



WICHITA STATE
UNIVERSITY

UNIVERSITY LIBRARIES

**Tactile-speech perception with gamified
training on the vibey transcribey**

Item Type	Dissertation
Authors	Canare, Dominic
Publisher	Wichita State University
Rights	© Copyright 2022 by Dominic Canare All Rights Reserved
Download date	2026-05-18 10:23:01
Link to Item	https://soar.wichita.edu/handle/10057/24970

TACTILE-SPEECH PERCEPTION WITH GAMIFIED TRAINING ON THE VIBEY TRANSCRIBER

A Dissertation by

Dominic Canare

Master of Arts, Wichita State University, 2017

Master of Science, Wichita State University, 2008

Bachelor of Science, Pittsburg State University, 2006

Submitted to the Department of Psychology
and the faculty of the Graduate School of
Wichita State University
in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

December 2022

© Copyright 2022 by Dominic Canare

All Rights Reserved

TACTILE-SPEECH PERCEPTION WITH GAMIFIED TRAINING ON THE VIBEY TRANSCRIBER

The following faculty members have examined the final copy of this dissertation for form and content, and recommend that it be accepted in partial fulfillment of the requirement for the degree of Doctor of Philosophy with a major in Psychology.

Rui Ni, Committee Chair

Quan Lei, Committee Member

Carryl Baldwin, Committee Member

C. Brendan Clark, Committee Member

Vinod Namboodiri, Committee Member

Cynthia Richburg, Committee Member

Accepted for the College of Liberal Arts and Sciences

Andrew Hippisley, Dean

Accepted for the Graduate School

Coleen Pugh, Dean

DEDICATION

To the late Dr. Duncan Rose for showing me what it meant to a modern, educated, caring person during
an important time in my life.

ACKNOWLEDGEMENTS

This work would not have been possible without so many wonderful people. I want to recognize Sci-Hub's founder Alexandra Elbakyan as well as the late Aaron Swartz for their dedication to accessible science. I want to thank the Drs. Barbara and Alex Chaparro for welcoming me into the field of Human Factors, Dr. Rhonda Lewis for her unwavering and enthusiastic support, Dr. Vinod Namboodiri for inspiring the work that became this project, and Dr. Rui Ni for his confidence and faith in me. I feel so fortunate to have been a part of the psychology department and graduate school at Wichita State, where I can count so many friends among the faculty, staff, and fellow graduate students. I will forever value the relationships I have made here.

Lastly, I'd like to express immeasurable gratitude to my partner, Roz. Her love, humor, companionship, and guidance have been a stable comfort throughout this journey of triumphs and failures.

ABSTRACT

Access to information is indispensable in this modern age, and more inclusive methods for communication should be explored so that all people may benefit from this information. Tactile communication systems have many potential applications in a variety of conditions and scenarios. Previous tactile-speech systems convey only phonemic or prosodic elements of speech, but both are important. In the current study, a novel method of transforming speech into vibrotactile sensations by encoding and displaying both segmental and suprasegmental features was implemented with a low-cost, sleeve-worn device. An experimental group of participants ($n = 6$) learned to differentiate tactile words by playing training games in their home while a control group ($n = 6$) did not. Although the games improved participants' ability to differentiate tactile words during training, results from tests taken before and after training indicated that participants did not improve in their ability to identify tactile phonemes, or their ability to perceive tactile prosody, and did not experience an increase in tactile-auditory sensory integration.

Achieving tactile-speech perception remains an ambitious goal with numerous parameters. This configuration of encoding dimensions, stimulation method, body site, and training techniques did not demonstrate tactile-speech perception, but the ideas and materials used here may be expanded and/or altered for future work. These materials are made available on the Open Science Framework (<https://osf.io/3f5nu/>), which includes the encoding scheme design, an implementation of that design, electronic schematics, firmware, 3D models, stimuli, training game source code and assets, evaluation software, and analysis software. Future studies may wish to employ these materials as the landscape of tactile communication merits further exploration to advance inclusivity.

TABLE OF CONTENTS

Chapter	Page
INTRODUCTION.....	1
1.1 Motivation	1
1.1.1 Universal Design	1
1.1.2 Multiple Resource Theory.....	3
1.1.3 Sensory Substitution	4
1.1.4 The Utility of a Tactile-Speech Display.....	6
1.2 Past Work.....	7
1.3 Designing a solution.....	8
1.3.1 Speech Perception	8
1.3.2 Tactile Perception	16
1.3.3 Research questions.....	30
METHOD.....	32
2.1 Materials	32
2.1.1 Hardware	32
2.1.2 Phonemic feature mapping	38
2.1.3 Prosodic feature mapping.....	38
2.1.4 Stimuli	39
2.1.5 Other Software and Data.....	41
2.2 Participants	41
2.3 Procedure.....	42
2.3.1 Evaluation	42
2.3.2 Training	46
RESULTS	59
3.1 Training	59
3.2 Phoneme Perception	61
3.3 Prosody Perception.....	63
3.3.1 Word Focus Matching.....	63
3.3.2 Phrase Boundary Matching.....	65
3.4 Perceptual Integration	67
DISCUSSION.....	70
4.1 Limitations	72
4.2 Future directions.....	75
4.3 Conclusion.....	76
REFERENCES.....	77

LIST OF TABLES

Table	Page
1. Mechanoreceptors.....	17
2. Vibrotactile actuators	21
3. Psychological Needs for Gamification	30
4. McGurk stimuli and options	41
5. Individual training reports	59

LIST OF FIGURES

Figure	Page
1. Multiple resource theory model	4
2. Equal loudness contours	12
3. Veridical and illusory tactile stimuli sites	19
4. Somatosensory nerve density	22
5. Somatosensory homunculus	22
6. Temporal discrimination thresholds across the body	23
7. Encoding flow and tactile arrays	33
8. Custom electronics to drive LRA arrays	34
9. Modelling a life-size jig for equidistant actuator spacing	35
10. An assortment of 3D printed jigs and electronics cases	36
11. 3D printed electronics and actuator housings	36
12. The Vibey Transcribey device as worn	37
13. Device display states	37
14. IPA Vowel Chart	38
15. Phoneme identification task screenshot	43
16. Focus discrimination task screenshot	44
17. Phrase boundary discrimination task screenshot	45
18. Perceptual integration task screenshot	46
19. Device check screen	47
20. On-screen device status indicator	47
21. Training software main menu	48
22. Leaderboard screen	48

LIST OF FIGURES (continued)

23.	Display settings	49
24.	Advanced device settings and troubleshooting screen	49
25.	Information screen	50
26.	Category selection screen	51
27.	Level select screen	51
28.	Pre-game and pause menu screen	52
29.	Level complete screen	53
30.	Secret Shopper screenshot	54
31.	Soop Loops screenshot	55
32.	Pin Pals screenshot	56
33.	Snack-a-mole screenshot	57
34.	Prevalence of phonemes during training versus conversation	60
35.	Training performance progression	61
36.	Phoneme identification before and after training	62
37.	Training effect on post-test phoneme identification	63
38.	Word focus matching before and after training	64
39.	Training effect on post-test focus matching	65
40.	Phrase boundary matching before and after training	66
41.	Training effect on post-test phrase boundary matching	67
42.	Perceptual integration before and after training	68
43.	Training effect on post-test perceptual integration	69
44.	A Vibey Transcribey with a modified pitch array band	74

CHAPTER 1

INTRODUCTION

This manuscript presents a novel method for, and implementation of, encoding tactile speech. The benefits of inclusive design are presented first, followed by a review of previous works in this domain, considerations for a tactile-speech device, and necessary tests for evaluating one. The following sections describe the details and development of materials for the current study as well as the procedure for evaluating the device and training, the results of those evaluations, and a discussion of how this body of work may inform future work.

1.1 Motivation

1.1.1 Universal Design

In this modern era of technology and resources, the design and creation of human-made environments, systems, and structures should provide access for every person as a minimum endeavor. Globally, 3.4% of the world's population has a moderate to severe visual impairment (Bourne et al., 2017), and 5.5% of the world's population has moderate or higher levels of hearing loss (World Health Organization, 2021b). The World Health Organization estimates that over a billion people "live with some form of disability, of whom nearly 200 million experience considerable difficulties in functioning". What's more, these types of challenges tend to be more common among women, older people, and households that are poor. Low-income countries tend to have a higher prevalence of disabilities too. Further, people with disabilities have poorer health, attain fewer academic achievements, have less economic participation, and experience higher rates of poverty (World Health Organization, 2011). Disability rights are human rights, and all humans are entitled to the same inherent dignity, autonomy, and full and effective participation and inclusion in society without discrimination (Convention on the Rights of Persons with Disabilities, 2006).

Human rights and basic dignities aside, excluding people with impairments or otherwise hindering their full participation in society is expensive. The annual cost of productivity losses due to blindness and moderate to severe visual impairments in the US alone is conservatively estimated to be \$7.8 billion USD (Eckert et al., 2015). Global productivity losses from just two visual conditions left untreated, myopia and presbyopia, cost \$269.4 billion USD annually (World Health Organization, 2021a). Hearing loss, which often forces people into unemployment or early retirement, has a global annual cost of \$105 billion USD, the majority of which is incurred outside high-income countries (World Health Organization, 2017).

These estimates are challenging to make and cannot account for everything. Perhaps one overlooked cost is the opportunity-cost of innovations failed to be realized. Examples of non-existent innovations are difficult to conjure, but when people with impairments are not excluded, all people stand to benefit. Audio books, for example, can be absolute necessities for some with visual impairments, but their utility is not lost for those without any impairments. Automatic doors can be an incredible benefit to those with limited mobility but also to those who simply have their arms full. The electric toothbrush was invented for patients with impaired motor skills, but they have become more effective than manual toothbrushes (Yaacob et al., 2014). It took 88 years after the first voice phone call for the teletypewriter to facilitate the first long-distance phone call between two people who were deaf, but today text messaging is ubiquitous among people of all hearing-abilities and is often preferred over voice communications. These innovations benefit everybody but came to exist only because of a diverse population which includes those with impairments.

People with impairments must not be excluded from accessing the world fellow humans create. They have a right to participate fully in society, excluding them is massively expensive, and all people stand to benefit from innovations that result from inclusive design. The impetus falls to architects, engineers, designers, and makers to be as inclusive as possible in their work. In the domain of

communications, information available to more perceptual systems can help reach a wider audience and improve accuracy and efficiency, as explained in the following section.

1.1.2 Multiple Resource Theory

Designing information and communication systems often drives an exploration of the array of human perceptual abilities. While some types of information are best suited for a particular perceptual channel (hearing music versus reading the score, seeing a portrait versus hearing a description of it, etc.), others can be translated for display in multiple perceptual channels, which would be obviously beneficial when accommodating the large ranges of perceptual abilities. A system alert, for example, may be presented as a flashing light, an audible tone, and a vibration simultaneously. Naturally, there are limits of these perceptual systems, and understanding how they work with and against each other can inform designs that benefit all users.

Researchers investigating multitasking have developed a model that accounts for performance decrements under specific conditions. This is known as multiple resource theory (Figure 1), which describes a finite set of available resources across different stages of processing (Wickens, 2002). The stages of processing are serial, starting with perception/cognition and ending in a response. In the perception/cognition stage, a subject attends to sensory inputs across different modalities (visual or auditory). The model proposes that visual and auditory modalities draw from different pools of resources. Those inputs contain information that is either spatially or verbally coded. Separate pools of resources are also available for information that is coded spatially versus verbally. The data support this model's prediction that two simultaneous tasks are more likely to interfere with each other (causing decrements in performance) if the demand for any of these shared resources exceeds what is available (Wickens, 1980, 2002). The same is true in the response stage, where the type of response may be manual or vocal and convey information that is either spatial or verbally coded. Dual-task performance is less likely to decrease if this interference is avoided. This can be accomplished by designing the tasks

utilizing different resources during the same stages. For example, making verbally coded information available in the tactile modality would provide an additional, otherwise underutilized tactile input channel (potentially with its own resource pool). Performance in these scenarios may be comparable to auditory signaling alone (Lu et al., 2011).

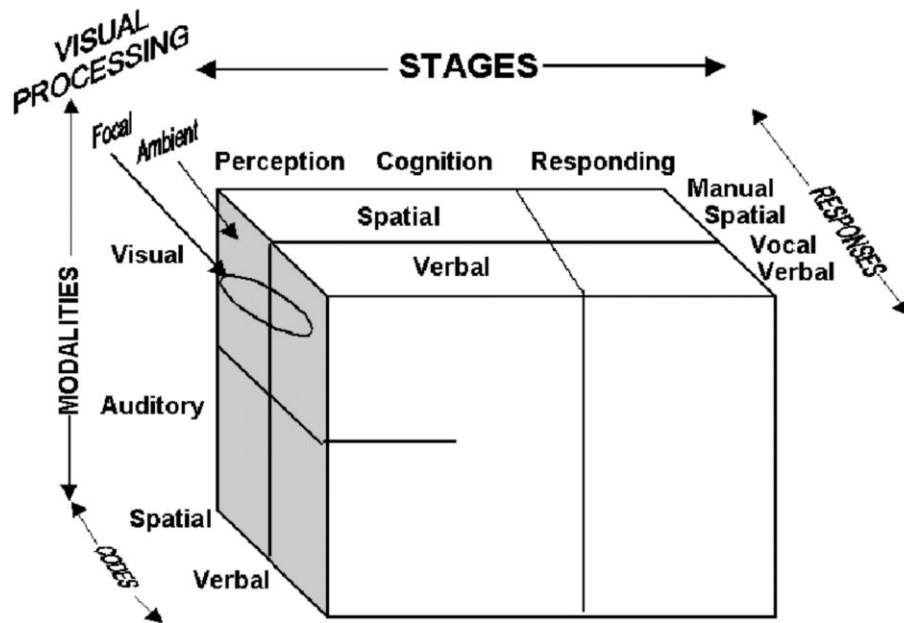


Figure 1. Multiple resource theory model.

1.1.3 Sensory Substitution

One mechanism to capitalize on the benefits gained from distributing resources across resource channels is sensory substitution. The initial idea of sensory substitution proposed that rendering visual information on the tactile sensory system allows the brain to learn how to "see" through the skin (Bach-y-Rita, 1967). In a first-of-its-kind experiment, Bach-y-Rita modified a dental chair to include an array of vibrating actuators which could be individually activated and controlled. The actuator array was connected to a remote video feed, and the image from the video feed was converted into actuations of the array which was positioned at the subjects' lower backs. After training, these visually blind participants were able to correctly identify objects in the video with only the vibration information of the actuator array. Since his subjects were interpreting this information in a visual way, Bach-y-Rita

believed that signals sent from the skin to the brain were being processed by the visual cortex (Bach-y-Rita et al., 1969).

At the time of Bach-y-Rita's pioneering work in sensory substitution, the scientific community was only beginning to understand and accept the adult human brain's plasticity—that is, the brain's ability to modify neural pathways and synaptic behavior to accommodate new information and knowledge (Pascual-Leone et al., 2005). Bach-y-Rita cited the ability to recover motor and sensory functions when related parts of the brain were damaged or removed as evidence of the plasticity of the adult brain and rationale for the feasibility of a sensory substitution device (Bach-y-Rita, 1967).

Bach-y-Rita demonstrated a successful attempt of mapping visual information onto tactile sensations—a visual-tactile substitution. Since then, a variety of different modality mappings have been examined, including tactile-vision (Bach-y-Rita et al., 1969; Zelek et al., 2003), auditory-vision (Brown et al., 2014; Moraru & Boiangiu, 2015; Striem-Amit et al., 2012), tactile-vestibular (Tyler et al., 2003), and tactile-auditory (Gescheider, 1965; Richardson & Frost, 1977; Rizza et al., 2018; Weisenberger & Percy, 1995; Yuan et al., 2005). Other researchers have looked beyond the natural abilities of the human sensory experiences. In "sensory augmentation", information which is not naturally perceivable by humans is mapped to a perceptual modality through the aid of technology. One example is magnetoreception, wherein the human perceptual experience is expanded to detect magnetic North through a tactile signal (Pielot et al., 2010; Schumann & O'Regan, 2017).

Sensory-substitution systems, in their purest form, translate and display a distal stimulus to a different sensory channel. These systems rely entirely on neural mechanisms to decode the information. The current study explores an alternative approach wherein some processing is performed on the original stimulus and a higher-order stimulus is presented to the user. Specifically, components of speech, rather than the acoustic signal itself, are presented to the user.

1.1.4 The Utility of a Tactile-Speech Display

Humans, generally, are privileged to have many ways of communicating information to each other. Among the richest of those, spoken language is acquired naturally and is arguably innate. Modern speech synthesizers make it easy to convert printed or digital text into speech, which enables more universal access to information (although much of the expression is lost). Unlike written languages that make use of a variety of different alphabets, the construction units of spoken languages, phonemes, are universal and finite, and the prosodic features of speech allow for an infinite range of expression. However, speech is not always a viable medium. Even observers with typical hearing may not be able to perceive speech correctly in noisy environments, in a discreet conversation, or when the auditory channel is occupied. These challenges are more pronounced for deaf people and the hard-of-hearing. In these situations, substituting auditory speech to a different sensory modality may enable access, facilitate environmental demands, accommodate personal or situational requirements, and even improve multitasking performance. Converting speech to the visual modality can be accomplished with printed text, but much of the expression is lost and there are no satisfactory translations for other sensory modalities. Braille, for example, is extremely verbose, expensive, rare, and cumbersome. Electronic refreshable Braille displays are prohibitively expensive. Reading Braille requires the use of one or both hands, such that the reader is limited in what they can do simultaneously.

A tactile-speech display can provide greater access to text-based information, and, if done correctly, can convey the same richness of expression that spoken word does. A more generalized sound-to-touch system could provide even greater utility, but as discussed later, the complexities of speech and limits of the somatosensory system make it difficult for such a device to be successfully used to interpret speech.

1.2 Past Work

Tactile languages have existed in various forms for some time. Braille, the prevailing tactile writing system, was first published in 1829 and based on a system developed by Charles Barbier de la Serre, who sought a writing method that could produce multiple copies simultaneously and be written and read silently and in the dark (Jiménez et al., 2009). Signed languages provide another example, despite being typically visual. One can learn to interpret signs when they are performed in the receiver's hand, as made famous by deaf-blind icon, Helen Keller. Another method for communication with deaf-blind people is Tadoma, where the deaf-blind person places a hand on the face of a speaker such that they can feel the motion of the lips and jaw, airflow, and vibrations on the speaker's throat (Reed et al., 1985).

These examples demonstrate some human capacity for tactile language and have helped motivate various attempts to augment or replace auditory speech perception with tactile cues and displays. Some of these attempts have been simple, like having "listeners" place their hand at the end of a long tube opposite (and out of audible range) of a person speaking into it (Gault, 1924; Gault & Crane, 1928) or simply hold an electronic speaker playing speech (Gault, 1927). But the human somatosensory system did not evolve in such a way that it could parse sounds as accurately, precisely, or quickly as the auditory system, so decoding speech from the original acoustic signal is difficult if not impossible. For this reason, other methods have sought to simplify or modify the signal to be more compatible with the somatosensory system. The simplest examples of these typically convey information about the amplitude envelope and little or nothing else, like the Minifonator and Minivib (Szeto & Christensen, 1988; Weisenberger & Russell, 1989). More complex examples isolate various parts of the speech signal for tactile display, perhaps inspired by a set of experiments by G. v. Békésy which compared perceptual phenomena facilitated by the organ of Corti to simulated sensations on the skin (v. Békésy, 1957). Some of these "tactile hearing aids", like the Tactaid II, the Tactaid 7, the Tacticon 1600, the DigiVoc, the Tickle

Talker, and the Queen's University vibrotactile vocoder build on this comparison using pitch-to-place displays (Blamey & Clark, 1985; Brooks & Frost, 1983; Eilers Rebecca E. et al., 1996; Saunders & Franklin, 1985). The acoustic signal is decomposed into constituent frequencies, each of which is assigned to a different place on the skin tonotopically, much like the organ of Corti. The most complex systems reencode speech in novel ways like "vibratese"—a tactile alphabet that distinguishes symbols (letters, common short words, and numerals) in a five-vibrator display by varying in the location, duration, and intensity of the vibrations (Geldard, 1957). Some researchers have developed systems that display the phonemic symbols that make up speech, rather than the acoustic properties (Ellis & Robinson, 1993; Reed et al., 2019; Rizza et al., 2018; Zhao et al., 2018). The VEST system borrows techniques from data compression methods to encode sound for tactile consumption (Novich, 2015). Although not evaluating the perception of any prosodic features of speech, one study found that participants using a symbolic/phonemic display significantly outperformed those using a device displaying acoustic signals in a word recognition task (Turcott et al., 2018).

1.3 Designing a solution

Some important decisions in the design of a new communication display include what features/components of the information will be displayed, how it will be displayed, how the device interfaces with its users, and how users will learn to interpret the display.

1.3.1 Speech Perception

A good display needs to balance information load, and this starts with feature selection and display format. First, it must convey enough information such that the receiver can correctly interpret it. Second, the display must be appropriate for the sensory channel without exceeding the limits of that sensory channel. Because speech perception is a subset of audition, it is more efficient to filter out the non-verbal information in the pure acoustic signal. More precisely, with an understanding of the speech perception process a tactile-speech perception system can limit its presentation to only the most

informative features of the acoustic signal. Likewise, complex transformations carried out by neural processes in the auditory cortex can be off-loaded to the device before tactile display.

1.3.1.1 Speech Components

The auditory signal produced by speech is complex but can be better understood by first understanding the building blocks of speech. These building blocks are separated into two categories: segmental components and suprasegmental components.

Segmental components of speech, known as phones or phonemes, are the smallest distinguishable units in a language. Phonemes are chained temporally to construct syllables and then words. Individual consonants and vowel sounds are simple examples of phonemes, but consonant and vowel sounds do not correspond one-to-one with the letters of a language's written alphabet. For example, the English letters "s" and "h" each represent two unique phonemes when pronounced, but "sh" produces a third, unique phoneme when pronounced. Likewise, the letter "c" can be pronounced as a "k" or as an "s", and vowel letters have multiple pronunciations. The phonemes used in any specific language is a subset of all phonemes known to exist in natural human languages. The International Phonetic Alphabet (IPA) has defined over 150 different phonemes, but American English, the native language of the author and focus of the current study, only uses about 45 of those—an imprecise number due to regional differences in speech and the inclusion or exclusion of borrowed words.

The suprasegmental components of speech, also known as "prosody", encompasses speech features which are not segmental—that is, everything about the conveyance of speech aside from the phonemes, like intonation, tone, stress, rhythm, etc. These prosodic features have a role in organizing speech recognition, as demonstrated by the perceptual advantage under prosodic presentation. For example, participants more easily recall nonsense syllables when presented with sentence morphology (Epstein, 1961), but when the words are presented vocally, the benefit is only realized when spoken with sentence prosody (Leonard, 1973; O'Connell et al., 1968). A similar benefit is realized on the

shadowing of grammatical strings versus non-grammatical ones—but only when they have sentence prosody (Martin, 1968). Prosodic cues for continuity are prioritized over semantic continuity (Darwin, 1975). There are several other lines of research which demonstrate how prosodic features facilitate speech processing (see Cutler et al., 1997 for a review).

Prosody undoubtedly plays an important role in the human perception of speech. Compared to segmental features and syllable structure, prosody has been found to be the most closely related to intelligibility (Anderson-Hsieh et al., 1992). This is further supported by evidence showing that differences in prosody have a larger effect on comprehension than differences in segments (Anderson-Hsieh & Koehler, 1988). It may then come as no surprise that prosody also has a larger effect on comprehension than segments (Derwing et al., 1998). Indeed, rhythm, stress, and intonation patterns are significant contributors to speech recognition (Grant et al., 1985), and any attempt to substitute speech perception to another sensory experience is incomplete without including prosodic elements.

Among those prosodic elements, linguistic focus and phrase boundaries are of particular interest. Focus cues listeners' attention to important words or parts of a sentence that carry more weight than normal, while phrase boundaries provide cues about sentence structure and inform expectations on otherwise ambiguous elements. Increases in loudness, increases in duration, and pitch excursions cue listeners to prosodic focus (Fry, 1958) and lexical stress (Chrabaszcz et al., 2014; Mattys, 2000). Similarly, phrase boundaries are cued by lengthening the syllable just before the boundary and changes to pitch (Choi et al., 2005; de Pijper & Sanderman, 1994; Lehiste et al., 1976).

1.3.1.2 Acoustic Dimensions of Speech

The acoustic signal generated by speech production manifests separate perceptual properties of sound: pitch, loudness, and speed. Each of these are encoded differently and affect speech perception in different ways.

Pitch and frequency are often used interchangeably, and although they are closely related, there is a distinction to be made. Frequency is an empirical, physical property of the sound wave, whereas pitch is an individual's subjective perception of the sound wave, which cannot be directly measured (Yantis & Abrams, 2014). This distinction becomes important when deciding how one might substitute a speech signal from an auditory signal to a tactile one. A Fourier transformation on the acoustic signal can determine and order dominant frequencies, but these frequencies may not match up to the listener's perception of the pitch. Different computational methods have been proposed and implemented to estimate pitch, and ongoing research in this domain produces increasingly accurate and efficient methods. One such method is a fully-convolutional neural network trained on high-quality synthesized speech with known ground-truth pitches (Ardaillon & Roebel, 2019). This method consistently provided more accurate pitch estimations for both synthetic and natural speech versus competing algorithms.

Loudness, like pitch, is another perceptual experience often used interchangeably with a physical counterpart: sound pressure (and its waveform correlate: amplitude). Intensity and power are also often used to describe the loudness of sound; while related to loudness, these are measurable physical attributes that are distinct from loudness, a perceptual experience that cannot be measured directly. Again, this distinction is important when transforming an acoustic speech signal to a non-auditory sensory modality. Loudness estimation is complicated by the fact that, besides being subjective, is a function of both the sound pressure level and the frequency of the sound. Larger sound pressure levels will result in louder sounds, but the effect of frequency on loudness is not monotonic (see **Error! Reference source not found.**). Additionally, spectral masking and temporal influences can affect the perception of loudness (Zwicker & Fastl, 2013). Broadcast media and other forms of entertainment have helped drive the standardization of methods to measure loudness. One such

method is known as the "Zwicker method" and is defined in DIN 45631/A1:2010 and ISO 532-1:2017 (Zwicker et al., 1991; Zwicker & Fastl, 2013).

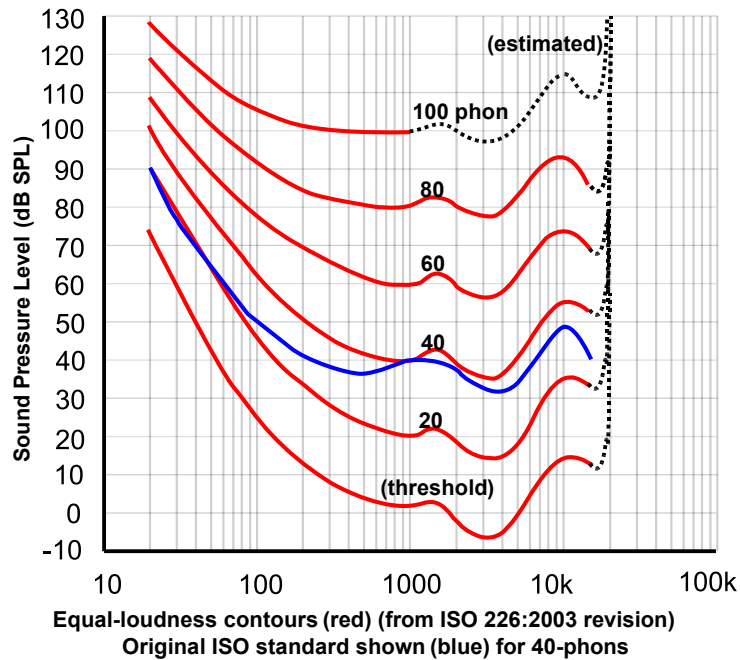


Figure 2. Equal loudness contours.

1.3.1.3 Auditory Transduction

Speech perception starts, primarily, as an auditory experience. Sound manifests as compression waves of air traveling from the speaker's mouth to the listener's ears. These waves are funneled by the outer ear, travel through the ear canal, and strike the tympanic membrane, more commonly known as the "eardrum", at the middle ear. The middle ear includes three bones—the malleus, incus, and stapes—referred to collectively as the ossicles. As the tympanic membrane vibrates sympathetically with the compression waves, these vibrations are mechanically transmitted and amplified through the ossicles to the oval window of the inner ear. As the oval window vibrates with the sound, it moves fluid within the cochlea—a spiral-shaped bony structure of the inner ear. As these vibrations traverse the cochlear fluid, the motion is transmitted to hair cells on the organ of Corti, positioned on the basilar membrane which runs the length of the cochlea. The mechanical motion of these hair cells as they press

against the tectorial membrane generates electrical impulses which travel down the auditory nerve to the primary auditory cortex in the temporal lobe of the brain (Yantis & Abrams, 2014).

The physical composition of the cochlea produces a tonotopic mapping. The highest-perceivable sound frequencies vibrate hair cells at the base of the cochlea. Decreasing frequencies vibrate hair cells further along the cochlea, with the lowest-perceivable frequencies vibrating hair cells at the apex of the structure. This tonotopic mapping is preserved along the auditory nerve and in the primary auditory cortex (Yantis & Abrams, 2014).

1.3.1.4 Theories of Speech Perception

While the transduction of sound waves in the auditory system is well-understood, the way the composition of those signals is eventually perceived as speech is less clear. It is understood that segmental aspects of speech perception are categorical: the acoustic features of speech vary along a continuum, but the perception of phonemes is not. While working on speech synthesis in the 1950s, researchers examined how the modification of acoustic features affected speech perception and showed that varying these features along their continuums still generated discrete percepts (Liberman, 1957). This led to the motor theory of speech perception. It has evolved over time with criticism and discovery, but maintains the position that speech perception is the perception of vocal tract gestures and involves access to the speech motor system (Galantucci et al., 2006). Put simply, the acoustic features of speech are the mere end-result of neuromotor intentions—and it is those intentions that are perceived when speech is heard, according to the motor theory (Diehl et al., 2004; Liberman et al., 1967). This theory is supported by a robust phenomenon known as the McGurk Effect (McGurk & MacDonald, 1976), wherein sensory information from the visual channel modifies speech perception, which was otherwise assumed to be strictly informed by the auditory channel. When an auditory signal of one consonant is synchronized with the video signal of a different consonant, humans often report having heard a third consonant, distinct from any that were displayed to them. For example, an auditory

/ba/, which is articulated at the lips (bilabial), synchronized with a visual /ga/, which is articulated at the back of the throat (velar), will often be perceived as a /da/, which would have been articulated at the center of the mouth (alveolar).

Motor theory is also supported by research on the functional activity of the nervous system. Functional imaging of the brain shows activation in speech-production areas (Wilson et al., 2004), and motor-evoked potentials measured around the lips are enhanced when listening to speech or lip-reading (Watkins et al., 2003). Similarly, transcranial magnetic stimulation applied to the pre-motor cortex disrupts the perception of plosives (Meister et al., 2007). Of course, there are valid criticisms of the motor theory. Under motor theory, speech perception should develop after speech production does, due to the dependency. However, infants show surprisingly early speech perception abilities (Eimas et al., 1971) and that early speech perception ability actually predicts later production ability (Tsao et al., 2004). Motor theory also fails to account for influences on perception from non-articulatory sources. For example, the context around which an utterance is observed can alter one's perception of it (Massaro & Palmer, 1998), but semantic information is not coded in the motor cortex. There are numerous other criticisms of motor theory (Lane, 1965; Massaro & Palmer, 1998; Ohala, 1996).

One challenge to all models and theories of speech perception is the lack of invariance in speech segmentation. The exact pronunciation and resulting acoustic signal of any particular phoneme can vary dramatically, but a listener will reliably perceive that phoneme, making it extremely difficult to uniquely identify phonemes based solely on the acoustic signal (Liberman et al., 1967). The acoustic manifestations of phonemes depend on speaker identity (individual differences), context (preceding and following phonemes), speech rate, and other prosodic features. This invariance in perception is a motivating factor for motor theory to seek an articulatory description, rather than an acoustic one, explaining that adjacent phonemes are merged or assimilated (Diehl et al., 2004).

The motor theory partly explains how individual phonemes are identified and distinguished. The cohort model of speech perception provides an explanation for how percepts are mapped to words and is supported by shadowing experiments, where participants repeat words they hear and are able to begin repeating a word before hearing it entirely (Marslen-Wilson & Welsh, 1978). Within this model, phonemes are processed linearly, like an assembly line. Once the first phoneme of a word is identified, an initial group of candidate words (the "cohort") is compiled. As each successive phoneme is identified, words that do not match the chain of perceived phonemes are dropped from the cohort until only one word, the perceived word, remains. The TRACE model of speech perception works in a similarly temporally-sequential way but works across multiple levels. As percepts are processed, cohorts among auditory features are pruned and selections on this level inform cohort pruning among phonemes which, in turn, informs cohort pruning among words. Connections between levels are bi-directional, such that word selection can inform phoneme selection, and it can inform auditory feature selection (McClelland & Elman, 1986).

1.3.1.5 Feature Selection

Related work with tactile-speech systems have generally taken one of two approaches in deciding what features from the speech stream to present to the user's skin: waveform functions or symbolic displays. Commercial devices have typically used the former: some function of the acoustic waveform which conveys only the amplitude envelope, the power in banks of frequencies, or something similar. Although these may be useful as speech aids, their shortcomings prevent them from being used alone for tactile-speech perception: the simpler displays do not convey enough information for intelligible speech, while the more complex ones exceed the psychometric limits of the skin. Despite that, there is value in this type of display. The perception of segmental and suprasegmental features of speech is improved with tactile displays conveying the amplitude envelope, fundamental frequency, formant frequencies, zero-crossing frequency, and voicing duration (Grant et al., 1985; Summers et al.,

1996; Summers & Gratton, 1995). This aligns with research on the contribution of these prosodic features on speech comprehension (see Cutler et al., 1997 for a review).

Symbolic displays more efficiently convey speech by presenting a higher-order representation. In other words, rather than displaying some acoustic property or function thereof, symbolic displays render phonemes or articulatory gestures. Staunch proponents of motor theory may argue that the auditory system possesses a unique speech decoder that is not available to other sensory channels (Liberman et al., 1967, 1968). Under this theory, it is necessary for the acoustic signal to be decoded before being presented to a non-auditory sensory channel. Motor theory aside, symbolic displays offer a simplified and more easily interpretable presentation of speech. One drawback of such a system is that a real-time speech stream must first be processed by a speech recognition system to determine which phonemes to display. However, automated speech recognition is imperfect, and the most accurate recognizers require significant delays. Further, symbolic displays fail to convey all or most of the nuance and richness that generally accompanies spoken word, like tone, inflection, timbre, accents, emphasis, etc.

The current study proposes a novel display which presents segments via a symbolic display as well as pitch and intensity information for suprasegmental cues.

1.3.2 Tactile Perception

1.3.2.1 Mechanoreceptors

Tactile perception is transduced through four classes of mechanoreceptors. These are further classified by their adaptation rate (fast or slow) and receptive field size (type I or type II). Slow-adapting (SA) mechanoreceptors produce a burst of action potentials when a stimulus is initially applied, and a lower but sustained response while the stimulus remains present. Fast-adapting (FA) receptors produce action potentials when a stimulus is applied or removed, but no sustained signal if the stimulus is unchanging. Type I receptors are more densely innervated and have small receptive fields, while type II

receptors are less densely innervated and have larger receptive fields (Yantis & Abrams, 2014). Each of the four types of mechanoreceptors is terminated with a specialized ending suited to a particular type of stimulus (see Table 1).

	Slow adapting		Fast adapting	
	Type I	Type II	Type I	Type II
Termination	Merkel Cell	Bulbous / Ruffini Corpuscle	Tactile / Meissner Corpuscle	Lamellar / Pacinian Corpuscle
Skin type	Glabrous	Connective tissue	Glabrous	Glabrous, Hairy
Depth	Upper Dermis	Dermis	Upper Dermis	Lower Dermis
Density	Dense	Sparse	Dense	Sparse
Spatial Acuity	High	Low	High	Low
Temporal Acuity	Low	Low	Medium	High
Stimuli	<ul style="list-style-type: none"> ● Indentation ● Low-frequency vibration ● Edge pressure 	<ul style="list-style-type: none"> ● Stretch 	<ul style="list-style-type: none"> ● Low-frequency vibration 	<ul style="list-style-type: none"> ● High-frequency vibration
Functions	<ul style="list-style-type: none"> ● Patterns ● Textures ● Shape 	<ul style="list-style-type: none"> ● Skin stretch ● Hand conformation 	<ul style="list-style-type: none"> ● Slip ● Grip control 	<ul style="list-style-type: none"> ● Fine textures
Field size (mm²) (Median)	2-100 (11.0)	10-500 (59)	1-100 (12.6)	10-1000 (101)
Frequency (Hz) (most sensitive)	0.4-100 (7)	7	10-200 (20-40)	40-800 (200-300)

Table 1. Mechanoreceptors. Adapted from: Vissel (2009); Yantis and Abrams (2014).

The SAI channel, terminated by Merkel cells, responds best to low-frequency vibrations (5-15 Hz) and deep touches. The small receptive fields of Merkel cells yield high tactile acuity and are most abundant in areas like the fingertips. The FAI channel (terminated by tactile / Meissner corpuscles) activates on moderate-frequency (10-50 Hz) vibrations. The SAII channel (terminated by bulbous / Ruffini corpuscles) activates upon the stretching of the skin. The FAII channel (terminated by Lamellar / Pacinian corpuscles) activates on vibration and pressure, such as detecting surface texture. Hairy skin also features hair follicle afferent fibers which sense low frequency (<80 Hz) vibrations (Mahns et al.,

2006). Higher-frequency vibrations likely depend on lamellar corpuscles which, in hairy skin, are only present in deeper tissue surrounding the joints and bone (Mahns et al., 2006).

1.3.2.2 Tactile Illusions

Perceptual illusions can have dramatic effects on a display, and thus need to be considered by display designers. Although visual and auditory illusions are better known, various tactile illusions should not be ignored. While some of these perceptual illusions limit what can be displayed and correctly perceived, others can be leveraged to enhance the display. A brief review of highly-relevant illusions is presented below, but a more comprehensive review can be found in Hayward (2008).

One such tactile illusion is the funneling effect, wherein two vibratory stimuli with some space between them will be perceived as a single stimulus located between the two veridical stimuli (v. Békésy, 1958). The spacing between stimuli, their relative amplitudes, and the temporal order all moderate the perceived location and intensity of the signal (Cha et al., 2008). The funneling effect limits the highest resolution of a tactile display within which the observer can reliably identify two adjacent, simultaneous stimuli. However, this effect can also be leveraged to increase the resolution of a tactile display where only a single stimulus needs to be perceived at a given time. For example, a hex-grid of only three vibratory actuators can create the perception of six stimulus sites, where tactile "pixels" can be perceived at the location of each actuator as well as midway between each pair of actuators (see Figure 3).

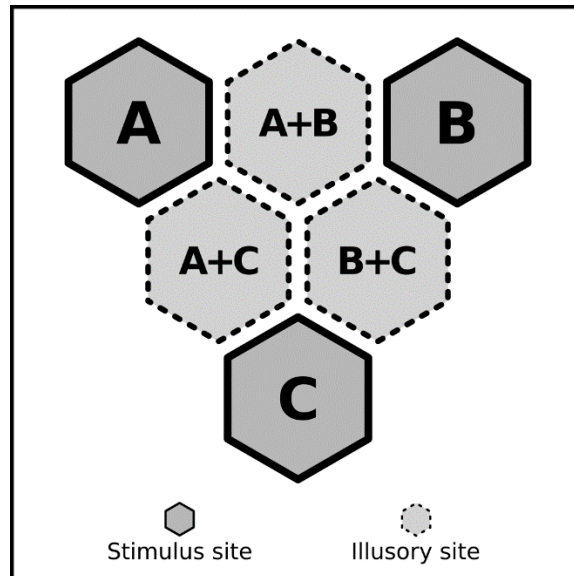


Figure 3. Veridical and illusory tactile stimuli sites.

Another better-known tactile illusion is the "cutaneous rabbit" or "cutaneous saltation" illusion, first described in the early 1970s (Geldard & Sherrick, 1972). This illusion can be experienced by producing a series of rapid, pulsed stimuli at one location of the body, followed by more at a second location (e.g., first at the wrist and then near the elbow). This works best in locations with low tactile acuity. Rather than perceiving taps in just the two stimulated locations, the observer will perceive the taps spaced across the skin, starting at the first location and ending at the second. It is likened to a tiny rabbit hopping along the skin—hence the name. Like the funneling effect, the cutaneous rabbit illusion can cause observers to misidentify the location of stimuli. Producing the "cutaneous rabbit" illusion is less reliable than the funneling effect; combined with its spatiotemporal nature, it is less ideal as an intentional means of displaying information.

The tau effect is another spatiotemporal illusion in tactile perception. When multiple, successive stimuli are presented in equal spacing across the skin, but in unequal temporal intervals, observers will report stimuli that have shorter temporal spacing to be spatially closer, and stimuli with longer temporal spacing to be spatially farther, despite all of the stimuli being equally distributed in space (Helson & King, 1931). Similarly, the kappa effect can be experienced when stimuli have equal temporal spacing,

but unequal spatial spacing. Rather than misestimating distance, observers will misestimate time (Goldreich, 2007). Space and time are intrinsic dimensions of tactile perception, so tactile displays are very likely to encode information in one or both dimensions.

1.3.2.3 Stimulators

The somatosensory system is responsible for the array of sensations that are often collectively referred to as the "sense of touch". This includes skin deformation/indentation, vibration, skin stretch, and proprioception (sense of self-movement and body position), but also includes nociception (pain) and thermoception (temperature). In a technical sense, any of these sensory inputs could potentially be used for a display, but some limitations are evident. Few people would voluntarily use a display that caused pain, and one that conveyed information by moving limbs or appendages would be usable in very few environments and situations and would likely be cumbersome. Thermoception is slow and has low acuity. Among all the sensations that can be presented on the skin, vibration seems to be the best suited for a dynamic, high-bandwidth display.

The most common vibratory actuator is the eccentric rotating mass (ERM) motor. When DC electricity is supplied to an ERM, the motor shaft spins an off-center weight which causes an undulation with each rotation. ERMs are very cheap and easy to integrate into a system. Altering the electrical supply will change the speed of the ERM and the intensity of the vibration—the speed and intensity are strongly coupled and will always present in a sinusoidal form. Linear resonant actuators (LRA) oscillate perpendicular to their mounted surface and generate a sinusoidal signal. Unlike ERMs, an LRA's intensity is not coupled to its frequency, which is fixed typically around 200 Hz. LRAs are also more responsive than ERMs and a little more expensive and complex. Piezoelectric transducers are highly responsive, but the necessary driving circuitry can become quite complex when employed in an array. In this consideration, voice coils can be likened to higher-quality piezo transducers, with similar pros and cons. Haptic piezoelectric actuators are highly responsive and can generate a high-fidelity signal but can be

extremely expensive. The current study uses a device equipped with LRAs, as they provide a good balance between cost, performance, and complexity. A summary comparison of actuators is provided in Table 2.

	Cost	Fidelity	Response Rate	Complexity
ERM	\$	Low	Low	Low
LRA	\$\$	Moderate	Moderate	Moderate
Piezo (buzzer)	\$\$\$	High	High	High
Piezo (haptic)	\$\$\$\$	Very high	High	Moderate
Voice coil	\$\$\$\$	Very high	High	Very high

Table 2. Vibrotactile actuators.

1.3.2.4 Body Site

Spatial acuity varies greatly across the body and thus the highest resolution of a tactile display may vary for different body parts. In other words, a region with higher spatial acuity can support a higher resolution tactile display over the same area size. Spatial acuity is typically measured using a two-point discrimination (2PD) task, where an instrument with two adjustable prods (e.g., calipers) is touched against a participant's skin who then reports whether they feel one or two points. Although alternative tasks have been presented as more accurate measures (Tong et al., 2013), the 2PD task is more prevalent. Data from 2PD studies has been reconciled with physiological measures to provide reasonable estimates of innervation density and tactile acuity for both slow- and fast-adapting afferents (Corniani & Saal, 2020). The hands and face are the most densely innervated regions, followed by the feet, neck/scalp, arms, legs, and trunk (see Figure 4 and Figure 5). It is worth noting that localization of a vibrotactile stimulus is the most accurate when presented near anatomical points of reference, such as the wrist, elbow or spine (Cholewiak & Collins, 2003), and that dynamic stimuli will enhance spatial acuity on at least some body sites (Jones & Sarter, 2008).

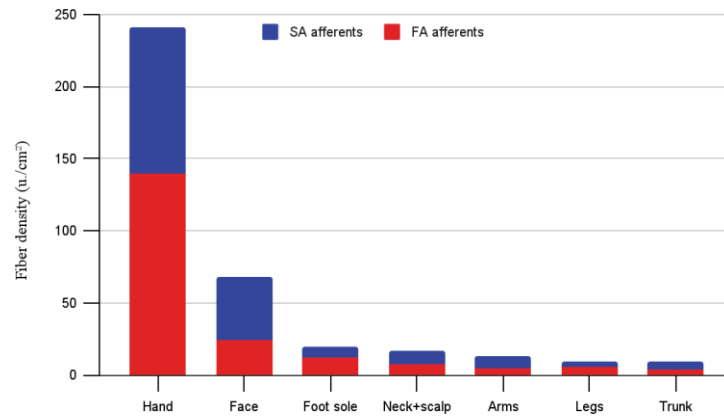


Figure 4. Somatosensory nerve density. Adapted from Corniani and Saal (2020).

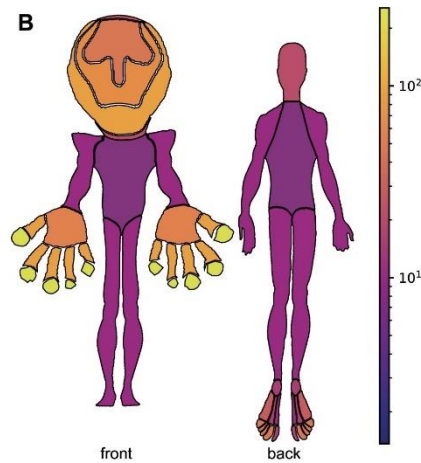


Figure 5. Somatosensory homunculus (Corniani & Saal, 2020).

The temporal sensitivity of a body region should also be taken under consideration; users are likely to misinterpret signals if information is presented too quickly. Temporal sensitivity is typically measured with a discrimination task which presents two successive pulses to the skin. A temporal discrimination threshold (TDT) is determined by changing the interstimulus interval until a minimum time is determined, below which the two pulses are perceived as a single pulse (Lacruz et al., 1991). Comparisons of the TDT across different parts of the body (see Figure 6) indicate that the face/forehead and forearm/hand/finger have lower thresholds than the rest of the body (Hoshiyama et al., 2004). The lower thresholds in these regions make them better candidates for faster tactile-displays.

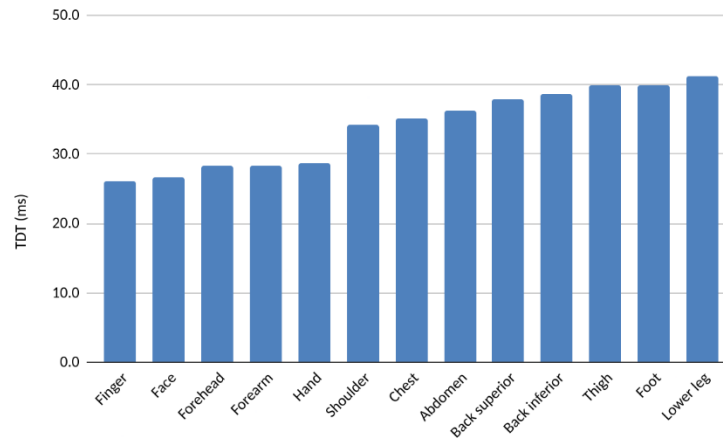


Figure 6. Temporal discrimination thresholds across the body. Adapted from Hoshiyama et al. (2004).

Rather than providing strictly discrete binary sensations (by enabling or disabling a stimulus), a tactile display may encode information in intensity by modulating the displacement of the actuator via the amplitude of the vibrational signal. Previous work have shown successful discrimination in four or fewer intensity levels (Craig, 1972), and that detection thresholds are higher on hairy skin than on glabrous skin (Mahns et al., 2006). Care should be taken when encoding information in a vibrating signal's intensity and frequency simultaneously, as the perceived intensity of the signal is a function of the frequency (Verrillo et al., 1969).

Another potential dimension for encoding information is in the frequency of a vibrating signal. Research suggests that observers could distinguish seven to ten frequency levels (Rothenberg et al., 1977), but need to take considerable training effort (Jones & Sarter, 2008). It should be noted skin is most sensitive to a signal in the range of 150-300 Hz (Jones & Sarter, 2008), which limits the number of distinguishable frequencies. In addition, a system that encodes information in both intensity and frequency need to account for the perceptive relationship between the two factors (Verrillo et al., 1969).

Although these physiological limits are useful for guiding the design of a tactile display at potential body locations, they should not be taken as rigid limits. Firstly, the human brain is plastic, as

training/experience can improve sensitivity (Elbert et al., 1995; Godde et al., 2000; Hodzic et al., 2004). Secondly, there are limitations and discrepancies in what conventional psychometric methods can measure. The existence of hyperacuity in multiple sensory channels (Altes, 1989; Loomis, 1979), for example, demonstrates that perceptual metrics are sometimes incomplete (Strasburger et al., 2018)(Altes, 1989; Loomis, 1979). Individuals can sometimes exhibit abilities that exceed their own well-established expectations. For example, participants were asked to identify which of two successive tones has a higher pitch in two ways: by humming the higher of the two tones or by verbally stating "first" or "second". The verbal responses of tone-deaf individuals are only correct at chance-level, but surprisingly, the hummed responses of those same individuals are significantly more accurate (Loui et al., 2008). Another surprising example is that of "blindsight"—individuals who are completely blind can unconsciously avoid obstacles during walking although they have no conscious awareness of the obstacles in their way (Celesia, 2010).

Besides the physiological abilities and limits of various body sites, social factors affecting the adoption must also be taken into consideration for the placement of a tactile display. Research on assistive-technology adoption is very informative in this regard, as these devices are abandoned by users at high rates. One study showed that 29.3% of the technologies were completely abandoned (Phillips & Zhao, 1993). Research in abandonment does provide some advice to help. For example, devices must be simple to set up and use. They must also be aesthetically pleasing, age-appropriate, fashionable, and culturally and socially acceptable. Devices must also avoid giving users the appearance of being "handicapped" (Kintsch & DePaula, 2002). Failing to accommodate these needs may create negative social stigma for users and become a significant barrier to adoption (Parette & Scherer, 2004). More broadly, Universal Design contributes basic principles that serve as meaningful guidelines in design (Story, 2011):

- Equitable Use: The design is useful and marketable to people with diverse abilities.

- Flexibility in Use: The design accommodates a wide range of individual preferences and abilities.
- Simple and Intuitive Use: Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.
- Perceptible Information: The design communicates necessary information to the user, regardless of ambient conditions or the user's sensory abilities.
- Tolerance for Error: The design minimizes hazards and the adverse consequences of accidental or unintended actions.
- Low Physical Effort: The design can be used efficiently and comfortably and with a minimum of fatigue.
- Size and Space for Approach and Use: Appropriate size and space is provided for approach, reach, manipulation, and use regardless of the user's body size, posture, or mobility.

The high acuity performance across multiple dimensions makes the hands or face seem like ideal body sites for a tactile display. However, a device on the hands would make it difficult or impossible to do anything with the adorned hand(s) while using the display. For example, it may interfere with the ability to use sign language, explore the environment manually, operate a keyboard, etc. Further, a device worn in such an obvious place would almost certainly generate social stigma. Tactile acuity across the rest of the body is relatively homogenous, but temporal sensitivity of the forearm is on par with the hand. A device worn on the forearm has the potential to be relatively quick and easy to apply and remove and easily concealed with a sleeve.

1.3.2.5 Perceptual learning

Perceiving non-trivial information on a tactile display will require its users to learn how to decode the stimuli of that display. This type of learning falls into the domain of perceptual learning which refers to "an increase in the ability to extract information from the environment, as a result of experience and practice with stimulation coming from it" (Gibson, 1969). Perceptual learning occurs

through low-level sensory experiences, and research often focuses on the low-level processes and mechanisms involved, but perceptual learning influences skilled behaviors and serves as the foundation for higher cognitive processes as well (Kellman, 2002). Perceptual learning is the mechanism through which an oncologist learns to recognize abnormal MRI scans, where a layperson would see meaningless smears and shadows; a person who is blind can read Braille while others just feel a bumpy surface; a piano tuner can do their job without any absolute reference sound; sommeliers can identify a wide variety of flavors and fragrances from a single sip of wine; etc.

Changes incurred by perceptual learning can be described as either changes of discovery, fluency, or a combination of the two. Discovery describes how one identifies, selects, or amplifies features required for discrimination. Fluency describes changes in the ease of information extraction (Kellman, 2002). Learning in the form of both discovery and fluency are important for a tactile display to be successfully used.

With perhaps the exception of Braille for the few who can read it, the tactile modality is not typically used for conveying complex and confusable information. Training focused on discovery will be necessary to improve sensitivity to distinguish between spatially and temporally proximal stimuli. With a tactile display that encodes information using multiple simultaneous stimuli, users will need to learn how to select the differentiating features between disparate pieces of information. A training environment which allows users to deploy their attention (deliberately or not) may facilitate faster learning (Ahissar & Hochstein, 2004; Byers & Serences, 2012; Goldstone, 1998).

The simple, mostly binary nature of most tactile stimuli requires little to no energy to interpret, but this will not be the case for a complex tactile display. Because of this, users will need training focused on improving fluency so that they may consume information at a reasonable rate. The ultimate goal is to achieve automaticity, the highest level of fluency where the presentation of a stimulus consistently activates the appropriate concept in long-term memory without needing the observer's

control or attention (Schneider & Shiffrin, 1977). This achievement may be best accomplished through training designed with the mechanisms of perceptual learning in mind: attentional weighting, stimulus imprinting, differentiation, and unitization (Goldstone, 1998).

A stimulus may carry many features across many dimensions, some of which may be signal, and others may be noise. An experienced observer can learn to selectively apply and weigh their attention to task-relevant features. This is known as *attentional weighting*. Attentional weighting has been used to help explain effects in categorical perception of speech. For example, classic research showed that listeners can more accurately distinguish speech sounds when they come from different categories, despite the physical differences between the sounds having been equated (Liberman, 1957). Since the categorical perception of speech sounds has been demonstrated in young infants, some argue that this ability is innate, rather than learned. If that is true, it certainly would not be the case for tactile-speech perception. As such, training for tactile-speech perception should include tasks that offer the opportunity or require the learner to learn how to weigh the relevant dimensions.

As a learner continues to receive stimulus exposure, they may develop specialized detectors that activate for specific stimuli or stimulus features—this is known as stimulus imprinting. Imprinting occurs through repeated stimulus exposure and can improve speed and accuracy. Differentiation occurs when an observer eventually learns to distinguish different stimuli, and, like imprinting, can occur between whole stimuli or between stimulus features. Novel stimuli to a naive observer may be indistinguishable initially, but repeated exposures to those stimuli will facilitate differentiation. Finally, unitization is a mechanism of perceptual learning through which separate features are perceived and decoded as a single unit. For example, a typed letter is perceived immediately as that whole letter, rather than being perceived as its constituent strokes and marks. See Goldstone (1998) for a more comprehensive review of perceptual learning mechanisms.

1.3.2.6 Language acquisition

As phonemes are the smallest building-block of a spoken language, it is logical that discriminating between phonemes is an important first-step of language acquisition. It follows naturally that acquiring a tactile language that encodes phonemes as its basic unit will also require phonemic discrimination. In spoken languages, very young infants demonstrate the ability to distinguish phonemes from all human languages but lose this ability with phonemes outside of their native language before their first birthday (Werker et al., 2012). This and other aspects of language acquisition and development may help inform the successful acquisition of a tactile language.

Statistical learning theories have advanced the understanding of language acquisition. These ideas were first put forward when 8-month-old infants in a lab study very quickly demonstrated the ability to learn the significance of the relationships between word sounds (Saffran et al., 1996). Researchers familiarized the infants with a constructed language consisting of four three-syllable nonsense pseudowords presented in a stream of synthesized speech. Words in the stream were continuous, containing no pauses between words. It was also monotonic with no stress differences or any other acoustic or prosodic cues to indicate word boundaries. The words were constructed such that the ordered syllable pairs occurred more frequently within words than across a word-boundary. This statistical relationship between syllables served as the only word-boundary cue in the speech stream. After only two minutes of familiarization, the infants successfully discriminated between words from the constructed language and words that were not part of the language but consisted of the same set of syllables. This indicated that, in a very short time, the infants were able to learn the statistical relationship between syllables, rather than just having become familiar with the syllables themselves. Although it is more difficult to acquire new languages in adulthood, this effect has been demonstrated in adults as well (Saffran et al., 1996).

Another statistically based learning mechanism for language acquisition is distributional learning. Distributional learning proposes that language learners track the relative frequencies of phonetic tokens to which they are exposed. Researchers created eight stimuli across a continuum between "da" and "ta" moderating the presence of voicing and formant frequency trajectories and familiarized infants with them. One group of infants were presented with stimuli sampled from the extreme ends of the continuum, while the other group was presented with stimuli from the center of the continuum. After only 2.3 minutes of familiarization, the infants who were familiarized with the divergent stimuli became better at discriminating between "da" and "ta" than the infants who were not.

1.3.2.7 Gamification

Motivating learners to put in the time and practice to acquire a new skill or ability can be a challenge. Academic studies in perceptual learning often disregard motivation, perhaps because perceptual learning can be demonstrated even under some of the most mundane training regimens. Surprisingly, perceptual learning can occur with merely the repeated exposure to a stimulus, even when it is irrelevant to the learner's task and below the threshold for detection (Watanabe et al., 2001), so is the effort/cost of finding a mechanism to motivate the user worthwhile? Perhaps more so outside of a laboratory setting, where learners control their own training schedule and operate in environments full of potential distractions. Intrinsic motivation can be created by applying game design elements to the task, a motivation strategy known as "gamification". This idea builds on the demand to satisfy certain psychological needs with different elements (see Table 3 for examples).

Psychological Need	Mechanism	Game Design Element
Competence	Granular feedback	Points
	Sustained feedback	Performance graphs
	Cumulative feedback	Badges
		Leader boards
Autonomy (decision freedom)	Choices	Avatars
Autonomy (task meaningfulness)	Volitional engagement	Meaningful stories
Social relatedness	Sense of relevance	Teammates
	Shared goal	Meaningful stories

Table 3. Psychological Needs for Gamification (Sailer et al., 2017).

Studies of gamification have shown that it can be used to improve motivation, behaviors, and learning outcomes (Hamari et al., 2014; Sailer & Homner, 2020), and it has been applied in contexts similar to the current study. For example, Deveau et al. (2014) successfully measured improvements across multiple visual tasks after training participants with a perceptual-learning-based video game. As a tool for second-language learning, gamification has been popular in apps like Duolingo, Class Dojo, Edmodo, Zondle, Socrative, and Brainscape (Flores, 2015). Gamification has even been applied in a context as niche as vibrotactile-speech learning (Martinez, 2019).

1.3.3 Research questions

The current study proposes a novel display which presents segments via a symbolic display as well as pitch and intensity information for suprasegmental cues. To adequately determine whether tactile-speech perception has been achieved through the training, the following hypotheses are considered:

- Participants will correctly identify more tactile-phonemes after training with the Vibey Transcribey and versus participants who do not train over the same period.
- Participants will correctly match more tactile-phrases with specific focused words after training with the Vibey Transcribey and versus participants who do not train over the same period.

- Participants will correctly match more tactile-phrases with varying phrase boundaries after training with the Vibey Transcribey and versus participants who do not train over the same period.
- Participants will integrate more tactile syllables with auditory syllables after training with the Vibey Transcribey and versus participants who do not train over the same period.

CHAPTER 2

METHOD

2.1 Materials

The software, schematics, and design files for the entirety of this project are available on the Open Science Framework website at <https://osf.io/3f5nu/>.

2.1.1 Hardware

A wearable tactile array was developed to convey tactile sensations to the wearer. The device was worn as a sleeve along the arm and featured 16 Jinlong G1040003D LRA linear resonant actuators to stimulate 42 cells and a linear array. Similarly to the array proposed by Ellis and Robinson (1993), the cells were arranged in a 6x7 hexagonal grid, with the six cells wrapped around the forearm and the seven cells along the length of the forearm. Within the hex-grid, twelve actuators were spaced in alternating cells such that actuator-cells were separated by an empty cell (see Figure 3 and Figure 7). In this configuration, 12 cell locations were stimulated by a single actuator, while the remaining 30 cells were stimulated via the funneling effect generated by the nearest two actuators. Four additional actuators, arranged in a line on top of the bicep, also employed the funneling effect. The hexagonal arrangement and distribution of directly stimulated cells allowed for an efficient and cost-friendly design. The directly stimulated cells were organized into three groups of four LRAs, with each group sharing a single driver through a multiplexer. These were arranged within the hex-grid such that every funneled cell was stimulated using two LRAs, each from different groups (i.e., LRAs from the same group never share a funneled cell). This setup allowed stimulation of 42 cells using only 12 LRAs. The actuators were attached to jersey knit fabric with 3d-printed housings. The fabric stretched to accommodate different arm sizes and shapes. Spacing between actuators varied between 4-8cm to induce the funneling effect (Cha et al., 2008).

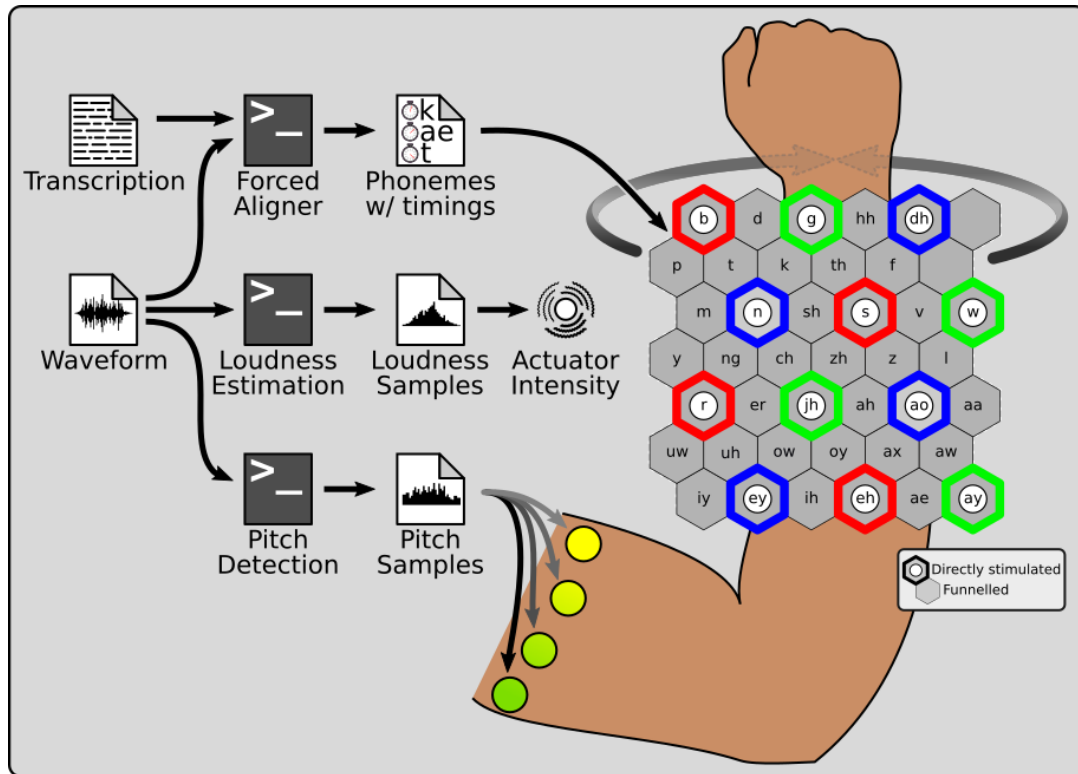


Figure 7. Encoding flow and tactile arrays.

The devices were controlled by an Espressif ESP32-WROOM-32D microcontroller in a DevKitC-v4 package mounted to a custom PCB (Figure 8). On each device, three Texas Instruments DRV2605L LRA drivers were multiplexed (using NXP NX3L4051PW's) to control 12 LRAs on the forearm array. Two additional drivers were multiplexed to control the four actuators on the bicep. Drivers were managed over an I²C bus, with SCL on each driver switched on another multiplexer. The LRA drivers featured overdrive and active braking to improve the rise and fall time responsiveness of the actuators in closed-loop mode. The microcontroller was connected to a PC via USB/UART operating at 115200 baud through a supplied powered USB hub which, aside from ensuring participants had a free USB port to connect the device, added a layer of separation between the device and the participants' hardware in the event of electrical failure. Powered hubs were specifically chosen to provide a more reliable current source during potential higher-amperage spikes in the device's usage.

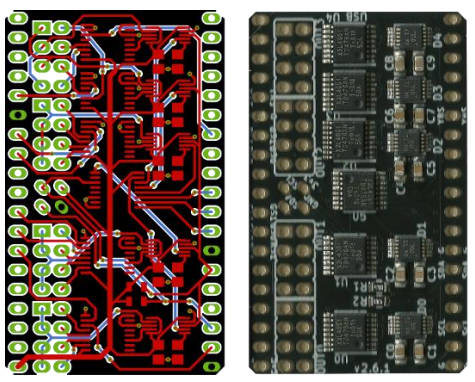
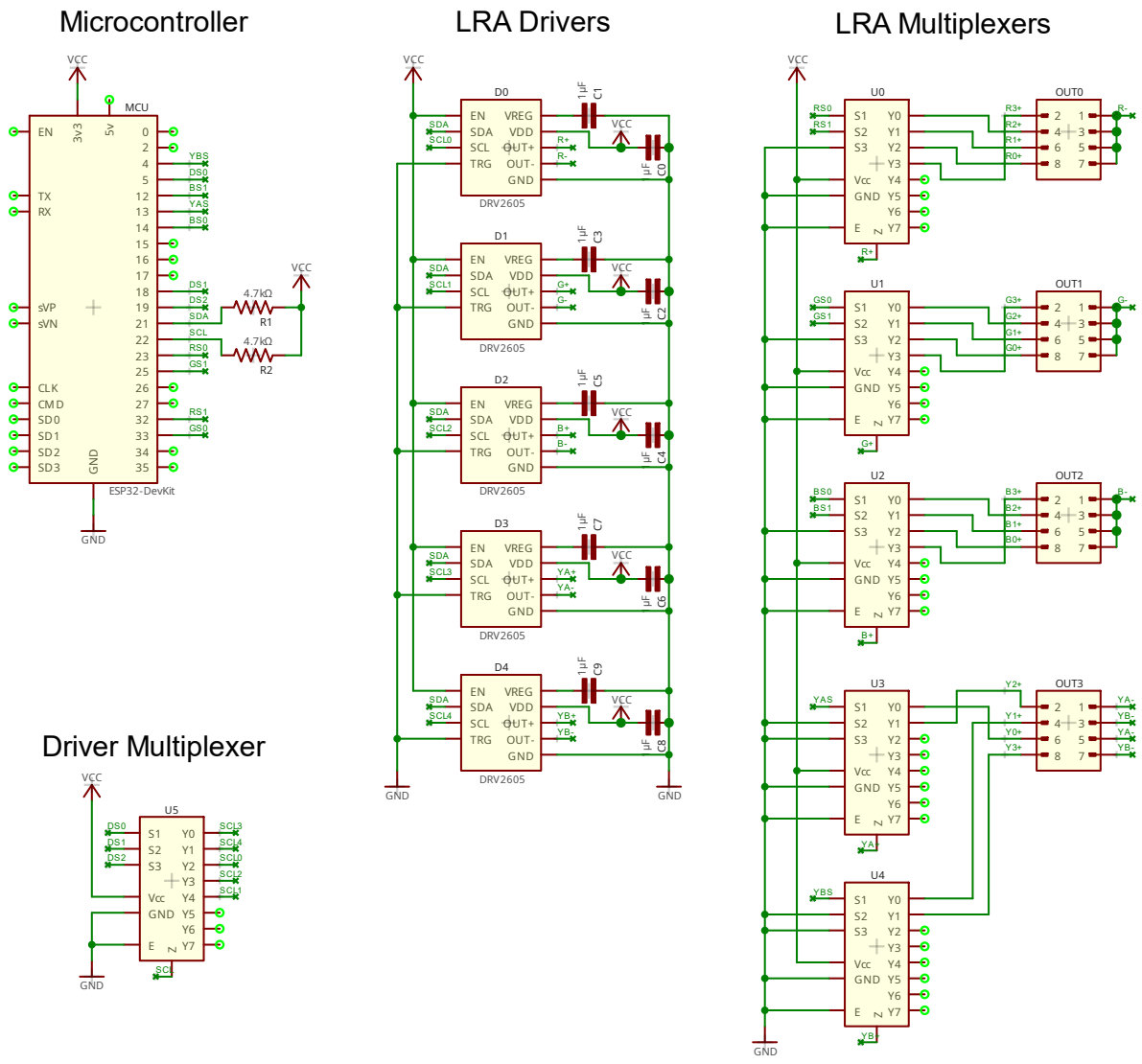


Figure 8. Custom electronics to drive LRA arrays.

The sleeves used for this study were cylindrical with a slight taper toward the wrist in their relaxed state. When worn, the fabric stretched in non-regular directions to accommodate the shape of the arm. Because of this non-regular stretching, points which were equidistant in the sleeve's relaxed state lost this property when worn. To approximate equal spacing between actuators when worn, life-size jigs were created using human anatomy models from the BodyParts3D project (Mitsuhashi et al., 2009). Another model was created to represent the desired equally-spaced locations of the actuators by extruding circles from a central axis in the forearm. The second model was geometrically subtracted from the model of arm, leaving holes in the arm model (see Figure 9). The result was 3D printed (see Figure 10) and fitted with a sleeve. Then, using the holes as guides, marks are made where actuators were to be positioned. The sleeve was then removed from the jig and actuators were attached in the marked locations. The actuators were attached to the fabric by first being fitted into a 3D printed casings (see Figure 11) which also provided stress relief for the connecting wires. The housings were affixed to the knit fabric with cyanoacrylate glue (specifically, Gorilla brand "Brush & Nozzle" super glue was found to work best). A 3D printed jig was designed to complement the housing and used to clamp pieces while the glue dried. The actuator casings were color-coded and numbered to make it easy to reconnect if they became disconnected during use. The device as worn is presented in Figure 12.

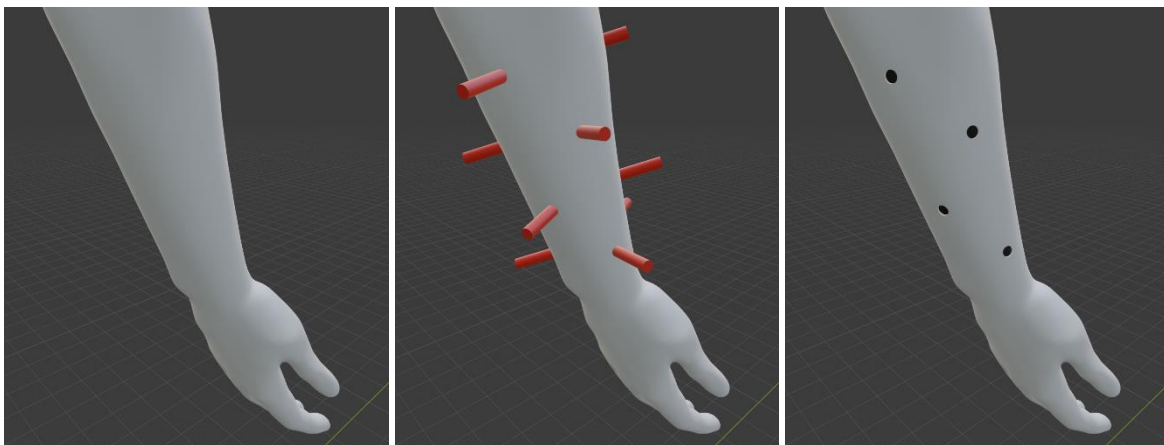


Figure 9. Modelling a life-size jig for equidistant actuator spacing.



Figure 10. An assortment of 3D printed jigs and electronics cases.

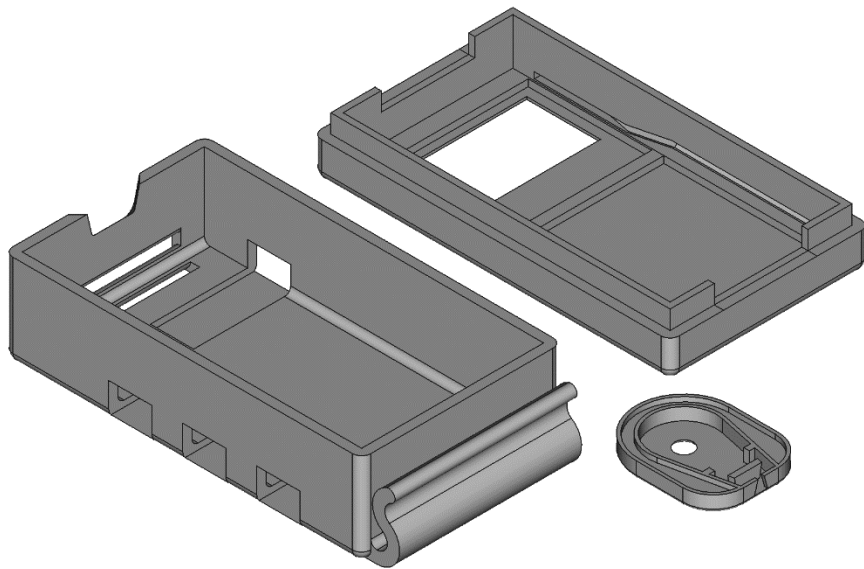


Figure 11. 3D printed electronics and actuator housings.

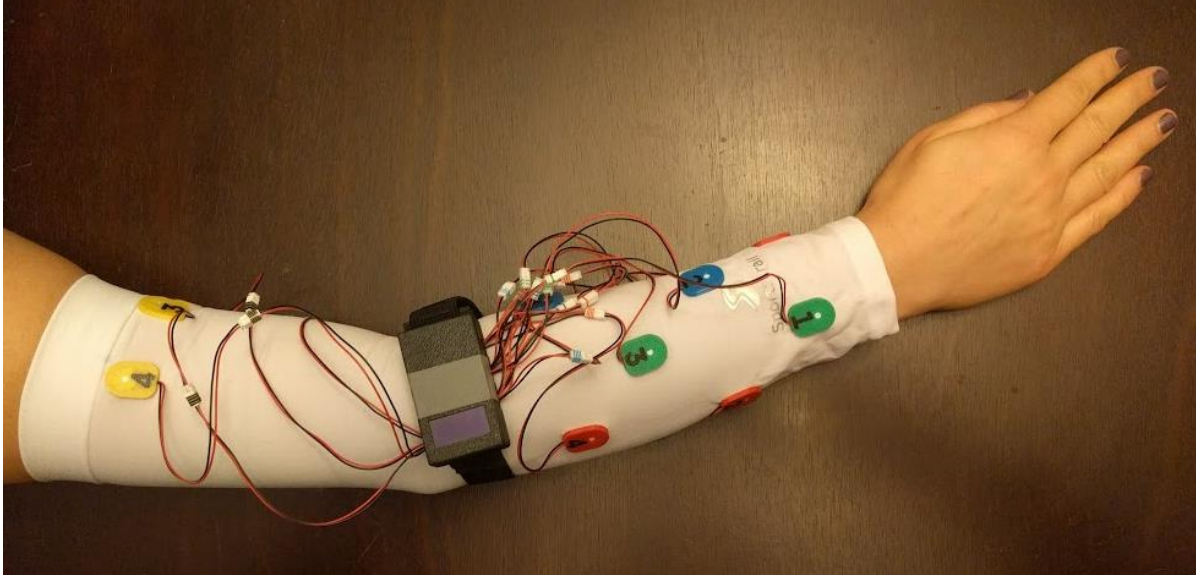


Figure 12. The Vibey Transcribey device as worn.

The electronics were housed in a two-part 3D printed enclosure. The body of the housing was designed to accommodate an elastic band on either end such that the band could be adjusted easily. The faceplate for the housing featured a small OLED display (Figure 13). This displayed a graphical representation of the device status as a character's face. When the device was successfully connected to the training or evaluation software, the display presented a smiling face with opened eyes. When vibrating a stimulus, the device displayed closed eyes with squiggled lips. When not connected, the display rendered Zs for eyes and a straight mouth.

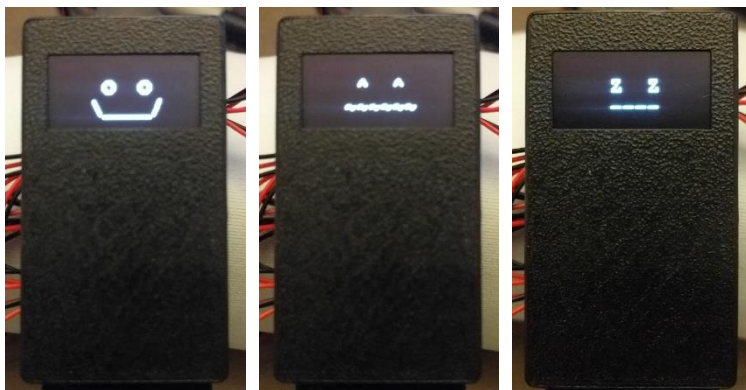


Figure 13. Device display states.

The total hardware cost to produce nine Vibey Transcribey units in their final form was approximately \$1,300 USD or about \$150 each. Costs were not tracked for the various revisions and prototypes of the hardware.

2.1.2 Phonemic feature mapping

Each of the hex cells in the tactile array were assigned to a unique phoneme, with similar phonemes loosely grouped together (see Figure 7), consonants near the wrist and vowels near the elbow. Starting at the wrist, consonants were grouped by plosives, fricatives, nasals, and approximates. Vowels were loosely arranged according to the IPA vowel trapezium / first two formants (see Figure 14).

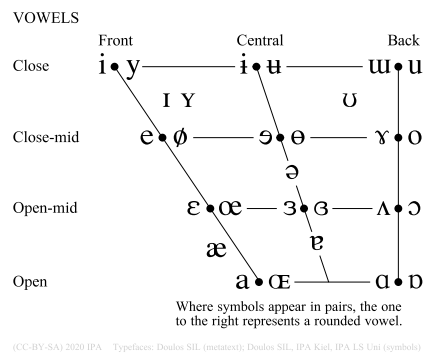


Figure 14. IPA Vowel Chart (International Phonetic Association, 1999).

2.1.3 Prosodic feature mapping

The four actuators positioned on the bicep displayed pitch and loudness. Pitch was mapped tonotopically in a frequency-to-place fashion: the lowest-frequency pitch was associated to the actuator on the inside of the bicep, while the highest-frequency pitch was associated to the actuator on the outside of the bicep. Pitches between the extremes were mapped to up to two of the four actuators, making use of the funneling effect to stimulate a continuous spectrum across the bicep. Loudness was matched to actuator intensity: louder speech yielded more intense vibrations. All features of the display were temporally rendered one-to-one with the source audio.

2.1.4 Stimuli

Pre-recorded auditory speech stimuli from previous research were converted to vibratory stimuli for both training and evaluation. First, audio samples were resampled if necessary to single-channel, 44.1 kHz, 16-bit PCM WAV files. Using SoX (Bagwell, n.d.), a gain was applied to normalize each sample to the EBU R128 standard of -18 LUFS. The gain level was calculated for each stimulus independently using loudgain (Hormann, 2019/2022)—short samples were padded with silence to fill the minimum 500ms prior to gain calculation. TextGrid files were generated using the Montreal Forced Aligner (McAuliffe et al., 2017) and an extended version of the Librispeech lexicon (Panayotov et al., 2015). These files contained a timestamped phonemic transcription of the stimuli. A convolutional neural network and pre-trained model (Ardaillon & Roebel, 2019) produced a CSV file of the estimated continuous pitch. Similarly, a CSV file of the estimated continuous loudness was generated using the Zwicker method (Zwicker & Fastl, 2013) for non-stationary signals (Green Forge Coop, 2021). Finally, a custom Python script combined the TextGrid transcript, pitch CSV, and loudness CSV into a single ".vtt" binary file. This file format included a short header containing a simple transcription of the speech, the CMUdict-encoded phonetic transcription, the sample period, and sample count. All stimuli in this study were sampled with a 1ms period. Each sample was made up of three bytes: one byte to identify the phoneme, one byte for the pitch (0-255), and one byte for the loudness (0-255). The encoded pitch value was first clamped to a 30-260 Hz range, converted to the Mel scale (Stevens et al., 1937), and then mapped to 0-255. Loudness was clamped at 50 LUFS and then mapped to 0-255.

All of the tactile stimuli used in training were adapted from audio files from the Angel Sound software (*Angel Sound - Interactive Listening Rehabilitation and Functional Hearing Test Program*, n.d.). Words were selected to represent four different categories. Each word had four speakers (two male, two female) whose recorded speech was converted to tactile stimuli using the process described above.

For segment identification, pre-recorded, naturally produced stimuli from previous research was used. Hillenbrand et al.'s (1995) samples of ten speakers (five women and five men) vocalizing each vowel in an /h/-VOWEL-/d/ context were used as vowel stimuli. The vowels included ten monophthongs and two diphthongs: /i ɪ ε æ u ʊ ɑ ɒ ɔ ɜ əʊ eɪ/. Shannon et al.'s (1999) samples of 10 speakers (five women and five men) presenting each consonant in a /a/-CONSONANT-/a/ context were used for consonants. The constants were: /b d g p t k m n l r y w f s ʃ v z ð tʃ dʒ/.

To evaluate prosody perception, phrase recordings from Jasmin et al. (2020) were used. To test focus perception, 47 phrases were each recorded twice—once emphasizing an early word and the second time emphasizing a later word. For example, one pair includes: "Mary likes to READ books" and "Mary likes to read BOOKS". Jasmin et al. created spectrums of stimuli across pitch and time dimensions, but this study used a single early-focus and single late-focus version of each sentence, where pitch- and time-warping are matched at 25% for early-focus stimuli and 75% for late-focus stimuli.

Jasmin et al. also provided stimuli for phrase boundary detection: 42 clauses were recorded twice—once with an early closure and another with a late closure. For example, consider this pair of clauses (note the location of the comma): "when a train passes, the station ..." and "when a train passes the station, ...". Like the focus-detection stimuli, this study used a single early-closure and a single late-closure version of each sentence, with pitch- and time-warping matched at 25% and 75% respectively.

A set of auditory and visual McGurk stimuli from Stropahl et al. (2017) were modified to measure sensory integration. Their stimulus set included eight speakers with auditory and visual recordings of eight syllables, but only an incongruent subset of each were paired, yielding twelve combinations per speaker. The current study replaced the visual stimulus with the auditory recording of that visual stimulus converted into a tactile stimulus using the process described previously. See Table 4 for a listing of the auditory-tactile stimulus pairs.

A-T Stimulus	4-AFC Options
--------------	---------------

	Auditory	Tactile	Fused 1	Fused 2
Ba-Da	Ba	Da	Ga	Pa
Ba-Ga	Ba	Ga	Da	Ma
Ba-Ka	Ba	Ka	Ga	Da
Ba-Na	Ba	Na	Ga	Da
Ba-Ta	Ba	Ta	Pa	Da
Ma-Ga	Ma	Ga	Na	Ba
Ma-Ta	Ma	Ta	Na	La
Pa-Da	Pa	Da	Ka	Ta
Pa-Ga	Pa	Ga	Ka	Ta
Pa-Ka	Pa	Ka	Da	Ta
Pa-Na	Pa	Na	Ka	Ta
Pa-Ta	Pa	Ta	Da	Ka

Table 4. McGurk stimuli and AFC options.

2.1.5 Other Software and Data

A combination of original and open-source software was utilized throughout the development and execution of this study. A custom Python program was used to convert WAV audio files into tactile stimuli for evaluations and training. Software used for pre- and post-test evaluations was written in Python with PySide. Training software was written in C++ with Unreal Engine 4.25.4 (*UnrealEngine 4.25.4*, n.d.), Psydekick for Unreal Engine (Canare, n.d.), and the Boost C++ libraries (*Boost C++ Libraries*, n.d.) to provide cross-platform serial port communication.

Transcripts from the Buckeye Corpus of Conversational Speech (Pitt et al., 2007) were used to determine the relative frequencies of phonemes in conversational English. During training, participants were incentivized to play levels that contained training words with phonemes to which they were relatively underexposed.

2.2 Participants

A convenience sample of 12 volunteers were recruited with a mean age of 35.69 ($SD = 9.93$). Each participant was right-hand dominant, was a native English speaker, self-identified as at least a

moderate video game player, and had direct and regular access to a gaming PC. All of the participants reported having no known language impairments. Five participants identified as female, six identified as male, and one identified as genderqueer. Half of the participants were randomly assigned to an experimental group to receive training (described below) between pre- and post-tests, and the remaining participants received no training. One participant in the experimental group was dismissed after having missed multiple weeks of training.

2.3 Procedure

2.3.1 Evaluation

This study loosely followed the closed-set testing procedure by Fu et al. (2005) to evaluate perception of segments using the previously mentioned vowel and consonant recordings from Hillenbrand et al. (1995) and Shannon et al. (1999). To evaluate vowel performance, participants completed 120 trials of a 9-AFC task (see Figure 15). Stimuli were randomly selected without replacement. The eight incorrect options in the AFC task were selected at random with replacement. Participants responded using the directional pad on their preferred video game controller to select the syllable they believed corresponded to the tactile-vowel, and then by pressing a button on the controller to confirm their selection. For consonants, participants completed 200 trials of a 9-AFC task, again with stimuli randomly selected without replacement. No feedback was provided to the participants.



Figure 15. Phoneme identification task screenshot.

This study also loosely followed Jasmin et al.'s (2020) procedure to evaluate prosodic detection of emphasis/focus and phrase boundaries, as well as their stimuli previously mentioned. Each participant completed 141 focus-detection trials and 126 phrase-boundary trials. At the start of each focus-detection trial, a randomly selected sentence was presented visually, with the first clause in the sentence underlined. The underlined clause was either early- or late-focused, and the emphasized word was presented in capital, italic letters (see Figure 16). Participants were asked to read the clause silently while imagining how it might feel when converted to vibrations. After four seconds, the early- and late-focus tactile recordings of the underlined portion were rendered sequentially, in random order, with a brief pause between them. While the tactile recordings were playing, a visual indicator simultaneously highlighted the corresponding button. After both options were presented on the device, the participant responded by selecting the button which they believed matched the clause printed on the screen, using a gamepad as described previously.

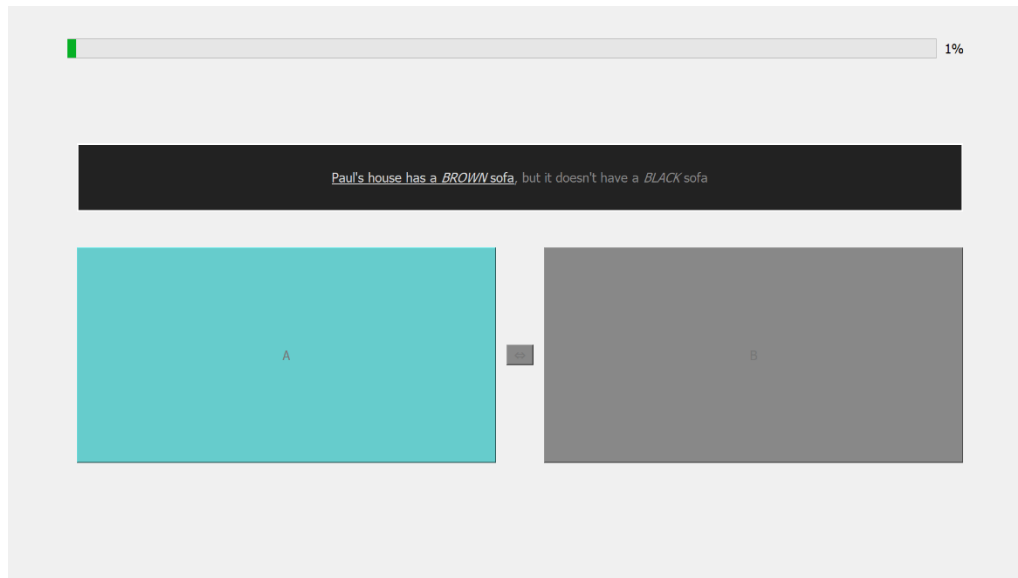


Figure 16. Focus discrimination task screenshot.

The phrase boundary detection task followed an identical procedure, but rather than identifying the position of an emphasized word, participants were tasked with identifying the position of a boundary between clauses (i.e., the position of a comma). Like the focus detection task, participants were visually presented with a randomly selected sentence and asked to consider how it should feel. After four seconds elapsed, early- and late-boundary versions of the underlined portion were displayed as vibrations in random order and the participant used their gamepad to choose whichever they believed matched the visually displayed sentence (see Figure 17). No feedback was provided to participants for either of the prosodic tests.

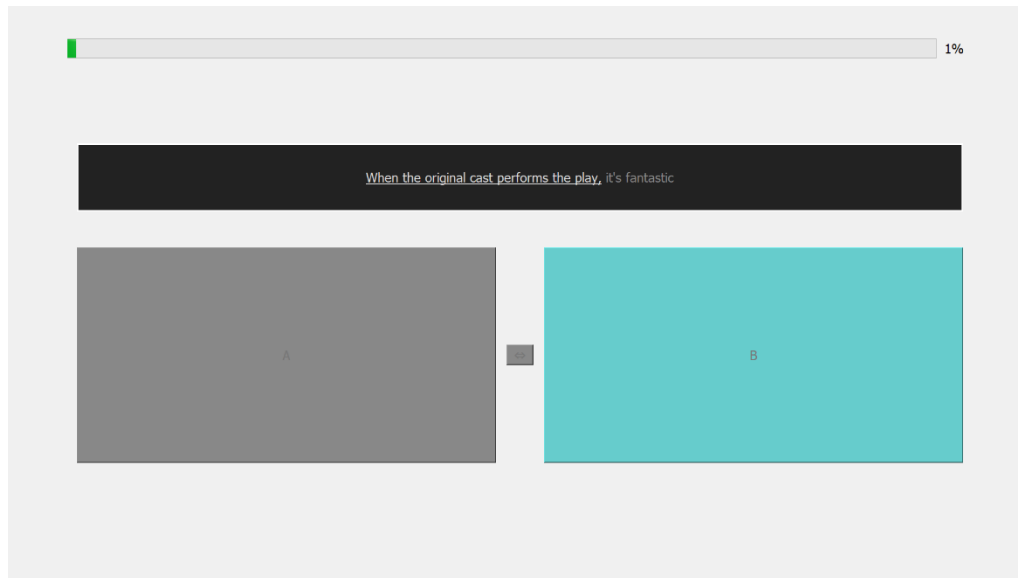


Figure 17. Phrase boundary discrimination task screenshot.

A procedure similar to Stropahl et al. (2017) was followed to measure perceptual integration. Participants complete 96 trials of a 4-AFC task wherein an auditory syllable was simultaneously rendered with an incongruent tactile syllable. Participants responded by indicating which of four randomly ordered options they heard (see Figure 18). One option matched the auditory stimulus, one matched the tactile stimulus, and the remaining two matched neither. The two options which did not match either stimulus were selected from the most-commonly selected perceptually-integrated options from the incongruent stimulus pair in the original MacDonald & McGurk study (1976) (see Table 4). Again, no feedback was provided to the participants.

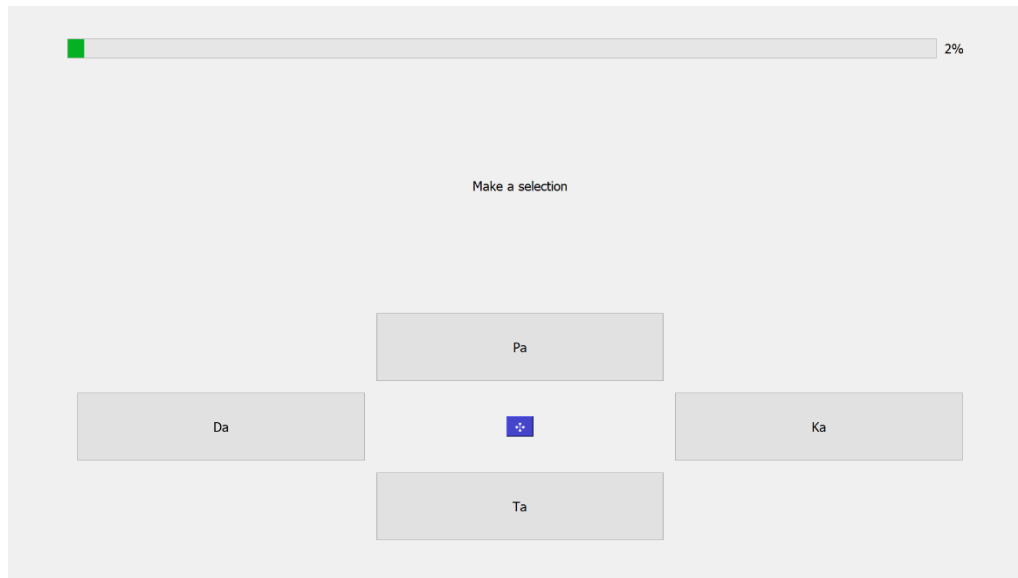


Figure 18. Perceptual integration task screenshot.

2.3.2 Training

At the conclusion of the pre-test evaluation, participants were sent home with the Vibey Transcribey device they used along with basic care and usage instructions and a flash drive containing the training software. Participants were asked to train two to four hours per week, 20–30 minutes per day, for a period of six weeks, with no more than one day of rest between training sessions.

At the start of each training session, participants slid their arms through the Vibey Transcribey sleeve, pulling it up beyond their elbow to a height that was comfortable. A logo on the wrist of the sleeve and virtual model (see Figure 19) helped ensure consistent orientation. The training software, upon launch, checked for updates and, if present, downloaded and installed them automatically. Following that, the training software established a serial connection with the participant's Vibey Transcribey device over a supplied Micro-USB cable. The software then presented the "Device Check" procedure to ensure that all actuators were working properly. During the device check, a 3D, virtual model of the participant's arm was shown on the screen with the actuators depicted in the same placement, color, and numbering scheme as the real-world device (see Figure 19). One-by-one, a virtual actuator would blink while its real-world counterpart pulsed vibrations. On-screen text instructed the

participant to ensure that the indicated virtual actuator and its real-world counterpart were indeed pulsing together, and to press a button on their gamepad to indicate a successful check and move on to the next actuator. If any actuators were not vibrating as they should, participants were instructed to contact the researcher for a repair.

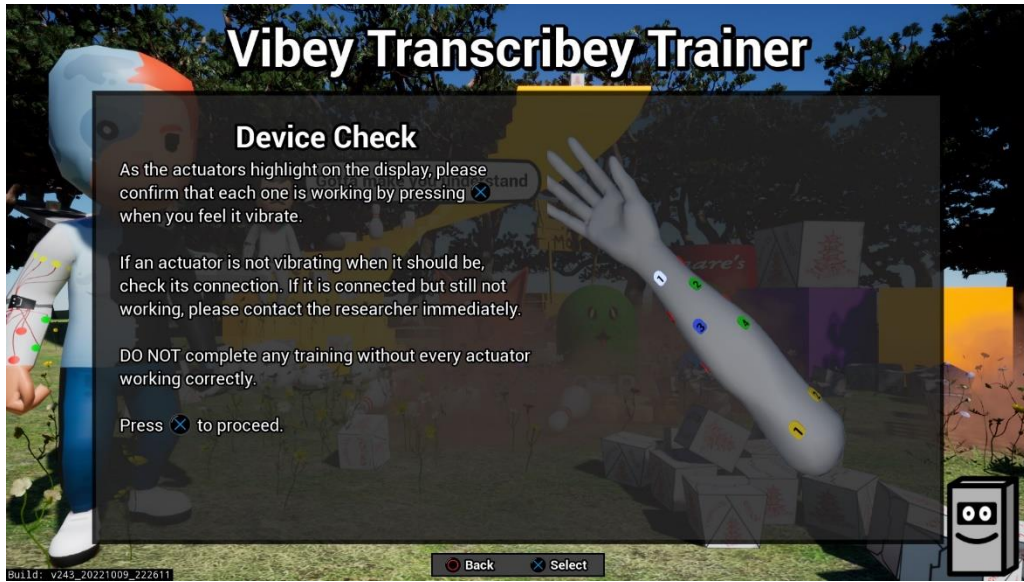


Figure 19. Device check screen.

A cartoon-styled character appeared in the bottom-right corner of every screen of the training software to provide visual feedback about the status of the device (Figure 20). When the device was connected but idle, the character displayed a smile. When the device was vibrating, its eyes and mouth changed to wavy lines. When the device was disconnected, the character became red, and its eyes became a pair of Xs.

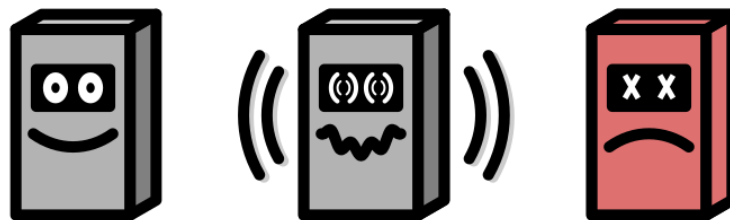


Figure 20. On-screen device status indicator.

After successfully finishing the device check, a main menu (Figure 21) was presented with the following options: Play (described below), Leaderboard (Figure 22), Display Settings (Figure 23), VT Device Settings (Figure 24), Information (Figure 25), and Quit.



Figure 21. Training software main menu.



Figure 22. Leaderboard screen.

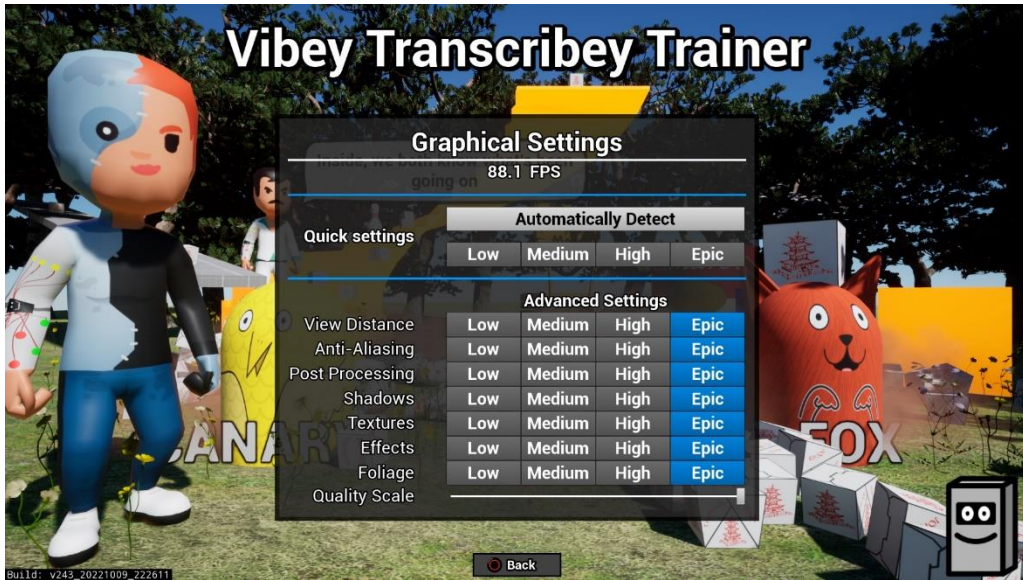


Figure 23. Display settings.



Figure 24. Advanced device settings and troubleshooting screen.



Figure 25. Information screen.

After selecting "Play" from the main menu, participants were presented with a menu where they could select one of four categories: Food, Letters, Numbers, and Animals (Figure 26). To the right of each category option was a count of the total number of "stars" the player had earned in that category. After choosing a category, the participant was then presented with a menu of levels for that category (Figure 27). Each level option included: the level name, a score multiplier, and up to three earned stars. Moving between level options updated a display on the right side of the screen which included a personal high score for the selected level.



Figure 26. Category selection screen.



Figure 27. Level select screen (food category).

When a participant selected a level, the training software loaded the game mode and map for that category and presented the user with the pre-game/pause screen (Figure 28). This screen included instructions, gamepad button layout, actions to begin/resume play and leave the level, and a menu of each of the terms that were part of that level. Using the gamepad, users could navigate to each of the

terms and select them to feel the tactile versions of each term without any of the pressures of the game in that moment.

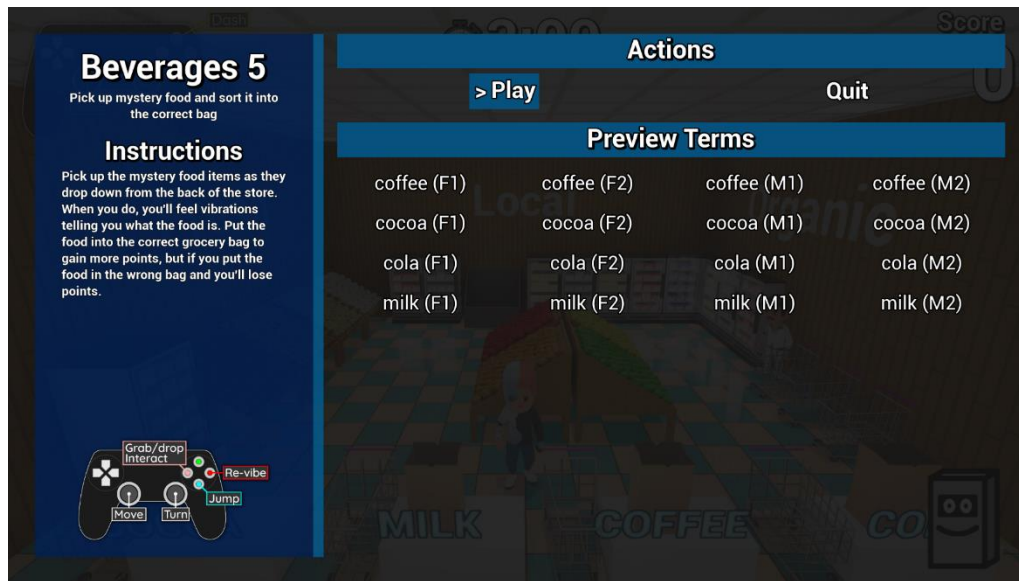


Figure 28. Pre-game and pause menu screen.

Each of the games required players to distinguish between different tactile words and react in the game-world accordingly. Each level had a three-minute timer, but players could pause the game at any time. During those three minutes, players moved a character throughout a 3D environment and interacted with some element in that environment that presented a target word. Upon interaction, they received a tactile stimulus through the Vibey Transcribey device, then they would match that stimulus to its onscreen counterpart (the target word represented as either an image or in written form), and finally they received visual and auditory feedback indicating whether their reaction was correct or not. A more detailed description of each game and player tasks are described in the following sub-sections. After receiving a tactile stimulus, one gamepad button allowed players to repeat stimulus as many times as they wanted. Stimuli were always played to completion and never interrupted.

After the level timer expired or when a player chose to quit the level from the pause menu, the "level complete" screen was displayed (Figure 29). This screen showed players how many trials they reacted to correctly, how many reactions were incorrect, their calculated score for that level, the

number of stars they earned based on that score, and their cumulative "Total Score" used in the leaderboard. Each correct response earned players 10 points, and incorrect responses cost them five, but scores were never allowed to be negative. The level complete screen also included a breakdown table indicating the number of correct and incorrect responses for each term. From this screen, players could choose to replay the level or return to the level selection menu.



Figure 29. Level complete screen.

2.3.2.1 Secret Shopper Game

The Secret Shopper game (Figure 30) featured food-related words and was set in a small grocery-store-like setting. At the far end of the grocery store, take-out boxes dropped from the ceiling into a bin. When a player moved their character close enough to a take-out box, they could press a button on the gamepad for their character to pick up the take-out box. Doing so would trigger a random word from that level's training word list to vibrate on the Vibey Transcribey device. At the opposite side of the grocery store were counters and bags, each labeled with one of the level's training words. The player would then move their character within reaching distance of the bag labelled with the word they believed matched the word they felt and would then press a button on the gamepad to cause their character to place the take-out box into that bag. If the word for the take-out box matched the bag, the

player's score would increase, an affirmative tone was played, and an animation of the take-out word (colored green) would float out of the bag. If the word did not match, the player's score was reduced (but never below zero), an explosion sound effect was played, and an animation of the take-out word (colored red) would fall out of the bag. The player repeated the process of acquiring take-out boxes and sorting them into the bags as quickly as possible, while navigating their character around level obstacles like loose shopping carts.



Figure 30. Secret Shopper screenshot.

2.3.2.2 Soop Loops Game

The Soop Loops game (Figure 31) trained participants on the names of the 26 letters of the English alphabet as their character swims around in a giant bowl of alphabet soup. A non-playable character (the "Boss") would indicate via speech bubble which letter the player should find and that all other letters should be ignored. The Boss's requested letter changed through to each of the letters in the training list at even intervals throughout the three-minute level timer. The player navigated their character around the bowl until they felt the name of a letter through the Vibey Transcribey. If they believed that letter matched what the Boss was asking for, they could press a button on the gamepad to pick up the letter. If the letter did not match, they were meant to ignore it. When the player correctly

grabbed a requested letter or correctly ignores a non-requested letter for five seconds, then the player's score was increased, a 3D noodle-version of the letter was drawn out of the soup above the character's head, and an affirmative tone was played. If the player incorrectly grabbed a letter they were meant to ignore or incorrectly ignored a letter they were meant to grab, their score was decreased (but never below zero), a green, mold-colored 3D noodle version of the letter was drawn out of the soup above the character's head, and an explosion sound effect was played. Shortly after the feedback, the noodle-letters fall into the soup and float at the surface for the remainder of the level, providing a visual representation of the player's correct and incorrect responses and creating obstacles to slow the character's movement within the soup.



Figure 31. Soop Loops screenshot.

2.3.2.3 Pin Pals Game

The Pin Pals game (Figure 32) was set in a bowling alley and trained participants on different number words. Bowling ball racks were positioned at the front-side of the bowling alley, and each was labeled with one of the level's training words. Several bowling lanes occupied the rest of the alley, with ball return mechanisms featured in the middle of the environment. Each ball return periodically produced a bowling ball which could be grabbed by the player's character. Upon doing so, a number

word associated with the ball was vibrated by the Vibey Transcribey, and the participants were then tasked with placing the ball on the rack with the corresponding number. When a ball was placed on the correct rack, the player's score was increased, an affirmative tone was played, an animation of the number word (colored green) floated up from the top of the rack, and the ball would appear on the rack. When a ball was placed on an incorrect rack, the player's score was decreased, an explosive sound effect was played, an animation of the ball's number word (colored red) fell out of the rack, and the ball disappeared from gameplay. The ball return mechanisms would eventually fill up if a player was too slow, and this caused balls to explode from the ball return and litter the gameplay area with bowling balls. To add to the gaming experience, the bowling lanes had pins positioned at the far end of the environment which could be knocked down by balls rolled by the player, with no effect on player score or other level mechanics.

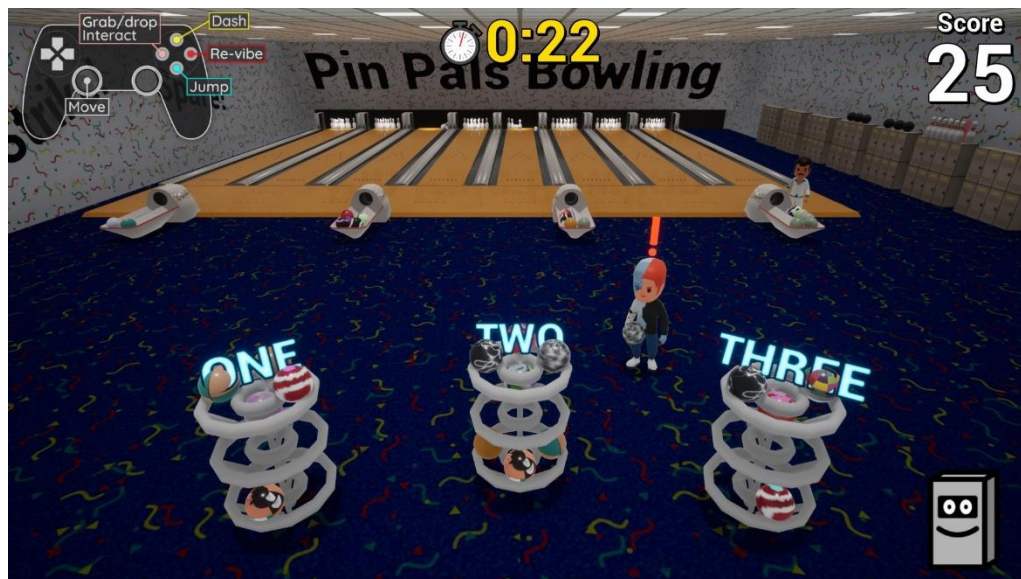


Figure 32. Pin Pals screenshot.

2.3.2.4 Snack-a-mole Game

The Snack-a-mole game (Figure 33) was set in a forest clearing and trained participants with animal names. The Boss character served up dishes of food at fast enough intervals that the player is never waiting. When a food dish was picked up by a player, an animal name was vibrated on the Vibey

Transcribey device, indicating which animal was meant to receive that dish of food. Meanwhile, different animal characters periodically sprouted from the ground in random locations briefly before digging back in. The player navigated their character to deliver food dishes to the correct animals. When a food dish was given to the correct animal, the player's score was increased, an affirmative tone was played, an animation of the animal's name (colored green) floated up from the top of the animal, and the animal briefly flashed and grew. When a food dish was given to an incorrect animal, the player's score was decreased, an explosive sound effect was played, an animation of the food dish's word (colored red) fell from the animal, and the animal would itself turn red and shake from side-to-side (as though it were shaking its head).

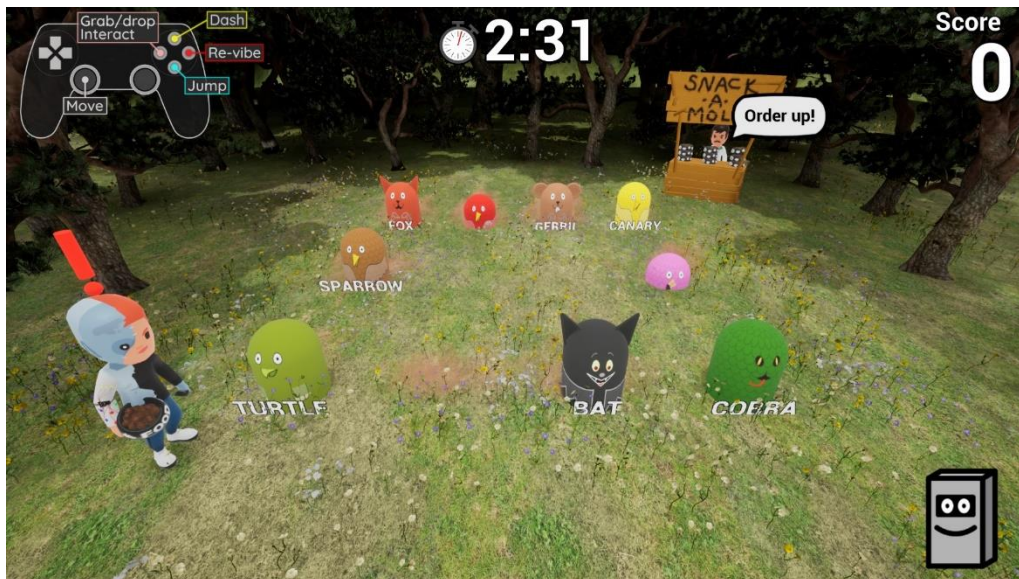


Figure 33. Snack-a-mole screenshot.

2.3.2.5 Leaderboard and Score Multipliers

From the main menu, players could access a leaderboard (Figure 22) which showed their cumulative score ranked against those of other (anonymized) experimental group participants. This leaderboard served two purposes: to motivate players who were competitively driven and to create a distribution of phoneme exposures that matched conversational English.

After completing a level, a player's cumulative score was increased by the number of points they earned in that playthrough and multiplied by an incentivizing bonus. Bonuses, shown to the player on the level select screen (Figure 27), were calculated by comparing the distribution of phonemes the player had received through the Vibey Transcribey to the distribution of phonemes in conversational English, calculated using the Buckeye Corpus of Conversational Speech (Pitt et al., 2007). When the proportion of exposures a participant received for any phoneme was lower than that phoneme's proportion in conversational English, levels containing that phoneme were given more weight. Likewise, when a phoneme was overrepresented in the training, levels containing it were given less weight. After completing any level, weights for all levels were recalculated and normalized to a range of 1.0x to 3.0x as the bonus multipliers.

CHAPTER 3

RESULTS

For each of the hypotheses tested, data were processed and analyzed using the R programming language (R Core Team, 2021) and the lme4 package for linear mixed-effects models (Bates et al., 2015). For some models, tests for significance was provided by the lmerTest package (Kuznetsova et al., 2017). The complete source code for these analyses, along with the data, are available with other project materials on the Open Science Framework.

3.1 Training

During the six-week training period, participants each completed multiple training sessions with a mean count of 13.60 ($SD = 8.62$), with each training session having a mean duration of 35.77 minutes ($SD = 17.29$). Participants played training levels between 54 and 224 times ($M = 123.60$, $SD = 78.76$), wherein they received whole-word tactile stimuli with a mean total count of 7,106.40 ($SD = 4,990.89$). Each of those stimuli were composed of multiple phonemes, such that the participants received a mean total count of 26,521.40 ($SD = 18,486.95$) tactile phonemes. Details of individuals' training reports is available in Table 5. One device stopped working in the first week and was replaced with a spare. All of the devices were returned at the end of the training period in working order, with no apparent signs of wear or damage.

PID	Training Sessions	Mean Session Time (m)	Total Training (m)	Level Plays	Phrase Exposures	Unique Phrases	Phoneme Exposures
0	7	29.08	203.56	62	2,966	62	10,787
1	19	37.44	711.36	192	10,229	74	35,273
2	26	34.56	898.56	224	14,324	137	54,963
3	10	37.71	377.10	86	4,852	67	18,725
4	6	40.32	241.92	54	3,161	58	13,359

Table 5. Individual training reports.

A Chi-square goodness-of-fit test was performed to determine whether the distribution of phoneme exposures that experimental group participants received was representative of conversational

English (Figure 34). The test showed a significant difference between distributions $\chi^2(37, N = 162,120) = 37,364, p < .01$, indicating that the distribution of phonemes participants trained on did not match the distribution of phonemes in conversational English.

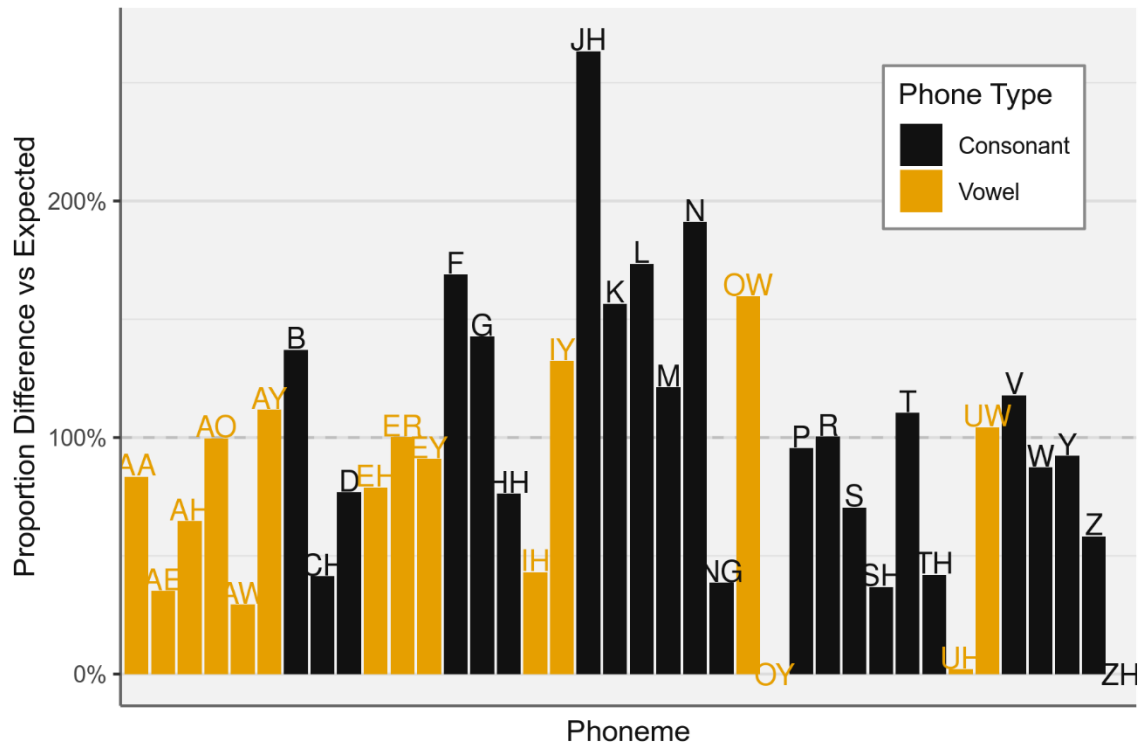


Figure 34. Prevalence of phonemes during training versus conversation.

To determine whether participants' performance improved as their training progressed (Figure 35), a series of mixed-effects models were constructed and evaluated for fit. The portion of correct trials during a level attempt was used as the dependent variable, and a normalized representation of the individual's attempts (progression) on each level was used as the only fixed effect. The record representing a participant's first attempt at a level had a progression value of 0.0. The record representing that participant's final attempt at that same level had a progression value of 1.0. Each of the participant's attempts between their first and last have progression values equally distributed between 0.0 and 1.0 accordingly. Levels that were not attempted at least twice by a participant were not considered in this dataset. The models were weighted using the trial count for each level

playthrough, as the number of trials in each level attempt varied on the game and player speed. Additionally, players could exit a level at any time—even after only a single trial.

Constructing the models by sequentially increasing complexity by adding terms showed that random effects for participant and level (nested within game) were meaningful contributors to the final, best-fitting model, which showed that a participant's continued attempts at each level yielded an improvement of 15.28% from their first attempt to their last attempt ($p < .01$).

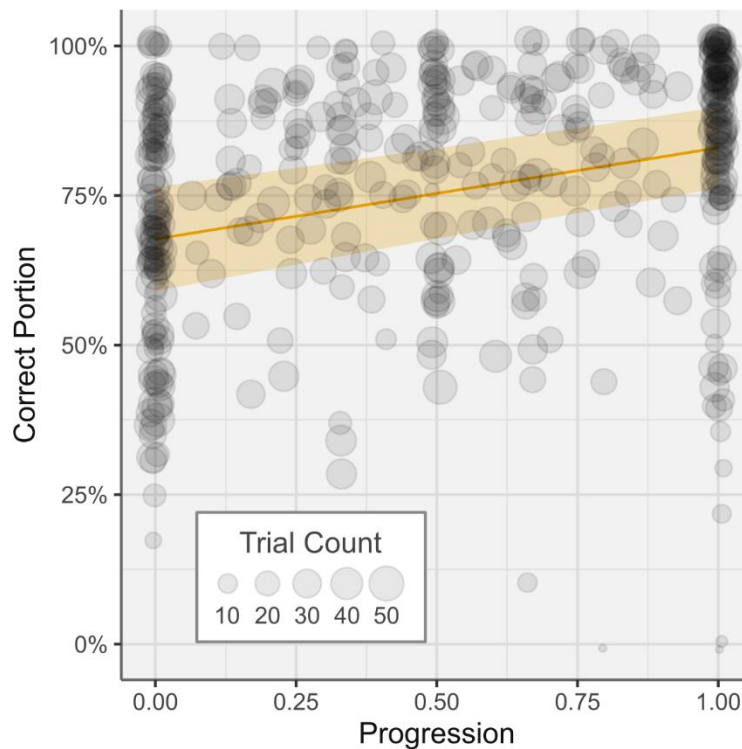


Figure 35. Training performance progression. Each point represents a level attempt.

3.2 Phoneme Perception

To test for an effect of training on phoneme perception (Figure 36), logistic generalized linear mixed models were constructed with trial correctness as the dependent variable, and fixed effects for testing session (pre- vs post-training) and group (control vs experimental) plus the interaction. Participant ID and phoneme (nested within phoneme type [vowel vs consonant]) were modelled as random factors to account for repeated measures/non-independent data. A comparison of AICs

between models showed that participant ID was not a meaningful contributor to the model, but phonemes were. The final model did not show a significant interaction between fixed effects ($p > .05$).

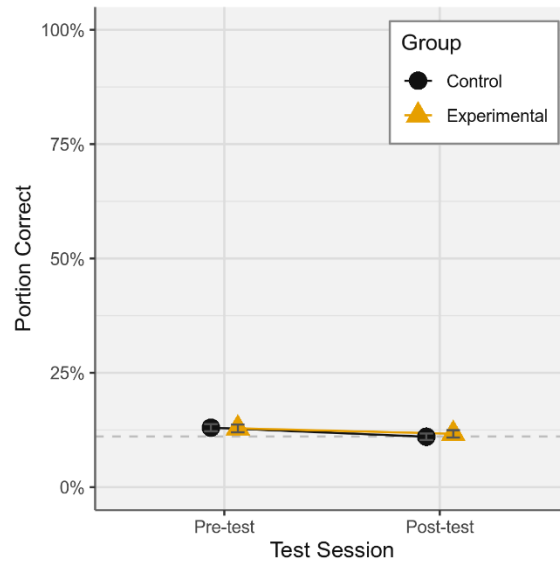


Figure 36. Phoneme identification before and after training.

To determine whether the amount of training exposures to each phoneme had an effect on phoneme perception (Figure 37), another logistic generalized linear mixed model was fit using only post-test data for the experimental group. In this model, the amount of training each participant received (the min-max normalized count of the number of exposures the participant received to the tested phoneme during training) was used as a fixed effect, while phoneme and participant ID were once again used as random effects. Training amount did not improve model fitness versus a basic random-intercepts model, indicating that no effect of training could be identified.

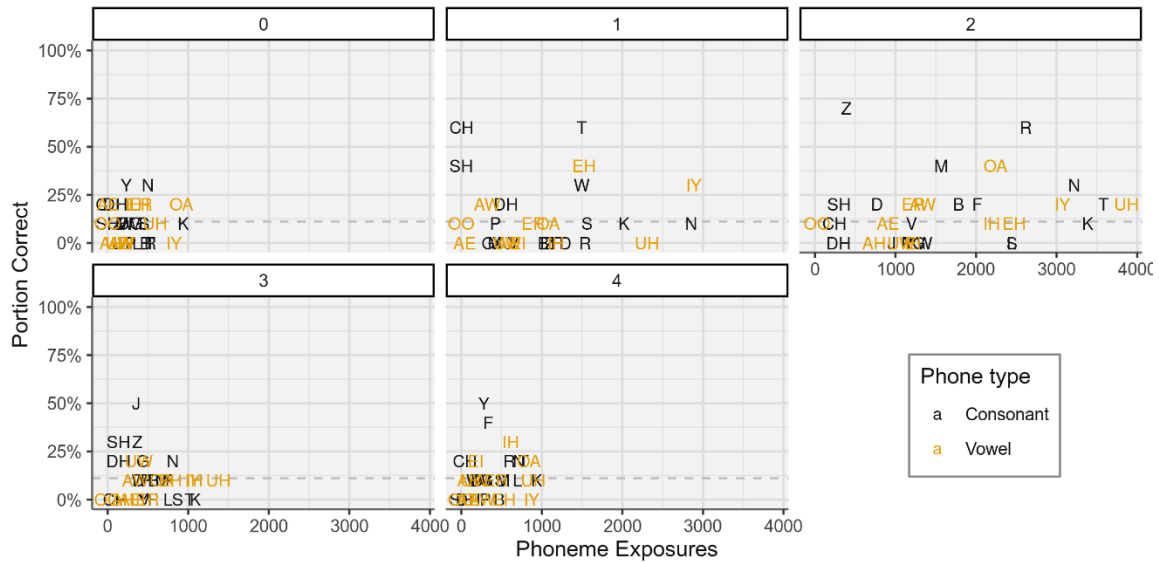


Figure 37. Training effect on post-test phoneme identification.

3.3 Prosody Perception

3.3.1 Word Focus Matching

To test for an effect of training on participants' ability to match focused words in phrases (Figure 38), a logistic generalized mixed effects model was used again. Trial correctness was used as the dependent variable, with group assignment and testing session as the fixed effects plus their interaction. Random effects for the stimulus and participant ID were modeled to account for repeated measures/non-independent data but were found to not improve model fitness. None of the fixed effects for group assignment, testing session, or their interaction were found to be significant effects ($p > .05$).

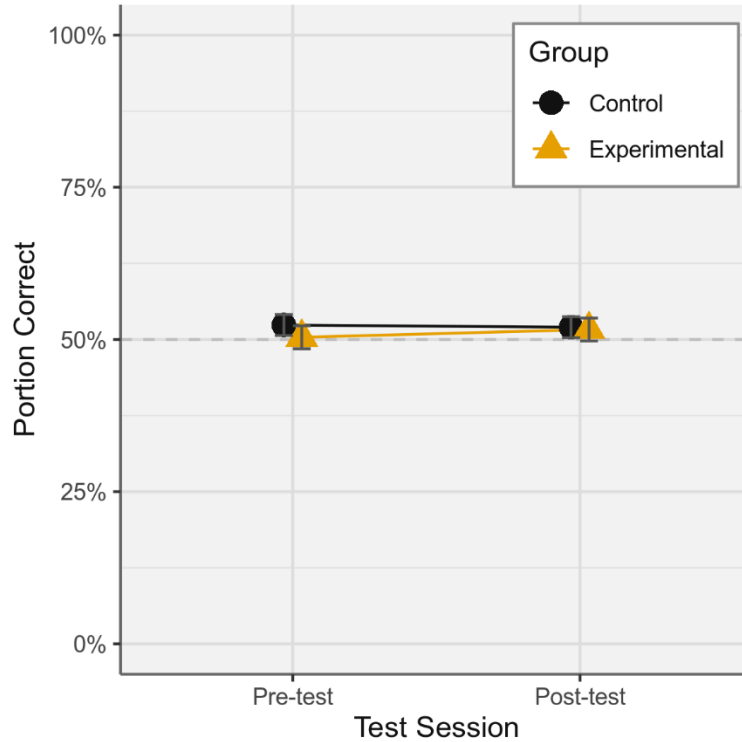


Figure 38. Word focus matching before and after training.

To test if the amount of training received between experimental-group participants could account for differences in their post-test performance (Figure 39), another logistic mixed-model was used. Trial correctness was again used as the dependent variable, with a count of the training word exposures used as the single fixed effect. Random effects for the stimuli and participant were modelled to account for repeated measures/non-independent data but did not improve model fitness. Training did not have a significant effect ($p > .05$).

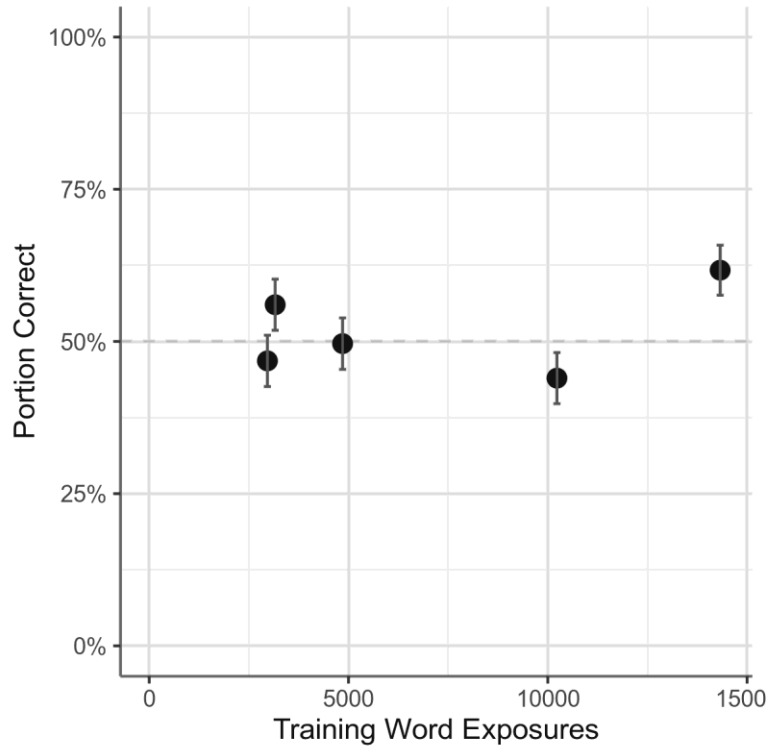


Figure 39. Training effect on post-test focus matching.

3.3.2 Phrase Boundary Matching

To test for an effect of training on participants' ability to match phrases based on phrase boundaries (Figure 40), a logistic generalized mixed effects model was fit again with trial correctness as the dependent variable and fixed effects for group assignment and training session, plus their interaction. Random effects for the stimuli and participant ID were modelled to account for repeated measures and non-independent data, but they did not improve model fitness. No significant effects were found for group assignment, training session, or their interaction.

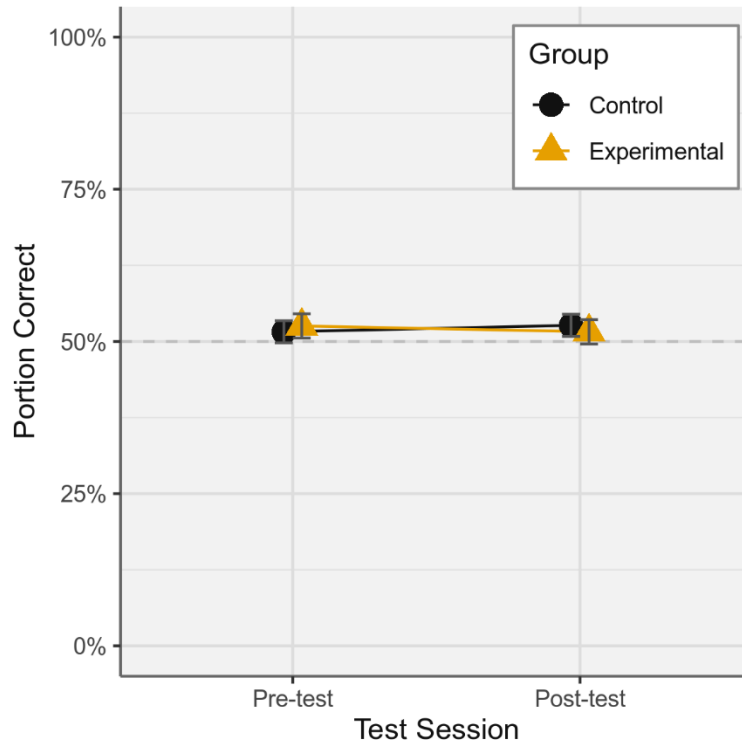


Figure 40. Phrase boundary matching before and after training.

To test if the amount of training received between experimental-group participants could account for differences in their post-test performance on phrase boundary matching (Figure 41), another logistic mixed-model was used. Trial correctness was again used as the dependent variable, with a count of the training word exposures used as the single fixed effect. Random effects for the stimuli and participant were modelled to account for repeated measures/non-independent data but did not improve model fitness. Training did not have a significant effect ($p > .05$).

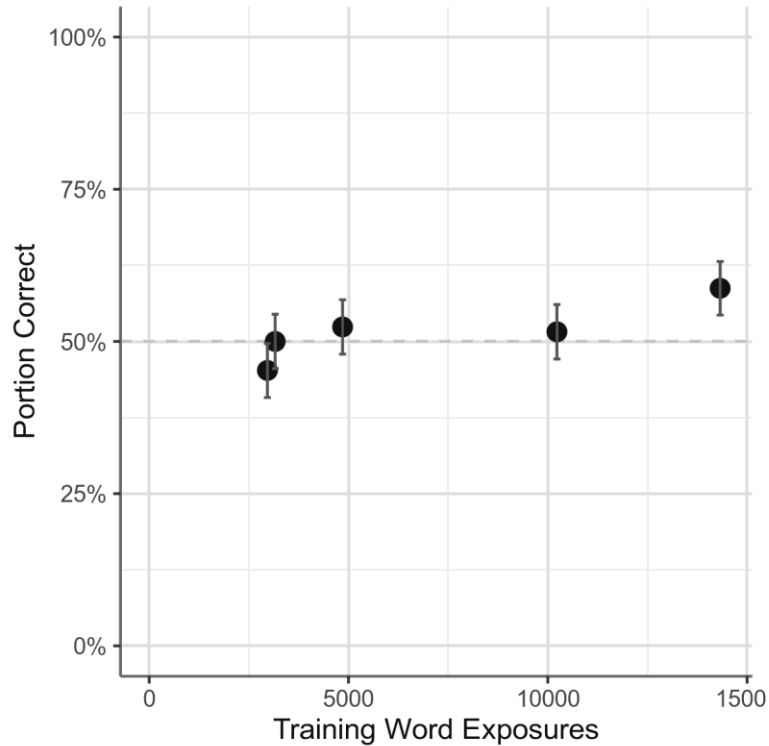


Figure 41. Training effect on post-test phrase boundary matching.

3.4 Perceptual Integration

Responses to the perceptual integration task were noisier than expected, and an analysis of pre-training responses revealed that three auditory stimuli were correctly identified on less than half of the trials (*m_Pa_TK07.wav*, *f_Pa_TK03.wav*, and *f_Ma_TK02.wav*). All trials in both pre- and post-training data using one of these stimuli were removed from the dataset and any further analysis.

To evaluate whether perceptual integration increased due to training (Figure 42), another logistic generalized mixed effects model was used. The stimuli for a trial were considered integrated when a participant selected neither of the auditory or tactile stimuli which were presented to them. This served as the dependent variable, with fixed effects for testing session and group assignment. Random effects were modelled for the tactile stimuli, auditory stimuli, and participant to account for repeated measures and non-independent data, but the random effect for the auditory stimuli did not contribute

to model fitness. Group assignment, testing session, and their interaction were all shown to not be significant factors ($p > .05$).

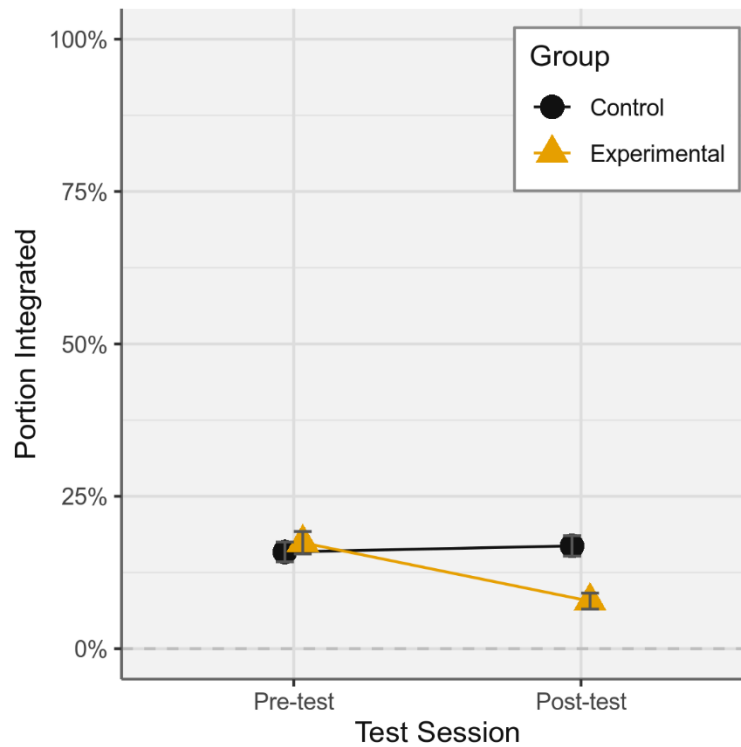


Figure 42. Perceptual integration before and after training.

To test if the amount of training received between experimental-group participants could account for differences in their perceptual integration results after training (Figure 43), another logistic mixed-model was used. Trial correctness was again used as the dependent variable, with a normalized count of number of exposures of each tactile-consonant as the single fixed effect. Random effects for the auditory stimuli, tactile stimuli, and participant were modelled to account for repeated measures/non-independent data, but only participant improved model fitness. Training did not have a significant effect ($p > .05$).

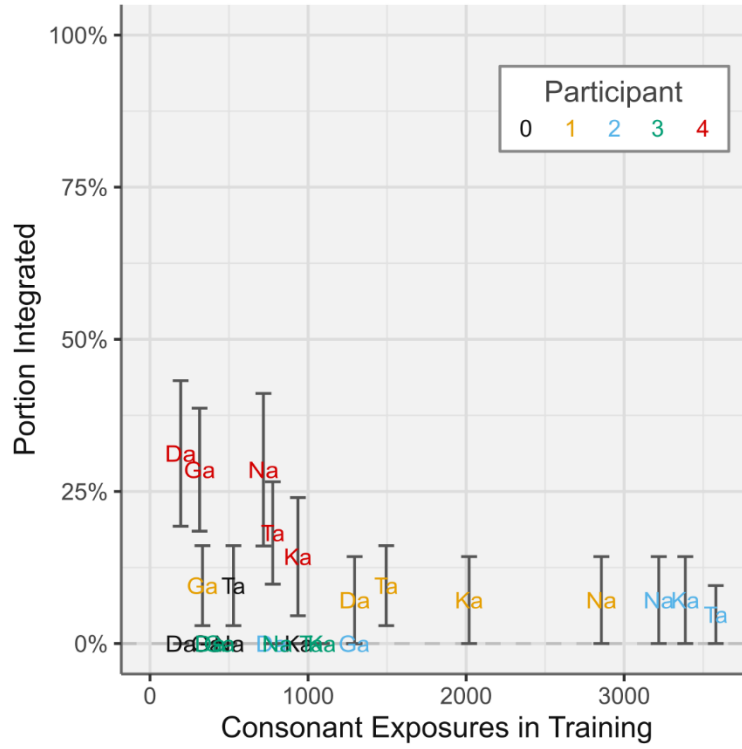


Figure 43. Training effect on post-test perceptual integration.

CHAPTER 4

DISCUSSION

The work presented here demonstrated the successful development of a tactile-speech device which rendered a complete set of English phonemes and elements of prosody, where previous work in this domain generally included only a very limited set of phonemes or some elements of prosody, but never both. If true speech perception is possible in the tactile domain, a demonstration of it is only valid if it conveys both, and any study which purports to achieve true tactile-speech perception without exhibiting perception of both phonemes and prosody is incomplete at best. Further, the inclusion of a test for perceptual integration can only bolster claims for new sensory and perceptual experiences. In this study, participants did not improve in their ability to identify tactile phonemes, did not improve in their ability tactile perception of prosodic speech features, and did not demonstrate any auditory-tactile speech integration.

However, data from participants training logs did demonstrate that an ability to discriminate tactically-encoded words can be learned through gameplay, even if the training was unable to induce speech-like perception using the Vibey Transcribey encoding scheme and device. A more generous evaluation could have been performed wherein participants were asked to discriminate among the same stimuli they were trained on. Although the results from such an evaluation may have been more encouraging, it would do little to advance the goal of speech perception in non-auditory modalities—arguably, such results would even be detrimental to the science. Rather than push encouraging but misleading results, this study showed that training participants on whole words and whole words alone is not a promising method to develop a reliable ability to discriminate phonemes in this configuration. This training method further failed to demonstrate that pitch and temporal cues could be perceived as prosodic features despite being mapped directly. Given these findings, it may be unsurprising that no

cross-modal perceptual integration could be detected, however, an inability to lip-read does not preclude the McGurk effect in auditory-visual speech perception.

High-quality stimuli that illicit reliable responses in single modality identification are an essential part of a valid study. Here, the auditory stimuli provided by Stropahl et al. (2017) may be found by some to be unreliable, although their willingness to create and share these stimuli is commendable.

Indeed, finding stimuli that is appropriate for a study's needs, high-quality, acquirable, and permitted for use can be a significant burden. Some journals make web-hosting available to researchers who wish to share their stimuli, but many journals do not, and of those that do, many authors do not participate. Some authors make their stimuli available on university-hosted servers or on their own personal websites, although these sometimes require a fee for use. Sometimes authors will respond positively to correspondences and requests for stimuli, but addresses change, and people retire. Likewise, recreating methods of previous work can be a challenge when researchers rely solely on a written description of the procedure. Certainly, a good scientist seeks to describe their method clearly, but there are unavoidable ambiguities in the written word that, in some cases, could be easily clarified by simply providing software source code from the study.

Like experimental stimuli, the software necessary to replicate or expand an experiment can be difficult or impossible to acquire and use. To avoid the scavenger hunt, many researchers find themselves creating their own stimuli and software, often at significant cost and diminished quality. Even worse, the fruit of that labor is often reduced to mere written descriptions, continuing the cycle rather than being freely shared. The upward progress of science has always relied on researchers building on the work and ideas of each other, and modern scientists do a great disservice to their fields by not making materials accessible, especially when digital assets like stimuli and source code are so easily shared.

To that end, the method and materials used in this study, provided freely, may contribute to future research in this or related domain. A method for extracting and encoding pitch, loudness, and phonemic information from an auditory speech signal is provided along with an implementation built entirely on other free and open-source software. A spatial map of English phonemes is provided and can be easily adapted to other parts of the body. A circuit diagram, its accompanying low-cost, two-layer, compact PCB schematic, and firmware source code are provided for a microcontroller platform supporting USB, Bluetooth, and Wi-Fi data to build a multi-channel LRA array. Source code and digital assets for the training software and evaluation software is also made available. Perhaps most importantly, the author hopes to further contribute to the ideas that:

1. Inclusive designs are worthy ventures that benefit all people
2. Technology can enable more accessible communication
3. Tactile-speech devices and their evaluation should include phonemes and prosody
4. Gamification can make for more enjoyable and engaging learning studies
5. Freely sharing methods, materials, designs, and source code empowers future work

4.1 Limitations

Perhaps the biggest limitation of the Vibey Transcribey as a tactile-speech device was its inability to render speech as it happens. The presented method relies on the availability of a plain-text transcript of the speech, and current tools for automated speech recognition are much too slow for the tactile translation to occur without significant, perceivable latency. The Vibey Transcribey was also limited to encoding and displaying only a single phoneme at a time, despite the prevalence of coarticulated phonemes in auditory speech.

The Vibey Transcribey's phonemic array relied heavily on the funneling effect to distinguish between different phonemes. The forearm has relatively low spatial and temporal sensitivities when compared to other regions of the body, like the hands and face, although these more sensitive areas

lack practicality. An inconsistent or unreliable funneling effect coupled with low tactile sensitivity may account for some poor performance. One participant reported experiencing an unintended illusory effect between the pitch actuators on the bicep and phoneme actuators on the forearm near the elbow, despite these actuators being spaced beyond the limits of previously reported funneling effect proximity. Alternative designs may consider using more actuators to eliminate the need for the funneling effect and having greater separation between the phoneme and pitch actuators.

Alternative designs might also consider different actuators. In this study, LRAs were selected for their balance between cost and responsiveness. One drawback of LRAs is that they can only operate at a single frequency. Although this frequency is optimized for tactile perception, more sophisticated actuators can respond to complex waveforms, creating unique sensations that change the quality of the sensation. This could be used as an additional display dimension. For example, the Vibey Transcribey encodes phonemes spatially on the forearm while also encoding pitch spatially on the bicep. While both spatial mappings have been demonstrated in previous work, using them together was novel and, potentially, confusing for users. With more sophisticated actuators, pitch could be mapped to the frequency of the actuators, for example, instead of being mapped spatially.

The forearm is a common body site for tactile-speech display but can be a bit cumbersome for a user to apply themselves, as the arm that would wear the device cannot be used to apply it. Further, slippage that occurs when the arm is rotated or moved may cause unreliable actuator placements. The current study used knit-fabric, athletic-style sleeves to create a more uniform spacing of actuators without the need for custom tailoring. Despite their marketing, the sleeves were not one-size-fits-all, and one participant asked for an accommodation after experiencing discomfort on their bicep where the sleeve was too tight. For this participant, the pitch array was relocated to a separate, adjustable band (Figure 44).

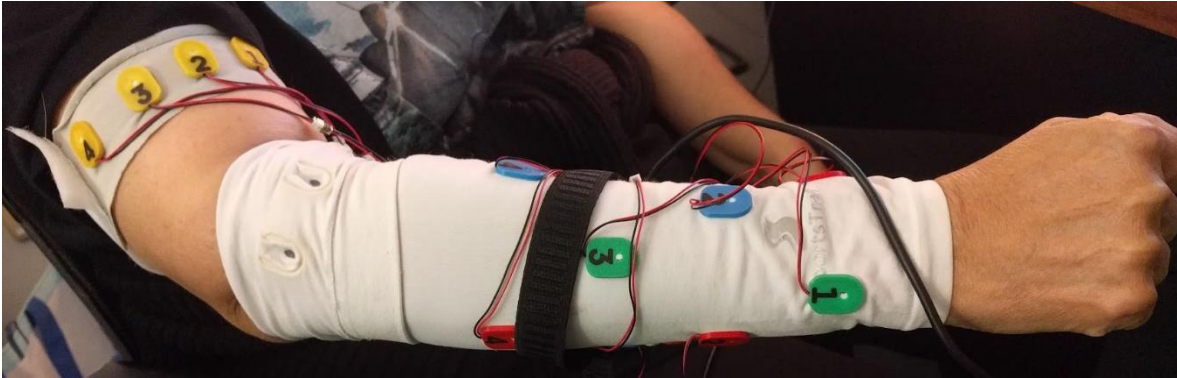


Figure 44. A Vibey Transcribey with a modified pitch array band.

Participants in the experimental group tracked their own training time to achieve the desired amount of training, which was two to four hours per week. However, only one participant met the minimum, but part of the discrepancy was likely due to poor communication. The training software logged session time from the first stimulus to the last stimulus, which only occurred while participants were actively playing a training game. Participants, however, counted training time from the moment they began donning the Vibey Transcribey until it was removed and the software closed. An in-game display, like a simple progress bar, could have been used to show their progress in real-time. An in-game tutorial system would also benefit the training experience. Some participants indicated that they did not understand how to play the Soop Loops game at first, and it took them a while to figure it out. At least one participant did not know that you could pause the game and preview all the terms. Another did not know that there was a dedicated button on the controller to re-activate the last vibrated word.

Evaluations in this study had participants identify phonemes and differentiate phrases, but participants were trained only on whole-words in a "middle-out" learning paradigm, rather than either top-down or bottom-up exclusively. One potential improvement to the training process would be granting users some level of control over the speed of the stimuli as they learn, perhaps even allowing them to self-regulate their top-down or bottom-up learning approach. Similarly, the prosodic evaluations presented participants with entire clauses, despite them having only been trained on whole

words. Although important prosodic cues like pitch, loudness, and time were all present in the training words, they were presented without potentially important contextual cues that may be necessary for learning to identify emphasis or phrase boundaries.

4.2 Future directions

The encoding scheme, cue mappings, hardware platform, training methods, and evaluation methods from this research can each be built upon for further exploration and study. This version of the Vibey Transcribey encoded vowels and consonants separately, building on the idea that these phonemes are the building-blocks of spoken languages. However, it would be possible to encode speech where the syllable is the building block, with consonants serving as the root of the unit and vowels as secondary components, not unlike abugidas. This idea lends itself nicely to spatiotemporal tactile gestures. For example, a consonant can be mapped to specific places on the skin, and any following vowel sounds could be felt as moving sensations emanating from the consonant location, with different directions corresponding to different vowels. With or without new encoding schemes, new phonemic mappings may be more easily learned and understood than what was presented in this study. Additionally, developing these mappings could be considered a dimension reduction problem, and one could apply dimension reduction techniques from machine learning contexts to create new mappings for tactile-speech.

This study explored the use of gamified training, but the games were a little simple and very similar to each other. Games in other genres may yield better training discipline, especially if players can choose from a variety that meets their preferences. Role-playing games often include significant amounts of dialog with other characters and narrative, which may be an especially avenue for tactile-speech perception training, like what was accomplished by Martinez (2019). The best training, of course, will be the one that participants diligently complete, so it may be important that we examine other forms of entertainment for training purposes. Books, movies, music, personal communications, etc. are

all potential candidate mediums. If acquiring tactile-speech perception works similarly to acquiring auditory speech perception, it may be important or even necessary that it be a primary method for communication, and one that even happens passively. A real-time system could facilitate training in many activities of daily life.

The tactile arrays of the Vibey Transcribey device are not limited to phonemes, pitch, and loudness. The hexagonal grid of the phonemic array can be used to render virtually any two-dimensional data source. Likewise, the pitch-array can be repurposed for one-dimensional data rendering. This flexibility gives the Vibey Transcribey the potential to be used in a variety of research. Navigational cues, for example, can be provided without any need for hearing or sight. Positional and gyroscopic feedback can be provided in real-time to improve a golfer's swing. The possibilities are bountiful.

4.3 Conclusion

Despite being unable to produce results that demonstrate tactile-speech perception, the work presented here contributes to the growing body of literature on vibrotactile learning. It seems that the exact configuration of speech encoding, tactile display, and training strategy presented here may be ruled out for viability, but each component, variations on those components, and innumerable combinations thereof can provide a basis for future studies in tactile communications, inclusive design, non-auditory speech perception, and perceptual learning. All the necessary materials and tools that were created in pursuance of this work is made available on the Open Science Framework at <https://osf.io/3f5nu/>.

REFERENCES

REFERENCES

- Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 8(10), 457–464. <https://doi.org/10.1016/j.tics.2004.08.011>
- Altes, R. A. (1989). Ubiquity of hyperacuity. *The Journal of the Acoustical Society of America*, 85(2), 943–952. <https://doi.org/10.1121/1.397566>
- Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentals, Prosody, and Syllable Structure. *Language Learning*, 42(4), 529–555. <https://doi.org/10.1111/j.1467-1770.1992.tb01043.x>
- Anderson-Hsieh, J., & Koehler, K. (1988). The Effect of Foreign Accent and Speaking Rate on Native Speaker Comprehension*. *Language Learning*, 38(4), 561–613. <https://doi.org/10.1111/j.1467-1770.1988.tb00167.x>
- Angel Sound—Interactive Listening Rehabilitation and Functional Hearing Test Program*. (n.d.). Retrieved September 9, 2021, from <http://angelsound.tigerspeech.com/>
- Ardailon, L., & Roebel, A. (2019, September). Fully-Convolutional Network for Pitch Estimation of Speech Signals. *Interspeech 2019*. <https://doi.org/10.21437/Interspeech.2019-2815>
- Bach-y-Rita, P. (1967). Sensory Plasticity. *Acta Neurologica Scandinavica*, 43(4), 417–426. <https://doi.org/10.1111/j.1600-0404.1967.tb05747.x>
- Bach-y-Rita, P., Collins, C. C., Saunders, F. A., White, B., & Scadden, L. (1969). Vision Substitution by Tactile Image Projection. *Nature*, 221(5184), Article 5184. <https://doi.org/10.1038/221963a0>
- Bagwell, C. (n.d.). *SoX - Sound eXchange* (14.4.2). Retrieved August 9, 2021, from <http://sox.sourceforge.net/>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Using lme4. *J. Stat. Softw.*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Blamey, P. J., & Clark, G. M. (1985). A wearable multiple-electrode electrotactile speech processor for the profoundly deaf. *The Journal of the Acoustical Society of America*, 77(4), 1619–1620. <https://doi.org/10.1121/1.392009>
- Boost C++ Libraries*. (n.d.). Retrieved December 1, 2022, from <https://www.boost.org/>
- Bourne, R. R. A., Flaxman, S. R., Braithwaite, T., Cicinelli, M. V., Das, A., Jonas, J. B., Keeffe, J., Kempen, J. H., Leasher, J., Limburg, H., Naidoo, K., Pesudovs, K., Resnikoff, S., Silvester, A., Stevens, G. A., Tahhan, N., Wong, T. Y., Taylor, H. R., Bourne, R., ... Zheng, Y. (2017). Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: A systematic review and meta-analysis. *The Lancet Global Health*, 5(9), e888–e897. [https://doi.org/10.1016/S2214-109X\(17\)30293-0](https://doi.org/10.1016/S2214-109X(17)30293-0)

- Brooks, P. L., & Frost, B. J. (1983). Evaluation of a tactile vocoder for word recognition. *The Journal of the Acoustical Society of America*, 74(1), 34–39. <https://doi.org/10.1121/1.389685>
- Brown, D. J., Simpson, A. J. R., & Proulx, M. J. (2014). Visual Objects in the Auditory System in Sensory Substitution: How Much Information Do We Need? *Multisensory Research*, 27(5–6), 337–357. <https://doi.org/10.1163/22134808-00002462>
- Byers, A., & Serences, J. T. (2012). Exploring the relationship between perceptual learning and top-down attentional control. *Vision Research*, 74, 30–39. <https://doi.org/10.1016/j.visres.2012.07.008>
- Canare, D. (n.d.). *Psydekick-Science-Kit/Psydekick at 0b62344130640a39ccf4f13bb1ca2e2c719fa348*. Retrieved December 1, 2022, from <https://github.com/Psydekick-Science-Kit/Psydekick/tree/0b62344130640a39ccf4f13bb1ca2e2c719fa348>
- Celesia, G. G. (2010). Visual Perception and Awareness. *Journal of Psychophysiology*, 24(2), 62–67. <https://doi.org/10.1027/0269-8803/a000014>
- Cha, J., Rahal, L., & El Saddik, A. (2008). A pilot study on simulating continuous sensation with two vibrating motors. *2008 IEEE International Workshop on Haptic Audio Visual Environments and Games*, 143–147. <https://doi.org/10.1109/HAVE.2008.4685314>
- Choi, J.-Y., Hasegawa-Johnson, M., & Cole, J. (2005). Finding intonational boundaries using acoustic cues related to the voice source. *The Journal of the Acoustical Society of America*, 118(4), 2579–2587. <https://doi.org/10.1121/1.2010288>
- Cholewiak, R. W., & Collins, A. A. (2003). Vibrotactile localization on the arm: Effects of place, space, and age. *Perception & Psychophysics*, 65(7), 1058–1077. <https://doi.org/10.3758/BF03194834>
- Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers. *Journal of Speech, Language, and Hearing Research : JSLHR*, 57(4), 1468–1479. https://doi.org/10.1044/2014_JSLHR-L-13-0279
- Convention on the Rights of Persons with Disabilities, Res 61/106, U.N. General Assembly, U.N. GAOR (2006). <https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html>
- Corniani, G., & Saal, H. P. (2020). Tactile innervation densities across the whole body. *Journal of Neurophysiology*, 124(4), 1229–1240. <https://doi.org/10.1152/jn.00313.2020>
- Craig, J. C. (1972). Difference threshold for intensity of tactile stimuli. *Perception & Psychophysics*, 11(2), 150–152. <https://doi.org/10.3758/BF03210362>
- Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. *Language and Speech*, 40(2), 141–201. <https://doi.org/10.1177/002383099704000203>
- Darwin, C. J. (1975). On the Dynamic Use of Prosody in Speech Perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and Process in Speech Perception* (pp. 178–194). Springer. https://doi.org/10.1007/978-3-642-81000-8_11

- de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America*, *96*(4), 2037–2047. <https://doi.org/10.1121/1.410145>
- Derwing, T. M., Munro, M. J., & Wiebe, G. (1998). Evidence in Favor of a Broad Framework for Pronunciation Instruction. *Language Learning*, *48*(3), 393–410. <https://doi.org/10.1111/0023-8333.00047>
- Deveau, J., Lovcik, G., & Seitz, A. R. (2014). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research*, *99*, 134–140. <https://doi.org/10.1016/j.visres.2013.12.015>
- Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech Perception. *Annual Review of Psychology*, *55*(1), 149–179. <https://doi.org/10.1146/annurev.psych.55.090902.142028>
- Eckert, K. A., Carter, M. J., Lansingh, V. C., Wilson, D. A., Furtado, J. M., Frick, K. D., & Resnikoff, S. (2015). A Simple Method for Estimating the Economic Cost of Productivity Loss Due to Blindness and Moderate to Severe Visual Impairment. *Ophthalmic Epidemiology*, *22*(5), 349–355. <https://doi.org/10.3109/09286586.2015.1066394>
- Eilers Rebecca E., Cobo-Lewis Alan B., Vergara Kathleen C., Oller D. Kimbrough, & Friedman Karen E. (1996). A Longitudinal Evaluation of the Speech Perception Capabilities of Children Using Multichannel Tactile Vocoders. *Journal of Speech, Language, and Hearing Research*, *39*(3), 518–533. <https://doi.org/10.1044/jshr.3903.518>
- Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech Perception in Infants. *Science*, *171*(3968), 303–306. <https://doi.org/10.1126/science.171.3968.303>
- Elbert, T., Pantev, C., Wienbruch, C., Rockstroh, B., & Taub, E. (1995). Increased Cortical Representation of the Fingers of the Left Hand in String Players. *Science*, *270*(5234), 305–307. <https://doi.org/10.1126/science.270.5234.305>
- Ellis, E. M., & Robinson, A. J. (1993). *A Phonetic Tactile Speech Listening System*. 17.
- Epstein, W. (1961). The influence of syntactical structure on learning. *The American Journal of Psychology*, *74*(1), 80–85.
- Flores, J. F. F. (2015). Using Gamification to Enhance Second Language Learning. *Digital Education Review*, *27*, 32–54.
- Fry, D. B. (1958). Experiments in the Perception of Stress. *Language and Speech*, *1*(2), 126–152. <https://doi.org/10.1177/002383095800100207>
- Fu, Q.-J., Galvin, J., Wang, X., & Nogaki, G. (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. *Acoustics Research Letters Online*, *6*(3), 106–111. <https://doi.org/10.1121/1.1898345>
- Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377. <https://doi.org/10.3758/BF03193857>

- Gault, R. H. (1924). Progress in experiments on tactual interpretation of oral speech. *The Journal of Abnormal Psychology and Social Psychology*, 19(2), 155–159. <https://doi.org/10.1037/h0065752>
- Gault, R. H. (1927). “Hearing” through the sense organs of touch and vibration. *Journal of the Franklin Institute*, 204(3), 329–358. [https://doi.org/10.1016/S0016-0032\(27\)92101-2](https://doi.org/10.1016/S0016-0032(27)92101-2)
- Gault, R. H., & Crane, G. W. (1928). Tactual Patterns from Certain Vowel Qualities Instrumentally Communicated from a Speaker to a Subject’s Fingers. *The Journal of General Psychology*, 1(2), 353–359. <https://doi.org/10.1080/00221309.1928.9920129>
- Geldard, F. A. (1957). Adventures in tactile literacy. *American Psychologist*, 12(3), 115–124. <https://doi.org/10.1037/h0040416>
- Geldard, F. A., & Sherrick, C. E. (1972). The Cutaneous “Rabbit”: A Perceptual Illusion. *Science*, 178(4057), 178–179. <https://doi.org/10.1126/science.178.4057.178>
- Gescheider, G. A. (1965). Cutaneous Sound Localization. *Journal of Experimental Psychology*, 70(6), 617–625.
- Gibson, E. J. (1969). *Principles of perceptual learning and development*. Appleton-Century-Crofts.
- Godde, B., Stauffenberg, B., Spengler, F., & Dinse, H. R. (2000). Tactile Coactivation-Induced Changes in Spatial Discrimination Performance. *The Journal of Neuroscience*, 20(4), 1597–1604. <https://doi.org/10.1523/JNEUROSCI.20-04-01597.2000>
- Goldreich, D. (2007). A Bayesian Perceptual Model Replicates the Cutaneous Rabbit and Other Tactile Spatiotemporal Illusions. *PLoS ONE*, 2(3). <https://doi.org/10.1371/journal.pone.0000333>
- Goldstone, R. L. (1998). Perceptual Learning. *Annual Review of Psychology*, 49(1), 585–612. <https://doi.org/10.1146/annurev.psych.49.1.585>
- Grant, K. W., Ardell, L. H., Kuhl, P. K., & Sparks, D. W. (1985). The contribution of fundamental frequency, amplitude envelope, and voicing duration cues to speechreading in normal-hearing subjects. *The Journal of the Acoustical Society of America*, 77(2), 671–677. <https://doi.org/10.1121/1.392335>
- Green Forge Coop. (2021). *MOSQUITO* (0.3.4). <https://doi.org/10.5281/zenodo.5639403>
- Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. *2014 47th Hawaii International Conference on System Sciences*, 3025–3034. <https://doi.org/10.1109/HICSS.2014.377>
- Hayward, V. (2008). A brief taxonomy of tactile illusions and demonstrations that can be done in a hardware store. *Brain Research Bulletin*, 75(6), 742–752. <https://doi.org/10.1016/j.brainresbull.2008.01.008>
- Helson, H., & King, S. M. (1931). The tau effect: An example of psychological relativity. *Journal of Experimental Psychology*, 14(3), 202–217. <https://doi.org/10.1037/h0071164>

- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. <https://doi.org/10.1121/1.411872>
- Hodzic, A., Veit, R., Karim, A. A., Erb, M., & Godde, B. (2004). Improvement and Decline in Tactile Discrimination Behavior after Cortical Plasticity Induced by Passive Tactile Coactivation. *Journal of Neuroscience*, *24*(2), 442–446. <https://doi.org/10.1523/JNEUROSCI.3731-03.2004>
- Hormann, M. C. (2022). *Loudgain* [C]. <https://github.com/Moonbase59/loudgain> (Original work published 2019)
- Hoshiyama, M., Kakigi, R., & Tamura, Y. (2004). Temporal discrimination threshold on various parts of the body. *Muscle & Nerve*, *29*(2), 243–247. <https://doi.org/10.1002/mus.10532>
- International Phonetic Association. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press. <https://go.exlibris.link/10Nsqk6D>
- Jasmin, K., Dick, F., & Tierney, A. T. (2020). The Multidimensional Battery of Prosody Perception (MBOPP). *Wellcome Open Research*, *5*, 4. <https://doi.org/10.12688/wellcomeopenres.15607.1>
- Jiménez, J., Olea, J., Torres, J., Alonso, I., Harder, D., & Fischer, K. (2009). Biography of Louis Braille and Invention of the Braille Alphabet. *Survey of Ophthalmology*, *54*(1), 142–149. <https://doi.org/10.1016/j.survophthal.2008.10.006>
- Jones, L. A., & Sarter, N. B. (2008). Tactile Displays: Guidance for Their Design and Application. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(1), 90–111. <https://doi.org/10.1518/001872008X250638>
- Kellman, P. J. (2002). Perceptual Learning. In *Stevens' Handbook of Experimental Psychology* (pp. 259–299). John Wiley & Sons Inc. <https://doi.org/10.1002/0471214426.pas0307>
- Kintsch, A., & DePaula, R. (2002). A Framework for the Adoption of Assistive Technology. *SWAAAC 2002: Supporting Learning through Assistive Technology*, 1–10.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. <https://doi.org/10.18637/jss.v082.i13>
- Lacruz, F., Artieda, J., Pastor, M. A., & Obeso, J. A. (1991). The anatomical basis of somaesthetic temporal discrimination in humans. *Journal of Neurology, Neurosurgery & Psychiatry*, *54*(12), 1077–1081. <https://doi.org/10.1136/jnnp.54.12.1077>
- Lane, H. (1965). The motor theory of speech perception: A critical review. *Psychological Review*, *72*(4), 275–309. <https://doi.org/10.1037/h0021986>
- Lehiste, I., Olive, J. P., & Streeter, L. A. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *The Journal of the Acoustical Society of America*, *60*(5), 1199–1202. <https://doi.org/10.1121/1.381180>

- Leonard, L. B. (1973). The Role of Intonation in the Recall of Various Linguistic Stimuli". *Language and Speech*, 16(4), 327–335. <https://doi.org/10.1177/002383097301600403>
- Lieberman, A. M. (1957). Some Results of Research on Speech Perception. *The Journal of the Acoustical Society of America*, 29(1), 117–123. <https://doi.org/10.1121/1.1908635>
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431–461. <https://doi.org/10.1037/h0020279>
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1968). Why Are Speech Spectrograms Hard to Read? *American Annals of the Deaf*, 113(2), 127–133.
- Loomis, J. M. (1979). An investigation of tactile hyperacuity. *Sensory Processes*, 3(4), 289–302.
- Loui, P., Guenther, F. H., Mathys, C., & Schlaug, G. (2008). Action–perception mismatch in tone-deafness. *Current Biology*, 18(8), R331–R332. <https://doi.org/10.1016/j.cub.2008.02.045>
- Lu, S. A., Wickens, C. D., Sarter, N. B., & Sebok, A. (2011). Informing the Design of Multimodal Displays: A Meta-Analysis of Empirical Studies Comparing Auditory and Tactile Interruptions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 55(1), 1170–1174. <https://doi.org/10.1177/1071181311551244>
- Mahns, D. A., Perkins, N. M., Sahai, V., Robinson, L., & Rowe, M. J. (2006). Vibrotactile Frequency Discrimination in Human Hairy Skin. *Journal of Neurophysiology*, 95(3), 1442–1450. <https://doi.org/10.1152/jn.00483.2005>
- Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, 10(1), 29–63. [https://doi.org/10.1016/0010-0285\(78\)90018-X](https://doi.org/10.1016/0010-0285(78)90018-X)
- Martin, J. G. (1968). Temporal word spacing and the perception of ordinary, anomalous, and scrambled strings. *Journal of Verbal Learning and Verbal Behavior*, 7(1), 154–157. [https://doi.org/10.1016/S0022-5371\(68\)80181-1](https://doi.org/10.1016/S0022-5371(68)80181-1)
- Martinez, J. S. (2019). *Tactile Speech Communication: Design and Evaluation of Haptic Codes for Phonemes with Game-based Learning* [Thesis, Purdue University Graduate School]. <https://doi.org/10.25394/PGS.8026403.v1>
- Massaro, D. W., & Palmer, S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Mit Press.
- Mattys, S. L. (2000). The perception of primary and secondary stress in English. *Perception & Psychophysics*, 62(2), 253–265. <https://doi.org/10.3758/BF03205547>
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *INTERSPEECH*. <https://doi.org/10.21437/INTERSPEECH.2017-1386>

- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. [https://doi.org/10.1016/0010-0285\(86\)90015-0](https://doi.org/10.1016/0010-0285(86)90015-0)
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746. <https://doi.org/10.1038/264746a0>
- Meister, I. G., Wilson, S. M., Deblieck, C., Wu, A. D., & Iacoboni, M. (2007). The Essential Role of Premotor Cortex in Speech Perception. *Current Biology : CB*, *17*(19), 1692–1696. <https://doi.org/10.1016/j.cub.2007.08.064>
- Mitsuhashi, N., Fujieda, K., Tamura, T., Kawamoto, S., Takagi, T., & Okubo, K. (2009). BodyParts3D: 3D structure database for anatomical concepts. *Nucleic Acids Research*, *37*(suppl_1), D782–D785. <https://doi.org/10.1093/nar/gkn613>
- Moraru, D., & Boiangiu, C.-A. (2015). Seeing without eyes: Visual sensory substitution. *Journal of Information Systems & Operations Management*, *13*.
- Novich, S. D. (2015). *Sound-to-touch sensory substitution and beyond* [PhD Thesis].
- O’Connell, D. C., Turner, E. A., & ONUSKA, L. A. (1968). Intonation, Grammatical Structure, and Contextual Association in Immediate Recall. *Journal of Verbal Learning and Verbal Behavior; New York*, *7*(1), 110–116.
- Ohala, J. J. (1996). Speech perception is hearing sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1718–1725. <https://doi.org/10.1121/1.414696>
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>
- Parette, P., & Scherer, M. (2004). Assistive Technology Use and Stigma. *Education and Training in Developmental Disabilities*, *39*(3), 217–226.
- Pascual-Leone, A., Amedi, A., Fregni, F., & Merabet, L. B. (2005). The Plastic Human Brain Cortex. *Annual Review of Neuroscience*, *28*(1), 377–401. <https://doi.org/10.1146/annurev.neuro.27.070203.144216>
- Phillips, B., & Zhao, H. (1993). Predictors of Assistive Technology Abandonment. *Assistive Technology*, *5*(1), 36–45. <https://doi.org/10.1080/10400435.1993.10132205>
- Pielot, M., Poppinga, B., & Boll, S. (2010). PocketNavigator: Vibro-tactile waypoint navigation for everyday mobile devices. *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services*, 423–426. <https://doi.org/10.1145/1851600.1851696>
- Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye corpus of conversational speech (2nd release). *Columbus, OH: Department of Psychology, Ohio State University*. <https://buckeyecorpus.osu.edu/>

- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Reed, C. M., Rabinowitz, W. M., Durlach, N. I., Braid, L. D., Conway-Fithian, S., & Schultz, M. C. (1985). Research on the Tadoma method of speech communication. *The Journal of the Acoustical Society of America*, *77*(1), 247–257. <https://doi.org/10.1121/1.392266>
- Reed, C. M., Tan, H. Z., Perez, Z. D., Wilson, E. C., Severgnini, F. M., Jung, J., Martinez, J. S., Jiao, Y., Israr, A., Lau, F., Klumb, K., Turcott, R., & Abnoui, F. (2019). A Phonemic-Based Tactile Display for Speech Communication. *IEEE Transactions on Haptics*, *12*(1), 2–17. <https://doi.org/10.1109/TOH.2018.2861010>
- Richardson, B. L., & Frost, B. J. (1977). Sensory Substitution and the Design of an Artificial Ear. *The Journal of Psychology*, *96*(2), 259–285. <https://doi.org/10.1080/00223980.1977.9915910>
- Rizza, A., Terekhov, A. V., Montone, G., Olivetti-Belardinelli, M., & O'Regan, J. K. (2018). Why Early Tactile Speech Aids May Have Failed: No Perceptual Integration of Tactile and Auditory Signals. *Frontiers in Psychology*, *9*. <https://doi.org/10.3389/fpsyg.2018.00767>
- Rothenberg, M., Verrillo, R. T., Zahorian, S. A., Brachman, M. L., & Bolanowski, S. J. (1977). Vibrotactile frequency for encoding a speech parameter. *The Journal of the Acoustical Society of America*, *62*(4), 1003–1012. <https://doi.org/10.1121/1.381610>
- Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Old Infants. *Science*, *274*(5294), 1926–1928. <https://doi.org/10.1126/science.274.5294.1926>
- Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior*, *69*, 371–380. <https://doi.org/10.1016/j.chb.2016.12.033>
- Sailer, M., & Homner, L. (2020). The Gamification of Learning: A Meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. <https://doi.org/10.1007/s10648-019-09498-w>
- Saunders, F. A., & Franklin, B. (1985). Field tests of a wearable 16-channel electrotactile sensory aid in a classroom for the deaf. *The Journal of the Acoustical Society of America*, *78*(S1), S17–S17.
- Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. <https://doi.org/10.1037/0033-295X.84.1.1>
- Schumann, F., & O'Regan, J. K. (2017). Sensory augmentation: Integration of an auditory compass signal into human perception of space. *Scientific Reports*, *7*(1), 42197. <https://doi.org/10.1038/srep42197>
- Shannon, R. V., Jensvold, A., Padilla, M., Robert, M. E., & Wang, X. (1999). Consonant recordings for speech testing. *The Journal of the Acoustical Society of America*, *106*(6), L71–L74. <https://doi.org/10.1121/1.428150>

- Stevens, S. S., Volkman, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3), 185–190.
<https://doi.org/10.1121/1.1915893>
- Story, M. F. (2011). Principles of universal design. In *Universal Design Handbook* (2nd ed., p. 4.3-4.12). McGraw-Hill Professional.
- Strasburger, H., Huber, J., & Rose, D. (2018). Ewald Hering's (1899) On the Limits of Visual Acuity: A Translation and Commentary. *I-Perception*, 9(3), 2041669518763675.
<https://doi.org/10.1177/2041669518763675>
- Striem-Amit, E., Cohen, L., Dehaene, S., & Amedi, A. (2012). Reading with Sounds: Sensory Substitution Selectively Activates the Visual Word Form Area in the Blind. *Neuron*, 76(3), 640–652.
<https://doi.org/10.1016/j.neuron.2012.08.026>
- Stropahl, M., Schellhardt, S., & Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychonomic Bulletin & Review*, 24(3), 863–872. <https://doi.org/10.3758/s13423-016-1148-9>
- Summers, I. R., & Gratton, D. A. (1995). Choice of speech features for tactile presentation to the profoundly deaf. *IEEE Transactions on Rehabilitation Engineering*, 3(1), 117–121.
<https://doi.org/10.1109/86.372901>
- Summers, I. R., Milnes, P., Cooper, P. G., & Stevens, J. C. (1996). Coding of acoustic features for a single-channel tactile aid. *British Journal of Audiology*, 30(4), 238–248.
<https://doi.org/10.3109/03005369609076771>
- Szeto, A. Y. J., & Christensen, K. M. (1988). Technological devices for deaf-blind children: Needs and potential impact. *IEEE Engineering in Medicine and Biology Magazine*, 7(3), 25–29.
<https://doi.org/10.1109/51.7931>
- Tong, J., Mao, O., & Goldreich, D. (2013). Two-Point Orientation Discrimination Versus the Traditional Two-Point Test for Tactile Spatial Acuity Assessment. *Frontiers in Human Neuroscience*, 7.
<https://doi.org/10.3389/fnhum.2013.00579>
- Tsao, F.-M., Liu, H.-M., & Kuhl, P. K. (2004). Speech Perception in Infancy Predicts Language Development in the Second Year of Life: A Longitudinal Study. *Child Development*, 75(4), 1067–1084. <https://doi.org/10.1111/j.1467-8624.2004.00726.x>
- Turcott, R., Chen, J., Castillo, P., Knott, B., Setiawan, W., Briggs, F., Klumb, K., Abnoui, F., Chakka, P., Lau, F., & Israr, A. (2018). Efficient Evaluation of Coding Strategies for Transcutaneous Language Communication. In D. Prattichizzo, H. Shinoda, H. Z. Tan, E. Ruffaldi, & A. Frisoli (Eds.), *Haptics: Science, Technology, and Applications* (pp. 600–611). Springer International Publishing.
https://doi.org/10.1007/978-3-319-93399-3_51
- Tyler, M., Danilov, Y., & Bach-y-Rita, P. (2003). Closing an open-loop control system: Vestibular substitution through the tongue. *Journal of Integrative Neuroscience*.
<https://doi.org/10.1142/S0219635203000263>

- UnrealEngine 4.25.4*. (n.d.). Retrieved December 1, 2022, from <https://github.com/epicgames/unrealengine/tree/4.25.4-release>
- v. Békésy, G. (1957). Sensations on the Skin Similar to Directional Hearing, Beats, and Harmonics of the Ear. *The Journal of the Acoustical Society of America*, *29*(4), 489–501. <https://doi.org/10.1121/1.1908938>
- v. Békésy, G. (1958). Funneling in the Nervous System and its Role in Loudness and Sensation Intensity on the Skin. *The Journal of the Acoustical Society of America*, *30*(5), 399–412. <https://doi.org/10.1121/1.1909626>
- Verrillo, R. T., Fraioli, A. J., & Smith, R. L. (1969). Sensation magnitude of vibrotactile stimuli. *Perception & Psychophysics*, *6*(6), 366–372. <https://doi.org/10.3758/BF03212793>
- Visell, Y. (2009). Tactile sensory substitution: Models for enactment in HCI. *Interacting with Computers*, *21*(1–2), 38–53. <https://doi.org/10.1016/j.intcom.2008.08.004>
- Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature*, *413*(6858), 844–848. <https://doi.org/10.1038/35101601>
- Watkins, K. E., Strafella, A. P., & Paus, T. (2003). Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia*, *41*(8), 989–994. [https://doi.org/10.1016/S0028-3932\(02\)00316-0](https://doi.org/10.1016/S0028-3932(02)00316-0)
- Weisenberger, J. M., & Percy, M. E. (1995). The Transmission of Phoneme-Level Information by Multichannel Tactile Speech Perception Aids: *Ear and Hearing*, *16*(4), 392–406. <https://doi.org/10.1097/00003446-199508000-00006>
- Weisenberger, J. M., & Russell, A. F. (1989). Comparison of Two Single-Channel Vibrotactile Aids for the Hearing-Impaired. *Journal of Speech, Language, and Hearing Research*, *32*(1), 83–92. <https://doi.org/10.1044/jshr.3201.83>
- Werker, J. F., Yeung, H. H., & Yoshida, K. A. (2012). How Do Infants Become Experts at Native-Speech Perception? *Current Directions in Psychological Science*, *21*(4), 221–226. <https://doi.org/10.1177/0963721412449459>
- Wickens, C. D. (1980). The structure of attentional resources. *Attention and Performance VIII*, *8*, 239–257.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. <https://doi.org/10.1080/14639220210123806>
- Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, *7*(7), 701–702. <https://doi.org/10.1038/nn1263>
- World Health Organization. (2011, December). *World Report on Disability*. <https://www.who.int/publications/i/item/9789241564182>

World Health Organization. (2017). *Global costs of unaddressed hearing loss and cost-effectiveness of interventions*. <http://apps.who.int/iris/bitstream/10665/254659/1/9789241512046-eng.pdf>

World Health Organization. (2021a, February). *Blindness and vision impairment*. <https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment>

World Health Organization. (2021b, March). *World Report on Hearing*. <https://www.who.int/publications/i/item/world-report-on-hearing>

Yaacob, M., Worthington, H. V., Deacon, S. A., Deery, C., Walmsley, A. D., Robinson, P. G., & Glenny, A. (2014). Powered versus manual toothbrushing for oral health. *The Cochrane Database of Systematic Reviews*, 2014(6), CD002281. <https://doi.org/10.1002/14651858.CD002281.pub3>

Yantis, S., & Abrams, R. A. (2014). *Sensation and perception*. Worth Publishers New York, NY.

Yuan, H., Reed, C. M., & Durlach, N. I. (2005). Tactual display of consonant voicing as a supplement to lipreading. *The Journal of the Acoustical Society of America*, 118(2), 1003–1015.

Zelek, J. S., Bromley, S., Asmar, D., & Thompson, D. (2003). A haptic glove as a tactile-vision sensory substitution for wayfinding. *Journal of Visual Impairment and Blindness*, 97(10), 621–632.

Zhao, S., Israr, A., Lau, F., & Abnoui, F. (2018). Coding Tactile Symbols for Phonemic Communication. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, 1–13. <https://doi.org/10.1145/3173574.3173966>

Zwicker, E., & Fastl, H. (2013). *Psychoacoustics: Facts and Models*. Springer Science & Business Media.

Zwicker, E., Fastl, H., Widmann, U., Kurakata, K., Kuwano, S., & Namba, S. (1991). Program for calculating loudness according to DIN 45631 (ISO 532B). *Journal of the Acoustical Society of Japan (E)*, 12(1), 39–42. <https://doi.org/10.1250/ast.12.39>