.11(b). An ANOVA analysis revealed that a significant difference in performance was not present between phases 1 and 2 for whistling and humming.

Another between subjects analysis was performed comparing the difference of average task times in categories 5 or 6 per participant for each modality (whistling and humming.) The box plot for this comparison is shown in 6.11(c). An ANOVA analysis revealed that a significant difference in improvement was not present between whistling and humming.

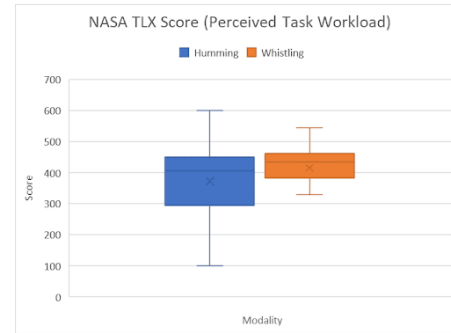
Table 6.2: A table containing the average scores for each NASA TLX question for both whistling and humming with the NVVI technique.

	Q1	Q2	Q3	Q4	Q5	Q6
Whistling	65.7	61.0	44.4	63.8	71.0	61.0
Humming	69.4	66.5	58.2	67.9	71.5	63.5

Figure 6.12: Box plots comparing NVVI performance using whistling and humming.



(a) NASA TLX scores by question of humming and whistling with the NVVI technique.

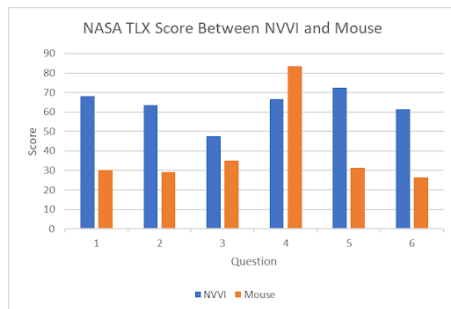


(b) Total task workload index between humming and whistling with the NVVI technique.

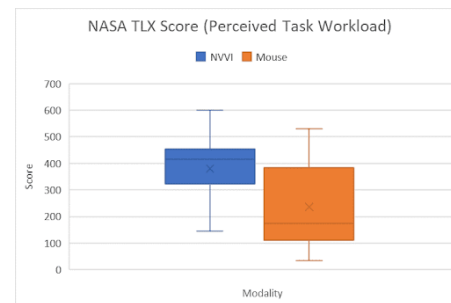
Table 6.3: A table containing the average scores for each NASA TLX question for using the NVVI technique and the computer mouse.

	Q1	Q2	Q3	Q4	Q5	Q6
NVVI technique	68.3	63.4	47.6	66.6	72.5	61.6
Mouse	30.2	29.2	35	83.6	31.3	26.4

Figure 6.13: Box plots comparing NVVI performance with the computer mouse.



(a) NASA TLX scores by question of using the NVVI technique and the computer mouse.



(b) Total task workload index between the NVVI technique vs. computer mouse.

6.2.3 Task Workload

After participants completed each phase in the study, they completed a NASA TLX (task load index) questionnaire to quantify and categorize the perceived task workload they experienced. For phase 2, participants chose to either hum or whistle for a comprehensive assessment with the NVVI technique. An analysis was performed comparing the NASA TLX scores for participants who hummed and whistled for phase 2.

Table 6.2 contains the average, normalized TLX score for each question per modality (whistling or humming.) Figure 6.12(a) shows a bar chart comparing the NASA TLX scores for each question. The horizontal axis of this figure provides a number corresponding to the NASA TLX question indicated in Table 6.1. As depicted, the NASA TLX scores between whistling and humming are similar but slightly lower in favor of humming for every question except for question 3, which is noticeably different between the 2 modalities.

An analysis was performed to compare the normalized and summed score (total perceived task load index) of the TLX questionnaire between whistling and humming. Figure 6.12(b) shows a box plot of this comparison. An ANOVA analysis did not reveal a significant difference of perceived task load index between the modalities of whistling and humming when used with the pitch based NVVI technique.

Another analysis was performed comparing the average and normalized NASA TLX scores of phase 2, a comprehensive assessment with the pitch based NVVI technique controlled with either whistling or humming, and phase 3, a baseline assessing performance with the computer mouse for each question. Figure 6.13(a) contains a bar chart comparing the scores between the two interaction modalities. The values are provided in Table 6.3. As depicted, the NASA TLX scores between the pitch based NVVI technique and the mouse are noticeably different for each question in favor of the mouse, but closer for questions 3 and 4.

A further analysis was performed to compare the normalized and summed score (total perceived task load index) of the TLX questionnaire between using the NVVI technique in phase 2 and the computer mouse in phase 3. Figure 6.13(b) shows a box plot of this comparison. An ANOVA analysis revealed a significant difference of perceived task load index between using the NVVI technique and using the computer mouse, with the use of the computer mouse being the lesser.



Figure 6.14: A box plot of the System Usability Scale (SUS) questionnaire scores for the pitch based NVVI prototype used in the study.

6.2.4 System Usability

After completing a comprehensive assessment with the NVVI technique in phase 2 and a baseline test with a computer mouse in phase 3, participants completed a system usability scale (SUS) questionnaire to provide a “quick and dirty” usability evaluation of the pitch based NVVI technique. A score above 68 indicates above average usability.

Figure 6.14 contains a box plot with the recorded SUS scores. The scores were normalized into the range of $[0, 100]$. The median score is approximately the average score of 48.31.

Therefore, the system has a below average usability score. The high and low scores were 0 and 100.

Chapter 7

Discussion

The results presented in Chapter 6 provide a large amount of data that can help answer the research questions listed in Chapter 3. User experience with the tool is discussed in Section 7.1. The research questions are addressed in Section 7.2. The limitations from the results and analyses performed also warrant discussion and are detailed in Section 7.3. Finally, future work is identified in Section 7.4.

7.1 User Experience

User experience with the pitch based NVVI technique is considered both quantitatively and qualitatively in this section. Section 7.1.1 explores user performance with the pitch based interaction tool, and Section 7.1.2 explores user feedback with the tool.

7.1.1 Performance

Task completion is an important metric in understanding user performance with the tool. Categories 5 and 6 were used for the analyses because they indicated completed tasks. Additionally, most of the analyses were performed with phase 2 because participants had undergone some training with the tool in phase 1. However, only 60.82% of attempted tasks were successful in these two categories for phase 2.

The pitch based NVVI technique used in this work received an average SUS score of 48.31, a below average usability score. Below average usability is supported by a 61.34% overall task attempted success rate.

As observed in Figure 6.5, the average path seemed to arrive close to the final position after approximately 15% of the time taken. Additionally, Figure 6.6(b) depicts an increase of over-adjustments and overshooting around 40-70% of the way through.

Vocal Scrolling

Figure 6.8(a) represents the “scrolling” phenomenon. This appears to be a logical approach to adjusting the slider knob with pitch, as it involves increasing or decreasing the pitch of either whistling or humming multiple times. This is likely because the end position is too far away to adjust pitch in one increment.

This design pattern should be considered when developing a pitch based NVVI system. This phenomenon alludes to a trade off made between precision and distance travelled. For this study, the amount of distance the slider knob makes per pitch adjustment (a weight) was an informally experimentally tuned value. If this value is large, this phenomenon would be less frequent, but it would likely come at the cost of precision (as informally supported via researchers’ experimentation.)

Overshooting

Figure 6.6(a) contains the aggregate paths of all attempted and successful tasks. Though both the time and distance axes are normalized, the distance axis is multiplied by a factor of 16. The initial peaks present indicate a significant overshooting using the tool, as some paths are clipped that are over 16 times the initial distance. This is indicated in Figure

6.6(b).

Figure 6.8(b) depicts the overshooting phenomenon in a single task as an example. Overshooting is visible in Figure 6.6(a) (the aggregate plot of paths) as well as Figure 6.5 (the average, interpolated path taken,) though only slightly visible. Figure 6.8(c) contains an example of severe overshooting. Severe overshooting is also noticeable in Figure 6.6(a), as the noticeable heights of the paths range up to 6 times the initial distance at around the 40-70% time markers (as the time scale is normalized.) In other words, overshooting appears to occur during the middle of the time using the pitch based NVVI technique.

7.1.2 Feedback

This section will explore survey data and feedback provided by participants of the study. Demographic and study feedback will be discussed.

Demographic The study was performed with a split amount of biological males and females (32 and 40, respectively.) Both sexes indicated they primarily have a medium voice type, though others were present as mentioned in section 6.2.1. Overall, a variety, but not an extensive amount, of vocal types were considered for analysis, as most were medium ranged individuals. 45 of the 72 considered participants were considered to be able to whistle by the metrics described in Section 5.1.5.

Most participants (39 of 72 or 54.17%) indicated they have a solid musical background by answering “Agree” or “Strongly Agree” to the Likert question “I have a strong musical background (e.g. playing melodic instruments, singing).” 9 participants had experience singing (e.g. in musical theatre or choir,) and 10 participants had experience playing instruments, some having both.

Learning the Tool. Some of the following feedback was recorded after phase 1. Though the

tool was thoroughly explained before phase one, it appeared that participants that struggled either skipped the reading and tried to learn the tool on the fly or they did not adequately understand the reading.

Some participants expressed difficulties learning the tool, though ultimately successful. One participant reported “Once I figured out how to hum to get the right frequency the humming became easy,” and another said “When I finally figured it out it was very easy to operate.” Another participant reported “The slider was a little more difficult to move than I expected. I did start to get a rhythm towards the end thou.” Many participants reported similarly, indicating that a learning period occurred and eventually they learned to use the tool successfully.

Expectations. Many participant indicated the the pitch based NVVI technique worked as expected. However, some participants indicated that their expectations of the tool differed from their reality using it. One participant said “Humming was easy but it was not as I expected to control the slider.” Another participant reported “Humming is a lot more difficult. I would’ve thought it would’ve been a lot easier. The slider didn’t go as much as I wanted it to.” Several more participants mentioned that the tool did not work as they expected.

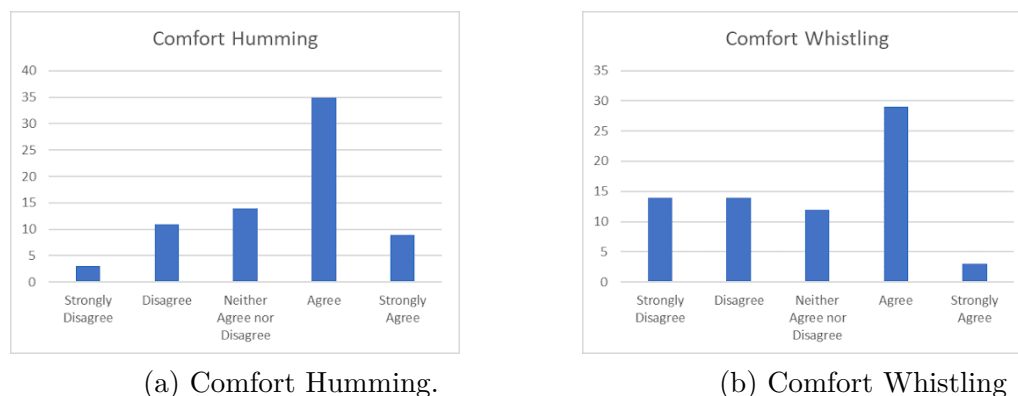
Understanding. Some participants expressed that they simply did not understand the tool. One participant said “I don’t really understand it but maybe that’s because I don’t have any musical training at all.” A lack of understanding of pitch and music could be correlated with lower usability. Another participant said “Humming is hard. It was hard to get the right pitch.”

Usability. All participants were able to use the tool to complete phase tasks. However, some participants indicated that they could not use the tool at different points during the

study. One participant said “It was really all over the place all in wasn’t consistent.” Another said “... it kept going down no matter the pitch I did.” Another said “I could not control my humming enough to move the slider back enough.”

One participant seemed disheartened after using the tool, as they said “I felt let down when I couldn’t get the [slider knob] to move where I wanted to. It was a little bit demeaning I thought.”

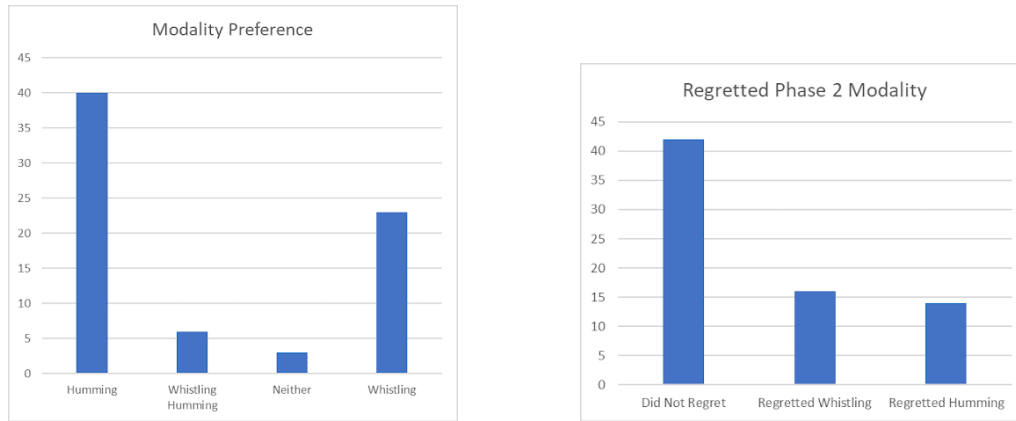
Figure 7.1: Bar charts depicting participants’ comfort using the NVVI technique. Answered according to the Likert statement “I felt very comfortable using the NVVI technique with ____.”



Humming vs. Whistling After completing the first part of phase 1, participants were asked to indicate their comfort humming. As depicted in Figure 7.1(a), the median response to the Likert question “I felt very comfortable using the NVVI technique with humming” was “Agree.” If participants could whistle, they were asked to indicate their comfort whistling. As depicted in Figure 7.1(b), the median response to the Likert question “I felt very comfortable using the NVVI technique with whistling” was “Neither Agree nor Disagree.”

Participants were also asked to compare their experience with using the tool with humming or whistling after phase 1. As indicated in Figure 7.2(a), the vast majority of participants preferred humming with the tool over whistling. One participant said “Humming was both more comfortable and more effective!” Another participant said “Humming was more com-

Figure 7.2: Bar charts depicting participants’ preferred modality to use with the NVVI technique and whether they regretted using the modality they used in phase 2.”



(a) Participants’ preferred modality to use with the NVVI technique. (b) Participants’ regret using the modality they chose for phase 2.

fortable as all I had to do was hum Mm” compared to whistling where I had to blow hard air.”

In phase 3, participants chose to either whistle or hum for a comprehensive assessment with the pitch based NVVI technique. After completing this phase, users were asked if they regretted their choice of modality. As depicted in Figure 7.2(b), most participants did not regret their choice, and around the same number of participants regretted their choice of whistling or humming (16 and 14, respectively.)

Selection. Some participants did not initially understand selection as implemented in our tool. One participant said “The humming part was easy. I got a little tripped up on the two succinct noises to turn it on and off but that may have been a mic issue.” Another participant said “...The tool did not move at all at first but only because I forgot to clap twice. Once I got the hang of that it became easier...”

Interactions.

After completing phase 1, users were asked to described the pitch control strategy they had

used to complete tasks. Many users reported using small increments or steps of pitch in either direction to move the slider, similar to the “scrolling” phenomenon depicted in figure 6.8(a). User reports were similar for both humming and whistling.

Some, but less, participants also reported using large increments. Some participants described switching from large to small increments or vice versa, though they seemed to agree that they felt they had more control with larger increments. One user said “I tried to move it in one Swift motion that found that task to be rather daunting so I ended up moving it mostly in small steps.” Overall, it appears smaller increments were more successful for participants.

One participant said “Depending on the distance I had to move the slider I would start much lower or higher,” indicating that the participant mentally and physically prepared for a longer distance by trying to expand the frequency range of their voice.

Physiological Feedback. Some participants reported physiological feedback, or their experience with the tool and how it physically affected them. One participant said “...I especially felt like I had to cough when I was doing it so I couldn’t do it as fully as I wanted.” Another user said “The tool worked as expected. My mouth.. eh not so much.” Another participant said “...you need to really focus and get some serious air in your lungs...”

One participant reported “I have very limited voice so this was hard. I cannot open my mouth and I have a paralyzed vocal cord. IT sucks. But I was able to hum and once I learned how to control the device it was easy enough.” However, this participant performed well, accurately completing every task. This shows that a participant with an impaired voice was able to successfully use the tool.

In phase 3 of the study, participants were comprehensively assessed with the pitch based NVVI technique. Multiple participants reported fatigue and discomfort after this phase.

One participant said “Phase three was ridiculously hard and now I have a sore throat and it wouldn’t work at all it was horrible,” implying that they could not use the tool for an extended period of time. Another user said “Wish I chose humming. I’m lightheaded with a headache now from whistling so much continuously.” After the study, one participant said “My palms sting from producing the loud claps. Ouch!”

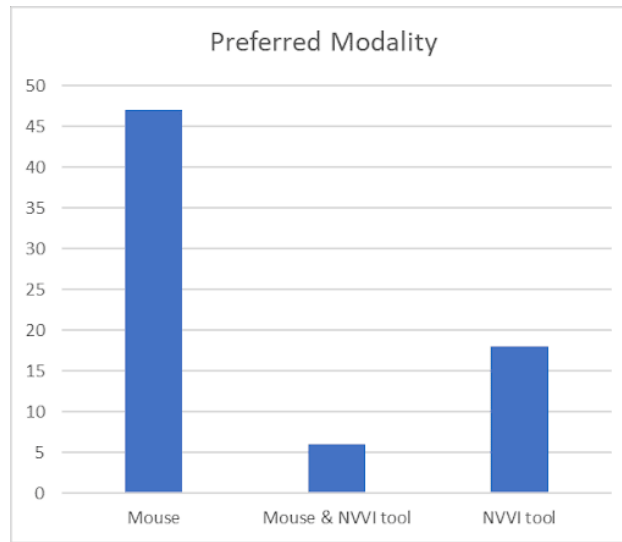


Figure 7.3: Bar charts depicting participants’ preferred modality to complete phase tasks (the pitch based NVVI technique or computer mouse.)

Pitch Based NVVI Versus the Mouse Participants were comprehensively assessed with the pitch based NVVI technique in phase 2 and then assessed with the computer mouse as a baseline in phase 3. After phase 3, participants were asked to compare their experiences using the NVVI technique and the mouse.

97% of participants said that the mouse was more comfortable to use than the pitch based NVVI technique and 91% said the mouse was more effective to use than the NVVI technique. Participants were asked to choose their preferred modality for completing phase tasks, the mouse, the pitch based NVVI technique, or both. As shown in figure 7.3, the mouse was significantly more preferred than the NVVI technique.

When describing their answers to the previous questions, many participants used words like “way” and “much” to qualify the improvement of experience with the computer mouse over the NVVI technique. Some users said that the mouse was more precise than the NVVI technique, and other mentioned that the mouse was quicker to use. When discussing the NVVI technique, some participants said the NVVI technique was more fun, but some followed up that it would be better suited for a game. One user said “It was less stressful [to] just use the mouse.”

NVVI technique and Study Feedback. After the study, users were asked if they thought that they could improve their skills with the tool with practice. Most participants said that they believed that they could become better, but many said they would still prefer to use an alternative tool with better performance, i.e. the mouse. One participant said “Maybe but I wouldn’t want to practice. Why have to practice it when I can already do it quickly and effectively with a mouse?”

Some participants said no because they believed the tool could use improvement, e.g. “No the tool just needs a bit of fixing before it can compare to the mouse.” Another user said “No. This tool needs to change before it can become better. There shouldn’t need to be practice involved. It is difficult and completely not user friendly.” Others said no because they were uncomfortable after using the tool, e.g. “I don’t know. It hurt my head after awhile all that humming.”

However, many participants were optimistic and said they believe they would improve after practice. Some participants described how practicing with the tool could actually benefit them, e.g. “Oh definitely! I could probably improve my pitch too. It might change my comparison.”

Participants were also asked to provide optional feedback on the study itself. One participant

said that it was one of the most difficult studies they had ever done. Some words used by participants in their feedback was “cool,” “unique,” “fun,” and “weird.” One participant said “This was meaningful and enjoyable!”

7.2 Research Questions

This section will discuss the implications of the findings regarding the research questions defined in Section 3.1.

7.2.1 RQ1: Does pitch-based NVVI follow Fitts’ Law?

A goal of this study was to determine if pitch based NVVI, using the relative pitch interaction technique used in the NVVI prototype, observes Fitts’ Law. To try and answer this goal, an analysis was performed with attempted successful tasks performed in phase 2, the comprehensive assessment using the pitch based NVVI technique. The top 7.5% of times were removed, as there were some considerable outliers. However, these outliers should be considered as a part of the overall the usability of this tool.

Figure 6.7(a) contains a plot of all of the normalized distances (start to end) of the start knob with the respective normalized time taken to complete the task from categories 5 and 6. A linear regression model was run on the data to analyse how closely the time taken and logarithm of distance traveled resembled a linear relationship (Fitts’ Law.) As described in the results section, the very low r^2 value implies that this relationship is not observed.

Another analysis was performed to observe a linear relationship between distance and time taken to complete the tasks. As indicated in the results section, another low r^2 value was calculated, implying that this relationship was also not reserved. This r^2 value was higher

than the logarithmic value, which means, sans errors, the relationship between time and distance appears to be more linear than logarithmic. Thus, we reject H_1 and claim that Fitts' law does not appear to be observed by the pitch based NVVI technique used in this study.

7.2.2 RQ2: How does a pitch-based NVVI technique compare with a computer mouse?

To answer RQ2 and understand how user performance varies using the pitch based NVVI technique as compared with a computer mouse, a within subjects comparative analysis was performed. The experimental hypothesis H_2 developed to answer this question was performance with a computer mouse and the pitch-based NVVI technique will not significantly differ.

An ANOVA was administered to validate this hypothesis. A significant difference was found between the average times of using a computer mouse and the NVVI technique, as shown in Figure 6.9(a). This is sufficient grounds to reject H_2 and conclude that participants performed significantly worse on average with the pitch-based NVVI technique than the computer mouse.

Overall, results relevant to this research question indicate that participants performed significantly better with the computer mouse than the pitch based NVVI technique for completing the tasks assigned in this study.

7.2.3 RQ3: With which mode of input for the tool do users perform better, whistling or humming?

To answer RQ3 and understand how user performance varies using the pitch based NVVI technique with humming as compared with to whistling, a between subjects comparative analysis was performed. The experimental hypothesis H_3 developed to answer this question was performance between whistling and humming will significantly differ.

An ANOVA was administered to validate this hypothesis. A significant difference was found between the average times of using humming and whistling the NVVI technique, as shown in Figure 6.10. This is sufficient grounds to reject the H_3 and conclude that performance difference was statistically significant in favor of humming.

Since there was a significant difference of performance measured in favor of humming, an analysis similar to the one done for the first research question was performed but with only humming instead of humming and whistling combined. The following experimental hypothesis was developed to address this analysis: performance with a computer mouse and humming will significantly differ.

An ANOVA was administered to validate this hypothesis. A significant difference was found between the average times of using a computer mouse and the NVVI technique with only humming, depicted in Figure 6.9(b). This is sufficient grounds to accept the experimental hypothesis and conclude that performance difference was statistically significant in favor of the computer mouse.

Overall, results relevant to this research question indicate that participants performed significantly better using the pitch based NVVI technique with humming over whistling, but performed significantly better with the computer mouse than humming or whistling with the pitch based NVVI technique for completing the tasks assigned in this study.

7.2.4 RQ4: Which modality (whistling or humming) do users experience less task workload?

To answer RQ4 and understand how the modality used with the pitch based NVVI technique to complete phase tasks varies, two analyses were performed.

A between subjects study was performed to consider the NASA TLX scores of all participants for phase 2. The experimental hypothesis H_4 developed to answer this question was the NASA TLX scores between whistling and humming will not significantly differ.

An ANOVA was administered to validate this hypothesis. A significant difference was not found between the NASA TLX scores recorded for humming and whistling. This is sufficient grounds to accept H_4 and conclude that participants experienced a similar task workload by whistling and humming.

A within subjects study was performed to consider NASA TLX scores of using the pitch based NVVI technique and the computer mouse (phase 2 and 3) of all participants. The experimental hypothesis developed to answer this question was the NASA TLX Score (perceived task workload index) will significantly differ between the modalities of pitch based NVVI and the computer mouse.

An ANOVA was administered to validate this hypothesis. A significant difference was found between the NASA TLX scores recorded for using the NVVI technique and computer mouse. This is sufficient grounds to accept the experimental hypothesis and conclude that participants experienced a significantly heavier task workload using the pitch based NVVI technique than with the mouse.

Overall, results indicate that participants experienced a significantly greater task workload whistling as opposed to humming with the NVVI technique. Additionally, it appears par-

ticipants experience a greater task workload overall when using the NVVI technique than when using the computer mouse.

7.3 Limitations

This work was ambitious in its conduction of a large scale study examining pitch based NVVI with a computer mouse baseline. It was conducted by introducing the topic of NVVI, pitch based NVVI, and the prototype for the study explicitly. The user then completed a series of tasks in phases. The user was trained with the tool in phase one for humming and whistling, the latter only if the study deemed them able. After this training, participants were more comprehensively assessed in phase 2 with the modality of their choice. Finally, participants completed the baseline using the computer mouse.

7.3.1 Methodology

The study procedure could be a limitation of interpretation of results. This study could have focused on only humming for the tool. The researchers believed that the work could provide a larger contribution if whistling was also analyzed and compared to humming. This came at a trade off for extending and complicating the study for both analysis and participants.

For multiple analyses, between subjects analyses were performed between participants who could and could not whistle, specifically phase 2 tasks that were attempted and successful (categories 5 and 6.) The number of studies actually considered is smaller than what researchers hoped for, though still offering valid insights. However, the number of whistling tasks completed in categories 5 and 6 from phase 2 was small enough to affect interpretation of results, even though the study was performed with 72 participants.

Additionally, the study performed analyses by considering the average time taken to complete tasks on a per participant basis. There may be better approaches for analyzing data, but this approach was chosen as it preserved the analysis' focus on participants instead of on individual tasks, as NVVI usability is likely consistent with a participant. A drawback of this approach is that for analyses like the one done for Fitts' Law, the top 7.5% of average times were removed, and results may have differed if they were considered. However, if task performance is consistent in participants, then high average times may be many consistently high times rather than a few that considerably affect the mean.

Categorization is another potential limitation of the work. For analysis, data was classified into 3 main groups, not attempted, attempted but unsuccessful, and attempted and successful. Each group had two subcategories, as defined in Section 6.2.2. This approach of categorization may have left out useful results or included bad data, though this approach was informally evaluated to be "accurate enough" by researchers performing the study. For example, one metric used for categorization was accuracy, specifically, if the final position of the slider knob was within 7% of the range ($|\text{start} - \text{end}|$) from the end position. It is possible that adjusting this value may provide better results.

Enforcement of particular tests was also limited due to the unproctored studies. For example, it is possible users lied about whistling when they actually hummed to complete phase tasks. Additionally, at least one user used a track pad instead of a mouse, though the study mentioned multiple times that a computer mouse was to be used.

7.3.2 MTurk

Amazon MTurk allowed this study to be performed at a much larger scale than studies performed in previous work. The study design was also asynchronous so that proctoring was

not required. There were drawbacks to using MTurk to conduct studies.

MTurk allowed for studies to be performed asynchronously, which meant proctoring was not necessary. This may have affected participants drive to complete the study to the best of their ability. Additionally, as described in Section 5.1.3, participants' primary motivation may have been money so they may not have tried as hard. On the opposite side, participants were rewarded only \$1 USD for the work, which may have affected their perceptions and motivation for completing the study to the best of their ability. For example, in some qualitative questions, participants would respond with "yes" instead of an actual answer to the question.

MTurk did not provide the requested 100 participants during the availability of the HIT before it closed. As a result, less participants than requested were able to attempt or complete the study. In hindsight, some additional HIT settings should have been configured differently. For example, the study should have been available for a longer time before it was closed off for participants (in this study it was only made available for 1 hour.)

7.3.3 Study Ecosystem

The study ecosystem also led to some limitations with the study. For example, an outage of the secure sever occurred that prevented some users from submitting their studies but still gave them the completion code for the effort they put in for attempting to complete the study. Additionally, a logic error on that server that was fixed after a participant reported an error did not store some submissions of participants that could not whistle.

The study was conducted asynchronously on a study website that was accessed from MTurk. During the study, a results file was continuously appended to with phase task results and surveys. In the pilot study, no results seemed to have any issues with missing data. After the

official study was conducted, some results were missing entire phase tasks and/or surveys. It is unclear why this data would go missing, but two results from the 74 collected were removed because they somehow were missing required data. This could have been an unsupported or partially supported browser issue, as mentioned in Section 5.1.3.

7.4 Future Work

In this study, 72 participants' data were collected and analyzed. 3,330 total tasks are collected and stored with participants' data. Demographic data and survey questions are also included in participants' data collected during this study. There are many individual facets of the data that were not explored, such as layered brushing. As an additional contribution, the study website (and analysis/filtering algorithms) are also included as contributions to this work. The entire study ecosystem (website and server) are also contributions for this work. Future work could, at a minimum, replicate this work to support these findings.

As described in the review of literature discussion section (Section 2.5.1,) one of the primary gaps considered in this research is the lack of literature and studies performed. Though there were drawbacks of this study's conduction mentioned in Section 5.1.5, the ecosystem used in this study permitted a large scale study of an NVVI system. Future work should consider the use of a similar and perhaps improved system that can be used to perform large scale analyses. Future work should also attempt to address some of the limitations described in the previous section.

In addition to larger scaled studies, future work should focus on conducting a higher quantity of studies and comparative studies. Another research gap mentioned in the discussion of the review of literature was the diversity of NVVI literature and studies without any overarching themes and connections. Future work should attempt to bridge findings from different forms

of NVVI to develop generalizeable guidelines and design patterns from synthesizing results from multiple studies.

As discussed in the limitations section, categorization may include bad data or exclude valid data that should have been considered. Future work should create a more robust classification system to maximize the amount of valid data included. The metric of average time, as described in section to do, may be improved upon in future work.

Participants were rewarded \$1 USD for completing the study. However, the NASA TLX task workload index scores described in Section 6.2.3, and user feedback in Section 7.1.2 indicate that the study was rigorous and challenging for some participants. In future work, participants should be rewarded more for challenging studies such as the one performed in this work.

Chapter 8

NVVI Framework

The research performed in this work has explored the field of NVVI.

The literature reviews presented in Chapter 2 have identified some existing systems and conclusions the authors made from their studies. A limitation discovered in the literature review was the scarcity of literature in NVVI and the lack of any generalizeable interaction patterns or design guidelines.

The study conducted in Chapter 5 and the results collected in Chapter 6 led to some interesting findings and discussion points detailed in Chapter 7.

With the knowledge gained from the research performed in this work, a general NVVI framework is developed in this chapter. In Section 8.1, the theoretical framework is described, along with design choices. In Section 8.2, an implementation of the framework is outlined.

8.1 NVVI Framework

The NVVI technique used in this study was built upon some general themes, as presented in this section. Section 8.1.1 provides a general discussion of results from the review of literature and user study to consider for designing the framework. Section 8.1.2 provides key insights to be considered for an NVVI system. Section 8.1.3 provides a conceptual structure and interaction flow for an NVVI system.

8.1.1 Discussion

NVVI Interaction Design

In Section 2.4.4, we explored three main types of NVVI interaction types: discrete/triggered input, continuous input, and vocal gestures that were usually implemented by analysing an individual's fundamental frequency created via voiced vocalization.

Vocal gestures were the most common of the explored NVVI interaction types, which occur by processing a user's pitch and mapping it to a set of commands associated with each gesture. Important insights found as a result to user studies was that gestures should not be too complicated complexity-wise or physically (think a deep voice trying to hit a high note), as they can cause vocal strain and fatigue. However, they can be effective.

Non-gestural continuous-based NVVI systems appears to be the second most common design of user input. Due to the limited number of studies found in this review that follow solely continuous input, more studies should be conducted so that a better analysis may be performed on continuous NVVI systems. Likewise, limited studies provided sole clicking functionalities, but perhaps this is because discrete interactions like clicking are not very functional without another, higher dimensional mode of navigation.

For any pitch related interaction techniques, the diversity of users' voices should be considered. Pathological and dysphonic voices were discussed in Section 3.2.5, and should be considered during NVVI system design. Additionally, some users may be quieter individuals than others, so any use of volume as a parameter to NVVI systems should be considered. Other social cues, such as embarrassment using NVVI systems, should be considered.

Though Poláček et al. report that users preferred absolute pitch gestures [58], the wide difference of vocal ranges between different genders and ages should be considered during

design, and relative gestures may seem to be more effective in future studies. This should be explored further.

Finally, NVVI may be better as an appropriate orthogonal controller scheme, depending on the application. For example, NVVI was used in conjunction with eye tracking to control in a game that users found engagement and enjoyment in [29], and it was not as successful as an alternative to keyboard compared to other alternative text input schemes [73].

NVVI Implementation Design

Once an interaction technique is chosen, users must design the high-level design of NVVI systems. For vocal gestural NVVI systems and continuous systems, some form of frequency detection and extraction should be used. In purely trigger based NVVI systems, other options to detect input may be considered, depending on the input type.

As described in Section 3.2.1, spectral analysis is a powerful tool for analyzing a window of audio data, and may be useful for analyzing more complicated sounds, such as vowels or sibilants. However, for basic gestural controls, a pitch detection algorithm will likely suffice. For trigger based NVVI input, machine learning algorithms were used to classify input as blowing or not blowing [82]. Regardless, the expected ambient noise must be considered, as this will likely effect the accuracy of any frequency extraction algorithm, as described in Section 3.2.6. This is on reason why *confidence* was developed as a part of the NVVI technique used in this study.

There may not be a specific best algorithm to use for pitch detection or frequency extraction, though several were described in this review, and many modern variations and optimized algorithms were cited as well. However, algorithms that take advantage of natural vocal characteristics will probably be a good choice. For example, the HPS algorithm described

in Section 3.2.2 takes advantage of harmonics naturally found in the human voice to derive a more confident pitch value. Experimentally, ACF was found to be proficient at classifying audio input as voiced or voiceless and pathological or not. Voicing identification plays an important part in identifying the beginnings and ends of vocal gestures. Depending on the expected or required type of input (e.g. singing, humming or whistling), as well as the target audience, the algorithm to use may vary and should be experimented with.

The transient detection used in this study was successful, as participants were able to use the tool with only minor feedback on its usability.

8.1.2 Key NVVI Implementation Insights

Real-time Interaction and Events

An NVVI system should be performing real-time audio processing to gauge user input, as that is what some other input modalities do. For example, when typing on the keyboard, each key press triggers an event that proceeds to interrupt the CPU so that a key press event can be sent to applications, along with other modes of CPU interrupts that draw similarities to event based programming [56]. Similarly, the NVVI technique should be sending information as close to real time as possible.

Therefore, a logical design pattern, similar to the one used in the NVVI prototype created for this study, should include an *event-based* system for updating any applications relying on the NVVI system for input. In the NVVI prototype described in this work, this event-driven behavior was implemented as a callback that would be called every time data was processed, and it contained a parameter indicating the confidence of the detected pitch and pitch information.

Multi-threading

Multi-threading was leveraged in the NVVI prototype used in the user study. Multi-threading can be particularly helpful when dealing with a large amount of data or complex algorithms.

This is additionally important for a web page environment like the one presented in the study performed in this work. By default, JavaScript code will execute on the same thread that renders all of the DOM elements and handles events. This can severely hinder the performance of the NVVI code and vice versa, so it makes sense to encapsulate most of the processing into another thread.

Processing Function Flexibility

When developing the pitch based NVVI prototype used in the study, many variants of processing functions were developed inside of the audio thread's scope. Developers will likely want to test a variety of functions or allow for some customizations of processing functions. Developers should also not have to reverse-engineer and understand all of the inner workings of the NVVI framework to implement a processing function.

An interface that allows the user to provide their own processing function can help alleviate these problems. This approach is taken in the implementation of the NVVI framework described in the next section. However, a challenge faced while implementing this approach was that the function had to be reconstructed after being transferred to the audio processing thread as a string, which to the developers' knowledge was the only way of achieving this functionality. As a consequence, any outside scope used by the processing function is lost. To attempt to make up for this, a dependency library is implemented in the framework described in the next section. But because any dependencies also lose their scopes, the

dependencies must also be serializable.

8.1.3 Structure and Interaction Flow

Based on these guidelines, an NVVI system that supports a variety of NVVI input types can be designed. The design should try to be accommodating to different forms of interactions and should support flexibility for the client to customize the NVVI system without modifying any integral framework code.

The NVVI system should receive an input stream of data to consider as input audio data. Processing should occur using this data, and then the output should be fed to an external application. For example, in the study performed in this work, an NVVI system used microphone audio as input, processed it with the processing function, and then sent the result to a callback function that ultimately controlled the sliders. In other words, three main interfaces should be provided to client programmers: **input**, **processing**, and **output**. Figure 8.1 depicts a high level overview of such a system.

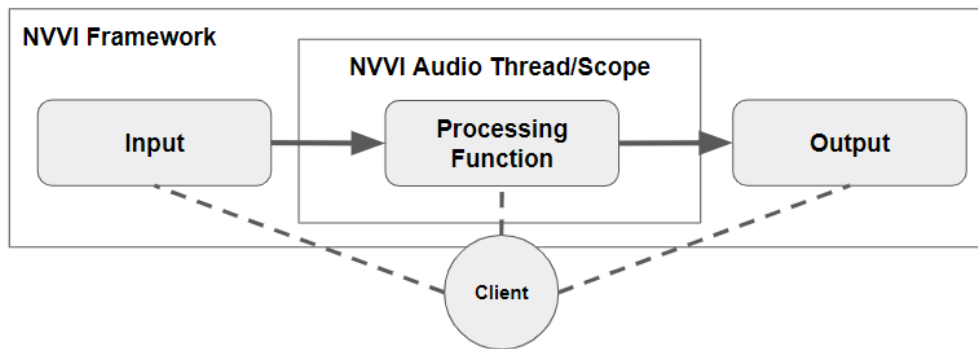


Figure 8.1: The flow and structure of three interfaces customizable by clients implementing an NVVI system using an NVVI framework.

Input The input to the NVVI system should be easy to configure, but may depend upon the language of implementation. Regardless of how it is implemented, the input should not

be restricted to one type or context. For example, in some languages, a *file stream* could be used as an input stream, as this allows not just direct microphone input to be used, but also data streaming over the web or from an input file.

Processing The NVVI system should allow the client to submit a processing function to the system without having to access or modify framework code. This can be, for example, a *function pointer* in some low level languages, or a callback in higher level ones. This approach widens the possible variety of processing functions that can be used (e.g. a variable function) and speed of development (e.g. in the user study performed in this work, the framework was modified constantly to adapt to different processing functions. Such flexibility could have sped up development here.) This processing should occur on another thread, if possible, for clients to take advantage of performance.

Output Finally, the NVVI system should provide output in a usable and flexible way. It should match or contain the results of the processing function. Extending the analogy of a file stream input, a corresponding file stream output would likely be appropriate. Flexibility would also be important to consider here. For example, in JavaScript, if the processing function returns an object with particular keys, some version or container of that object should be returned.

8.2 Framework Implementation

The NVVI framework was implemented with the JavaScript programming language. Examples and demonstrations are available via a website and its source code. The intended use of the framework is to be implemented into a customizable NVVI technique that can run

on a JavaScript environment like a web page. The specifications of the tool are detailed in Section 8.2.1. Examples and demos are provided in Section 8.2.2.

8.2.1 Specifications

The NVVI technique is built as a collection of JavaScript ES6 modules. All modules are implemented using functional programming in JavaScript. The primary NVVI module is located at “/js/NVVI/nvvi.js” and is referenced in every NVVI demonstration provided.

The NVVI technique should be instantiated with the following arguments passed in a single object parameter to the constructor.

- **sampleRate:** the sample rate for the NVVI technique to process audio.
- **callback:** The callback that will be activated after the NVVI technique processes a frame of audio. The single parameter to this function, *e*, will contain data returned from the audio processing function in *e.data*.
- **frameSize:** The integer size of one audio frame.
- **internalBufferSize:** The integer size of the internal buffer.
- **processRate:** The rate (in samples) at which the tool should process audio.
- **processAudioFunction:** The processing function for the NVVI technique. This function should not rely on any scope or library except for as defined in the next parameter. This function should return an object containing data to be sent to the callback function.

Some optionally present options are also available for configuration.

- **initiallyActive:** Denotes if the tool should activate immediately or require the client to activate it. False by default.
- **dependencies:** An object that contains functions or constants. This allows the user to pass in additional information for the processing function to be able to access. This may be an object but it must have a constant value and not be a deeply nested object. This object is ultimately string-ified and cloned into another thread's web worker, elaborated in another section.

The client should pass in a **processAudioFunction**, as this will perform customizable audio processing in another thread. The scope of the function should be limited to the parameters passed to it. The **callback** function's scope may be considered, as demonstrated in the examples.

Calling the constructor of the NVVI module will return an object with the following four functions used to control the NVVI technique.

- **start:** starts the NVVI technique and it will begin audio processing.
- **stop:** stops (pauses) the NVVI technique from processing audio until resumed with start.
- **push(<audio data>):** adds audio data to the internal buffer. Also increments the internal buffer's processing counter for each sample read into the internal buffer until it matches the processing rate. At that point, the process audio function will be called.
- **destroy:** Destroys the NVVI technique and cleans up the audio processing thread and state.

Microphone Input

One aspect of this implementation was to abstract away the WebAudioAPI interface for accessing user input data. For example, what if the client wants to send streamed data into the NVVI technique? For this reason, a separate entity was developed, the MicrophoneInput module.

The MicrophoneInput module creates an AudioContext instance and uses the WebAudioAPI to ask for access to the user's microphone. It can then pass data directly to the NVVI technique. The constructor for the MicrophoneInput module takes no parameters and returns an object containing the following functions.

- **initialize:** Initializes the microphone stream. This triggers a call to the WebAudioAPI to ask the user for microphone access. If permission is granted, will begin recording user input and passing it to the NVVI technique if configured.
- **connect(<NVVI technique>):** Connects a the MicrophoneStream to the NVVI technique by routing the MicrophoneStream's microphone data to the NVVI technique by an internal call to its push method (so the client doesn't have to keep making this call manually.)
- **stop:** Stops the microphone stream and therefore pauses the NVVI technique if it is connected to the MicrophoneStream. The NVVI technique's state is still active after this call.

Internals

The NVVI technique is implemented as a JavaScript ES6 Module function. Its state is almost entirely stored in a WebWorker [52]. This WebWorker is essentially an audio thread with

state. The NVVI technique can be thought as just a messenger to the WebWorker (audio thread.) A more elaborate explanation of the technical internal components, including the state and all variables, can be found in the source code comments of the NVVI framework.

8.2.2 Examples and Demonstrations

The implemented NVVI framework includes several examples demonstrating a small variety of approaches and applications for the tool. It is recommended to read the source code comments to understand the programming flow when using the NVVI technique. The provided examples are listed below.

- **Oscilloscope** (</oscilloscope.html>): An SVG oscilloscope is implemented with the NVVI technique. The NVVI technique is provided a processing function that normalizes the audio data from $[-1, 1]$ to $[0, 1]$ and returns this data. The NVVI technique is provided a callback function that updates the oscilloscopes SVG path with the normalized audio data returned from the processing function. A `MicrophoneStream` is created and connected to the NVVI technique. When the user clicks start, the oscilloscope begins to process audio and display the audio frame.
- **Spectrogram** (</spectrogram.html>): An SVG spectrogram is implemented with the NVVI technique. An FFT function is provided as a part of the **dependency** parameter. The NVVI technique is provided a processing function that performs an FFT using the provided FFT function. The NVVI technique is provided a callback function that updates the spectrogram's SVG path with the spectrum returned from the processing function. A `MicrophoneStream` is created and connected to the NVVI technique. When the user clicks start, the spectrogram begins to process audio and display the frequency spectrum of the audio frame.

- **Slider Controller (/slider_controller.html):** A relative pitch based NVVI technique similar to the one used in this work's user study is implemented with the NVVI framework. The NVVI technique is provided a processing function that performs pitch and confidence estimation described in the user study NVVI prototype. The NVVI technique is provided a callback function that updates the position of an HTML slider based on the resulting pitch and confidence values. A MicrophoneStream is created and connected to the NVVI technique. When the user clicks start, the tool becomes active.
- **Blow Controller (/blow_controller.html):** An application that uses a form of blowing as an NVVI modality is implemented with the NVVI framework. A simple SVG environment is created with a ball that adheres to gravity, and updates its position and velocity frame by frame. The NVVI technique is provided a processing function that returns the RMS of the input audio frame. The NVVI technique is provided a callback that lifts the ball such that it fights gravity dependent on the RMS calculated in the processing function. A MicrophoneStream is created and connected to the NVVI technique. When the user clicks start, the application begins to allow the user to blow to lift the ball.

Chapter 9

Discussion

The primary aim of this work was to explore the niche field of NVVI. In this work, a review of literature was performed to explore the gaps in the field of NVVI. Two major gaps that were identified were a lack of literature and studies, and a missing bridge connecting the few existing NVVI systems in terms of design guidelines, interaction patterns, and a design framework.

Since the studies conducted in prior work were often small in participant count, a large scale study was planned. A pitch based NVVI system was chosen to be explored because of its ability to act as a continuous input source based on the input frequency. Research questions were designed to explore the functionality of pitch based NVVI. They were:

1. Does pitch-based NVVI follow Fitts' Law?
2. How does a pitch-based NVVI technique compare with a computer mouse?
3. With which mode of input for the tool do users perform better, whistling or humming?
4. Which modality (whistling or humming) do users experience less task workload?

Relative Pitch Based NVVI Design A design was introduced in Section 4, modeled in part from the conclusions made in the review of literature. A relative pitch based interaction technique is detailed and incorporated into the design. Trade-offs between this and other

approaches are discussed, but the primary benefit of this approach is described to be the lack of configuration necessary to immediately use the tool, as relative pitch is based on a change of pitch and is not bound to any particular frequency.

Additionally, a mechanism we called *selection* was detailed as to how it could help improve user intent with the tool. In our design, selection is performed by creating two *transients*, or short, loud impulses such as clapping, to enable or disable the tool.

In the pitch based NVVI design, we describe *confidence* as a useful metric to understand when the user is intending to produce a pitch versus when they aren't.

Large Scale and Extensive User Study The user study conducted in this work is outlined in section 5 to address the research questions developed in this work. Amazon MTurk was chosen as a service to conduct our large scale study. The pitch based NVVI prototype and implementation used for the study is detailed in this section. The input of the tool is intended to be a person's whistling or humming, as each was individually explored in prior work.

Phase tasks are described as essentially using relative pitch to move a slider knob from one position to another. The key insight behind this design decision is to leverage the continuous, one dimensional input provided by a pitch based NVVI system.

Results Results from the study are provided from a diverse set of 72 participants. Demographic information such as biological sex, vocal range, and musical background are discussed. Aggregate and exemplary task paths and phenomena are explored. Data and figures relevant to the research questions are provided. *Task categorization* is described, as there were cases in which users did not attempt to perform tasks, and for analyses only attempted and successful tasks were considered, which are brushed by this categorization process.

The results are discussed and contextualized. User performance is considered first, with an attempted task completion rate of 60.82%. The average System Usability Scale (SUS) score was 48.31, indicating a below average usability score.

The paths taken of the slider during the study were recorded and analyzed. The phenomenon of vocal *scrolling* is depicted in Figure 6.8(a), where the user seems to raise or lower the pitch of their voice in fixed increments. This phenomenon is supported by user feedback.

Overshooting and *severe overshooting* are explored. Figure 6.8(b) depicts the aggregate of all attempted and successful normalized graphs (both time and distance normalized,) though the distance is multiplied by a weight of 16. The fact that some paths are still clipped in the beginning is an example of severe overshooting. Additionally, it seems that another area for overshooting is around 40-70% through the path taken.

Findings Over 50% of participants had a fair musical background, 9 with experience singing and 10 with experience playing instruments (not disjoint.) 45 of 72 participants were able to whistle for the study. Though the biological sexes were split almost evenly among participants, most indicated they had “medium” voices, as opposed to low or high.

Some participants had difficulty initially learning the pitch based NVVI technique in phase 1, but eventually got the hang of it by practice and “trial and error.” Some participants noted that their expectations of how to use the tool differed from actually using it. Some participants indicated they did not understand the tool, one alluding to their lack of musical background. Learning *selection* with the NVVI technique was difficult for at first for some participants.

All participants were able to use the tool to complete phase tasks. Some participants did face difficulties using the tool, citing factors like consistency and lack of control. At least one

participant said they felt “let down” that they couldn’t use the tool the way they wanted to. Participants strongly preferred humming to whistling when using the NVVI technique, citing comfort and effectiveness. A participant noted that whistling required blowing “hard air.” After being comprehensively assessed with the modality of their choice in phase 2, most participants did not regret their choice of modality, but around the same number did for whistling as humming.

Participants described using small and large increments to control the slider in phase tasks. Some participants switch from large to small and vice versa, but generally agreed that smaller increments were more effective or comfortable, indicating they used the *scrolling* method as previously described. One participant noted that they prepared for larger distances by preparing to start at a lower or higher pitch than normal.

Some participants also experienced some fatigue and discomfort during and after using the pitch based NVVI technique. Some reports included soreness of the throat, coughing, light-heartedness. One participant reported their palms stung after performing claps for selection. One participant indicated that they had a vocal impairment and that they felt discomfort using the tool, though they were able to successfully complete tasks.

Participants strongly preferred using a computer mouse over the NVVI technique, many citing comfort and effectiveness was better with the former. Some participants noted that the tool was fun, with some saying it would be a fun input modality for a game. Game contexts for NVVI were attempted in prior work [29, 30].

Some participants said they could improve their ability to use the tool with practice. Some participants mentioned that even though they believed they could improve with the tool, they would still use a mouse primarily. Some said they do not think they could improve with practice because the tool itself needed improvement.

No participants in the official study mentioned embarrassment while using the tool, while this phenomenon was present in prior work[29] and the pilot study.

Results from the study indicate that relative pitch based NVVI as used in this study does not follow NVVI. Additionally, the time taken to complete tasks and distance do not appear to have a direct linear relationship. However, future work should explore performing more large scale studies to record more data points and attempt to find a correlation between distance and time.

With support from participant responses, the computer mouse outperforms the pitch based NVVI technique used in this study. Results from an ANOVA analysis support the claim that the average performance times were significantly different between the computer mouse and NVVI technique, in favor of the computer mouse.

As most participants indicated a preference of humming over whistling, results from an ANOVA analysis supported the findings of a significant difference of performance of using the NVVI technique with humming and whistling, in favor of humming.

Participants reported that they thought that the tool was fairly intensive to use, especially as compared to the mouse. Participants tool a NASA TLX (task workload index) questionnaire to gauge the task workload after they were comprehensively assessed with the NVVI technique and the computer mouse. Figure 6.12(a) indicates a slight difference of task workload experienced between humming and whistling with the NVVI technique, in favor of humming. Figure 6.13(a) indicates a significant difference of task workload experienced between using the pitch based NVVI technique and the computer mouse. Altogether, participants seem to experience significantly higher task workload using the NVVI technique than the computer mouse and a slight higher task workload whistling over humming.

Limitations There are potential limitations of this work of the methodology of the study, including its length, ecosystem and task design. Many tasks were unsuccessful and many were not even attempted (categories 1 and 2,) limiting the pool of considered tasks. The metrics used for categorization (*local variance*, *global variance*, and *accuracy* were detailed and tested by researchers, but it is possible that the algorithm included or excluded some tasks. For example, the metric of *accuracy* was a final slider knob position within 7% of the designated end position.

Additionally, the methodology of considering the *average time* taken for attempted and successfully completed tasks from phase 2 narrowed the pool even further. The concept of average time taken may also not have been the most representative metric to use for analyses such as one way ANOVA, though it did preserve the focus of participants. The top 7.5% of average times were also removed for some analyses, which may affect interpretation of the results.

Regarding the study ecosystem, an outage of a server that communicated with the study website to capture results, as well as a logic error in the code running on the server are two limitations of the study. Additionally, two results missed essential data somehow and were removed from the analysis pool of results.

Amazon MTurk allowed for a large scale study to be performed asynchronously, but came with some drawbacks. For example, there may have been a primarily monetary incentive to complete the study. Additionally, the lack of oversight may have discouraged participants from trying. In some qualitative feedback sections, participants responded to open ended questions with “yes” or “good” when these terms did not make sense as answers. MTurk also did not provide the 100 participants requested, instead closing the study HIT off at a total of only 74 successful submissions.

NVVI Framework An NVVI framework is proposed in Chapter 8 synthesizing study results with prior literature. Interaction designs including vocal gestures, relative and continuous pitch gestures, and implementation areas are discussed. Implementation designs, including ecosystems, algorithm decisions, event driven systems, multi threading, and flexibility for client programmers are discussed.

A major contribution of this work is provided as an NVVI framework implementation. The implementation is coded in JavaScript and intended for website interactions. The design and specifications are detailed regarding the implemented tool, which offers a simple interface involving:

- A (scopeless) *processing function* that the client wants the NVVI system to run every time a frame of audio is processed.
- A *callback function* that allows the client to use the results calculated (returned) by the processing function within the scope and context that the NVVI technique is being used in.

Several simple examples are provided, including an oscilloscope, a spectrogram, a slider controller, and a blow controller. They are simple web pages that use the provided NVVI framework. A MicrophoneController module is also introduced that allows for microphone input to the NVVI technique while decoupling it from the NVVI technique itself.

Future Work Future work should re-examine the existing literature of NVVI, as the field may be growing at the time of writing of this work.

More large scale studies should be conducted to validate findings from this work. If possible, the number of results of attempted and successful tasks should be maximized for the great-

est understanding of NVVI usage, as many tasks completed in this user study were either unattempted or attempted and failed.

The study performed in this work may also be replicated or improved upon, as all components are provided as a contribution. For example, the classification system may be improved upon to include more valid results and less invalid results. Future studies should also provide a higher compensation to participants, as only \$1 USD was provided to participants in the study performed in this work.

A higher quantity of studies, not just in participant count, should be performed with a variety or variations of different NVVI systems, which may lead to an understanding of overarching interaction patterns and design guidelines when synthesized.

The NVVI framework proposed and implemented for this work should be tested and explored to evaluate its functionality and expressiveness for a variety of NVVI modalities.

Chapter 10

Conclusion

The primary aim of this work was to explore the niche field of NVVI. The contributions provided by this work are:

1. A discussion of the current state of NVVI resulting from a systematic literature review conducted following PRISMA guidelines.
2. A detailed explanation and implementation of a pitch based NVVI technique coded in JavaScript.
3. A modular framework for conducting user studies and questionnaires, coded in JavaScript, HTML and CSS.
4. An evaluation of a pitch based NVVI technique with the presentation of both qualitative and quantitative results from the conducted user study.
5. A discussion of relevant usability phenomena observed using the study and design guidelines considering these phenomena and results from the user study evaluation.
6. An implementation of a general NVVI framework coded in JavaScript, as well as a variety of implemented demonstrations using the framework.

In Chapter 1, the background of NVVI and the motivation for this work are explained. Research questions to guide the work and the approaches to answering them are proposed. The contributions this work provides are also enumerated in this chapter.

In Chapter 2, both a narrative and systematic literature review are detailed. The guiding research questions for both reviews are provided, along with the guidelines used to conduct the systematic review. The results from each review are presented along with a discussion about their impacts and contributions to the field of NVVI. Research gaps are also identified.

In Chapter 3, the problem definition is detailed, including the research questions developed to guide the work and foreseen research challenges.

In Chapter 4, the approach to answering the research questions from Chapter 3 are provided. The relative pitch interaction technique is detailed and the pitch based NVVI technique design is outlined.

In Chapter 5, the Study Design is detailed. The chapter includes how the research questions and challenges are addressed by the user study, as well as the limitations with addressing the research challenges. The pitch based NVVI prototype is detailed, and the study ecosystem and procedure is described. The statistical methods used in the study, along with the hypotheses to be tested, are provided in this chapter.

In Chapter 6, the results from a pilot and official user study are presented.

In Chapter 7, the results presented in Chapter 6 are discussed. The research questions and challenges described in this work are addressed. Patterns and insights into usage of the tool are explored.

In Chapter 8, results from the literature reviews performed in Chapter 2 and results from the study provided in Chapter 6 are synthesized into a theoretical NVVI framework. An implementation of this framework is also detailed.

In Chapter 9, the research done in this work is filtered and the most important conclusive findings are extracted and presented.

Bibliography

- [1] —. Ieee recommended practices for burst measurements in the frequency domain. *IEEE No 265-1966*, pages 1–20, Feb 1966. doi: 10.1109/IEEESTD.1966.7393399.
- [2] Sound Intensity, sep 12 2022. [Online; accessed 2023-11-20].
- [3] S. Ahmadi and A.S. Spanias. Cepstrum-based pitch detection using a new statistical v/uv classification algorithm. *IEEE Transactions on Speech and Audio Processing*, 7(3):333–338, May 1999. ISSN 1558-2353. doi: 10.1109/89.759042.
- [4] Sadeen Alharbi, Muna Alrazgan, Alanoud Alrashed, Turkiayh Alnomasi, Raghad Almojel, Rimah Alharbi, Saja Alharbi, Sahar Alturki, Fatimah Alshehri, and Maha Almojil. Automatic speech recognition: Systematic literature review. *IEEE Access*, 9:131858–131876, 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3112535.
- [5] Rafael George Amado and Jozue Vieira Filho. Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes. In *2008 International Conference on Audio, Language and Image Processing*, pages 449–454, July 2008. doi: 10.1109/ICALIP.2008.4590188.
- [6] Amazon.com. Amazon mechanical turk. <https://www.mturk.com/>. Accessed: 2023-11-28.
- [7] B. S. Atal and M. R. Schroeder. Adaptive predictive coding of speech signals. *The Bell System Technical Journal*, 49(8):1973–1986, Oct 1970. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1970.tb04297.x.

- [8] He Ba, Na Yang, Ilker Demirkol, and Wendi Heinzelman. Bana: A hybrid approach for noise resilient pitch detection. In *2012 IEEE Statistical Signal Processing Workshop (SSP)*, pages 369–372, Aug 2012. doi: 10.1109/SSP.2012.6319706.
- [9] Roland Badeau, Valentin Emiya, and Bertrand David. Expectation-maximization algorithm for multi-pitch estimation and separation of overlapping harmonic spectra. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3073–3076, April 2009. doi: 10.1109/ICASSP.2009.4960273.
- [10] J. Baker. The dragon system—an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 23(1):24–29, 1975. doi: 10.1109/TASSP.1975.1162650.
- [11] Jeff A. Bilmes, Xiao Li, Jonathan Malkin, Kelley Kilanski, Richard Wright, Katrin Kirchhoff, Amarnag Subramanya, Susumu Harada, James A. Landay, Patricia Dowden, and Howard Chizeck. The vocal joystick: A voice-based human-computer interface for individuals with motor impairments. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, page 995–1002, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220700. URL <https://doi.org/10.3115/1220575.1220700>.
- [12] Tom Brill, Laura Munoz, and Richard Miller. Siri, alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. *Journal of Marketing Management*, 35, 11 2019. doi: 10.1080/0267257X.2019.1687571.
- [13] John Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189:4–7, 1996.
- [14] Francisco Maria Calisto and Jacinto C. Nascimento. Nasa-tlx survey, 2018. URL <http://rgdoi.net/10.13140/RG.2.2.26978.79044>.

- [15] JANE P. CLENDINNING. *COMPUTER RECOGNITION OF PITCH FOR MUSICAL APPLICATIONS*. PhD thesis, 1983. URL <http://login.ezproxy.lib.vt.edu/login?url=https://www.proquest.com/dissertations-theses/computer-recognition-pitch-musical-applications/docview/303183213/se-2>. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-02-19.
- [16] W.T. Cochran, J.W. Cooley, D.L. Favon, H.D. Helms, R.A. Kaenel, W.W. Lang, G.C. Maling, D.E. Nelson, C.M. Rader, and P.D. Welch. What is the fast fourier transform? *Proceedings of the IEEE*, 55(10):1664–1674, Oct 1967. ISSN 1558-2256. doi: 10.1109/PROC.1967.5957.
- [17] K. H. Davis, R. Biddulph, and S. Balashek. Automatic Recognition of Spoken Digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 06 2005. ISSN 0001-4966. doi: 10.1121/1.1906946. URL <https://doi.org/10.1121/1.1906946>.
- [18] Patricio De La Cuadra, Aaron S Master, and Craig Sapp. Efficient pitch detection techniques for interactive music. In *ICMC*, 2001.
- [19] Susan Denham. Pitch detection of dynamic iterated rippled noise by humans and a modified auditory model. *Biosystems*, 79(1):199–206, 2005. ISSN 0303-2647. doi: <https://doi.org/10.1016/j.biosystems.2004.09.008>. URL <https://www.sciencedirect.com/science/article/pii/S0303264704001480>. Selection of Papers from the 5th International Workshop on Neural Coding.
- [20] J. Dubnowski, R. Schafer, and L. Rabiner. Real-time digital hardware pitch detector. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(1):2–8, 1976. doi: 10.1109/TASSP.1976.1162765.

- [21] Steven J. Eliades and Xiaoqin Wang. Contributions of sensory tuning to auditory-vocal interactions in marmoset auditory cortex. *Hearing Research*, 348:98–111, 2017. ISSN 0378-5955. doi: <https://doi.org/10.1016/j.heares.2017.03.001>. URL <https://www.sciencedirect.com/science/article/pii/S037859551630586X>.
- [22] Yi FanChiang, Cheng-Wen Wei, Yi-Le Meng, Yu-Wen Lin, Shyh-Jye Jou, and Tian-Sheuan Chang. Low complexity formant estimation adaptive feedback cancellation for hearing aids using pitch based processing. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(8):1248–1259, aug 2014. ISSN 2329-9290. doi: 10.1109/TASLP.2014.2327300. URL <https://doi-org.ezproxy.lib.vt.edu/10.1109/TASLP.2014.2327300>.
- [23] Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47(6):381–391, 1954.
- [24] D. Jake Follmer, Rayne A. Sperling, and Hoi K. Suen. The role of mturk in education research: Advantages, issues, and future directions. *Educational Researcher*, 46(6): 329–334, 2017. doi: 10.3102/0013189X17725519. URL <https://doi.org/10.3102/0013189X17725519>.
- [25] GitHub. Github pages. <https://pages.github.com/>, 2023. Accessed: 2023-11-27.
- [26] Kang Guangyu and Guo Shize. Improving amdf for pitch period detection. In *2009 9th International Conference on Electronic Measurement & Instruments*, pages 4–283–4–286, Aug 2009. doi: 10.1109/ICEMI.2009.5274087.
- [27] Mohamed Handosa, Archi Dasgupta, Mark Manuel, and Denis Gračanin. Rethinking user interaction with smart environments—a comparative study of interaction modalities. In Norbert Ztreitz and Shin’chi Konomi, editors, *Distributed, Ambient and Pervasive Interactions. HCII 2020*, volume 12203 of *Lecture Notes in Computer Science*, pages 39–57, Cham, 19–24 2020. Springer.

- [28] Susumu Harada, Jacob O. Wobbrock, Jonathan Malkin, Jeff A. Bilmes, and James A. Landay. Longitudinal study of people learning to use continuous voice-based cursor control. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 347–356, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605582467. doi: 10.1145/1518701.1518757. URL <https://doi.org/10.1145/1518701.1518757>.
- [29] Ramin Hedeshy, Chandan Kumar, Mike Lauer, and Steffen Staab. All birds must fly: The experience of multimodal hands-free gaming with gaze and nonverbal voice synchronization. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 278–287, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393904. doi: 10.1145/3536221.3556593. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/3536221.3556593>.
- [30] John Heerema and J. R. Parker. Music as a game controller. In *2013 IEEE International Games Innovation Conference (IGIC)*, pages 72–76, Sep. 2013. doi: 10.1109/IGIC.2013.6659147.
- [31] Chao-Ling Hsu, DeLiang Wang, and Jyh-Shing Roger Jang. A trend estimation algorithm for singing pitch detection in musical recordings. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 393–396, May 2011. doi: 10.1109/ICASSP.2011.5946423.
- [32] Li Hui, Bei-Qian Dai, and Lu Wei. A pitch detection algorithm based on amdf and acf. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I, May 2006. doi: 10.1109/ICASSP.2006.1660036.
- [33] Denis Jouviet and Yves Laprie. Performance analysis of several pitch detection algorithms on simulated and real noisy speech data. In *2017 25th European Signal Process-*

- ing Conference (EUSIPCO)*, pages 1614–1618, Aug 2017. doi: 10.23919/EUSIPCO.2017.8081482.
- [34] R.D. Kent. Vocal tract acoustics. *Journal of Voice*, 7(2):97–117, 1993. ISSN 0892-1997. doi: [https://doi.org/10.1016/S0892-1997\(05\)80339-X](https://doi.org/10.1016/S0892-1997(05)80339-X). URL <https://www.sciencedirect.com/science/article/pii/S089219970580339X>. The Voice Foundation’s 22nd Annual Symposium.
- [35] Hae Young Kim, Jae Sung Lee, Myung-Whun Sung, Kwang Hyun Kim, and Kwang Suk Park. Pitch detection with average magnitude difference function using adaptive threshold algorithm for estimating shimmer and jitter. In *Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vol.20 Biomedical Engineering Towards the Year 2000 and Beyond (Cat. No.98CH36286)*, volume 6, pages 3162–3164 vol.6, Nov 1998. doi: 10.1109/IEMBS.1998.746164.
- [36] Ji-Sun Kim, Denis Gračanin, Krešimir Matković, and Francis Quek. The effects of finger-walking in place (FWIP) for spatial knowledge acquisition in virtual environments. In Robyn Taylor, Pierre Boulanger, Antonio Kruger, and Patrick Olivier, editors, *Proceedings of the 10th Symposium on Smart Graphics (SG 2010)*, volume 6133 of *Lecture Notes in Computer Science*, pages 56–67, Berlin, 24–26 2010. Springer.
- [37] Karishma Kolhatkar, Mahesh Kolte, and Jyoti Lele. Implementation of pitch detection algorithms for pathological voices. In *2016 International Conference on Inventive Computation Technologies (ICICT)*, volume 1, pages 1–5, Aug 2016. doi: 10.1109/INVENTIVE.2016.7823210.
- [38] Sebastian Kraft and Udo Zölzer. Polyphonic pitch detection by matching spectral and autocorrelation peaks. In *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1301–1305, Aug 2015. doi: 10.1109/EUSIPCO.2015.7362594.

- [39] Sri H. Kurniawan and Adam J. Sporka. Vocal interaction. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '08, page 2407–2410, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605580128. doi: 10.1145/1358628.1358695. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1358628.1358695>.
- [40] Y.-H. Kwon, D.-J. Park, and B.-C. Ihm. Simplified pitch detection algorithm of mixed speech signals. In *2000 IEEE International Symposium on Circuits and Systems (IS-CAS)*, volume 3, pages 722–725 vol.3, May 2000. doi: 10.1109/ISCAS.2000.856162.
- [41] Everton B. Lacerda and Carlos A.B. Mello. Automatic classification of laryngeal mechanisms in singing based on the audio signal. *Procedia Computer Science*, 112: 2204–2212, 2017. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2017.08.115>. URL <https://www.sciencedirect.com/science/article/pii/S1877050917314710>. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 21st International Conference, KES-20176-8 September 2017, Marseille, France.
- [42] Daniel Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in Psychology*, 4, 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00863. URL <https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00863>.
- [43] J. Laver, S. Hiller, and R. Hanson. Comparative performance of pitch detection algorithms on dysphonic voices. In *ICASSP '82. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 7, pages 192–195, May 1982. doi: 10.1109/ICASSP.1982.1171723.
- [44] Rensis Likert. A technique for the measurement of attitudes. *Arch Psychology*, 22(140): 55, 1932.

- [45] Iker Luengo, Ibon Saratxaga, Eva Navas, Inmaculada Hernaez, Jon Sanchez, and Inaki Sainz. Evaluation of pitch detection algorithms under real conditions. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4, pages IV–1057–IV–1060, April 2007. doi: 10.1109/ICASSP.2007.367255.
- [46] Murni Mahmud, Adam J. Sporka, Sri H. Kurniawan, and Pavel Slavík. A comparative longitudinal study of non-verbal mouse pointer. In Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa, editors, *Human-Computer Interaction – INTERACT 2007*, pages 489–502, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-74800-7.
- [47] Diana Markova, Louis Richer, Melissa Pangelinan, Deborah H. Schwartz, Gabriel Leonard, Michel Perron, G.Bruce Pike, Suzanne Veillette, M. Mallar Chakravarty, Zdenka Pausova, and Tomáš Paus. Age- and sex-related variations in vocal-tract morphology and voice acoustics during adolescence. *Hormones and Behavior*, 81:84–96, 2016. ISSN 0018-506X. doi: <https://doi.org/10.1016/j.yhbeh.2016.03.001>. URL <https://www.sciencedirect.com/science/article/pii/S0018506X16301271>.
- [48] Ghulam Muhammad. Noise robust pitch detection based on extended amdf. In *2008 IEEE International Symposium on Signal Processing and Information Technology*, pages 133–138, Dec 2008. doi: 10.1109/ISSPIT.2008.4775670.
- [49] Kurnia Muludi and Aristoteles Aristoteles. Chord identification using pitch class profile method with fast fourier transform feature extraction. *International Journal of computer science issues*, (3), 2014.
- [50] Mozilla Developer Network. Web audio api. https://developer.mozilla.org/en-US/docs/Web/API/Web_Audio_API, . Accessed: 2023-11-28.

- [51] Mozilla Developer Network. Websockets - web apis | mdn. https://developer.mozilla.org/en-US/docs/Web/API/WebSockets_API, . Accessed: 2023-11-28.
- [52] Mozilla Developer Network. Web workers. https://developer.mozilla.org/en-US/docs/Web/API/Web_Workers_API/Using_web_workers, . Accessed: 2023-11-28.
- [53] A. Michael Noll. Cepstrum pitch determination. *The Journal of the Acoustical Society of America*, 41(2):293–309, 1967. doi: 10.1121/1.1910339. URL <https://doi.org/10.1121/1.1910339>.
- [54] M.S. Obaidat, C. Lee, B. Sadoun, and D. Nelson. Estimation of pitch period of speech signal using a new dyadic wavelet algorithm. *Information Sciences*, 119(1):21–39, 1999. ISSN 0020-0255. doi: [https://doi.org/10.1016/S0020-0255\(99\)00055-9](https://doi.org/10.1016/S0020-0255(99)00055-9). URL <https://www.sciencedirect.com/science/article/pii/S0020025599000559>.
- [55] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL <https://www.bmj.com/content/372/bmj.n71>.
- [56] David A. Patterson and John L. Hennessy. *Computer Organization and Design MIPS Edition: The Hardware/Software Interface*. Morgan Kaufmann, Santa Clara, CA, 6th edition, 2020. ISBN 9780128201091. URL <https://www.elsevier.com/books/computer-organization-and-design-mips-edition/patterson/978-0-12-820109-1>.

- [57] Ondrej Polacek, Zdenek Mikovec, Adam J. Sporka, and Pavel Slavik. Humsher: A predictive keyboard operated by humming. In *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '11, page 75–82, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450309202. doi: 10.1145/2049536.2049552. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/2049536.2049552>.
- [58] Ondřej Poláček, Adam J. Sporka, and Pavel Slavik. A comparative study of pitch-based gestures in nonverbal vocal interaction. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 42(6):1567–1571, Nov 2012. ISSN 1558-2426. doi: 10.1109/TSMCA.2012.2201937.
- [59] Ondřej Poláček and Zdeněk Míkovec. Hands free mouse: Comparative study on mouse clicks controlled by humming. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 3769–3774, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781605589305. doi: 10.1145/1753846.1754053. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1753846.1754053>.
- [60] Jikke Reinten, P. Ella Braat-Eggen, Maarten Hornikx, Helianthe S.M. Kort, and Armin Kohlrausch. The indoor sound environment and human task performance: A literature review on the role of room acoustics. *Building and Environment*, 123:315–332, 2017. ISSN 0360-1323. doi: <https://doi.org/10.1016/j.buildenv.2017.07.005>. URL <https://www.sciencedirect.com/science/article/pii/S0360132317302913>.
- [61] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley. Average magnitude difference function pitch extractor. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 22(5):353–362, October 1974. ISSN 0096-3518. doi: 10.1109/TASSP.1974.1162598.

- [62] Disha Sardana, Sampanna Yashwant Kahu, Denis Gračanin, and Krešimir Matković. Multi-modal data exploration in a mixed reality environment using coordinated multiple views. In Sakae Yamamoto and Hirohiko Mori, editors, *Human Interface and the Management of Information. Information Presentation and Visualization*, pages 337–356, Cham, 24–29 2021. Springer International Publishing.
- [63] Lukas Schickhofer and Mihai Mihaescu. Analysis of the aerodynamic sound of speech through static vocal tract models of various glottal shapes. *Journal of Biomechanics*, 99: 109484, 2020. ISSN 0021-9290. doi: <https://doi.org/10.1016/j.jbiomech.2019.109484>. URL <https://www.sciencedirect.com/science/article/pii/S0021929019307341>.
- [64] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4): 829–834, 1968. doi: 10.1121/1.1910902. URL <https://doi.org/10.1121/1.1910902>.
- [65] David A Schwartz and Dale Purves. Pitch is determined by naturally occurring periodic sounds. *Hearing Research*, 194(1):31–46, 2004. ISSN 0378-5955. doi: <https://doi.org/10.1016/j.heares.2004.01.019>. URL <https://www.sciencedirect.com/science/article/pii/S037859550400139X>.
- [66] C. Shahnaz, W. P. Zhu, and M. O. Ahmad. A pitch extraction algorithm in noise based on temporal and spectral representations. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4477–4480, March 2008. doi: 10.1109/ICASSP.2008.4518650.
- [67] Miloš Simonović, Marko Kovandžić, Ivan Ćirić, and Vlastimir Nikolić. Acoustic recognition of noise-like environmental sounds by using artificial neural network. *Expert Systems with Applications*, 184:115484, 2021. ISSN 0957-4174. doi: <https://doi.org/10.1016/j.eswa.2021.115484>.

- 1016/j.eswa.2021.115484. URL <https://www.sciencedirect.com/science/article/pii/S0957417421008952>.
- [68] M. Sondhi. New methods of pitch extraction. *IEEE Transactions on Audio and Electroacoustics*, 16(2):262–266, June 1968. ISSN 1558-2582. doi: 10.1109/TAU.1968.1161986.
- [69] Adam J. Sporka. Pitch in non-verbal vocal input. *SIGACCESS Access. Comput.*, (94):9–16, June 2009. ISSN 1558-2337. doi: 10.1145/1595061.1595063. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1595061.1595063>.
- [70] Adam J. Sporka and Pavel Slavík. Vocal control of a radio-controlled car. *SIGACCESS Access. Comput.*, (91):3–8, jun 2008. ISSN 1558-2337. doi: 10.1145/1394427.1394428. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1394427.1394428>.
- [71] Adam J. Sporka, Sri H. Kurniawan, and Pavel Slavík. Acoustic control of mouse pointer. *Universal Access in the Information Society*, 4(3):237–245, Mar 2006. ISSN 1615-5297. doi: 10.1007/s10209-005-0010-z. URL <https://doi.org/10.1007/s10209-005-0010-z>.
- [72] Adam J. Sporka, Susumu Harada, and Sri H. Kurniawan. Striking a c[h]ord: Vocal interaction in assistive technologies, games, and more. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '07, page 2869–2872, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595936424. doi: 10.1145/1240866.1241098. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1240866.1241098>.
- [73] Adam J. Sporka, Torsten Felzer, Sri H. Kurniawan, Ondřej Poláček, Paul Haiduk, and I. Scott MacKenzie. Chanti: Predictive text entry using non-verbal vocal input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 2463–2472, New York, NY, USA, 2011. Association for

- Computing Machinery. ISBN 9781450302289. doi: 10.1145/1978942.1979302. URL <https://doi-org.ezproxy.lib.vt.edu/10.1145/1978942.1979302>.
- [74] Arend Sulter, Harm Schutte, and Donald Miller. *Standardized Laryngeal Videostroboscopic Rating: Differences between untrained and trained male and female subjects, and effects of varying sound intensity, fundamental frequency and age*. 1996.
- [75] Arend Sulter, Hero Wit, Harm Schutte, and Donald Miller. *A structured approach to voice range profile (phonetogram) analysis*. 1996.
- [76] Stephanie Watson. The unheard female voice. *ASHA Lead.*, 24(2):44–53, February 2019.
- [77] G. White. Dynamic programming, the viterbi algorithm, and low cost speech recognition. In *ICASSP '78. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 413–417, 1978. doi: 10.1109/ICASSP.1978.1170385.
- [78] Karlton Wirsing. Time frequency analysis of wavelet and fourier transform. In Somayeh Mohammady, editor, *Wavelet Theory*, chapter 1. IntechOpen, Rijeka, 2020. doi: 10.5772/intechopen.94521. URL <https://doi.org/10.5772/intechopen.94521>.
- [79] In-Chul Yoo, Hyeontaek Lim, and Dongsuk Yook. Formant-based robust voice activity detection. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 23(12):2238–2245, dec 2015. ISSN 2329-9290. doi: 10.1109/TASLP.2015.2476762. URL <https://doi-org.ezproxy.lib.vt.edu/10.1109/TASLP.2015.2476762>.
- [80] Sang-Hyu Yoon, Kwang-Bock You, and Myung-Jin Bae. Experiments on the acoustics of whistling. *International Journal of Engineering Research and Technology*, 12(3): 378–384, 2019.

- [81] T. Yoshinaga, A. Van Hirtum, and S. Wada. Multimodal modeling and validation of simplified vocal tract acoustics for sibilant /s/. *Journal of Sound and Vibration*, 411: 247–259, 2017. ISSN 0022-460X. doi: <https://doi.org/10.1016/j.jsv.2017.09.004>. URL <https://www.sciencedirect.com/science/article/pii/S0022460X17306673>.
- [82] Daniel Zielasko, Neha Neha, Benjamin Weyers, and Torsten W. Kuhlen. Blowclick 2.0: A trigger based on non-verbal vocal input. In *2017 IEEE Virtual Reality (VR)*, pages 319–320, March 2017. doi: 10.1109/VR.2017.7892305.

Appendices

Appendix A

IRB Approval Letter



Division of Scholarly Integrity and
Research Compliance
Institutional Review Board
North End Center, Suite 4120 (MC 0497)
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-3732
irb@vt.edu
<http://www.research.vt.edu/sirc/hrpp>

MEMORANDUM

DATE: October 27, 2023
TO: Denis Gracanin, Sam Williams
FROM: Virginia Tech Institutional Review Board (FWA00000572)
PROTOCOL TITLE: The Usability of a Pitch-based Non-verbal Vocal Interaction Prototype and a Comparison with the Mouse
IRB NUMBER: 23-1098

Effective October 27, 2023, the Virginia Tech Human Research Protection Program (HRPP) determined that this protocol meets the criteria for exemption from IRB review under 45 CFR 46.104 (d) category(ies) 2(ii),3(i)(B).

Ongoing IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities impact the exempt determination, please submit an amendment to the HRPP for a determination.

This exempt determination does not apply to any collaborating institution(s). The Virginia Tech HRPP and IRB cannot provide an exemption that overrides the jurisdiction of a local IRB or other institutional mechanism for determining exemptions.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

<https://secure.research.vt.edu/external/irb/responsibilities.htm>

(Please review responsibilities before beginning your research.)

PROTOCOL INFORMATION:

Determined As: **Exempt, under 45 CFR 46.104(d) category(ies) 2(ii),3(i)(B)**
Protocol Determination Date: **October 27, 2023**

ASSOCIATED FUNDING:

The table on the following page indicates whether grant proposals are related to this protocol, and which of the listed proposals, if any, have been compared to this protocol, if required.