Enhancing the Perception of Speech Indexical Properties of Cochlear Implants through Sensory Substitution

by

Austin McRae Butts

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Biomedical Engineering

Approved July 2015 by the
Graduate Supervisory Committee:

Stephen Helms Tillery, Chair
Visar Berisha
Christopher Buneo
Troy McDaniel

ARIZONA STATE UNIVERSITY

August 2015

ABSTRACT

Through decades of clinical progress, cochlear implants have brought the world of speech and language to thousands of profoundly deaf patients. However, the technology has many possible areas for improvement, including providing information of non-linguistic cues, also called indexical properties of speech. The field of sensory substitution, providing information relating one sense to another, offers a potential avenue to further assist those with cochlear implants, in addition to the promise they hold for those without existing aids. A user study with a vibrotactile device is evaluated to exhibit the effectiveness of this approach in an auditory gender discrimination task. Additionally, preliminary computational work is included that demonstrates advantages and limitations encountered when expanding the complexity of future implementations.

I dedicate this to my family,

without whom I would not be where I am today.

Thank you all for your love and support.

ACKNOWLEDGMENTS

Funding and Resources:

This work was supported in part by the ASU Graduate and Professional Student Association (GPSA) Jumpstart Research Grant for the Spring 2015 Semester.

This work was supported in part by the ASU Center for Cognitive Ubiquitous Computing (CUbiC).

Personal Assistance:

The author wishes to thank S. Bala for his help in planning and conducting the human studies. The author also thanks the members of his supervisory committee for their guidance and insights.

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

I. BACKGROUND

A. Introduction to the Problem

1. History of Cochlear Implants

Hearing loss is one of the most common sensory disorders in the world. As of the year 2000, over 248 million people suffer from moderate (41+ dB) reductions in their hearing thresholds, with over 53 million having severe loss (61+ dB) [1]. At least 14 million people have profound hearing loss (81+ dB), enough to be classified as clinically deaf. Hearing loss and deafness are prevalent in both children and adults, with different etiologies. Total financial burden imposed due to severe hearing loss or greater is estimated at $297,000 a year per person on average, mostly due to loss of productivity and special education for child-onset patients [2]. Additionally, total disability impact is estimated to be 27.4 million in total years lost due to disability (YLD) [3]. While any amount of hearing loss can have a detrimental impact on quality of life, profound deafness has the largest toll, completely isolating individuals from their sound environment and impairing their access to conventional oral communication.

The most successful biomedical device to assist the deaf is the cochlear implant. It has been clinically indicated for a variety of conditions which lead to deafness, including presbycusis (age-related hearing loss), some inherited diseases, trauma, noise-induced hearing loss, and some complications of contagious diseases [4]. Since the approval of the first iteration of the cochlear implant by the FDA in 1984, it has seen widespread usage. Even over the last decade, the number of patients using a cochlear implant as risen dramatically. Older statistics from 2008 estimate the total number of recipients to be over 120k persons worldwide [5], and more recent figures from 2010 place the number at

219k [6] and 324k in December of 2010 [7]. Of the latest estimates, 58 thousand adults and 38 thousand children in the US have received an implant. Annual revenue for the industry is estimated to be in the billions of dollars [5].

Cochlear implants are most often the only viable treatment for profound hearing loss. Hearing aids are often prescribed to assist the hearing impaired up to severe disability, but the profoundly deaf generally cannot be helped by hearing aids. Even with its currently wide popularity among possible treatment options, the history of the cochlear implant began with modest in clinical successes. The relatively simple single-electrode devices were typically limited to aid lip-reading and provide awareness of sound in a patient's environment. In the early years, they could not provide open speech recognition independent of visual cues [5]. However, beginning in the 1990s, more sophisticated designs and algorithms enabled open-set speech recognition. This level of performance supports conversational speech with minimal interruption [5]. This is cited as the most major milestone thus far in cochlear implant research, and is a testament of the potential to meaningfully improve the lives of patients with biomedical implants.

2. Shortcomings of Cochlear Implants

Despite the hard-fought successes, many aspects of hearing remain challenging to cochlear implant patients. Reception of the linguistic aspects of speech often rely on a clear listening environment, one of the reasons why the highest levels of performance are often reported under "speech in quiet" conditions. Interfering noise [5] often reduces intelligibility below the conversational threshold, even at levels that would only be moderately distracting to normal hearing persons. Normal hearing listeners typically perform around 90% for vowel, consonant, and sentence scores in noise with a signal-to-

noise ratio of 0dB [8]. In contrast, cochlear implant patients typically perform below 50% for vowel and consonant sets [8] and exhibit similar performance in open-set word recognition [9]. Noise of this magnitude also drops sentence level recognition to marginal levels, barely performing above 0% for most users [8]. The issue with interfering noise also extends to filtering out the voices of multiple extraneous speakers. In those with normal hearing, the "cocktail party effect" [10] enables a listener to isolate the speech of one person with surrounding speech noise. Cochlear implant patients usually do not have this ability to the same extent, often suffering greatly in comparison.

Some languages provide other challenges that do not fare well in speech supported by a cochlear implant. Languages such as Chinese rely on tones, or specific inflections in pitch, to convey semantic meaning [5]. Pitch detection is cited [11] as a problem for cochlear implant patients, and not surprisingly, the difficulties in honing in on these cues carry over to this aspect of language.

Electric hearing has also proven to be inferior in aesthetic aspects of hearing, such as music perception [5]. Although there is some normalization in the chronically implanted, patients self report a significant drop in the quality of the music they listen to and its impact [11]. Cochlear implant patients have trouble identifying well-known songs by melody and rhythm compared to the nearly perfect performance of normal hearing subjects [12]. Music perception relies primary on auditory cues, as demonstrated to be subject to most confusion in a congruency task with tactile cues [13]. Some proposed solutions include controlling the listening environment, being selective about the music chosen, and using a contralateral HA when possible [11]. These ideas are mostly non-

technological, and those that are do not involve improvements in the design of the cochlear implant.

a. Indexical Properties

Most of the focus in improving the deficits of cochlear implants has related to linguistic aspects of speech. However, fundamental research in speech and hearing also investigates what are called indexical properties [14]. These qualities expand the notion of speech beyond the meaning of the language of the message. They can be more rigorously defined in abstract terms, but for a simple explanation, they pertain to the properties of the speaker. Indexical qualities are often categorized into subsets: group membership, speaker characteristics, and changes in the state of the speaker [14].

In the absence of visual cues, cochlear implant patients often have difficulty discriminating or identifying the speaker of sample utterances. In a study of pediatric listeners with a two-answer forced-choice (2AFC) paradigm [15], listeners could distinguish speakers of the same sex with 67% accuracy using the same sentences. The performance dropped further to 57% correct with different sentences. These performance levels are above chance, but indicate the task is incredibly difficult with unfamiliar speakers, especially in contrast to an 89% performance level in normal hearing subjects. Further studies have examined the role of frequency parameters, specifically the fundamental frequency (F0) and formant frequencies (F1 and F2), to highlight disparity in performance [16]. Researchers used a single sample and speech resynthesis to keep all but the selected parameters consistent. Most cochlear implant patients had difficulty with the task for much of the range of parameters, whereas normal hearing subjects could detect differences between speakers with as little as 2-2.5 semitone difference on average.

However, the effects on identifying are softened by familiarity, such as being the mother of the child cochlear implant patient, but not to normal hearing levels [17]. This familiarity effect can also be stymied by the alternative speaker imitating the mother's speech patterns and inflections, also called prosody in the literature. One additional effect not noted in the prior source, but apparent from the collected data, is a tendency for patients to identify competing woman's voice as their mother's over the opposite case. This bias also appears to be stronger in the imitation condition.

Correctly discriminating the gender of the speaker from auditory information alone is relatively difficult for cochlear implant patients. For most normal hearing persons, this drop in performance is often difficult to understand, as most people are thought to perform at ceiling in their daily utility of this skill. Of a sample of 11 patients, cochlear implant users ranged from 70% to near-ceiling, with a median around 90%, in a stimulus set of vowel tokens [18]. When the contrasts in F0 are reduced between the two groups by combining high-pitched male and low-pitched female speakers, the success rate drops further. In a study of 10 patients, all users performed between 60% and 80% accuracy with a median below 70% [19]. This drop in performance is also observed in normal hearing subjects using audio processed through a cochlear implant simulator. With additional tests in temporal resolution and spectral mismatch in simulations, the study suggests the results observed are best explained by CI patients utilizing periodicity cues to distinguish speaker gender. This prediction originates from normal hearing subjects suffering larger drops in performance with reduced envelope resolution [19]. Simulations confirm the notion both spectral and temporal information contribute in

normal hearing, but temporal information becomes especially important with fewer spectral cues [20], as is strongly indicated to be the case in cochlear implant patients.

Environments with multiple consecutive speakers have a detrimental impact on cochlear implant patients [9]. This applies not just in correctly pairing the subtle aspect of "who said what" in conversation, but also interference in rudimentary linguistic content tasks as open-set word tasks. The interactions between indexical qualities and linguistics of speech are much deeper and will be discussed to some extent later, but who the speaker is can provide information on what was actually said, and provide overall meaning in the progression of a conversation.

b. Telephone Conversations

Ease of conversational speech over the telephone has also become a research interest, particularly as the availability of cell phones has increased. More cochlear implant patients are using phones according to more recent surveys [21] in comparison to older surveys [22]. Usage is especially problematic when users must contend with unfamiliar talkers and subjects. There is a high amount of variability among users regarding success in using phones, with significant portions forgoing phone use altogether. There is also a large variability among surveys in reported statistics of performance and comfort in using phones. The explanation provided is that the surveys took place in different times with different, potentially generational attitudes to technology [21]. While the trend appears to suggest an overall improvement in usage of phones for everyday contexts, there are still observable gaps in cochlear implant patients compared to the population at large.

c. Speech Intonation

In terms of speech intonation, both production and perception qualities are lower in cochlear implant patients compared to people with normal hearing. Perception accuracy in a two-answer format for distinguishing interrogative and declarative statements is reported at 70.13% accuracy in children [23]. This study also shows production issues in these two prosodic formats as judged by adult normal hearing speakers. In a more specific examination of contrasts in production quality between the normal hearing, hearing impaired broadly, and cochlear implant patients [24], proper stress and resonance quality remains problematic in CI patients. This is in contrast to pitch and phrasing, which had little negative effect. Production quality is, however, still higher compared to severely hearing impaired peers without cochlear implants.

d. Expense and Risk

Aside from the issues in performance, both in remaining linguistic challenges and largely unaddressed indexical qualities, cochlear implant technology is fairly expensive. Cost for the device, procedure, rehabilitation, and follow-up visits typically total over $40,000 [25] and have been reported as high as $100,000 [26]. This generally limits the accessibility of the treatment to patients in developed countries. The procedure is also invasive, requiring access to the temporal boney structures in the cranium. Complications can often occur [27], and surgical procedures may not be recommended for those with higher risk of complications.

B. Possible Solution: Sensory Substitution

One promising field of research to improve the performance of the hearing impaired, in both linguistic and indexical qualities, is that of sensory substitution. The

study of sensory substitution aims to provide information about one sensory modality through another, usually accomplished by a machine interface [28]–[30]. Sensory substitutions register input related to target modality through sensors, process the information through a coupling system, and provide artificial stimuli through transducers at the receiving sensory organs for the substituted modality.

1. Forms of Sensory Substitution

Sensory substitution has been established in research since pioneering work began in the 1960s. Although designs are highly varied, they can be classified by the target sensory system - that is, the one with the deficit - and the receiving sensory system. The prevalence and notoriety of solutions are not independent between these categories. As will be seen, some specific implementations have received more attention due to a higher perceived need and accessibility and relevance of actuator systems.

The most commonly targeted sensory modality for sensory substitution is vision [28]. This bias should not be surprising considering vision is an overwhelmingly large part of sensory experiences in humans without impairments, and blindness is associated with a very high quality of life impact on the disabled. Tactile targets also have a fair representation, and most notable tactile-tactile systems assist patients with loss of peripheral sensation [30]. Sensory substitution can also assist the vestibular system with balance aids [28], again with mostly tactile actuators.

Variations exist in the receiving modalities that interact with the transducers of a machine interface. Tactile systems, also termed haptics in this and other applications, play a leading role [30]. These consist of slowly-varying stimuli either in the form of mechanical static touch, vibrotactile stimuli, or electrotactile stimulations provided by

electrodes placed on or inserted into the skin. Also noted in the literature are auditory systems that convert information in the form of spectrograms into sound [28], [31].

2. Notable Examples

Some of the founding work in sensory substitution was accomplished by Bach-y-Rita [32] in developing the tactile vision substitution system, or TVSS. In this system, a camera converted the pixels of an image into a height-map for a spring system contacting the skin of the user, typically on the back. Users can distinguish basic shapes using this system, and also score above chance for more complex tasks like facial recognition. Performance also extended to managing tasks that require motor interaction and coordination with the artificial visual field, such as catching a ball rolling on a table.

Another Bach-y-Rita device converts images from a camera into electrotactile sensations on the tongue [32]. Although the tongue has a smaller surface area, it has greater sensitivity to resolving spatial details than typical areas of application for mechanically tactile systems. The electrically conductive boundary that saliva offers with the mucosal lining also ameliorates issues surrounding reliable electrical contact with skin, which normally limits the performance and comfort of electrotactile devices. Abilities attained in this system are similar to those with the traditional tactile arrays.

An audiovisual substitution system developed to aid the blind is the vOICe [31]. This translates images taken by a camera into sound by scanning over the image and converting the pixels in each scan line into particular frequencies. This scanning approach overcomes the limitations imposed by converting a 2D image into a one dimensional signal, at the expense of a lower frame rate. Subjects originally naive to the

device learned how to use it for spatial interaction, including locomotion and obstacle avoidance, and for discriminating between objects in a closed set.

Exploratory usage of these devices has led to some interesting findings, many of which could not be predicted from structured hypothesis-based research alone. With use, the substituted sensations eventually manifest as perceptions that are distinguishable from just a simple redirected sensation. Based on the context of the task, users would know whether the sensation is salient to the task, and most often correctly interpret it if it is, or acknowledge it is just a normal sensation that has no relation to the task. Users in visual systems also discovered optical effects such as parallax and shadows that were, up until they used the device, completely outside their understanding of the world [29]. Most telling about the utility and potential of these devices are the abilities of users to overcome the resolution limitations of the device [28], [29]. Many of these devices have limited spatial and temporal resolution, but extended use in conjunction with an ability to manipulate the visual field is associated with increased perception aspects that theoretically fall below the resolution of the device. This is especially true in cases with a more limited number of actuators like the tongue system. Sensory substitution systems have also been used in isolated cases for work situations. Although performance in these tasks is improved with the devices, it is still substantially below that of a full-sighted person [29].

C. Discussions on Sensory Substitution

1. General Considerations

Despite the potential of sensory substitution systems, the field suffers from several ongoing issues. The research into sensory substitution is well established,

reaching approximately half a century in age as of this writing. Yet for the number of proposed sensory substitution systems we see, there is an almost equally matched failure to adopt these approaches. These systems generally do not adequately or fully replace the deficit system, not just in form, but from a functional standpoint. Simply providing a greater device resolution, which would be a typical engineering approach to mitigate this issue, does not resolve these deficits. As previously discussed, users can often overcome the resolution limitations, proving this approach unhelpful after the initial stages of development. The unmet promises of sensory substitution have striking resemblance to the difficulties cochlear implants have in fully restore normal hearing perceptions to patients. Partial restoration of function is possible with either device, some more dramatic than others, but the advanced usage bordering on full restoration has yet to be uncovered.

Part of the issue relating to understanding what sensory substitution can do involves how it fits with psychophysics theory. Devices of this variety, although commonly explained as conveying sensory cues, actually have the most meaning in the framework of discussing perceptions. Sensory is a rather deceiving term in this context, as the devices do not act like a plug-and-play adapter for sensory organs [29]. Users that expect the information conveyed to eventually turn into direct sensation, tantamount to synesthesia, are often disappointed by the lack of an accessible and ubiquitous experience. Neurological studies that show the adaptation of areas in cerebral cortex under sensory deprivation can mislead in certain interpretations. Response of a traditional modality-specific area to a different modality does not imply all the aspects of the original sensation are preserved. Reviews of sensory substitution devices also note that full utilization requires meaningful experiences in the substituted modality to have the

profound impact users expect. One source [29] illustrates this phenomenon as a congenitally blind patient using the device to see their spouse. This experience does not have the same emotional impact that other forms of interaction would, at least not initially. Broader associations must be drawn for someone to feel they are gaining something meaningful from the device.

Descriptions of sensory substitution devices as a purely sensory input to the nervous system omit a crucial component in how they are successfully used. Several reviews [28], [29], [32] make the case that aids are best used in a sensorimotor regime. Users often explore with the device, and by manipulating them, can infer more about their environment or objects than static usage suggests. They can also overcome alleged resolution limitations of the device itself. As an analogy, in a person with normal vision, the full range of motions of the eye and head allows him to move the high-acuity fovea to other areas of interest in the visual field, or maintain focus on an object as it moves [33]. The ability to manipulate the visual field improves performance in visual tasks above a fixed-eye position where regions of high acuity have a limited range. Users of sensory substitution devices, given a well-functioning system that allows for camera movement, can perform more complex tasks that would not be possible to accurately interpret with a static sensor position [29]. Open-loop sensory data models in this regime would not account for actual task performance in the highest functioning users.

The philosophical issues also pertain to whether sensory substitution is actually substituting information, or providing a separate additional stimulus. Some reviews [29], [32] discuss observations from user studies that demonstrate systems add to the overall experience instead of simply occupying a different sensory modality. Artificial stimuli do

not appear to detract from classic perceptions. The stated example shows if a user is touched by someone on a spot occupied by a device, the user will not mistake it for a cue from the device. Devices do not unconditionally mimic a sense, as it is more accurate to describe them as providing cues that users can integrate into a perceptual model depending on the source. Moreover, creative approaches can provide cues for things not even remotely related to the classic human senses [34], [35]. These newly formed "sixth-senses" are eventually associated with patterns of behavior to improve performance in a specified range of tasks. Information access alone may be a misguided approach, as designers should ask the question of what new things the user can do with the system.

Aside from the philosophical issues surrounding sensory substitution, several practical concerns exist that prevent realistic implementations from reaching the target population [29]. Although they are sometimes meant to serve as alternatives to expensive biomedical implants, the devices can also impose a large expense. Designs are also not always usable in practice due to excessive bulk or power demands. Combined with these and other factors, the lack of ubiquitous usage in a trial period hinders the learning experience and can prevent users from utilizing the full potential of the device.

2. Applications for Cochlear Implants

For a device intending to supplement the cues provided by a cochlear implant, perceptual frameworks allow for more meaningful discussions over constructions of raw sensory information alone, just as they do for single modalities. Perception is likely the most correct way to describe the process of a sensory substitution aid. Treating the contribution of a second modality as a perceptual gain clarifies the way questions are framed in the context of signal detection theory. One important question is asking if each

modality contributes unique sensory information or simply reinforces a weak signal from the other modality. From the perspective of signal detection theory, this is a dubious question, as the process is more accurately described as increasing the sensitivity to perception discrimination. The overall perceptual process that results in different actions in a psychophysical task requires prior integration from the individual modalities, albeit in a potentially complex way subject to immense amounts of mechanistic research [36]. Treating the additional information as advancing perception sensitivity instead of sensory cues also indicates whether or not the "addition" of auditory cues through another modality in a patients with electric hearing through an implant is any different than the process for a deaf sensory substitution user. From a sensory perspective, the two cases are different, as the former user population already has some auditory sensations while the latter has little to none. Again, this framework does not address the underlying perceptual processes, which suggests that both groups utilize whatever sensory information is available to arrive at a meaningful construction of the outside world. The distinction in the addition is not from a fundamental difference in source, but how clear the overall integrated result is, manifesting in potentially different performance measures across the two cases based on the strength and accessibility of these cues.

In principle, an auditory sensory substitution device can act in a sensorimotor regime. Using the existing sense organs as a guide, systems can contain receiver microphones that operate as a means to localize sound sources. However, aside from assistance in localization, some of the information received cannot be adequately modulated in terms of classic motor feedback. Systems might have potential in enhancing a perception in the same way that directly facing a source can, but usage is admittedly

more subject to ambient exposure with less emphasis on mechanically active use. Unlike the visual and tactile modalities, the spatial locations of objects in the human auditory system are not represented in spatial arrangement of pathways, indicating fundamentally different processes to extracting this information [33]. Much of what the auditory system extracts in regards to speech is not spatial in nature. This aspect of the auditory system would lessen the impact of sensorimotor feedback in this application, especially in contrast to the striking effects with visual devices. Intentionally omitting discussion of sensorimotor interaction for each of the major modalities reveals that audition appears to represent a class separate from vision and tactition, as interaction effects would be glaringly missing in how people actually utilize their senses in the latter, but much less obvious in the former. Nonetheless, possible methods might exist for user actions to modify the way the system processes sound, in effect changing modes at will. This requires expanding the possible effects of feedback to ways that do not occur in the natural auditory system.

3. Aspects of Multisensory Integration

Putting the discussion on the true nature of sensory substitution aside, the tasks involved with aiding cochlear implant patients with cues from another modality are clearly multisensory in nature. The underlying principle for this approach is to take an existing sensory modality with degraded perceptions intimately linked to that modality, and attempt to strengthen the perceptions through an additional sensory system. Literature that demonstrates or even advocates this form of enhancement is notably lacking. The probable causes, although highly speculative, include a traditional framework that cochlear implants and sensory substitution aids are in direct competition,

and a lack of predicted benefit by requiring patients to utilize yet another device inferior to natural organs. Although there is a broad understanding that the goal of a sensory aid in speech applications can be to provide assistance in a multisensory way, this is not often the stated framework. In addition to practical concerns, the literature has minimal discussion on the theoretical questions that arise, especially how this combined regimen differs, if at all, from utilizing either a cochlear implant or a tactile aid individually.

Speech is already a multimodal phenomenon in its natural, unsubstituted state. Lip-reading often plays a major role in acquiring additional speech information in the hearing impaired, both aided and unaided [37]. The role of lip reading in patients with cochlear implants in particular demonstrates that sound and vision, each providing inadequate or incomplete speech cues, can sum together to yield performance metrics that are higher than the task with either modality individually. Setting these cues in conflict further exemplifies the multisensory nature of speech. The McGurk effect, as classically explained, is when one phoneme in sound ("ba") and second phoneme in vision ("ga") yield a perception of a different third phoneme ("da") [38]. This third phoneme is an inescapable perception that is not influenced by conscious awareness of the process, and cannot be adequately explained by framing aspects of speech as modality specific. In another context, different modalities can elicit the same or very similar perceptual cues. For example, cues on the emotional state of others can be provided through either vision or sound [39]. This demonstrates the relevant information is often not specific to a specific modality, and depending on the access of cues, one modality may play a greater role.

4. Pertinent Neural Mechanisms

Although the primary focus of the discussion here takes a psychophysics perspective, some explored and proposed neural mechanisms for different sensory regimes germane to the work of auditory-tactile sensory substitution are discussed. These provide possible avenues for neurological investigation of the processes underlying natural and modified perception.

One of the considerations is how neural responses to tactile stimuli in deaf individuals differ from those with normal hearing, as this would suggest processes that might underlie information extraction when adding tactile cues. Auditory cortex in deafened adults experiences the archetypical patterns in neuroplasticity [40], [41]. Cortex that is normally described as auditory has an increased response to tactile cues in patients with prolonged auditory deprivation. Generalizing to other forms of sensory deprivation, as verified in studies on other forms of sensory deprivation across species [42], the cortex is "reassigned" to other modalities by becoming responsive to cues in those modalities. Note that the studies on deaf individuals are not definitive in terms of what cues are actually extracted, although they did hypothesize the possible role of high-powered hearing aids in conveying these cues. In particular, Auer [41] contrasts the responses in full hearing and deafened adults to fixed vibrotactile frequency and the fundamental frequency of speech without providing justification on what properties in the signal-listener system lead to this interaction effect.

Much of what is known about indexical properties of speech, in particular evaluations of quality and speaker identification, relates to direct observations of patients with lesions in specific areas of cerebral cortex [43], [44]. Strong evidence in correlating

lesion site and task performance suggests that discrimination of qualities and speaker recognition are dissociated tasks. Damage to the temporal lobe of either hemisphere interferes with speaker discrimination tasks, while damage to the right parietal lobe results in poorer recognition of familiar voices. These studies do not identify particular regions within these broad descriptions, nor necessarily rule out other regions that can influence task performance. Nonetheless, these works demonstrate that perception of sounds, even of seemingly related aspects of speech, actually function in different ways and have different representations in cortex. Further studies in blood-oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) have also identified voice-selective regions of auditory cortex in both speech and non-speech. The region with the most activation specifically to human voices appears to be the upper bank of superior temporal sulcus (STS), with middle regions of the STS also correlated [45]. While these are possible candidates for involved regions in a auditory task for speaker characteristics, the study examined speech broadly without considering tasks to isolate speaker characteristics specifically. Results might differ substantially when implementing a task requiring subjects to distinguish speakers instead of passively listening to vocal segments.

In line with the results for tactile activation of auditory cortex in the deaf, some results have also shown a weaker but significant activation in peripheral regions of auditory cortex in full-hearing subjects. This evidence, if it proves sufficient, would provide a possible mechanism for haptics aid in general auditory tasks. Schürmann [46] describes interactions of auditory and tactile stimuli in regions of secondary and association cortex, in particular the auditory belt, secondary somatosensory cortex, and

posterior parietal cortex. Most notably different between this study and the studies on deafened adults is the scope of cortical areas. Only deafened adults, in contrast to those with normal hearing, appear to have cross-modal activation of primary auditory cortex. This indicates chronic deprivation and exposure to other modalities is a likely prerequisite for robust associations, with cochlear implant patients occupying an unexplored intermediate group.

Extreme cases of cross-modal activation come from patients with specific forms of synesthesia. Classically, synesthesia involves at least one abstract domain with little topographical representation in specific cortical areas, such as associations between numbers and colors. However, one rare case of stroke-induced synesthesia had evidence of cross-modal activation [47]–[49]. Stroke specifically affecting the right ventrolateral nucleus of the thalamus resulted in immediate loss of tactile sensation on the left side of the patient, followed by a delayed onset of certain sounds resulting in tingling sensations, especially on the arm and hand. Follow-up measures of BOLD fMRI confirmed that the patient's somatosensory cortex had lower than expected response to tactile stimuli and higher responses to auditory stimuli, while activation patterns of auditory cortex were unaffected. The authors propose that, following the stroke to the thalamus, the now deprived regions of cortex experienced an unmasking of latent connections with regions of active primary sensory cortex, resulting in the acquired synesthesia. What distinguishes this phenomenon from the perceptual linkages found in sensory substitution applications needs to be established, whether it is simply by degree or if the two represent fundamentally different processes.

D. Prior Work in Auditory Sensory Substitution

The literature on sensory substitution has isolated examples of auditory targets and applications. There are not many recent comprehensive reviews on notable works, thus a brief review on select cases found in the literature will be completed here. Works generally fall into one of two categories based on the processing employed: spectral energy bands, where the energy content in ranges of frequencies is converted to strength of activation of particular actuators, and fundamental frequency and formant isolation, where further extraction is performed by detecting periodicity in the signal and filtering properties of the vocal apparatus. Each of the discussed examples will mention, when available, the technical specifications of the mapping employed, collected results or metrics, and nuances of the target applications as available.

1. Direct Spectral Mapping

De Filippo [50] used a combination of single vibrotactile and electrotactile transducers on the tongue. Vibrotactile activation used a direct speech signal low-pass filtered speech at 1000 Hz. The electrode component utilized a signal high-pass filtered at 4 kHz to modulate the number of fixed-rate pulses in a burst. The device was used in a linguistic task, measuring speech tracking rates in a lip-reading task for the simulated deaf using syllables, words, and sentences. The study utilized two subjects, and the long duration permitted alternating the unaided and aided conditions, resulting in comparable trials where observation could be made on both differences in aided/unaided performance and general learning. The results revealed modest gains from the aid in terms of accuracy and words per minute. As a fraction of ceiling words per minute in “normal” condition, a

gain of 10-20% was observed by adding the device. Through the experiment, it was observed that introducing a new talker resulted in a temporary reduction in rate.

A device entitled the Multipoint Electrotactile Speech Aid, or MESA, was developed by Sparks and colleagues [51] around the same time De Filippo's work was conducted. It consists of an array of 36 equally-spaced electrotactile channels with frequency bands extending from 86 to 10900 Hz placed on the abdomen. The intervals between electrodes are divided according to a tonotopic (frequency-to-space) model of the cochlea for a normal inner ear. Amplitudes of each electrode are quantized into eight 5dB steps, resulting in a theoretical 40 dB of dynamic range. The device was used to test accuracy of determining phonetic contrasts, as well as word tracking in lip-reading. The device initially appeared to help with moderate usage compared to a lip-reading only condition. However, with more extensive training, the performance with and without the device converged to similar levels.

Work on audio-tactile aids in the 1990s produced system displays with more sophisticated mappings, including displays that utilized a second spatial dimension to express temporal characteristics. Wada [52] developed several iterations, most notably an 8x4 fingertip vibrotactile matrix that utilizes columns for spectral features (from 150-4700 Hz) and rows to rows to change the positions of spectral columns with an experimentally-optimized velocity. This approach was developed to avoid the issue of "backward masking" on consonant contours, where tactile cues would reduce the perceptual strength of prior activations. The style of activation is spatially very similar to what a Braille reader might sense with a constant finger velocity. In a task to distinguish

a selection of five monosyllables, the tactile device improves all tested syllables over lip-reading alone, and most over a static display.

A comprehensive test of TACTAID models by Galvin [53] compares and contrasts the performance of two models. The TACTAID II+ consists of two vibrators activated with a 250 Hz carrier and modulated in strengths by the assigned spectral bands. The TACTAID 7 utilizes a row of seven vibrators with frequency bands in approximately logarithmic divisions from 200Hz to 7kHz. The TACTAID II+ is worn primarily on the wrist, while the TACTIAD 7 is usually worn on the trunk or abdomen. In phonetic contrasts without visual cues, users performed slightly above chance for both models, and also showed minor improvements in word tracking while lip reading.

2. Fundamental and Formant Isolation

Several models of tactile aids have been produced that aimed to isolate the fundamental frequency and occasionally formant components of speech. Rothenberg and Molitor [54] produced a system that converted the fundamental frequency into a different frequency using a linear function to activate an Electrodyne AV-6 vibrotactile transducer. They tested a variety of parameters, including vibration center frequency and modulation amount, to improve determining intonation or stress patterns of syntactically identical, but semantically different, phrases. Results showed that it is possible to gain information, but the rate of speech needs to be slowed down to have substantial gain. They cited the issue as indeterminacy in timing of modulation of the frequency. Users tended to perform better on the lower vibration frequencies (50 Hz) and with greater changes per unit of F0 change in speech. The device was demonstrated to have these effects in both deaf subjects and simulations with normal hearing subjects.

Some patent work in tactile aids supplements the research performed in journal articles, albeit usually without metrics of effectiveness. One patent [55] describes a device that maps fundamental frequency to the frequency of an array of vibrotactile actuators. The fundamental frequency also controls the site of activation on the array. The claim does not mention number of channels, however the diagrams suggest it is intended to run with eight vibrators. A later patent [56] uses formant processing to determine place activation on vibrotactile array of eight cantilevered beams. The fundamental frequency, also termed the glottal pulse rate, drives the vibrators. Tasks or usage recommendations are not mentioned for these systems, but they were likely designed for linguistic applications, especially for the system with formant isolation as the diagrams show distinct regions of vowels.

3. Contemporary Work

This review of aids for auditory applications is not complete without discussing ongoing contemporary work. Neuroscientist David Eagleman has engaged in a recent effort to revive the efforts of this field. At present, publications are limited to plenary talks [57] and conference posters [58] with minimal in-print discussion of results. The application areas in speech are anticipated to be linguistic, at least initially, but Eagleman has suggested the aid can also be applied to pattern detection in abstract data like stock-market movements and emotional language on websites [34]. Researchers savvy in sensory substitution should play close attention to upcoming publications from this development.

4. Discussion

Sensory substitution to replace hearing does provide some sort of information in nearly every respect. It is rare to see a category in the literature that is at chance. Because of this, these devices offer a possibility when the hearing aid, a superior option for patients with residual hearing, is not viable as a treatment [59]. In the early days of single electrode cochlear implants, tactile aids appeared to be on par in most respects. Several sources and reviews also highlight that the non-invasiveness of these aids is a key advantage. Sensory substitution aids provide a means to try a device without commitment in the event it fails to work effectively.

However, these devices, and to some extent the literature which reviews them, have several shortcomings. Earlier reviews [60] show relatively mild improvements with multichannel cochlear implants, but subsequent developments and reviews [59] have made the advantage of the best performing cochlear implants more strident. Tactile aids such as the Tactaid II are now clearly below the multichannel electrode in terms of linguistic metrics, and they are now comparable only to single electrode CIs [59]. Although performance limitations are also noted for cochlear implants, users hit a ceiling much quicker for sensory substitution aids by comparison. It often takes substantial repetition to see performance gains.

Prospects for multisensory usage of substitution devices in linguistic applications also appear dim. The exception to ties observed with single electrode cochlear implants did not hold in multimodal tasks (i.e. with visual cues), where cochlear implants offered a distinct advantage [59]. Moreover, applications that attempt to aid cochlear implant patients with supplemental aids have additional challenges. Cochlear implant users

already have performance gains on all metrics. It is a more difficult requirement to develop devices and algorithms with a demonstrable utility, compared to the relative ease with which it is to show improvement when the unaided performance is chance or near-chance. However, these doubts may not apply to indexical properties. The literature understandably focuses on speech intelligibility, largely because the motivations for designing the reviewed devices also do so. Research in auditory sensory substitution rarely has metrics on indexical properties aside the occasional inclusion of speaker gender. Additionally, when reviews do include relevant metrics, they are on devices that are not explicitly designed for that application, such as the TACTAID II. No substantial conclusions can be drawn in regards to performance on devices with indexical properties explicitly considered in the design. This highlights the need to develop and test for non-linguistic contexts of speech and hearing.

E. Approach: Receiver and Target Domains

Having discussed the problem of enhancing the strength of indexical perceptions in cochlear implant patients and the need for sensory substitution solutions geared to these properties, a specific approach to address these deficits should be considered. Modality-specific concerns will be addressed, as well as global issues here relating to both auditory and somatosensory systems. These modalities have some similarities, but as the roles are different in this application as target and receiver respectively, they will be discussed separately after this section.

One common approach in haptics is to generate a small set of patterns that can be intuitively associated with an aspect of the cue they are deriving, such as a leftward moving pattern prompting a turn to the left [61]. For a specified task with a narrow range

of cues, this method is convenient and not too difficult to learn how to use. It is also a viable approach in context of a gender discrimination task, albeit it would be fairly trivial to categorize two distinct patterns. However, this can become an intractable method with other indexical properties such as speaker identity. If the number of adequately discriminable speakers is approximately equal to the amount of friends and associates people have, otherwise known as Dunbar's number and predicted to be approximately 150 [62], that would represent the number of unique patterns the device would need to utilize. Practical usage of the device indicates users do not even perceive discrete patterns with haptic device, but instead exhibit smooth excursions with intrinsic variability in interpretation of direction throughout the stimulus [61]. Additionally, this possible approach would require the device to be reprogrammed every time the user encounters a new person. By comparison, the native process of acquiring representations of new speakers in the auditory system is more fluid and dynamic. Despite these complications, this approach does appear to function well in the context of some constrained problems.

An alternative to the discrete pattern approach is to utilize cues from each domain that can be represented on a continuum. Processing information in this way permits more intuitive and fluid use of the input features as it does not require classifying or abstracting. This approach also has the potential to enable users to utilize cues that are not explicitly considered in the design. While literature on haptic aids in general often considers discrete patterns, past developments in aids for audition clearly utilize continuum-based metrics and output. The design of a developed device guides the subsequent quantization of the dimensions in the receiving modality. The process of quantization can be thought of as a necessary but undesirable process to make physical

construction possible, as is the case for actuators in archetypal examples of sensory substitution systems.

Utilizing multidimensional spaces of cues also has the added benefit of not requiring explicit consideration of how these cues correspond to the desired outcome of a task. We will see that modality-specific dimensions have complicated relations with more abstract notions such as speaker identity.

The lack of concern with finding explicit cue-feature relations, however, does not imply the best solution is to directly map the power of spectral information to the functional dimensions of the device, as seen with several approaches to tactile aids in this application. Indexical properties, even those that are less abstract, rarely directly relate to raw spectral information. Fully utilizing the information presented typically requires intense training, if it is even possible at all. Considering it appears the existing literature favors the direct approaches by volume and notoriety, and preprocessing is less considered, there is a disconnect with applying indexical property detection with existing processing methods. Some information can be left for the brain to interpret, but that does not imply all information should be treated this way, as certain coding of information may be intractable without special modification. Some initial feature extraction will be considered in both modalities to present cues that are more salient, minimizing the complexity of pattern detection that is required of the subject.

F. Auditory Aspects

1. Cues in Speech and Hearing Literature

The literature on human speech yields a volume of characteristics and methods relevant to distinguishing the quality and identity of various speakers. Speech features

which have been identified as relevant to this area of research are rhythm [17], speaking rate [17], breathiness [17], nasality [17], pitch and intonation [16], [17], [63], formants [16], [17], [63], and dynamic articulatory cues, or the timing of phoneme production [17], [63]. These aspects, although varying in how directly they reflect features of the raw signal, can be linked to mathematical processes that make these objectively definable. For example, formants are defined as the peaks in a spectral envelope, representing the physical characteristics of the vocal apparatus for phonemes like vowels. Many of these cues have importance in both indexical and linguistic properties of speech [9], [14], [17], [64]. This is the central tenet of the idea that the two properties of speech are likely not independent, and that advances in researching cues in one application might be applicable to the other application.

Several methods exist to identify these cues and their relative importance in the context of speech and hearing research for linguistic and indexical properties. The most direct way to determine if a particular cue is relevant is to diminish or eliminate the information and observe significant changes in abilities of subjects. If a cue is modified across a series of samples and the samples become less distinguishable, indicated by a drop in accuracy, then that cue provides relevant information to the task. Information content in a cue can be reduced either by computer processing or, in select cases like rhythm and intonation [17], by directly instructing participants for sample collection to mimic a target sample. Alternatively, resynthesis techniques allow modifications of select samples so all other information except the targeted cue remains constant [16]. In a carefully designed experiment that requires a subject to discriminate scenarios with same and different speaker segments, the relative tendency to respond "different" to two

segments with the same source indicates how much the cue leads to a distinction. Some cues such as pitch and formant structure allow for complete isolation through artificial stimuli. Stimuli like sinewave speech, which isolates the formants and synthesizes sine waves at the peak frequencies, can be used to test if subjects can still use the segments to accurately distinguish aspects of speech, including speakers [65]. Above chance levels determine if the cue is useful, as is the case for the reduced sinewave speech [14].

Other attempts to understand the perceptions of indexical properties forgo examining subject performance by varying the signal properties and seek to understand abstract characteristics as interpreted by the listener. One method of defining a perceptual space is to use factor analysis [14], [66]. Factor analysis procedures require subjects to listen to stimuli and rate the voices' perceived qualities on a comprehensive list of factors. These factors are usually two antonym word pairs, such as masculine-feminine or flat-animated. The model analyzes these dimensional ratings by assuming each point reflects a linear transformation of a smaller number of underlying dimensions. This process is similar to that of principal component analysis (PCA), which also seeks to explain the most amount of variance in the data with as few dimensions as possible.

Another frequently utilized method is that of multidimensional scaling (MDS) [14], [67], [68]. This process, like factor analysis, assumes the test samples are located on a multidimensional representation of features. Confusions in the stimulus-response pairs are converted into similarity indexes, or quantities that describe how close two samples are in this hypothetical space [69]. The locations of these samples are fitted to minimize the error in the predicted and actual similarity values. This process maps results based on quantifiable evidence of psychophysical similarity alone without having semantic

intermediates. The discovered layout can then be correlated with known cues to determine likely feature correspondence. Note that the MDS procedure does not require the cue to be explicitly known to find that a dimension exists. In fact, studies regularly find dimensions that do not correlate well with any tested quantity [14]. Many of the correlated features correspond to the influential cues tested in the isolation and elimination procedures.

The extracted factors or dimensions from factor analysis and MDS are quite extensive, and often vary in selection and importance between studies [14]. Complicating this picture, especially for factor analysis, is that the resulting list of highly correlated factors can only be from the initial selection. Factors on a word-pair rating are usually not defined in the experiment, implying there is room for discussion on both what they mean to the listeners and as an objective quantity.

2. Cues in Computer Science Literature

Research in computer analysis of human speech yields separate interpretations of speech cues. These are generally purely mathematical without much consideration for the physical speech apparatus. They derive numerical gross features from the frequency spectra of small time windows, also referred to as short time [70]. Segments of the signal are converted to spectral information using a discrete Fourier transform. The resulting spectra are divided into a number of overlapping bands. The mathematical function to describe divisions bands can be logarithmic or Mel frequency [70], the latter commonly referred to as Mel frequency cepstral coefficients (MFCCs). They are similar in computation, but the Mel-scale utilized in computing MFCCs is supported by perceptual studies in the difference between frequencies as well as frequency models of the cochlea

[70]. The coefficients describe the patterns in the spectra by using a discrete cosine transform of the logarithm of spectral power in each band. Alternatives such as linear predictive cepstral coefficients [71] predict the future values of the time-domain signal based on linear combinations of past values. Plotting the spectrum of the approximation and original signal shows the fits are similar. There are also alternatives that use wavelet transforms instead of cosine transforms [72]. Although the specific calculations in all these methods vary, the overriding theme is that they describe overall features in the frequency spectrum using a comparatively small set of numbers.

In addition to the variations of computational procedures to extract features, there are other related features that can be derived which concern overall power of the signal and changes of the cepstral values over time. Typically cepstral coefficients are numbered 1 through n, but a 0th coefficient that describes the overall power of the spectral window can be included. This can be excluded from analysis, but when included can form a criteria for exclusion, such as omitting quiet frames [73]. Vectors equal in size to the original cepstral space can describe the change in values relative to time. These delta coefficients are akin to velocity when describing the change of the static cepstral coefficients, and the subsequent differentiation, termed delta-delta coefficients, are analogous to acceleration. Information provided in deltas supplements that of the state alone, and is crucial for reflecting patterns of changes in the physical system, such as distinguishing forward and backward speech [74].

Following extraction of these coefficients, there must be a computational framework that builds on the mathematics of features to model the underlying processes [75]. One approach, and one of the most studied, is to model the patterns in the cepstral

features as multivariate Gaussian distributions. Although some literature uses only single Gaussians [76], more often the coefficients are described using mixtures of Gaussian distributions, appropriately called Gaussian mixture models (GMMs) [77]–[80]. To determine an aspect of speech, such as the identity of a speaker, competing models are compared by using maximum likelihood procedures to pair test data with the probable model source distribution. GMMs can also be strengthened by what is called a universal background model (UBM) of imposter data, which is important to reliably define the rejection strength in speaker verification.

Another type of model, the hidden Markov model (HMM), predicts the changes in the state of a system based on the observable output data [81]. The state represents the "hidden" component of the model, which in this context, corresponds to the speaker generating the content. The observable data are represented in the cepstral coefficients. This type of model has significant use when describing processes that can transition from one regime to another [82].

Support vector machines (SVMs) identify the boundaries between clusters of data in a multidimensional space. They identify a "maximal margin hyperplane," with the support vectors defining the transition from one cluster to the boundary on the margin. They can also be retooled to handle non-linear separations between regions. Typically SVMs categorize novel data between two categories. Of particular relevance in problems with potentially multiple categories, expansion to more than two classes requires more sophistic optimization, which is termed the multiclass problem. Multiclass solutions, as well as more complicated two-class problems, utilize a "kernel trick" to transform data with non-linear boundaries into spaces with linear demarcations. SVMs have been shown

to perform rudimentary text-dependent speaker classification without additional transformations [83], [84].

Artificial neural networks have also seen some exposure in the research literature. The constructs themselves are highly flexible, identifying abstract relations between features to categorize highly non-linear data without explicit representations. Some approaches have utilized collections of neural networks that hierarchically classify speakers in a large number of categories with a high level of performance [73]. However, unlike GMMs and other models that provide generative parameters, neural networks are not transparent regarding what the discovered features represent. This makes them difficult to apply as a tool to understand the mathematical basis of speaker-channel models of speech, however useful they are in performing novel categorizations.

Mathematical constructs called identity vectors, or i-vectors, further build on the GMM-UBM framework. I-vector computation uses the parameters from a GMM and reduces the dimensional representation of these parameters by estimating an appropriate transformation matrix. Although they are powerful in reducing complicated data to a small set of parameters, they require subsequent classification methods, such as cosine distance scoring [85] or SVMs [85], [86], to fully utilize.

3. Reconciling the Two Perspectives

Through a variety of techniques, results from computational frameworks demonstrate computers can correctly identify speakers using cepstral coefficients. However, it should be noted the reduced dimensions or patterns in these coefficients do not clearly correspond directly to variations in the physical speech apparatus. This disconnect between speech and hearing science and computational approaches is not so

much theoretical as it is practical. Experts in the field tend to realize the abstract hierarchy that exists in speech applications [75]. However, the mathematically simple cepstral coefficients are convenient to calculate, especially for those that have in-depth computer science knowledge but are less knowledgeable of speech and hearing studies. In computational applications, it is often preferred to establish simple features and use sophisticated algorithms to extract the categorization desired. This preference in sophistication in the cues versus algorithms is reversed for those in speech and hearing science. In this field, there is a significant amount of fundamental research on the salient cues, but less contribution to developing algorithms to automate the process.

For a sensory substitution application involving machine interfaces, the goals do not completely align with that of speech and hearing science, nor with automatic classification. The salient features with attached meaning are those found in speech and hearing science, but extracting these features requires individualized algorithms. Additionally, there is no guarantee they contain comprehensive information to the task at hand. Conversely, computational classifiers have a simple way to capture the raw information in the signal in a holistic framework, but only the underlying transformations are desired, without the explicit classification. This becomes apparent when distinctions are drawn between generative models, which simply represent the parameters of the observed distributions, and discriminative models, which actively form boundaries in the feature space to aid a classification system [75]. Generative models will ultimately best serve this application, while discriminative models can serve as mediating tools to simulate a human classifier.

G. Tactile Aspects

1. Neural and Psychophysical Descriptions

Discussion of tactile cues from the somatosensory system generally begins with the different receptor types, particularly on glabrous (or hairless) skin. Classic work in electroneurography describing the types of receptors from their neural response characteristics dates back to Johansson and colleagues [87], [88]. In this seminal work, the authors discuss four different response types based on combinations of the speed of adaptation (slow or rapid) and the relative size of the receptor fields (I - small, II - large). Throughout and following this work, mechanistic studies aligned these responses to candidate mechanoreceptors in the skin [89]. Designations for these conduits based on either neural response or mechanoreceptor are often interchangeable, particularly in the critical years where mechanisms were being discovered.

Although the information obtained from electroneurography is critical and demonstrates the influence of multiple channels, the task at hand requires an understanding of the more abstract perceptions that occur for an arbitrary tactile stimulus. A general understanding of abstracting receptor properties into psychophysical properties, even for simple tasks like threshold detection [90] and localization [91], is that the relationship between the two are very complex. Multiple channels are activated in varying patterns for different stimuli and differ from one region to another [90]. This is further complicated by the change in type and density of some receptors when transitioning from glabrous to hairy skin. These activation complexities apply equally to the variety of mechanical actuators used in sensory substitution devices, as any physical stimulus will not elicit one type of receptor exclusively. The question "how do the

different mechanoreceptors contribute to a holistic perceptual model?" lacks a sufficient and complete answer. While the question is critical and represents a valid area of research, in the context of sensory substitution devices providing perceptual cues, it is more meaningful to examine psychophysical discriminability of different artificial stimulation patterns.

Lines of inquiry in psychophysics have yielded information that is important to the design of a sensory substitution device. Early work by Weinstein [92] demonstrates the general sensitivity properties of different regions of the skin to pinpoint tactile stimuli. Narrowing the focus, applications in vibration-based haptic devices show an interest in the perceptual and biomechanical frequency response of the skin [93]. Investigation into vibrotactile perceptual properties in the congenitally deaf [94] suggest increased abilities of frequency change detection and possibly discrimination, validating the mechanism-focused literature on increased tactile representation in subjects without access to the auditory modality. This last point also suggests a sensory-deprived individual operates under an increased, or at least a different, perceptual regime for unaffected modalities compared to people without any sensory deprivation.

2. Potential Dimensions

Investigations into tactile perceptual spaces generally follow similar models and procedures as research into auditory perception. Tasks establish the perceptual proximity of different stimuli in a restricted set using the confusion properties and correlate the largest dimensions with known properties of the stimuli. For considerations in this application, there are two different sources that describe the scope to either general somatosensory theory or specific haptic applications. For stimuli through static materials,

roughness and hardness are two recurring properties, with the possibility of compressional elasticity [95]. Examinations into vibrotactile stimuli identify mechanical pitch and loudness [96], and even the temporal envelope [97]. This last property in particular has an apparently circular perceptual representation, a complexity which does not often arise in psychophysics studies. Efforts to focus on specific vibrotactile systems identify the distinguishing roles of spatial location [98]–[100], frequency [98], [99], intensity [98], [99], waveform (as in sine, square, triangular, etc.) [98], duration [98], rhythm [99], and temporal envelope [99]. While this second area of focus may have problems generalizing conclusions to other devices and certainly in a holistic touch perception theory, it does provide valuable information on how common it is to discover certain perceptual distinctions for a variety of systems.

As with examination of dimensional representations of indexical qualities, some of the dimensions for tactile perceptions are not well defined and vary according to the application and task. It is difficult to infer how discriminable two stimuli are for a novel device from existing literature, as no holistic, rigorous perceptual framework exists. However, a more immediate concern is that devices provide a more limiting interface than what is imposed by the limits of knowledge of psychophysics alone. Devices cannot utilize all of the possible perceptual outcomes, most immediately obvious between different classes of actuators. For example, most vibrotactile actuators cannot provide the static rigidity needed in slowly-varying patterns. Moreover, the idea that these device dimensions are truly independent from a perceptual standpoint has yet to be adequately proven. For the sake of providing a manageable implementation, each of these functional dimensions can be treated as though they are independent with the physical limitations in

mind. Complex interactions or confusions would signal a perceptual confounding between the dimensions, and necessitate investigation should they ever occur.

H. Linking Dimensions of the Modalities

Once the modality-specific dimensions are determined for an implementation, the link between them can be established in a fairly straightforward way. Each dimension is translated directly, or by using mathematical functions such as logarithms when necessary, from the source parameters to an independent factor of the device. In the framework of each modality, each property is independent, though there might be underlying correlations based on the nature of the population distribution, such as that between formants and fundamental frequency. Corresponding one dimension of auditory space to a dimension of the device-skin system preserves this independence. Utilizing multiple dimensions has the nontrivial process of determining which assignment permutation to use. What seems like an intuitive relationship may not be what leads to the highest performance levels in all users or even in particular users. Besides direct comparisons, the process can be facilitated by identifying dimensions that have experimental evidence for existing multisensory perceptions between them [37]. In the initial stages where a low number of dimensions are used, it is not highly important to get the exact process correct. The testing is more concerned with cue discriminability using some sort of computational approach.

Using a dimensional approach to relate dimensions in sensory substitution applications involves two notable issues [14]. These problems arise more from the modality-specific process of determining salient cues. Most studies that report representations of perceptual dimensions rarely take individual variation into account,

combining the results into a single factor analysis or multidimensional scaling problem of similarity measures. This eliminates potentially differing representations for individuals when combining similarity measures of groups. It is known that individual perceptual systems can be highly idiosyncratic, and ignoring these differences introduces large amounts of error and variability in the results. Secondly, the results obtained are often highly dependent on the experimental framework used to collect the data. Some of these may correspond to genuinely different mechanisms, such as speaker quality and identity as previously discussed. However, generalizable conclusions become problematic when trying to determine rudimentary transformations in a different application. These issues can be resolved with a more sophisticated group and individual dimensional parameters framework. This is supported by improved reporting style as researchers in perceptual psychophysics suggest [14], but that is outside the scope of this thesis.

I. Experimental Framework

Having discussed the motivation and theory, the specific approach that is tested in this thesis should be considered. A single quality is utilized in each modality: audition and tactition. For the tactile cues, the spatial dimension of up/down is considered. Utilizing an array of vibrators and not modulating patterns within an actuator allows simple motors to be used. The construction can be performed with eccentric rotating mass (ERM) motors, as these assemblies are simpler to construct assemblies and use with inexpensive microcontrollers. This spatial dimension is also commonly explored in haptic research, along with the second, currently unutilized spatial dimension of azimuth.

For the auditory perceptual space, the device will utilize one aspect that has been identified in literature as meaningful in speech and hearing science and that can be

mathematically defined. For this application, the fundamental frequency of the signal will be calculated [101]. This cue in the context of perceptual information can be related to the pitch of a voice. Although these cues are not the same, as not all sounds that have pitch have a fundamental frequency in nuanced discussions of signal periodicity, they are closely related for the unaltered human voice.

The task to test the effectiveness of the device must align with some distinctive quality of a speaker. Here, the gender of the speaker will be the category to be considered. In terms of fundamental frequency, voice gender has a bimodal distribution with little overlap [102]. There are also other cues that can indicate gender, such as formant frequencies, that are not explored in the context of this study, but are viable for future research. Additionally, pitch temporal changes and patterns can indicate more than just gender, such as prosody and emotional state. This crossover is fairly typical in contexts where abstraction from low-level sensory cues to perceptions occurs.

After the empirical user study, computational simulations are completed to test an expansion to identifying the speaker of an utterance and of the expressed dimensions. The feasibility of a multidimensional approach will be tested using an algorithmic classifier as a simulated human decision maker. The mean value of MFCCs for a speech segment will be categorized based on a linear classifier utilizing training segments of dimensionally reduced cues. In particular, the effect of number of dimensions implemented on predicted performance will be quantified. Earlier literature, as previously stated, uses more sophisticated procedures to identify the speaker. However, the simulations will serve as a starting point to identify how satisfactory linear feature selection is before considering more mathematically demanding methods. Additionally, the patterns in cepstral

coefficients will be contextualized with the empirically tested cues of fundamental frequency.

II. GENDER DISCRIMINATION TASK: EXPERIMENTS 1 & 2

A. Methods

1. Experiment 1

a. Materials: Equipment

The sensory substitution device consisted of a chair with a vibrotactile haptic system, depicted in Fig. 1. It was constructed for use in prior studies by Arizona State University Center for Cognitive Ubiquitous (CUbiC) students and faculty [103]. The chair itself is a standard office chair. Fabric was cut in the front and back to allow custom printed circuit board (PCB) tactor strips to be passed through in assembly and individual tactors to be attached to front. The vibrotactile array consisted of eccentric rotating mass (ERM) motors approximately 1.2 cm across arranged into a repeating two-dimensional

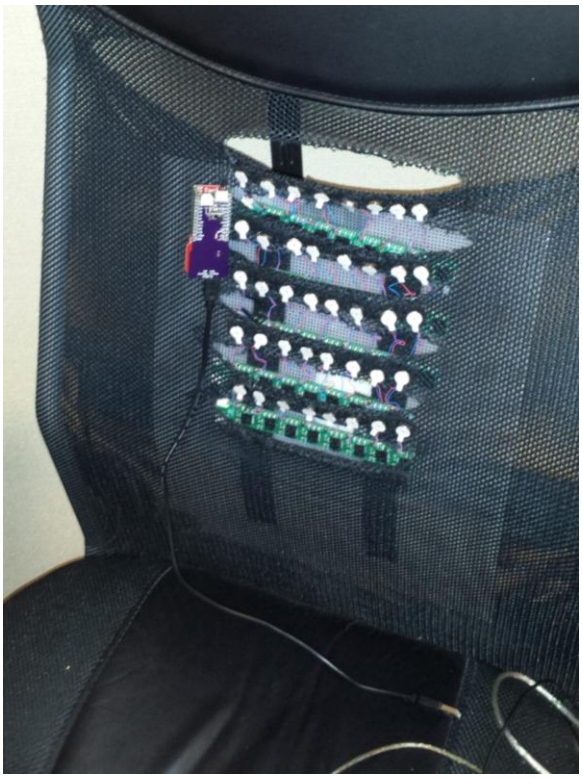

Fig. 1. Picture of the front of the vibrotactile chair. The Arduino board and USB cables are fed forward for presentation and are normally located behind the back.

pattern. There were 8 columns with a variable number of rows, based on how many of the tactor strips were attached. Initial implementations had six rows, but usage in this study had five rows to provide additional swappable redundancy after issues with strip failures. As measured from the centers of each tactor, approximately 4.1 cm separated each row and 1.8 cm separated each column. There was some variability in these values, as attachment to the chair was performed manually on deformable fabric.

The microcontroller for the assembly was an Arduino FIO with a custom shield design. The software and firmware utilized was from the open-source haptics project downloaded from Github [104]. A USB dongle provided the connection to the computer. In addition to power through the USB, a 3.7V 2Ah battery provided supplemental power. The Lenovo Z580 laptop computer running Ubuntu v.14.04 was equipped with Arduino drivers, the Python IDE, and additional Python libraries to serve as the interface and data storage. The programmed interface was handled through the Chrome web browser.

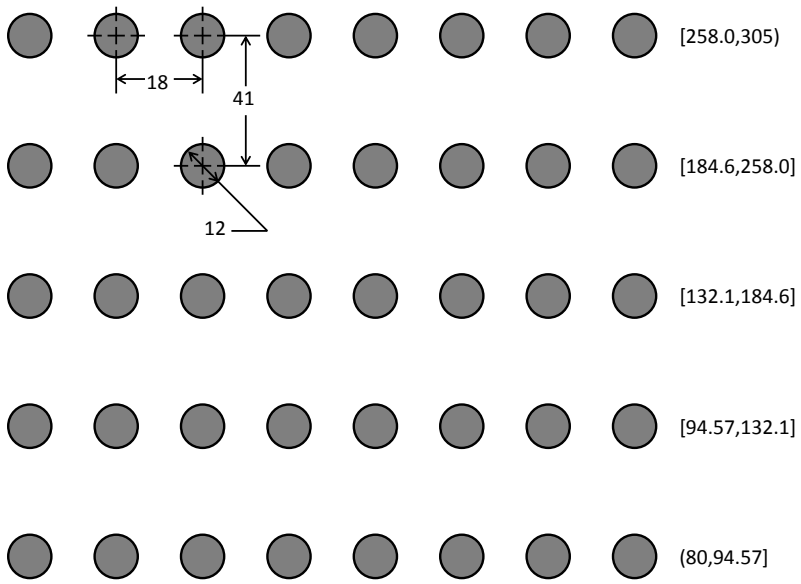

Fig. 2. Diagram of the vibrotactile array. Frequencies ranges in Hertz are designated on the left of each row. Diameters of the actuators and approximate spacing are also depicted.

Headphones were employed for subjects to listen to auditory stimuli while isolating them from environmental noises.

b. Materials: Stimuli Processing

The original sentence-length audio segments were obtained from the TIMIT database [105]. For ease of use in the study, the entire database was converted to the wave file format. One hundred fifty-six speakers were randomly selected with the following constraints: the gender composition was balanced overall (78 male, 78 female) and within dialect regions. However, no balancing in number occurred between dialect regions. After speakers were selected, one of the three SI segments was randomly selected for each speaker. These segments provide unique sentences with random contextual properties. Segments may have been removed and replaced if comments in the database indicated there were unusual occurrences in the recording, such as a nasally congested speaker or glottal fry.

The simulations of a cochlear implant were performed by AngelSim [106] using its noise vocoder. All segments were produced from the subset of TIMIT wave files using identical settings. The analysis and carrier bands consisted of eight separate channels. Divisions between channels were determined based on Greenwood's function for cochlear tonotopicity [107] from 200 Hz to 7 kHz and matched between the analysis and carrier bands. The filters for the analysis stage were Butterworth bandpass filters with 24 dB/octave roll-off. The envelope extraction for each band was performed with a 160 Hz cutoff frequency and 24 dB/octave of roll-off. The Butterworth bandpass carrier filters for the modulated white noise matched the analysis bands in cutoff frequency and also had 24 dB/octave of roll-off. Because the analysis and carrier bands matched, there was

no spectral shift in the output signal, maintaining natural frequency correspondence. The file specifying these parameters was saved for file reproducibility.

One option that was considered in the experimental design was to implement a spectral shift in the signal to simulate the effects of variations in insertion depth. Although meaningful metrics can be obtained from imaging data [108], there is evidence long-term plasticity reduces the impact of the physical misplacement of an electrode array [109]. Neural plasticity invalidates the justification to modify the shift in the spectrum in acute experiments in terms of actual performance drop. If spectral shift were to be implemented, the honest approach would be to nest participants by an amount of spectral shift so it is consistent and not constantly changing. However, this would have required many more subjects to obtain significant metrics.

Vibrotactile patterns corresponding to the fundamental frequency, represented as a series of entries in comma separated value (CVS) files, were obtained by performing several processing steps on the original wave files.

The fundamental frequency was extracted from the audio using Praat and its built-in procedure [101] based on autocorrelation with a 50-ms time window. The resulting data was graphed and manually inspected in MATLAB to remove inaccurate points.

Mapping the fundamental frequency to the device was designed around statistical properties of actual male and female speakers [102]. The original 6-strip implementation of the chair was designed to capture 95% of the excursions of the 1st or 99th percentile of the respective speaker groups. This resulted in minimum and maximum strip center frequencies of 70 Hz and 350 Hz. The scaling was selected to be logarithmic due to the overwhelming presence of logarithmic or near-logarithmic scales in natural pitch

perception, prior audio-tactile device designs, and reporting of speaker variability and excursions in semitones.

When reducing the design to five rows for this study, consideration had to be taken in whether to preserve either the resolution or dynamic range of the device. Resolution was selected to be preserved at the expense of dynamic range because, in the context of the selected task, it is not as important to capture the overall excursions of the speakers at the extremes of the device, whereas a reduction in resolution might adversely affect the ability to distinguish speakers in the boundary region. This does not mean data points outside those ranges are discarded, as anything beyond them is mapped to the closest extreme tactor strip. The range was changed to 80 Hz to 305 Hz based on these calculations.

Vibrations are displayed by row-wise addresses. Each tactor is independent in how it is coded at the Python interface, but row-wise operations are used to make the stimulus robust. For this experiment, two sides of each row are utilized for stimulation with different properties. Observing from the back with orientation comparable to how it

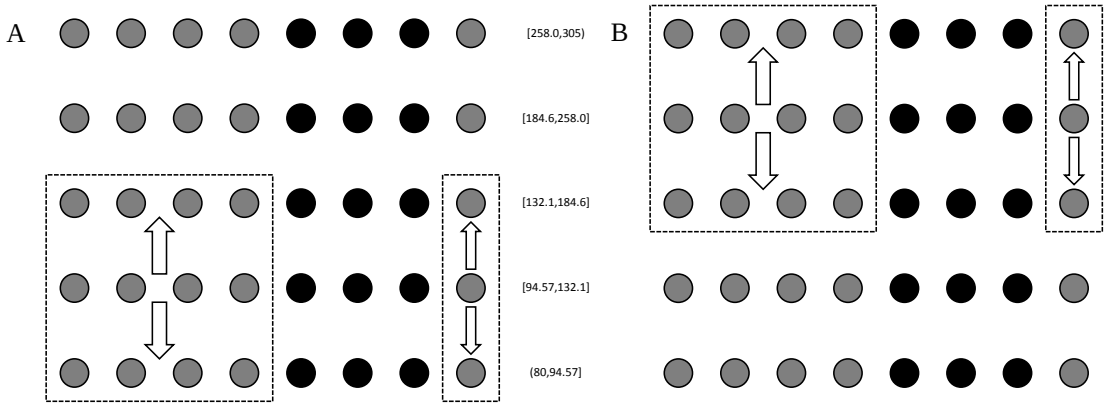

Fig. 3. Representative stimulation regions for Experiment 1. The patterns for male speakers (A) are typically lower than the patterns for female (B) speakers. The region on the left side of the array operates with four concurrent tactors for the larger time windows, while the region on the right operates with just one tactor with the raw segments produced in the fundamental frequency extraction.

is used, the left side of the chair has a larger 0.5-second time window refreshed at 50 ms of the geometric mean. This is again to make the stimulation more robust and less subject to rapid excursions. Four vibrators are assigned to each active window. The right side consists of a single column with the raw row assignments of the original 50-ms window. Both of these patterns play simultaneously. The entire pattern was post-processed to condense into a smaller number of commands, reducing the demand on the bandwidth of the device.

c. Materials: Participant Session File

Batches of step-by-step instructions for the program were generated in separate CSV files with three blocks of 52 files per block. The blocks for this experiment, normal audio, simulated cochlear implant audio, and cochlear implant audio combined with vibrotactile feedback, were randomized and counterbalanced among participants. Counterbalancing the block order is required because exposure order might affect outcome through learning. The original batch produced 18 spreadsheets, corresponding to the six possible orders with three participants in each. Note that this experiment did not complete the full run of the generated files. The result is that four permutations had two participants, one had three participants, and one had one participant. This imbalance was considered in the subsequent analysis. Each file name was randomly assigned, with gender balanced overall to blocks of equal length. Each file appears only once in the experiment overall for each subject. Six gender-balanced segments began each block and were designated as "training". This designation did not explicitly change how the questions were displayed. The batch was stored in a separate directory from the program files, and a file was retrieved and copied into the proper directory before the session.

d. Participants

Twelve normal hearing participants, ten male and two female, were recruited for this experiment. They had no hearing or tactile impairments and no prior exposure to cochlear implant simulations. Their informed consent was obtained and each was compensated 10 USD for their participation in a maximum 1-hour session. They were instructed on the nature of the task. Only the last four participants were informed on the pitch to vertical position mapping used.

e. Session Procedure

Following instruction, each participant was told to adjust volume to comfort level as needed before beginning the task. Each audio segment and/or vibrotactile pattern was presented in the order according to the CSV file. At the conclusion of each clip, the page immediately refreshed and presented the statement "Please select the gender of the speaker" and unnumbered, bulleted answer choices (Male, Female) horizontally below the question. The layout of the answer choices was fixed; "Male" appeared consistently on the left, and "Female" was consistently on the right. The participant answered the question using the left and right arrow keys, directly corresponding to the layout of the question. The page refreshed after a brief pause without providing feedback on the correct response, and then prompted the participant to press enter to proceed to the next question. Following the experiment, the participants provided scores on how difficult they perceived each of the block to be using a 7-point Likert scale, one indicating very easy and seven indicating very difficult.

2. Experiment 2

a. Materials: Equipment

Equipment usage in the second experiment was similar to the first experiment. A different laptop was available at the time: Dell Latitude E6530 running Windows 7. This implies that certain parameters such as response time may not be comparable between the two setups.

b. Materials: Stimuli Processing

Processing, unless otherwise stated, was performed with identical procedures and parameters to the first experiment. A different, broader selection of TIMIT sentences was utilized, producing an initial selection of 260 total speakers and subsequently reduced to 240 segments, with balances identical to those in the first experiment. In addition to the cochlear implant simulations, another set of audio clips was produced. The original wave files were converted to the AMR file format using WavePad at 4.75 kbits/s, then saved back to wave file format. This process imitates the compression effects on the voice over phone networks, such as predictive coding and elimination of high-frequency bands. These files were then processed using the cochlear implant simulator with the same parameters as the unmodified CI simulation segments. Because the AMR-processed files were perceptually quieter than the unprocessed CI files, both sets were volume balanced so each file had the same root mean square (RMS) value. The selected RMS was the loudest possible point where no file experienced clipping.

The information content of the vibrotactile patterns utilized the same number of rows and was processed in the same way. However, the column-segmented approach of dual streams was discarded for a single segment along the four middle columns. Two sets

of files were generated, one without windowing the Praat output, and the other with the 0.5-s windows. These corresponded to the patterns of the left and right side respectively of the first experiment, but this time they are presented separately and only one version in a given segment. As with the first experiment, the logarithmic moving average of the fundamental frequency was used in the windowed data.

c. Materials: Participant Session File

The batch of program instructions was similar to the first experiment. Because the experiment dropped the normal audio block in favor of a set of haptic-only stimuli, the three randomized and counterbalanced blocks were simulated CI audio only, haptics only, and a combination of the two. More overall segments were used for the training and general response segments of each block, 16 and 64 respectively. In both the training and general response of each segment, the trials were split evenly between the two types of audio (simulation only or AMR preprocessed) and haptic type (unwindowed and windowed) where applicable based on stimulus type, as well as balancing for gender. In the combined stimuli block, each of the four possible combinations were perfectly

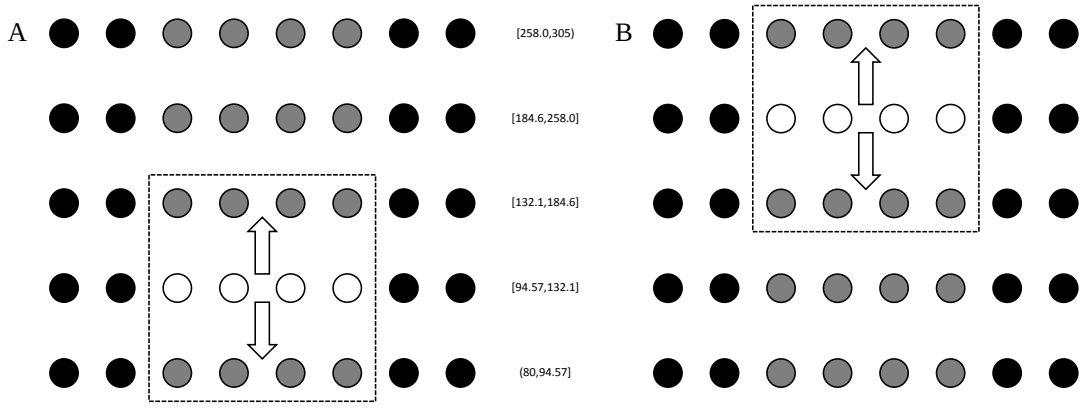

Fig. 4. Representative stimulation regions for Experiment 2. Like in Experiment 1, the patterns for male speakers (A) are typically lower than the patterns for female (B) speakers. Instead of multiple regions, the patterns occupy just one region with four concurrent tactors. Depending on the particular stimulus file, the device utilizes either the raw fundamental frequency data or time windows of the data.

balanced within themselves and with the speaker gender, yielding two training segments and eight general response segments for each of the eight audio-haptics-gender combinations.

d. Participants

Eighteen normal hearing participants, ten male and eight female, were recruited for this experiment. They had no hearing or tactile impairments, and only one had brief prior exposure to cochlear implant simulations. Their informed consent was obtained and each was compensated 10 USD for their participation in a maximum 1-hour session. They were instructed on the nature of the task. All participants were informed on the pitch to vertical position mapping used.

e. Session Procedure

The procedure used in the session was similar to the first experiment. However, the answer choices were arranged in a randomized order. The training that was implicitly segmented out in the prior experiment were fully implemented, thus the interface gave feedback on the correct response for the training segments on the window that prompts the user to continue.

B. Methodology of Analysis

1. Factors and Attributes

The experimental design has a total of five factors, three that examine holistic or between-stimulus influences, and two that are within-stimulus considered exclusively in the second experiment. Each factor is coded in the analysis to abbreviate the terms to which it corresponds, especially when discussing interaction effects. The stimulus order [A] refers to the particular permutation that the types of stimuli are presented. Both of the

empirical experiments have three types of stimuli, and thus six permutations. However, the particular types of stimuli are not equivalent between the experiments, nor are the orders. The order of presented stimuli has implications of learning effects, and can inform counterbalancing measures for future experiments. The subject [B(A)] serves as a random factor and, as indicated by the notation, is nested to the stimulus order. Each subject can only be exposed to one possible order, but the subjects between order levels do not have any relation to each other. Significance in factors with subject terms has implications on the variability in performance in the population, especially when characterizing performance in degraded conditions for an otherwise simple task. The type of stimulus [C] refers to the overall combination of stimuli for an experiment. This factor has implications on performance in each modality.

In the second experiment, two within-stimulus factors are included. The first corresponds to the type of auditory stimulus [D], i.e. CI simulations that have either been preprocessed with AMR compression or are in their original form. This factor dictates how susceptible performance is in further compressed audio conditions, especially those, like phone networks, that involve eliminating higher-frequency bands and use predictive coding to simplify spectral characteristics. The type of vibrotactile stimulus [E] indicates if either the default resolution or 0.5-s symmetric windows are used in displaying the haptic patterns. Performance changes regarding this factor show how much temporal information assists or interferes with the task. For example, finer temporal resolution might provide more information on excursion amount, but it could also interfere with the task if velocity does not contribute well to place information.

Although excluded from the general analysis of variance, a brief investigation of within-block learning effects considers the halves of a block as levels of a separate factor, [H]. This is used exclusively in a separate analysis for the second experiment. Additionally, only H and its interactions with between-stimulus factors - [A], [B(A)], and [C] - are considered. This is due to difficulty in analyzing these interactions and potential partial confounding with the within-stimulus effects.

Refer to TABLE I for the table of factor codes and their meaning. For either experiment, the type of design is a nested mixed-effects model. The subject factor is nested to stimulus order. Additionally, the subject and all of its interactions are random factors, but there are still fixed factors in the analysis, such as stimulus type.

The parameters of the files used in the stimuli are also considered for some of the analyses in the second experiment. The file durations are calculated using the sample rate and number of samples instead of relying on the file metadata. The log-mean fundamental frequency is calculated based on the manually checked Praat output. This is then subsequently translated to a cross-modal "distance" from the center frequency of the device, scaled so that each unit corresponds to the frequency or physical distance between adjacent rows. These parameters are excluded from ANOVA analysis of the response

TABLE I
CODES FOR FACTORS IN EXPERIMENTS 1 AND 2

Code	Factor
A	Order of Stimuli
B(A)	Subject
C	Type of Stimulus
D	Type of Auditory Stimulus
E	Type of Haptic Stimulus
H	Block Halves

variables, and are thus treated as a random uncontrollable effect. However, they are used in some of the subsequent correlations.

2. Response Variables

Decomposing the log files from each session allowed the collection of several response variables: accuracy, response time, bias metrics, and Likert self-assessment scores.

The accuracy is computed by averaging the series of matching stimulus-response pairs, 1 corresponding to a hit and 0 for miss. The chance level for a gender discrimination task is 50%. Although each data point represents an average, the repetition of questions enables deriving an approximate sum of squares for error for the ANOVA based on the confidence interval of the proportion. A detailed explanation of this will be presented later in the methodology.

The response time for each question is recorded in milliseconds, then subsequently converted to seconds for analysis. It indicates the time from completion of stimulus, when the question is presented, to a key press indicating answer choice. Because each combination of stimulus conditions for each subject (i.e. cell) has multiple trials, the error can be derived in the ANOVA using the standard process.

Quantifying bias in the context of a 2AFC paradigm in this study is accomplished using a metric published by Donaldson [110]. The original formula can be seen in Eq. 1, and Eq. 2 presents the form containing general categories and confusion matrix values used in this study. The sign of the final equation is switched from its original form so it corresponds to the over-answered choice instead of the position of the separating criterion

used in signal detection theory. Therefore a negative value indicates a bias towards category A, and a positive result towards category B.

Note that there are several alternative metrics for bias, each one having different properties and advantages for analyses in signal detection theory. Some other options include beta, C, and C', each of which rely on ordinate z-scores for estimated normal distributions of the two answer choices [111]. Unlike the results for accuracy, it is difficult to derive a theoretical basis for the standard error of this bias metric. Therefore, analysis will proceed only on fixed between-stimulus factors, [A], [C], and [AC], which have derivable denominator mean square values.

The Likert scores collected from the end survey are also considered. This metric, in addition to carrying information on influence of certain factors, also has implications on which of the primary performance metrics - accuracy or response time - are most reflected in perceived performance.

3. Transformation Techniques

a. Accuracy

Prior to any analysis of variance, transformations of the raw response metrics are carefully considered. The raw data behind accuracy measurements consists of Bernoulli trials, represented either as "1" for a hit and "0" for a miss. For each desired cell in the

$$\frac{(1-H)(1-FA) - H \cdot FA}{(1-H)(1-FA) + H \cdot FA}$$

Eq. 1. Original equation for bias as stated in [110]. H represents the hit rate (probability of a response given a true stimulus) while FA represents a false alarm (probability of a response given no stimulus).

$$\frac{-\big(P(A|B)P(A|A) - P(B|B)P(B|A)\big)}{P(A|B)P(A|A) + P(B|B)P(B|A)}$$

Eq. 2. Original equation for bias as stated in [REF]. H represents the hit rate (probability of a response given a true stimulus) while FA represents a false alarm (probability of a response given no stimulus).

ANOVA, the recorded measure can be statistically represented by the predicted accuracy $\hat{p}$ and standard error of the estimate $se(\hat{p})$. If these attributes are calculated, the comparisons between calculated accuracy can be treated similarly to estimated means from approximately normal distributions. However, the standard error of the mean for Bernoulli trials is not normally distributed, nor is it uniform as a function of the mean [112]. Compilation of confidence interval estimations [113] has produced a set of transformations that can be used. For the analysis of accuracy results in this and other experiments outlined here, the transformation used will be the arcsine-square root [114] as depicted in Eq. 3. This transform has a variance-stabilizing property that enforces a uniform standard error of the mean, as seen in Eq. 4, making it ideal for an ANOVA that assumes the standard deviations of each cell are approximately uniform. Although the variance-stabilizing transformation does have its limitations [112], it handles data variance better than the normal approximations of proportion confidence intervals in the unaltered domain, and is also much improved over ignoring the unequal confidence intervals in an ANOVA altogether. This effect does not have much influence over the holistic conclusions of the ANOVA, which is fairly robust for unequal variances, but is critical in producing accurate estimations for the post-hoc comparisons.

As an aside for the derivation of this particular transform, the calculus method to derive the transformations following a traditional Box-Cox procedure can be applied to

$$\bar{y}^* = \arcsin\sqrt{\bar{y}}$$

Eq. 3. Variance stabilizing transformation on averages of Bernoulli trials, i.e. accuracy.

$$\sin^2\left(\bar{y}^* \pm \frac{c}{2\sqrt{n}}\right)$$

Eq. 4. Approximate confidence interval for data in the variance stabilizing transformation. The uncertainty component follows a similar form to the standard error of the mean for normal distributions. In an instance the bounds of the sine argument fall either below or above the domain of [0,pi/2], the values are rounded to either 0 or 1 respectively.

the known properties of the standard deviation. The integral of $1/\sigma(\mu)$ as a function of μ is proportional to the transformation needed on μ [115]. Because the standard deviation for a Bernoulli distribution is $\sqrt{\hat{p}(1-\hat{p})}$ [116], the resulting integral is proportional to the aforementioned transformation.

To demonstrate this transformation produces data that is well-suited, a Box-Cox test [114] is performed on $1-\hat{p}$ and shown in Fig. 5. Because the theoretical standard deviation is closely approximated by $\sqrt{1-\hat{p}}$ for the variance-stabilizing transform, the confidence interval should yield that a square root transformation is well-suited. Indeed, this and the logarithmic transformation are both found to be viable candidates. The arcsine-square root transform is favored here over the square root because of its support from mathematical derivations. Additionally, the presence of data points near or below chance make the traditional Box-Cox set of transforms more ill-suited.

Following calculation using the accuracy data, the ANOVA table entries are modified to handle the inclusion of errors in accuracy estimations. The sum of squares for the error term is the product of the total number of trials for the entire analysis and 0.25,

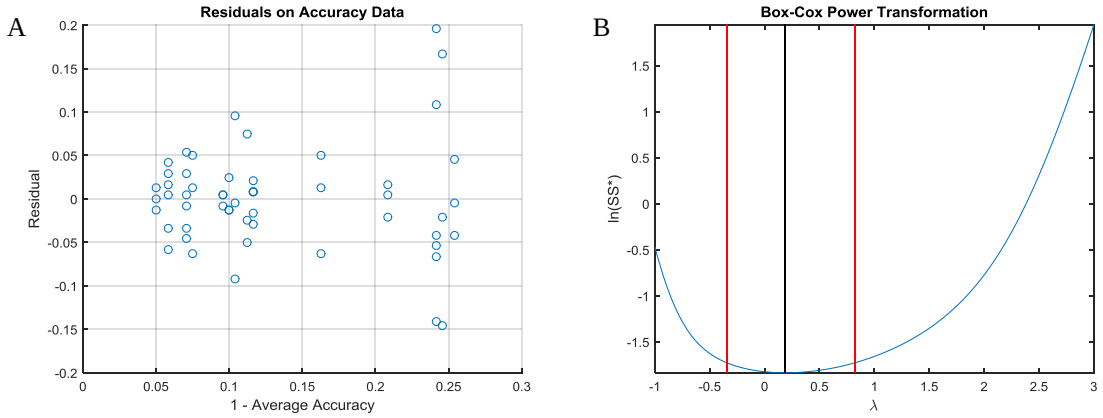

Fig. 5. Demonstration of best-suited transformations of the accuracy data for power Box-Cox transforms. The approximate distribution of residuals in (A) shows a non-uniform pattern as the fraction of misses, i.e. 1-accuracy, grows. A plot of adjusted sum of squares for the power transformations (B) reveal either a log or square root transform are well suited without prior knowledge of the true variance stabilizing transformation.

the mean square error of transformed proportion estimate. This yields a mean square error that is approximately equal to 0.25, but is slightly larger due to increased uncertainty by removing degrees of freedom for the existing coefficients. Additionally, the sum of square values for the existing terms are multiplied by the number of trials for each cell. The mean square values are calculated as normal. For F-tests including the mean square error term, the numerator, proportionally large to the number of trials, is compared against the mean square error denominator. This is equivalent to the scaling seen with replicates in numerical data where the expected numerator value is scaled for an approximately constant MS error value. This scaling does not affect the F-tests which do not use the MS error term, as these values are scaled by the same factor.

b. Response Time

Data collected for the response time does not have the same a priori knowledge for variance stabilization as the accuracy data. However, it does have replicates of numerical data, so the traditional Box-Cox procedure [114] can be used without major modifications. The transform is restricted to increments of 0.5 for lambda. This is for ease of use and to link candidate underlying mechanism with algebraic units to the resulting data. The data collected from the two experiments should have the same overall result if the mechanism is consistent between them. Following determining the transformation, points are iteratively eliminated if they are greater than 4 standard deviations from the mean of the transformed data [114].

c. Bias

For the sake of consistency, a potential Box-Cox transformation is tested for bias measurements. However, as the data could potentially be non-normal and not rectifiable

with these transforms on a relatively small number of points, inclusion of the null hypothesis of no transformation needed will not be taken as tacit indication the data is normal.

4. ANOVA Stages

In a typical analysis of variance without aliasing, all factors can be crossed where all possible combinations of levels are considered. In the first experiment, factors [A], [B(A)], and [C] can be handled in a single ANOVA. However, the within-stimulus factors of experiment 2 are only applicable for certain levels of the experimental factors, which limits the use of a single multi-way ANOVA. Note this is not the same as confounding, where possible combinations are not considered for the data collection. The unexplored cells in this case are truly nonsense, such as imposing a type of within-audio stimulus for haptic-only stimuli. Standard analysis of variance, even one that considers confounding or imbalanced data, will not work in this instance.

The solution implemented is as follows. The meaningful levels of the factors and their interactions are determined, verifying the total degrees of freedom to ensure no errors were made. Every stage of the analysis consists of a separate ANOVA. The data set for each stage is constructed so all contained factors are orthogonal to all other factors. Thus, cells are not imbalanced between levels of other factors. These effects are essentially removed from the subsequent steps where other factors are considered. Repeated factors that have already been derived are discarded, as these represent either whole factors or partial tests with certain contrasts. In the second experiment of this study, where analysis in multiple stages is necessary, the first stage consists of analyzing the effects including [A], [B(A)], and [C], and excluding the within-stimulus factors [D]

and [E]. The second stage is completed by including [D] with the first three factors, and in a separate ANOVA, including [E] with the first three factors. The last stage considers only the interactions of [D] and [E]. The sum of squares, and thus the mean square estimates, for the effect coefficients are the same as they would be through an automated regression analysis. The sum of squares for the error may be slightly overestimated for early stages, as some of the error is removed by significant terms in subsequent stages, but this impact is slight due to much lower sum of square values in later stages, and affects the significance of only the borderline random factors by making the estimates more conservative.

5. ANOVA F-values and Tests

Once the mean square values for the ANOVA tables are completely derived, the comparisons to compute F statistics are established using the restricted model for nested mixed effects. Prior research [114] has noted utilizing a restricted or unrestricted model doesn't appear offer a clear advantage in terms of reliability of hypothesis tests. As the process to derive the F-test indicates synthetic denominators are required for the unrestricted model [114], the simplest method of using the process for the restricted model was utilized. To briefly highlight the results of the procedure used to identify comparisons, each fixed factor is compared using the corresponding factor, i.e. the matching term that also includes [B(A)]. In the instances where the original factor contains [A], the denominator mean square term simply drops the [A] in favor of [B(A)]. Random factors, which already include [B(A)], compare to the mean square error for that stage of the analysis.

6. Post-hoc Tests

Depending on the comparison demands, one of several post-hoc procedures is used to find significant contrasts within or between significant factors. Means of single fixed factors are compared through Tukey's test, which utilizes the Studentized range statistic. This is essentially a modified t-test, and it is used as such while controlling for family-wise error. All other contrasts of fixed factors, including those of interactions, are compared with Scheffé's Method for comparing all contrasts. This simultaneously controls for family-wise error and considers the standard error for the contrast being tested.

Correlations between random factors have importance in this study. Relations between them indicate underlying dependencies that are otherwise treated as independent effects. For example, one might ask if the general performance levels of subjects are truly independent from particular individual strengths, or if the two relate to each other through some underlying trend. In context of the ANOVA, drawing correlations between the raw data collected simply confirms the significance of factors already derived. For example, showing scores in two different conditions form a strong linear relation can be explained by significant [B(A)] for the slope and significant [C] for the intercept.

Correlations of random independent factors, such as [B(A)C] vs. [B(A)], are performed with bootstrap t-statistic comparisons that are adjusted for family-wise comparisons. To calculate the critical value, one million iterations for the appropriate number of points and nesting are considered. Trials for the first experiment correspond to tests with 12 points and five degrees of freedom, and the second to 18 points and 11 degrees of freedom. These comparisons apply to family involving three fixed levels

versus families of either one or two fixed levels, as the families in the case of two levels would exactly mirror each other and have the same degrees of freedom as the family with one fixed level. The critical t-statistic for significance level of 0.05 was found to have a 95% confidence interval of 3.4584 - 3.4741 in the first experiment, and 2.7629 - 2.7741 in the second. The larger values of 3.4741 and 2.7741 respectively are taken to be the true value to err on the side of making the test more conservative.

When considering correlations of contrasts within a random interaction, specifically the difference between two levels of the fixed factor versus the remaining level, a separate bootstrap t-statistic calculation was performed to adjust for family-wise comparisons. One million iterations for are performed, this time only for the second experiment (18 points with 11 degrees of freedom) and a significance level of 0.05. The resulting statistic has a 95% confidence interval of 2.6917 - 2.7019. Again, the larger value of 2.7019 is used as the true value.

For correlations involving one response variable but multiple random independent variables, a linear model is produced with corresponding significance tests for the first-order coefficients and an intercept. Because there are no family-wise comparisons or adjustments in the degrees of freedom for the generated models, the t-statistics and p-values are taken as given.

Some tests require dropping the assumption of non-normal data to conduct accuracy analysis. This is especially true for the bias metric, as its distribution has not been well characterized outside detailed explorations in the relevant literature. Additionally, the means of biases and other data distributions are often excluded from the tools when performing an ANOVA. Some comparisons also require a different test when

taking an alternate model to a linear mixed effects explanation by comparing the relative performance between cells for individual participants. In these cases, sign tests are implemented to test data with potential failures to meet the assumptions underlying a linear model or ANOVA. A one-tailed test is used when the alternate hypothesis is only concerned with greater performance, and a two-tailed is used when a difference of any sign is considered.

C. Analysis

1. Experiment 1

a. Accuracy

The ANOVA for the accuracy data in the first experiment, with full details in TABLE II, yielded that effects for [B(A)] ($p < .001$), [C] ($p < .001$), and [AC] ($p = 0.011$) are significant with alpha = 0.05. No effects were found to be marginal with alpha = 0.1. Note that because the experiment was stopped before completing the randomized and counterbalanced batch, the data is unbalanced. However, utilizing Type III sum of squares process enabled calculation of the sum of squares for each term, as one of the advantages is that the results are not dependent on the sample size for each group.

TABLE II
ANOVA FOR ACCURACY DATA IN EXPERIMENT 1

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	16.29888	5	3.259776	1.246406	0.39248
B(A)	15.69204	6	2.61534	10.26018	3.37E-11
C	58.75532	2	29.37766	118.4185	1.26E-08
A*C	10.39272	10	1.039272	4.189205	0.01105
B(A)*C	2.977	12	0.248083	0.97325	0.47223
Error	468	1836	0.254902		
Discarded		0			
Total		1871			

Post-hoc analysis of the fixed effect for [C] requires a difference of .0752 in the transformed means. Both CI simulated audio alone (C2) and combined stimulus (C3) are associated with significantly less accuracy than with the normal audio (C1), C2-C1 = -0.3774 and C3-C1= -0.4125. This is not true between C2 and C3 (C3-C2 = -0.0351). However, it is noted that the effect of the combined stimulus trends lower than the CI simulation alone. Several contrasts for the factor [AC] are considered, largely highlighting the influence of learning in particular stimulus blocks. Blocks for the first and second exposure of any CI simulation audio are set in opposition, corresponding to learning conferred by a second exposure. Second, an effect of having the CI simulation alone before or after the combined stimuli on performance in CI simulations alone is calculated, indicating possible advantage by having one variety before other. The absolute orders of the blocks are also compared. Finally, a contrast is considered between first block in absolute order and the combined second and third blocks. None of these contrasts are significant, with p-values of 0.9596 for the first contrast and approximately 1 for all others. No correlations can be considered from [B(A)] alone.

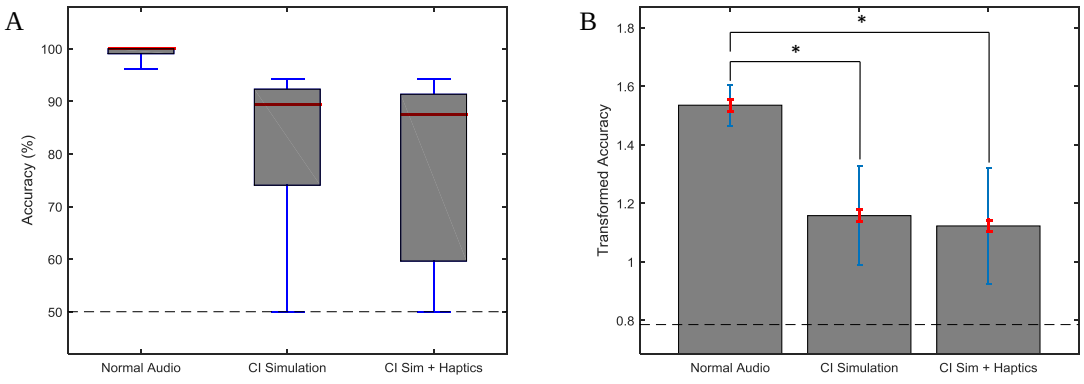

Fig. 6. Plots for the accuracy scores in Experiment 1. (A) Box and whisker plots representing quartiles show the general distributions of subject scores. (B) Bar graphs show the value of mean scores in the transformed space, with ±1 standard deviation in blue and the standard error of the mean in red. Single asterisks represent significant ($p<.05$) post-hoc comparisons. Black lines represent chance level in both graphs.

b. Response Time

The Box-Cox procedure for determining transformation power has an optimum at -0.81. The confidence interval ranges from -0.88 to -0.75, and the rounded power of -1 is used. An effect of using the Box-Cox procedure is that most points which initially appeared as outliers are actually not. The number has decreased substantially in the transformed data. However, the presence of some outliers implies an imbalanced ANOVA must be used. For purposes of deriving meaning from post-hoc comparisons, this also reverses the signs, i.e. a smaller value indicates a larger response time.

Results from the ANOVA indicate that, with respect to response time, the effects for [B(A)] ($p<.001$), [AC] ($p=.003$), and [B(A)C] ($p<.001$) are significant. Additionally, the effect for [C] ($p=.088$) is marginally significant.

Post-hoc comparisons for fixed effects pertain exclusively to contrasts of [AC]. The exact same contrasts are drawn as with the accuracy data. None of these contrasts are significant, with p-values of approximately 1 for all but the last contrast, which has a p-value of 0.9996. For the correlations of random effects, those between the families of [B(A)C] and [B(A)] are considered. None of the correlations are significant. With a

TABLE III
ANOVA FOR RESPONSE TIME DATA IN EXPERIMENT 1

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	11.2984	5	2.25968	0.670837	0.660957
B(A)	20.2107	6	3.36845	63.45833	9.91E-72
C	1.1781	2	0.58905	2.995423	0.08806
A*C	11.1141	10	1.11141	5.651716	0.003174
B(A)*C	2.3598	12	0.19665	3.704695	1.5E-05
Error	97.0326	1828	0.053081		
Discarded		0			
Total		1863			

critical value of 3.4741, the t-statistics for each comparison - B(A)C1 vs. B(A), B(A)C2 vs. B(A), and B(A)C3 vs. B(A) - are -0.1930, 0.3411, and 0.0394 respectively.

c. Bias

Bias is calculated and compared to a null hypothesis of no bias for the cells of stimuli without normal audio using a two-tailed sign test. The test statistic is four out of 20, which is below the left-side critical value of five. The participants are significantly biased towards responding with "Male" with p = 0.012. In the ANOVA that considers differences between factors, all of the tested factors - [A], [C], and [AC] - have p-values greater than 0.05. No post-hoc tests are necessary due to the lack of significant factors.

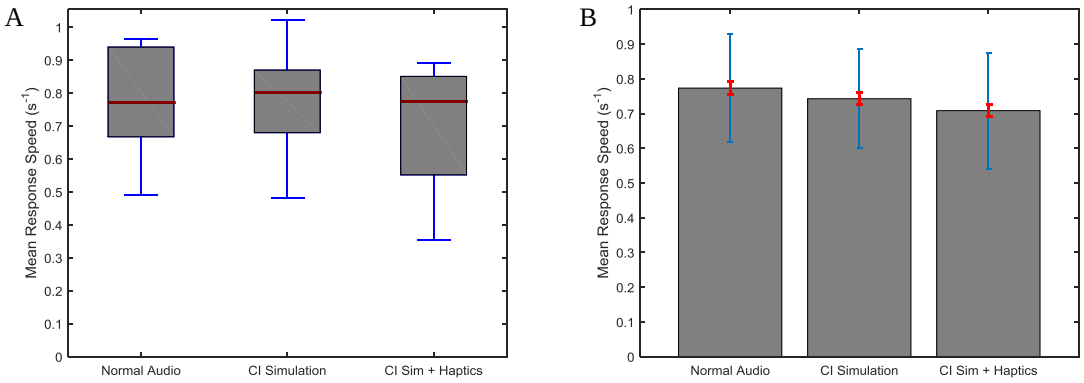

Fig. 7. Plots for the response speeds in Experiment 1. Speed represents the unit of the transformed response time. (A) Box and whisker plots representing quartiles show the general distributions of average subject scores. (B) Bar graphs show the value of mean scores, with ±1 standard deviation in blue and the standard error of the mean in red. There are no significant differences between these levels for this data.

TABLE IV
ANOVA FOR BIAS DATA IN EXPERIMENT 1

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	0.61887	5	0.123774	0.244091	0.928367
B(A)	3.04249	6	0.507082		
C	0.19503	1	0.19503	2.530283	0.162784
A*C	0.58577	5	0.117154	1.519934	0.310306
B(A)*C	0.46247	6	0.077078		
Error					
Discarded		0			
Total		23			

d. Possible Bimodality

In visual observations, the data appears highly skewed and suggests there might be multiple modes in the data. Plotting the results in the degraded stimuli as independent variables and fitting to a Gaussian mixture model revealed potential bimodality in the data. However, the significance of this model is not taken into consideration, as the experiment was changed for the last four subjects. This indicates the observed trends are due to the different experimental conditions and not within the subject population. Fig. 8 demonstrates the distributions are nearly identical between the two models. In practice, maximum likelihood tests can be used to get significance values of differences in models, but it is highly unlikely these models can be distinguished, especially with only 12 data points.

2. Experiment 2

a. Accuracy

The ANOVA on the accuracy data for the second experiment yielded that the effects from factors [B(A)] ($p<.001$), [C] ($p = .001$), [B(A)C] ($p<.001$), and [B(A)D]

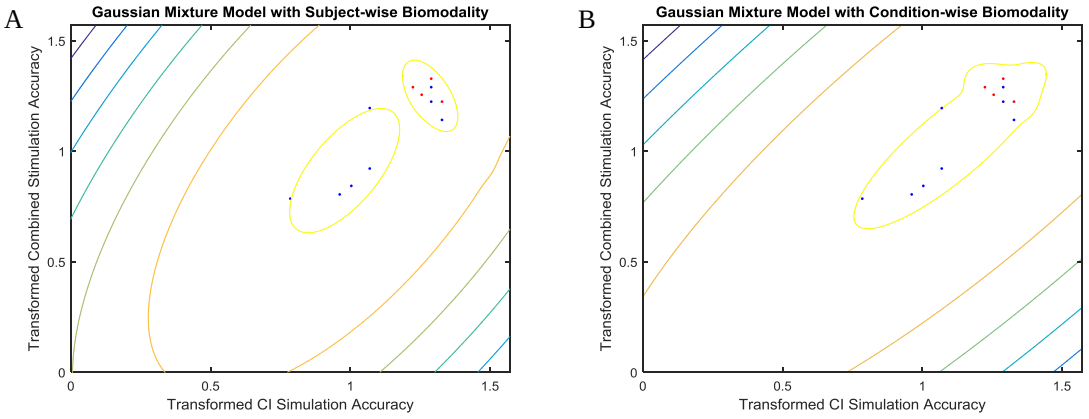

Fig. 8. Comparison of proposed Gaussian mixture models. Each one offers a different account of the bimodality in the first experiment's accuracy data. The representations are contour plots representing the log height of the probability distribution. Blue points represent subjects in the trials with no verbal description of device mapping, and red points represent subjects in trials that are given a description of the mapping method. Both (A) and (B) show distinct similarities with only a few portions of the domain forming distinguishing regions.

(p=.009) are significant. Additionally, the effects of [A] (p=.074) and [D] (p=.070) are marginal.

Post-hoc analysis of the estimated coefficients for [C] require the contrasts to exceed the critical value of 0.0810. The contrasts of C2-C1, C3-C1, and C3-C2 are 0.0565, 0.1378, and 0.0813 respectively. Thus, the accuracy in the combined exposure condition is significantly different than both of the separate stimulus conditions. However, the two separate conditions are not significantly different from each other, though they trend with the haptic stimuli resulting in a higher accuracy than the CI simulations. Although it is not significant, the results for factor [D] trend such that the accuracy for AMR preprocessed audio tends to be lower than the simulations without preprocessing. The difference between the two groups is 0.0560 in transformed space across the applicable stimulus blocks, and comparing the means in untransformed space compares an expected accuracy of 88.61% without AMR with an accuracy of 85.21% with AMR. Again, this effect is not significant, but because of its marginal significance, it is worth gauging the strength of the effect it proves to be significant in subsequent experiments.

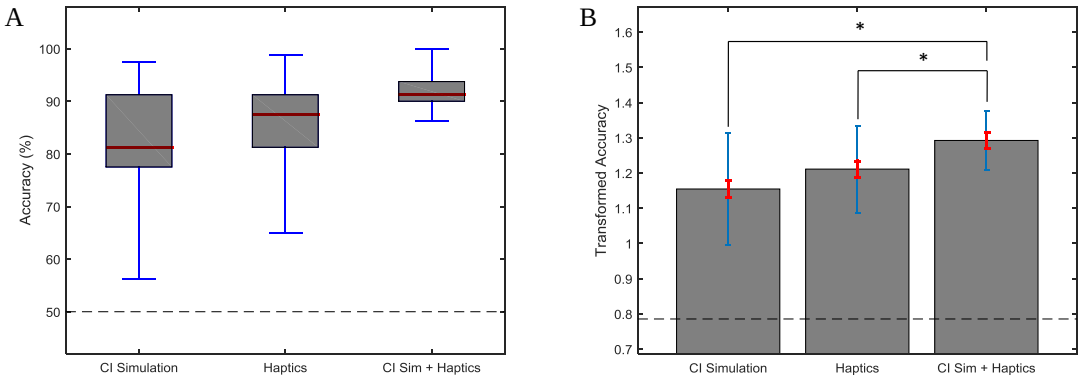

Fig. 9. Plots for the accuracy scores in Experiment 2. (A) Box and whisker plots representing quartiles show the general distributions of subject scores. (B) Bar graphs show the value of mean scores in the transformed space, with ±1 standard deviation in blue and the standard error of the mean in red. Single asterisks represent significant (p<.05) post-hoc comparisons. Black lines represent chance level in both graphs.

Correlations between the families of [B(A)C] versus [B(A)] must reach a critical value of 2.7741 to be considered significant. The comparisons involving the first, second, and third families - B(A)C1, B(A)C2, and B(A)C3 - have t-statistics of 0.2887, 0.3813, and -1.0103 respectively. This implies none of them are significant. By the same token, comparisons within families of [B(A)C] are also not significant. The relative contribution of added CI simulation B(A)(C3-C2) relative to the strength of CI strength alone B(A)C1 has a t-statistic of 0.9257. The relative contribution of added haptics B(A)(C3-C1) compared to the relative haptics strength alone B(A)C2 has a t-statistic of 2.0688. Against a critical value of 2.7019, the relative advantages in a single modality stimulus block cannot be said to confer an advantage in the combined stimuli block for accuracy, although it weakly trends that way for haptics. The B(A)C1 family of coefficients is compared to B(A)D2 for a t-statistic of -1.7688. Against a critical value of 2.7741, a relative advantage in accuracy in AMR audio does not appear to be correlated with the overall relative advantage in the CI simulation only block.

b. Response Time

The Box-Cox procedure for determining transformation power for response time data has an optimum at -1.19. The confidence interval ranges from -1.25 to -1.13, and the rounded power of -1 is used. This approximately matches the resulting power of the first experiment. Similarly, this reduces the amount of outliers in the data, the remaining of which are subsequently eliminated.

The ANOVA of response time produces significant effects associated with [B(A)] ($p<.001$), [C] ($p = .011$), [B(A)C] ($p<.001$), [AD] ($p = 0.049$), and [ADE] ($p = 0.002$), with no marginal effects. All other factors have p-values exceeding 0.1.

Significant post-hoc comparisons of fixed effects in [C] should exceed the critical value of 0.0495. The comparisons of means - C2-C1, C3-C1, and C3-C2 - have values equal to -0.0548, 0.0038, and 0.0587 respectively. Subjects are slower to respond to haptics alone than CI alone, but also quicker to respond to the combined stimulus compared to haptics alone. No difference can be concluded comparing CI simulations alone to the combined stimuli, which are approximately equal.

Concerning the contrasts of [AD], potential effects are drawn from having the CI simulations before or after the combined stimuli in combination with the preprocessing condition of CI simulations. Three specific coefficients of [A] in one group of [D] comprise each of these contrasts. Any significance found may pertain to improvements in reaction time of later blocks with a specific kind of audio. The contrast of -0.0186, compared to threshold of 0.0495, is not significant with a p-value of 0.8134.

Similarly to [AD], a significant contrast in [ADE] would indicate a specific order dependant change in response time involving both the CI simulation and type of haptic stimulus. This also requires contrasts that consider the order of either the CI and haptic

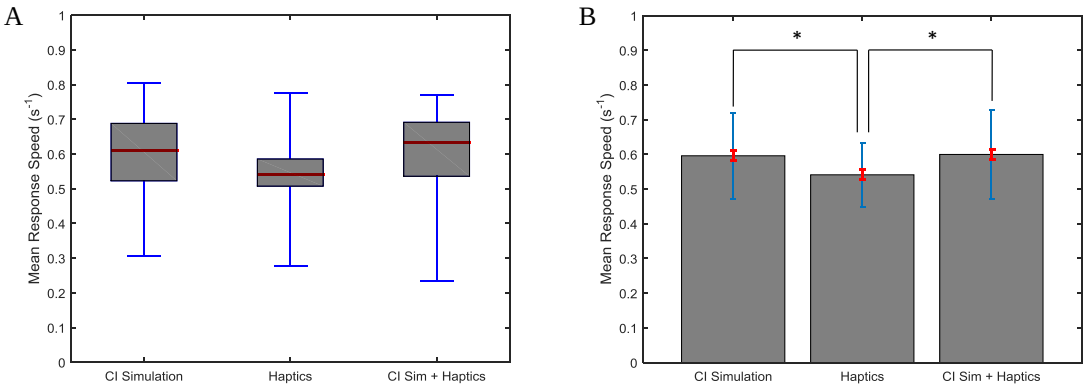

Fig. 10. Plots for the response speeds in Experiment 2. (A) Box and whisker plots representing quartiles show the general distributions of average subject scores. (B) Bar graphs show the value of mean scores, with ±1 standard deviation in blue and the standard error of the mean in red. Single asterisks represent significant ($p<.05$) post-hoc comparisons.

respectively relative to the combined stimulation. Contrasts of 0.0372 and -0.0021 are assigned to effects of CI and haptics order respectively. These do not exceed the critical value of 0.0944, with p-values of 0.7825 and approximately 1 respectively.

As with the data for accuracy, potential correlations of random effects are tested. Correlations of the subgroups of [B(A)C] versus [B(A)] do not reach the significance value of 2.774, with t-statistics of 0.6905, -1.2355, and 0.2291 respectively. Within [B(A)C], B(A)(C3-C2) versus B(A)C1 has a t-statistic of -1.0451, and B(A)(C3-C1) versus B(A)C2 has a t-statistic of -0.2036. Neither exceeds the critical value of 2.7019.c.

Bias

The change in the interface to randomize the presentation order for the answer choices allows extraction of both choice content and choice layout biases. For the bias in selecting speaker gender, a two-tailed sign test on all stimuli blocks has a test statistic of 20 out of 48, which is within the critical values of 16 and 32 ($p = 0.312$). An ANOVA shows all the tested factors - [A], [C], and [AC] - have p-values greater than 0.05, making post-hoc analysis unnecessary. Bias in selecting the position of the answer choice is also calculated, balancing for actual response to control for bias introduced by the experiment itself. The result of a two-tailed sign test on all stimuli blocks results in a test statistic of 24 out of 53, which is within the critical values of 18 and 35 ($p = 0.583$). In the ANOVA for both types of bias, outlined in TABLE V and TABLE VI, all the tested factors have p-values greater than 0.05.

d. Correlations with Random Attributes of Stimuli

The final linear model for correlating subject accuracy with file parameters includes coefficients for the intercept, duration, and absolute distance from center. The

intercept has a p-value less than .001, although it is not meaningful in this context. Additionally, the duration (p = .014) and distance (p<.001) are also significant.

e. Likert Scores

With a confidence interval for λ from -0.59 to 1.55, the Box-Cox procedure fails to reject the case where no transformation is needed. Proceeding with the ANOVA itself, [C] (p<.001) is significant, whereas [A] and [AC] are not. Post-hoc analysis requires contrasts of [C] to exceed the critical value of 1.0860 to be significant.The three possible comparisons between means - C2-C1, C3-C1, and C3-C2 - have values of -1.2222, -1.9444, and -0.7222 respectively. The CI simulated audio alone (C1) has significantly higher Likert scores, and thus higher perceived difficulty, than the scores of the other two

TABLE V
ANOVA FOR ANSWER CHOICE BIAS DATA IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	2.2965	5	0.4593	0.932352	0.49391
B(A)	5.9115	12	0.492625		
C	0.7374	2	0.3687	1.455778	0.253087
A*C	2.6395	10	0.26395	1.042182	0.440255
B(A)*C	6.0784	24	0.253267		
Error		0			
Discarded		0			
Total		53			

TABLE VI
ANOVA FOR LAYOUT SELECTION BIAS DATA IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	1.09857	5	0.219714	1.662003	0.218038
B(A)	1.58638	12	0.132198		
C	0.01086	2	0.00543	0.031892	0.968653
A*C	2.07606	10	0.207606	1.219317	0.328122
B(A)*C	4.08634	24	0.170264		
Error		0			
Discarded		0			
Total		53			

blocks. The difference between haptics alone and combined stimuli, though trending that haptic blocks have higher scores, is not significant.

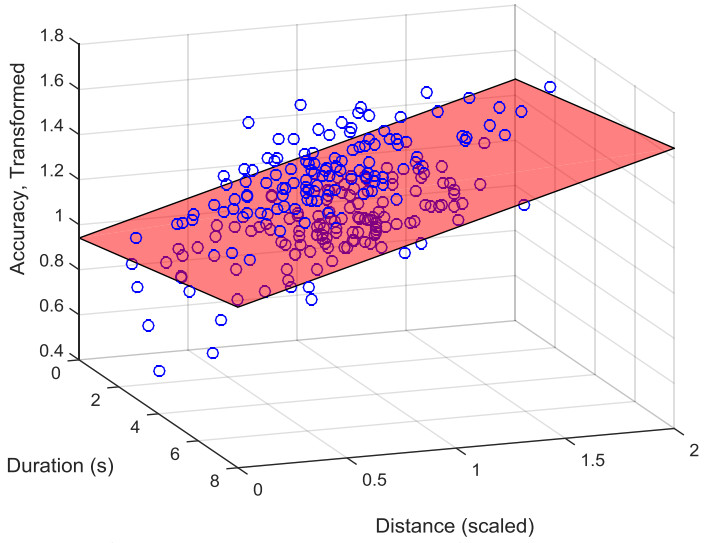

Fig. 11. Graph of the linear model for predicting correct answer choices on file parameters. Input variables are file duration and mean mode-independent distance from the center of the segments. Blue circles represent data associated with particular speech samples, and the red plane represents the best fit model.

TABLE VIII
LINEAR MODEL FOR ACCURACY VS. PARAMETERS OF SPEECH SEGMENTS

Source	Estimate	St. Error	t-statistic	p-value
(Intercept)	0.94088	0.041263	22.802	1.0843 E-61
Distance	0.26559	0.02958	8.9789	8.6729 E-17
Duration	0.021475	0.0087065	2.4665	0.014352

TABLE VII
ANOVA FOR LIKERT DATA IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
A	6.889	5	1.3778	0.609848	0.694504
B(A)	27.111	12	2.25925		
C	34.778	2	17.389	10.20656	0.00062
A*C	31.667	10	3.1667	1.85871	0.103523
B(A)*C	40.889	24	1.703708		
Error		0			
Discarded		0			
Total		53			

The linear model to establish correspondence of Likert scores to objective metrics considers the intercept, accuracy, and response time coefficients. Only the intercept has a significant effect ($p<.0001$), which again is not meaningful in this context. The accuracy and response time coefficients are not significant, with $p=0.057$ and $p=0.491$ respectively. However, the term for accuracy is marginal, and each has a negative coefficient as expected.

f. Performance in Half-Blocks

Measuring the effects of first or second half of a stimulus block on accuracy reveals no significant effects in any of the considered factors through an ANOVA: [H], [AH], [B(A)H], [CH], [ACH], and [B(A)CH]. However, [H] ($p=0.011$), [B(A)H]

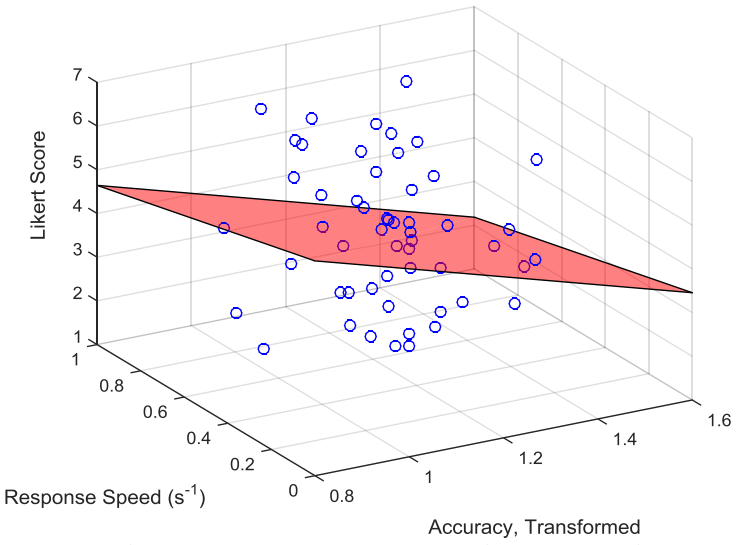

Fig. 12. Graph of the linear model for predicting Likert scores on subject performance. Inputs for the model are transformed accuracy and response speed scores. Blue circles represent data associated with particular subjects, and the red plane represents the best fit model.

TABLE IX
LINEAR MODEL FOR LIKERT SCORES VS. ACCURACY AND RESPONSE TIME

Source	Estimate	St. Error	t-statistic	p-value
(Intercept)	8.3753	2.2305	3.7549	0.00044586
Accuracy	-3.0754	1.5776	-1.9494	0.056754
RespTime	-1.2722	1.8331	-0.69402	0.49082

(p<.001), and [B(A)CH] (p<.001) are significant with respect to response time. Post-hoc comparisons on the significant factors for response time show the second half is globally faster than the first half. With a critical value of 2.7741, the t-statistics of 1.0226, -1.9738, and 0.9532 for the first, second, and third families of [B(A)C] respectively are not significant. Further correlations with currently and previously discovered random factors can be considered based on the available significant random factors. However, the lack of evidence to suggest dependence in general, complexity of the comparisons, and lack of controls for omnibus false positives make it undesirable to proceed without more informed hypotheses.

TABLE X
ANOVA FOR ACCURACY DATA SPLIT BY HALVES IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
H	0.038162	1	0.038162	0.122274	0.732641619
A*H	1.3796	5	0.27592	0.884076	0.520831804
B(A)*H	3.7452	12	0.3121	1.21719	0.264141725
C*H	2.164	2	1.082	2.85664	0.077113029
A*C*H	5.5324	10	0.55324	1.460635	0.214497679
B(A)*C*H	9.0904	24	0.378767	1.47719	0.062737187
Error	1080	4212	0.25641		
Discarded		53			
Total		4319			

TABLE XI
ANOVA FOR RESPONSE TIME DATA SPLIT BY HALVES IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
H	0.6793	1	0.6793	9.16116	0.010530176
A*H	0.2187	5	0.04374	0.589885	0.708230423
B(A)*H	0.8898	12	0.07415	4.247767	1.04075E-06
C*H	0.0816	2	0.0408	0.343495	0.712717437
A*C*H	0.9446	10	0.09446	0.795257	0.634194265
B(A)*C*H	2.8507	24	0.118779	6.804399	3.41998E-22
Error	73.386	4204	0.017456		
Discarded		53			
Total		4311			

g. Multimodal Enhancement Prevalence

An alternative to the model for subject performance, that is that the varying but overall tendency is for subjects to integrate information from both modalities, holds that users would more typically disregard the less meaningful stimulus and exclusively utilize the one that is advantageous for them. Prior forms of analysis demonstrate overall effects of modality use and existence of some subjects whom employ both modes, but a separate question to consider is if most subjects utilize both modes. One way to express this quantitatively would be that the combined stimuli would be above either of the other stimuli in subjects where that stimulus is above chance. This method has two different routes, one using CI simulations alone as the intermediate and one using haptics instead. Both are tested to confirm that the route taken does not change the results. An alternative way to express the null hypothesis is the combined stimulus would not be above the maximum of the other two stimuli. The analysis considers both forms of the expression.

Stepping through the CI simulation block, all subject CI blocks are above chance (18 of 18), exceeding the critical value of 13 ($p<.001$). Fourteen of the 17 non-zero differences show a combined stimuli accuracy above the CI simulations alone, all of which also had CI simulation alone performance above chance. This exceeds the critical value of 13 ($p=.0064$), indicating this route shows it is typical for both modalities to be utilized. In the case of using haptics as the route to the combined stimuli, all of the haptics accuracies are above chance (18 of 18), exceeding the critical value of 13 ($p<.001$). Fifteen of the 17 non-zero differences have combined stimuli accuracies above haptics alone and also have accuracies in haptics blocks above chance. This exceeds the

critical value of 13 (p=.0012), also indicating it is typical for both modalities to be utilized through this route.

The second method requires that the accuracies of the combined modalities be above those of both of the individual stimuli. Eleven of the 16 non-zero differences had combined block accuracies that exceed those of both the CI simulation and haptics block, falling short of the critical value of 12 (p=0.1051). The null hypothesis that less than 50% have higher combined stimuli performances than either isolated modality cannot be rejected.

D. Discussion

1. Experiment 1

a. Box-Cox

The results from the Box-Cox analysis of the first experiment (and, as will be seen, the second experiment) indicate a power transform of -1 is well-suited to stabilize the variance of the response time data. This power suggests the reaction speed, not time, is normally distributed. This observation should be noted in discussions of neural mechanisms for decision making, which, while outside the scope of this study, do provide a worthwhile field of research with implications on how psychometric data should be collected and processed.

b. Performance Statistics

Accuracy results demonstrate the task of gender discrimination given ideal stimuli is not too difficult for subject performance. On normal audio segments, accuracy performance is at ceiling for all participants. It is most typical for subjects to answer all of the 52 presented segments correctly, with a few participants answering only one or two

incorrectly. Additionally, both of the conditions with processed audio are significantly less in overall accuracy than under normal audio. This shows the perceptual accuracy is degraded to distinguish increased difficulty in the task from baseline performance. However, the vibrotactile system did not improve the accuracy of interpreting the degraded audio. In fact, the results show there is a trend in haptics presenting a slightly adverse condition. Overall, this implementation does not appear to adequately improve the accuracy in interpreting gender of a speaker above the degraded audio alone. The tacit conclusion is that subjects are unable to use, and possibly ignored altogether, the vibrotactile cues.

This experiment does, however, demonstrate that the performance of subjects can vary significantly in regards to overall performance levels and task-specific advantages. The significance of the effects of [B(A)] in both accuracy and response time indicates some subjects performed better overall in the task than others. This applied in particular to the degraded stimuli as the normal condition results are trivial and consistent between subjects. The significance of [B(A)C] in the response times, though not the accuracy, suggests subjects also had individualized advantages in how fast they responded to the different stimuli. Generalized to the population as a whole, one subject might have faster response times than predicted for the CI simulated audio alone, while others have faster responses in the combined stimuli.

Although the factor [AC] has statistical significance with respect to both accuracy and response time, there is a distinct lack of meaningful interpretations for block order or order-based interactions. The results appear unrelated to the order of trials in a

meaningful way. Presumably, any significant contrasts are unrelated to specific stimuli-related learning effects that involve complicated conditional statements on block order.

c. Difficulty in Utilizing the Chair

Questions arise as to why utilizing the vibrotactile output of the chair is difficult for the subjects of the first experiment, though the chair presumably provides distinguishable stimuli for the task. It is very likely a combination of several factors leads to poor utilization. The first eight subjects were not provided with instruction that the vibrotactile patterns signified pitch information. Moreover, the patterns consisted of an amalgam of two processing methods that, while dependent due to originating from the same data, present cues with different properties, namely temporal resolution, location, and number of utilized actuators. Subjects may find these dual streams difficult to interpret when conveyed holistically. Additionally, no direct feedback on the correct response is given at any point in the experiment. This potentially creates an artificial ceiling on subjects depending on their relative strengths. If a subject has poor discrimination in interpreting the cochlear implant simulations, then the subject will not properly associate any additional cues that might be provided without first understanding what they mean. In contrast, the subjects that perform well on the simulated audio would tend to find the vibrotactile patterns superfluous, even if they can properly form associations. Thus, the chair does not improve the performance of users, and the experimental design hides any relative advantages in interpreting types of stimuli have as an overall, stimuli-independent variation in performance.

d. Bias in Answer Choices

A sign test on degraded stimuli suggests there is a consistent bias in answering male over female in individual blocks of stimuli. Notably, no other discernible patterns related to the fixed factors in the experiment are significant. Although reasons for this bias are purely speculative without further experimental considerations, this could also be the result of several flawed choices in design and procedure. Recall that subjects are not instructed to the nature of the algorithm that produces vibrotactile patterns, and thus form their own associations on the nature of the output. These associations can also be flawed, especially in the context of a short experiment without direct feedback on the correct response. Subjects that are already not gaining much information from the audio itself could be swayed by extraneous spatial cues in the experiment. As the more robust windowed stimulation is chosen for the left side, and the answer choice for male is always on the left, there could be a subtle tendency to use the azimuth of the stimulus to indicate gender and lead to a disproportionate response for male. However, as the presentation order of answer choices is identical for every question, this bias in resolving speaker gender is confounded with using the left arrow key more often when making guesses on less discriminable stimuli. Randomization of the answer choices, as implemented in the second experiment, allows these two potential biases to be calculated independently. However, this might change the association and prevent observation of any bias. If it is desired to observe the bias and simultaneously determine its cause, the program can reverse either the stimulus or answer presentations, with particular bias remaining or changing proving indication of layout or choice bias respectively. However,

as the primary goal is to eliminate bias, randomization of answer choices takes precedence over experimentally explaining the bias.

2. Experiment 2

a. Performance Statistics

The second experiment reveals the vibrotactile array, through the combined changes in stimuli and procedure, has the potential to provide information for a gender discrimination task. Overall, the combination of both modalities has statistically greater accuracy than that of either modality alone. The primary comparison of concern, improvement of accuracy on CI simulations through addition of the device, has the greatest difference. Although smaller, the addition of CI simulations appears to improve the overall accuracy of users already utilizing the chair. It trends that the haptic mode is slightly more favorable to accuracy measures than CI simulation, but this is not significant. Introduction of response time results shows different and unexpected effects. The response time is slower for haptic stimuli than either of the other two types, and the combined stimuli blocks are on par with the CI simulations. Taking these two response variables together, on average, the device improves accuracy in the task over CI simulations alone without a significant drop in response time. Introducing the auditory modality to the device has an effect of both improving response time and accuracy. Utilizing one of the modalities individually might have a compromising effect between accuracy and response time. The results regarding performance in different stimuli support the hypothesis that indexical perceptual cues can be conveyed in a multisensory regime, with each modality providing meaningful information individually or in combination.

Reiterating the theme established with the first experiment, subjects have statistically significant variability in performance measures, both for the overall experiment and in relative advantages for certain stimuli. This has large implications for the application of theory in multisensory integration to the experiment and multimodal tasks in general. By studying the influence of stimuli with altered or unfamiliar qualities, it can be shown subjects derive information from the available sensory modalities differently with variable emphasis on each input. Aggregating the contribution of sensory modalities and generalizing results to all individuals is demonstrably not reflective of the nature of this task, and realistically the nature of others. Caution should be taken in extrapolating these results as the stimuli are unfamiliar to the subject. However, the same principal of this experiment can be applied to future experiments by taking extra steps to familiarize subjects with the stimuli, or identifying a natural multisensory task where performance is below ceiling for individual modalities. Holistic examinations of multisensory integration for perceptual cues in a variety of tasks have the potential to reveal if this variation in the population is typical, and if not, for what specific classes of tasks it is expected.

As with the first experiment, no meaningful interpretation of factors involving order of stimuli blocks could be discovered. For accuracy, the factor [AC] is significant, suggesting certain orders of stimuli might confer an advantage in particular blocks through learning. However, no meaningful and significant post-hoc contrasts could be constructed.

Two factors with significance confined to the response time are interactions involving the order and within-stimulus changes in the audio, [AD] and [ADE].

Significant contrasts could indicate a specific combination of stimuli saw improvement depending on the order. These would exhibit more significance in the interaction where the error term is relatively small, but would be muted when combined with overall changes for the lower-order factors. However, contrasts which pertain to learning between stimuli blocks do not show significant results.

The variety of audio presented shows variable effects on accuracy among subjects and also suggests a potential overall effect. The effect of [D] is marginal, with the data trending on the side of AMR preprocessing having a detrimental impact on accuracy. However, this hypothesis is not significant. Nonetheless, [B(A)D] is significant, indicating variations in subjects have relative performance strengths regarding the different kinds of processing. If the effect of [D] was significant, this effect would indicate that there is between-subject variation in how susceptible each is to the detrimental effects of preprocessing. If the results from these simulations translate to actual CI patient performance on phone networks, the significance of [B(A)D] implies not all patients react similarly in terms of performance metrics to identical changes in sound conditions. This indirectly confirms the finding of surveys of CI patients in which there remains a substantial variability in conveyance of indexical properties over telephones.

With the exception of the factor [ADE], changes in the type of haptic pattern parameters do not appear to affect the accuracy nor the response time for relevant trials. This effect, while not impacting this experimental setup, can have implications on later designs if a different task requires more temporal details. Confirming that temporal cues, as produced in this experiment, do not significantly affect results allows for

manipulations in temporal resolution without much fear of adversely affecting the performance gains already established. The results also suggest, for the tested parameter sets, users do not have much more difficulty in isolating general mean location of patterns when the tracking has less resistance to rapid excursions.

Analysis of the data from this experiment, and the first experiment to a lesser extent, do not reveal meaningful correlations between or within the random factors. It is plausible the influences of these factors on subject performance are truly independent, but the discrete and probabilistic nature of the accuracy data introduces additional variance in the data that weakens any potential correlations. However, these issues in variability do not apply to response time, which also lacked significant correlations. In any case, all the tests fail to reject the notion that these random effects, corresponding to specific individual advantages overall [B(A)], in specific stimuli [B(A)C], or specifically with the type of audio presented [B(A)D], are not related to each other.

b. Biases

In this experiment, no statistically significant bias exists in the answer choice nor the position chosen as the correct response. This implies the problems with bias from the first experiment are resolved, regardless of whether the problem is an issue in perceptual bias or a preference of location independent of answer choice. Since the cause of the bias cannot be determined, it is up to speculation, but it can be subject to further testing and scrutiny should it ever appear again.

c. Influence of File Parameters on Accuracy

The duration and normalized distances from the center point of each domain have a significant and positive impact on the accuracy within a file. The distance has a greater

influence in terms of magnitude and significance, and it is more apparent that the effects of the duration plateau in the actual data. The fact this technique does not balance files between the types of stimuli limits it ability to control sources of error. File distributions are random, which minimizes confounding, but still creates a source of unmeasured variance. Because each point also consists of only 18 trials, a large part of error could be from the discrete nature of the response variable. Nonetheless, the explained variance and number of files are large enough to produce significant effects in both of the tested file parameters.

d. Likert Scores

Analysis of variance of Likert scores shows subjects consistently agreed on which blocks are difficult based on stimulus type. There is a statistically significant consensus the CI simulation blocks are more difficult than the others. It is not significant that the combined stimuli block is easier than the haptics alone, though it trends that way. Interestingly, the lack of significance of this last comparison is exceeded by the objective improvements in accuracy and response time moving from haptics alone to combined stimuli. The most likely explanation in this discrepancy is a selective reduction in statistical power from a relatively larger standard deviation for this metric. Of the two remaining fixed factors, [A] and [AC], neither has a significant influence on Likert scores.

Correlations with the Likert scores and the transformed accuracy and response time metrics show Likert scores tended to improve with both improved accuracy and more rapid response time, but neither of these coefficients is significant. The term for accuracy is marginal and larger in magnitude, indicating it might be the more germane

variable if significance is eventually demonstrated. Significance for the present number of subjects may not be achievable due to a larger variance in the Likert scores. Users might give different scores, even if the accuracy and response times are identical, if they have different expectations of themselves or the task, or different conceptions what the scores correspond to in objective difficulty.

e. Splitting Trials in Half

Comparing the first and second halves of stimuli blocks does not reveal anything notable in terms of accuracy, but shows significant effects in regards to response times. Subjects showed an overall quicker response in the second half of blocks over time. This effect is independent of any other factor, including presentation order and type of stimulus, indicating an inherent process of familiarization increasing the rate of response within a block. Additionally, subjects showed an individualized improvement that applies to the block as a whole [B(A)H] and to particular stimuli [B(A)CH]. This pattern of within-block learning with respect to response times matches what is consistently seen in previous stages of analysis. Also consistent with past observations are a lack of correlation between [B(A)C] and [B(A)H]. However, the lack of significant correlations in the previous analyses limited the prospects for finding a notable trend at this more specific stage. Because of the lack of control for an omnibus error for these correlations, no further tests or conclusions are drawn for these random factors. The lack of significance in response times associated with the remaining factors - [AH], [CH], [ACH] - show the improvement within blocks is apparently not specific to particular blocks or orders. This also corroborates the previous results, that learning between stimuli, if it's even present at all, is more limited than the changes in response times within blocks.

f. Multimodal Enhancement Prevalence

No stimulus block for any subject in experiment two falls at or below chance, though it is possible some subjects may have been at chance in one block because scores fall within the 95% confidence interval of a proportion of 0.5 for 80 trials (39.3%-60.7%). The collected data overwhelmingly indicates subjects can utilize both CI simulations and haptics information to some degree when each is presented alone. Additionally, the haptics typically improve accuracy on CI simulations, and CI simulations typically improve performance with haptic patterns. This variation of the test, regardless of the route, is significant.

The alternate test of the multimodal stimuli, however, is typically strictly above both stimuli alone is not statistically significant, though it trends in that direction. The cause of different results originally presented a confusing conundrum. It is now known, despite having identical interpretations on the overall prevalence of multisensory strategies, that the two tests represent different relations between the accuracies of individual stimulus blocks. The test where each modality is significant in stages includes the groups where the modality added subsequently to the initial modality actually has higher accuracy scores than the multimodal block. The test where differences are taken only between the combined stimuli and each of the constituent parts excludes this possibility. Some subjects do indeed have combined stimuli scores above one modality but below another, and this occurs for both permutations of which single modality had the highest accuracy. Often their top two scores, though in the counterintuitive order, are very close and are not significantly different based on a chi-squared proportion test. Experimental error is the likely contributor to these different conclusions. The second

variation of this test has less statistical power as the conclusions rely on comparing three scores, as opposed to two scores and a constant presumed 50% accuracy for the first test. This would be true in either the null hypothesis or alternate hypothesis.

Statements can be made about the existence of broad categories for multimodal stimuli utilization based on the tests performed thus far, in particular the consistent significance of the factor [B(A)C] demonstrating modalities do not have uniform contributions to all subjects. However, it is more difficult to make even gross statements of overall size of these categories, as the experimental error complicates the picture for these comparisons to a much higher degree. Reducing the scope of the problem further, most individuals do not have unambiguous designations for processing the cues in this experiment with a primarily auditory, tactile, or multimodal strategy. Separate experiments regarding proportions of these groups in the general population, and in strategies for individual subjects, can probably examine this line of inquiry. This can be accomplished by using either a different design or more degraded stimuli with the explicit purpose of determining these effects for most individuals. Regardless of the conclusions of typical usefulness of any modality, it is demonstrable that at least a small group of subjects benefited, sometimes greatly, from the addition of one of the modalities to the other.

III. SIMULATION OF SPEAKER IDENTIFICATION: EXPERIMENT 3

A. Translation from Gender Discrimination

The process of identifying features to map to device for speaker identification took a step in the direction of literature on computational classification schemes. The goal is to work with Mel-frequency cepstral coefficients (MFCCs) to determine how an implementation from the previous experiment would work with these features. The first twelve coefficients for the audio files processed for the second experiment are extracted. Each file is represented in the MFCC parameter space as a vector where each component is averaged across time. In other words, if the MFCCs are represented as a 12-by-n matrix, where n is the number of frames, the row averages are computed to produce a 12-element matrix. A linear discriminate analysis (LDA) classifier takes the 260 vectors, each one representing a sentence speech segment from unique speakers, and constructs a hyperplane that has the highest accuracy in classifying male and female speakers.

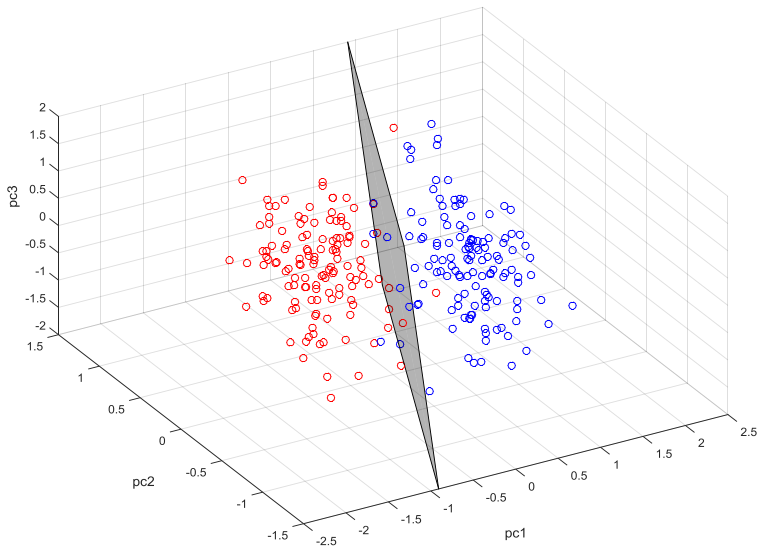

Fig. 13. Mean MFCC vectors of single speech segments of individual speakers. The three depicted dimensions represent the first three principal components of the distribution of speaker means. The plane in the graph represents the hyperplane of a linear discriminant model between male (blue) and female (red) speakers in the full vector space.

The linear classifier produces a hyperplane that accurately separates 129 of both the 130 male and 130 female speakers, resulting in a total accuracy of 99.2%. This near-ceiling performance is consistent for other random selections of gender-balanced speakers.This result is surprisingly strong for a blind introduction into machine learning, especially considering the method is technically text independent as the files have different linguistic content. Fig. 13 depicts projections of the hyperplane and speaker representations in principal component space to display the most amount of variation in this classification scheme.

The translational analysis also considers the strength of the correlations between the mean logarithmic fundamental frequency and the mean MFCC values for each file. For each speaker vector, the Euclidean distance to the hyperplane with sign preserved is calculated. Fig. 14 shows what appears to be a strong linear correlation between these two variables. Indeed, the amount of explained variance in this linear model is 85.3% of the total variance, resulting in a correlation coefficient of 0.924. However, the amount of explained variance drops for both groups, especially for male speakers, when separate linear models are computed. The amount of explained variance totals 52.7% for female speakers and only 12.3% for male speakers. This indicates that, although gender discrimination can be performed successfully using MFCCs and corresponds well to the fundamental frequency feature, extending the process with only linear transformations could be problematic for within-gender speaker discrimination.

Expansions into higher dimensional representations might resolve some of the issue in distinguishing speakers within gender categories. In a parameter space, it is assumed that the variation between speaker means would also correspond to components

with the greatest potential for discrimination. This may not necessarily be true as the most variable within-speaker components could also line up with largest between-speaker components, but it does serve as a starting point for moving into a computational-oriented approach to speaker dimensions.

B. Methods

Two-hundred sixty speakers were randomly selected from the wave-formatted TIMIT database. Each SI and SX-series file was processed using Voicebox [117] to extract the first 12 Mel-frequency cepstrum coefficients (MFCCs). Each frame was 8 milliseconds in duration.

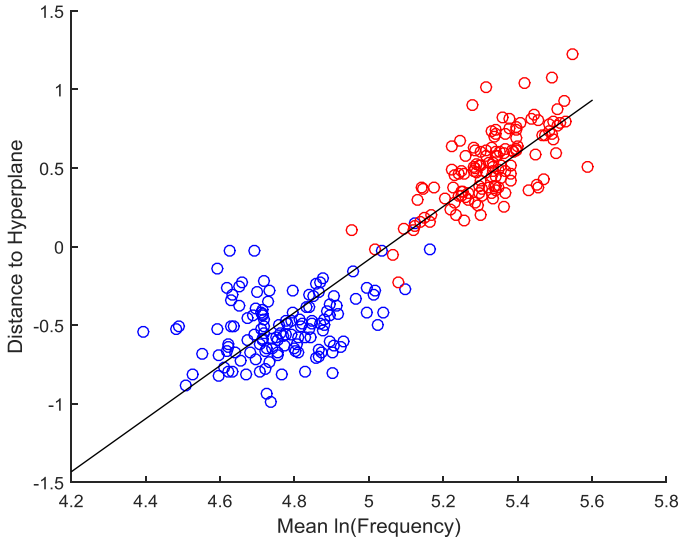

Fig. 14. Correlation of Euclidean distance of mean speaker vectors against the mean log-fundamental frequency. The black line represents the best fit model, blue points represent male speakers, and red points represent female speakers.

TABLE XII
SLOPE COMPONENTS OF LINEAR MODELS FOR HYPERPLANE DISTANCE VS. MEAN LOG FUNDAMENTAL FREQUENCY

Model	Slope Estimate	Slope Std. Error	t-statistic	p-value	Adjusted R^2
Both Genders	1.6882	0.04354	38.773	1.3038E-109	0.853
Male Only	0.54248	0.12418	4.3686	2.5581E-05	0.123
Female Only	1.5207	0.12636	12.035	8.8334E-23	0.527

The problem of building and testing the classification system was parameterized to three variables: number of speakers (2, 3, 5, 7, 10, 15, and 20), dimensionality of the space (integers one through twelve), and duration (zero exclusive to five seconds in increments of 0.25 seconds). Before each classifier was built, the training MFCC data was dimensionality reduced, based on the mean values of each speaker of the reduced speaker selection and using principal component analysis (PCA). This is to maximize the amount of variation between the means of each speaker, which are assumed to provide the most discriminating power. The classifier was built using linear discriminant analysis (LDA). The raw MFCC vectors for each frame of the SI files, dimensionally reduced based on the PCA between speaker means, were used to train the LDA classifier.

The classifier was tested using the SX series of TIMIT files, which provides an additional corpus of speech with random text. To build a test sample based on the desired duration, random sentences of the five choices are selected, with replacement, and parsed until the required duration is reached. Each value was dimensionally reduced, again using the same function derived from PCA, and then averaged across samples to a single vector. The process was iterated 1000 times over every parameter combination. This reduced the maximum error of the 95% confidence interval to approximately +/-3%.

C. Analysis

Analysis for the speaker identification simulations examines several aspects of the parametric space. This includes plotting accuracy against duration while holding either the number of dimensions or number of speakers constant, and a surface plot of the accuracy at the maximum duration (5 seconds) against both number of dimensions and number of speakers.

Fig. 15 depicts the accuracy plots of different number of speakers against duration with the number of dimensions equal to three. This number of dimensions was selected because it is a reasonable number to implement in a future design while also being reflective of the overall trends for other numbers of dimensions. The darkness of the curve corresponds to the number of speakers with 2 corresponding to the darkest and 20 to the lightest. Each curve appears to increase with longer durations but quickly reaches a unique horizontal asymptote. Increasing the number of speakers appears to cause a drop in the curve as a whole. Plots for other number of dimensions have the same pattern. Estimated values and 95% confidence intervals are presented for each curve at the terminating duration in the appendix.

Using a constant number of ten speakers, Fig. 16 shows the accuracy plots of different numbers of dimensions against duration. The value of ten speakers was selected because it is in the middle of the tested range and is representative of the trends seen in

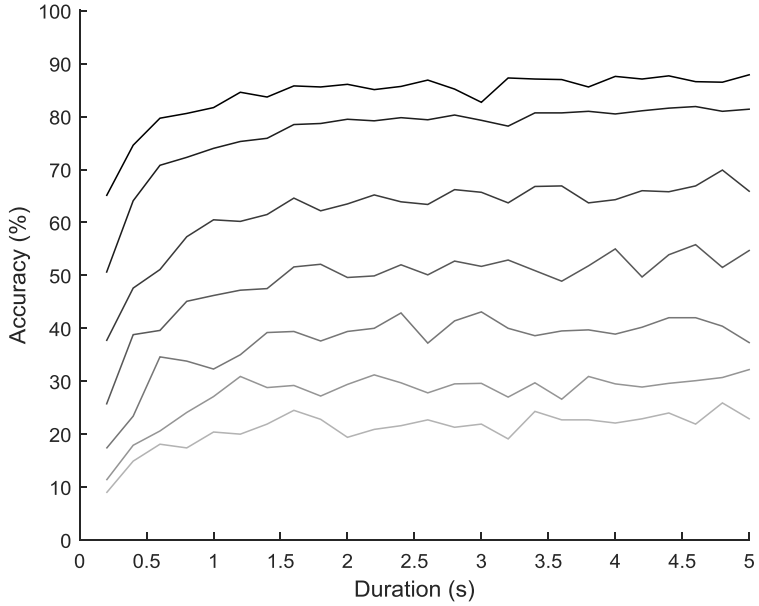

Fig. 15. Classifier accuracy with different number of speakers. The accuracy is plotted against the total segment duration for a variable number of speakers (2, 3, 5, 7, 10, 15, and 20) for three dimensions. Darker lines represent a smaller number of speakers, and lighter lines represent a larger number of speakers.

plots of other number of speakers. The darkness of the curve corresponds to the number of dimensions with 1 corresponding to the darkest and 12 to the lightest. This plot also exhibits the same behavior over time as curves for varying numbers of speakers, increasing with unique horizontal asymptotes. Increasing the number of dimensions causes a rise in the curve as a whole. This pattern appears to occur for any number of speakers. Estimated values and 95% confidence intervals are presented for each curve at the terminating duration in the appendix.

Having established the overwhelming tendency for the accuracy to approach different asymptotes for extended durations, the parametric behavior is examined across different numbers of speakers and dimensions at the maximum duration. The surface graph of data at the maximum duration (five seconds) is shown in Fig. 17 using number of speakers and dimensions as variables. Each cross section appears to represent a function with asymptotic relationships for increasing number of speakers or dimensions.

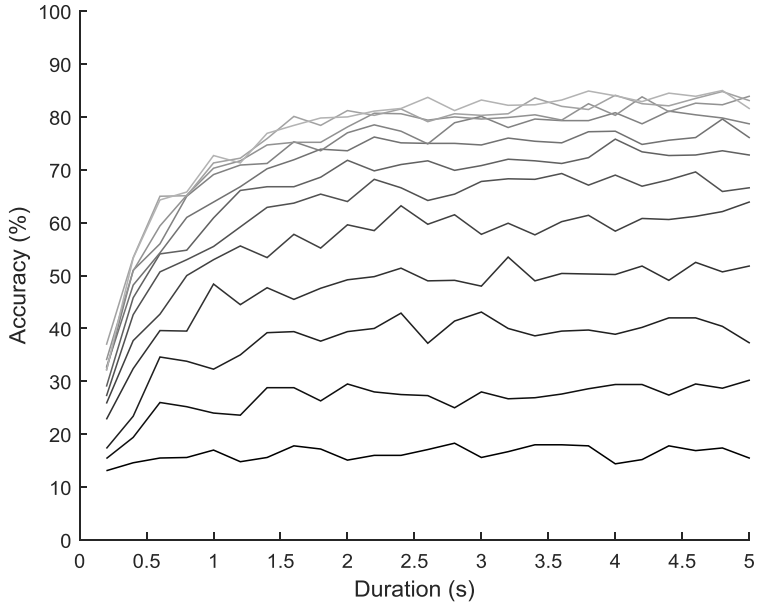

Fig. 16. Classifier accuracy with different number of dimensions. The accuracy is plotted against the total segment duration for a variable number of dimensions (1-12) with 10 speaker choices. Darker lines represent a smaller number of dimensions, and lighter lines represent a larger number of dimensions.

As shown in the previous figures, an increasing number of speakers results in a decreasing function, and an increasing numbers of dimensions results in an increasing function, each with asymptotes in the stated directions.

Extreme points in this figure are also quantified in TABLE XIII with estimated values and 95% confidence intervals. In regards to all points, tests of significance above chance can be performed using a chi-squared proportion test with a right-tailed rejection interval for alternate hypothesis being greater than chance. Each of the points for this surface is significantly above its respective chance value. The extreme points display

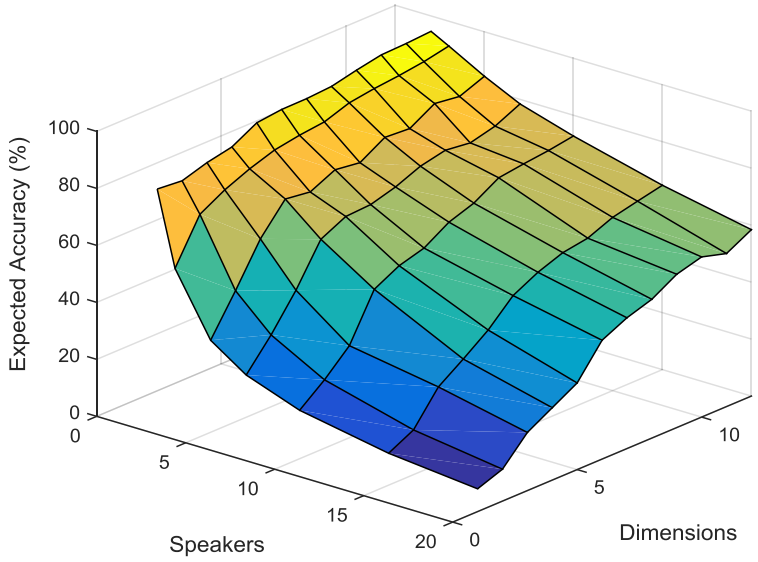

Fig. 17. Accuracy of classifiers against the number of dimensions and speakers. Time points are fixed at the maximum duration (5s).

TABLE XIII
CLASSIFIER ACCURACY AT EXTREME POINTS FOR MAXIMUM DURATION

Number of Dimensions	Number of Speakers	Estimate	95% CI
1	2	0.7980	0.7720 - 0.8217
1	20	0.0800	0.0647 - 0.0985
12	2	0.9460	0.9302 - 0.9584
12	20	0.5840	0.5532 - 0.6142

dramatically different accuracies. To display the different effects, each proportion is compared using a pooled two-proportion chi-squared test to verify the values are different. Each of the extreme values are significantly different from the others ($p<0.001$).

The potential to convey information in a certain parameter combination is quantified by transforming the accuracy data on the maximum duration surface. This quantity is determined by first scaling each result by a linear equation such that chance performance scores zero and perfect performance scores one. Each of these quantities is then scaled by the total number of speakers, with the resulting graph presented in Fig. 18. The greatest amount of information is conveyed by the full dimension classification system with a large number of speakers. Each of the other extreme value combinations displays significantly lower performance levels. This demonstrates both the number of speakers and dimensions interact to influence the capabilities of the classification system.

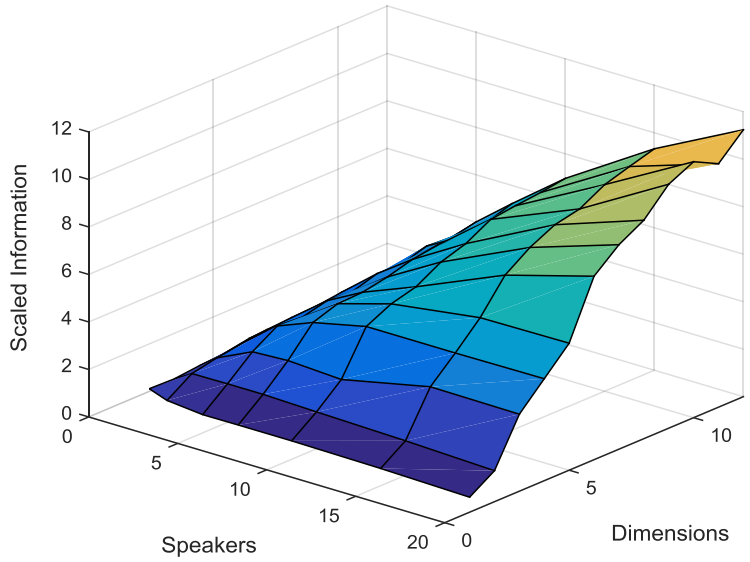

Fig. 18. Scaled accuracy of classifiers against the number of dimensions and speakers. Time points are fixed at the maximum duration (5s). The plot height for each point is determined by the amount above-chance performance and total number of speakers, roughly representing the amount of information conveyed in the transformation.

D. Discussion

The simulation results indicate computers utilizing a full dimensional representation of MFCC feature space can identify speakers with relative ease, even with a simplistic linear transformation and discrimination model. Accuracy in the task appears to be more robust when keeping the number of original dimensions intact, which is possible and desirable for an automated recognition system. However, there is a definite performance drop when adding complexity of the task by increasing the number of speakers.

The task becomes much more difficult when reducing dimensions, especially for larger numbers of speakers. The most dramatic drop in performance for the maximum allowed time can be seen in the case of a single dimension. While performance for this curve is always above chance, as it is for all the other tested points, it suffers the most dramatic drop when adding additional speakers. Other numbers of low dimensions, although less dramatic, also display the same fundamental problem.

Allowing for the limitations in usability when distinguishing a small number of speakers, it appears there is a fundamental information transfer bottleneck for smaller numbers of dimensions. Scaling for amount of above-chance performance demonstrates classification systems with a larger number of dimensions perform consistently better than those with low dimensional representations. Although systems with a small number of dimensions exhibit asymptotic behavior for information transfer in the tested range, this cannot be said for systems with more dimensions. Moreover, other functional relations cannot be ruled out, as different properties may emerge by changing the scope of the data or parameters. In particular, the results may differ by utilizing other MFCCs

(higher-order coefficients or deltas) or by considering much larger number of speakers. This last change in particular will determine the information plateaus for large dimensional classifications, but will require large increases in computational demands to build and test models.

As this combination of cue extraction and dimensionality reduction demonstrates serious shortcomings, a different mathematical approach is required. The information bottleneck imposed prevents utilizing this method for a small number of dimensions, which is a central requirement in this application.

IV. CONCLUSIONS

A. Overall Remarks

Results from the experiments presented in this study, guided by background research, demonstrate that sensory substitution devices can support the perception of indexical properties found in speech. Although this study limited empirical exploration to a gender discrimination task, the robust response of the second implementation gives credence to the idea that some of the potential of sensory substitution aids is in applications that have not been well considered.

The effectiveness of these aids does not apply equally to every individual. A consistent theme of variation of subjects, both overall and within a stimulus regime, demonstrates the sensory tools provided to a user are not utilized in the same way. Some subjects rely more heavily on one modality than another, and significant variation exists in the ability to utilize any of these stimuli. Variability in participant utilization of the changes in auditory cues also demonstrates within-modality differences in cue sensitivities as well. The demonstration of an effect relates to holistic changes in processing and not individual cues involved in compression, but it is very likely variations can be found in some of the constituent changes, such as eliminating high frequency portions of the spectra or utilizing linear predictive coding.

The possibility that unfamiliar cues are typically extracted in a multisensory way, with each modality contributing to the overall perception, has considerable implications on multisensory integration in learned tasks. Overall results show a multisensory condition, on average, confers an advantage in accuracy of perceptions over individual modalities regardless of individual modality strengths. This claim, however, is subject to

some debate as to what extent it applies to the population in general, as one of the tests yielded inconclusive results. Nonetheless, this effect definitely occurs in some people under their particular testing regime. This opens up the question of what makes these individuals different from those that displayed a definitive performance advantage for only one modality.

Initial tests in establishing a computational framework to handle cepstral coefficients in speaker identification tasks show more sophisticated mathematical models are required to relate speaker properties in a meaningful multidimensional space. Linear separations and transformations, especially in reduced dimensions, of raw MFCCs are not sufficient to allow consistently successful speaker identification. This is not a surprising finding considering no literature has been found to date with detailed explorations and good results using this method. All established procedures require decidedly non-linear transformations and models to construct speaker models. In contrast, it is interesting that speaker gender is well represented in a linear transformation, indicating not all indexical qualities require a complicated mathematical model. The good results in voice gender, however, do not appear to translate well to the property of speaker identity.

B. Limitations of Experimental Design

Using the simulation instead of actual cochlear implant patients presents a major limitation in generalizing the conclusions reached to real-world gains. Though simulations have been demonstrated to bring normal hearing participants to similar performance levels as CI patients in hearing tasks, there really is no complete substitute, especially when considering indexical considerations are not as strongly validated. Nonetheless, at the risk of becoming too invested in a design before validating on actual

patients, much of the prototype testing can still be handled beforehand. The approach in altering the confusing original display patterns shows pilot tests can succeed in bringing significant improvements to the design, even with users utilizing simulations.

Additionally, the subjects were not familiarized, or "stepped down", through the CI simulations. This is a procedure found quite often in the literature [18]–[20] to help acquaint people with the unfamiliar qualities of the vocoder. This could have an effect of artificially lowering the performance of subjects that, given enough time, would utilize the audio modality quite well. This practice should be uniformly implemented in a session as time will allow.

C. Future Directions

1. Device Components

Greater exploration of tactile sensory substitution will require testing different assemblies and actuators, such as linear resonant actuators (LRAs). Vibrotactile dimensions, especially those allegedly expressing "primacy" [96], will require reliably separating frequency from amplitude. The robustness of spatial dimensions to reduction of the utilized surface and alternate locations should also be tested. Relocation of arrays has been claimed to minimally affect performance changes [28], [32], but reproduction and verification of these results would be desirable given the claims are in regards to visual-to-tactile systems, which have more intuitive spatial relations between modalities.

The addition of a microphone pickup and implementation of the system in a live setting should also be considered. From a hardware standpoint, the location of the microphone in position relative to both the user and the display system are important design aspects. Because of emphasis on sensorimotor in effective use for other

modalities, at the very least the receiver microphones should work in much the same way as normal ears do for sound focusing. Attention should also be given to how user actions can modify the way the system processes sound, in effect changing modes on the fly. The signal processing will require a single environment to run with no manual interventions. Although the algorithms for each stage are readily available and can be implemented as a single process in principle, handling the errors occasionally produced, especially by fundamental frequency estimation, will require special attention.

2. Mapping Algorithms

An underlying assumption for the present design is that using functional relations to objective parameters might confer an advantage to cue utilization. However, this may not be the case. The present approach should be tested against both user-selected and random categorical approaches. This would require the algorithm to perform the categorization, which necessitates wholesale implementation and testing of a computational method.

The computational experiment demonstrated a need to improve the utilization of cepstrum coefficients in this application. For starters, the delta coefficients can be added for greater differentiation of place-based MFCC cues. A filter can also be established to selectively remove uninformative frames based on the 0th power coefficient. Additionally, testing can use one of the oft-cited methods in computational speech literature. Utilizing a support vector machine in a multi-class problem can definitively show if contiguous speaker-specific regions in MFCC parameter space exits. A suitable kernel-trick function for this problem would greatly substantiate a workable transformation for raw MFCCs. If this approach fails, i-vectors might provide a solution

for complexities in speaker models. A cursory glance shows i-vectors can reduce Gaussian mixture model parameters of speakers to a dimensionally reduced parameter space.

3. User Study Tasks

Following a suitable computational framework, expanding the device to a speaker identification task becomes a high priority. As a prerequisite independent of the computation, a prior experiment using the task as intended should be conducted with normal-hearing and normal utterances. This would ensure the experiment is not too difficult when introducing the simulated aspects. Assumptions cannot be made based on the ease of describing qualities of speaker voices, as these tasks are distinct based on taxonomical and neurological considerations [44], [118]. The question of sources for clips also presents a major item of concern. A potential database can contain clips of well-known voices, comprise an existing corpus of stranger's voices, or utilize a user's own familiar acquaintances. Each of these approaches would have tradeoffs in quality of results and ease in preparing for the experiment. Moreover, there is debate on speaker identification versus verification [75], as these tasks are also not identical. Real-world applicability of these processes might be dependent on the presence of prior expectations verification and require careful consideration. Verification is more representative when there is some expectation of a specific speaker, such as when a person in a contact task is listed as the caller for a phone conversation. Identification might be associated in scenarios where there is no expectation of a certain speaker, or as a subsequent test for a rejected speaker verification. These will require different experimental paradigms because they are fundamentally different tasks.

To add an additional level of realism, the task should also consider a simultaneous intelligibility task for CI simulations. The primary concern would be that attending to cues from the device, however beneficial for indexical properties, might increase cognitive load and thus prove detrimental for interpreting linguistic content. As a counterpoint, literature describing linkages between linguistic and indexical properties supports the idea that knowing the speaker aids with intelligibility. Overall detriments or improvements require empirical verification. Other indexical qualities or non-semantic linguistic cues like prosody can also be simultaneously tested. Although testing with this device occurred on a short to moderate term, longer-term studies and experiments are needed to demonstrate possible effects for extended use.

REFERENCES

[1] C. Mathers, A. Smith, and M. Concha, “Global burden of hearing loss in the year 2000,” World Health Organization, 2000.

[2] P. E. Mohr, J. J. Feldman, J. L. Dunbar, A. McConkey-Robbins, J. K. Niparko, R. K. Rittenhouse, and M. W. Skinner, “The societal costs of severe to profound hearing loss in the United States,” *Int. J. Technol. Assess. Health Care*, vol. 16, no. 04, pp. 1120–1135, Oct. 2000.

[3] C. Mathers, T. Boerma, and D. M. Fat, “The global burden of disease: 2004 update,” World Health Organization, Geneva, Switzerland, 2008.

[4] K. H. Lee, P. S. Roland, J. W. Kutz, and B. Isaacson, “Indications for cochlear implants: Etiologies of severe to profound hearing loss,” *Medscape*, 20-Jun-2015. [Online]. Available: http://emedicine.medscape.com/article/857164-overview#a3. [Accessed: 08-Jul-2015].

[5] F.-G. Zeng, S. Rebscher, W. Harrison, X. Sun, and H. Feng, “Cochlear implants: System design, integration, and evaluation,” *IEEE Rev. Biomed. Eng.*, vol. 1, pp. 115–142, Nov. 2008.

[6] “NIDCD fact sheet: Cochlear implants.” National Institute on Deafness and Other Communication Disorders, Mar-2011.

[7] “NIDCD fact sheet: Cochlear implants.” National Institute on Deafness and Other Communication Disorders, Aug-2014.

[8] L. M. Friesen, R. V. Shannon, D. Baskent, and X. Wang, “Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants,” *J. Acoust. Soc. Am.*, vol. 110, no. 2, pp. 1150–1163, Aug. 2001.

[9] M. S. Sommers, K. I. Kirk, and D. B. Pisoni, “Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format,” *Ear Hear.*, vol. 18, no. 2, pp. 89–99, Apr. 1997.

[10] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, Sep. 1953.

[11] V. Looi and J. She, “Music perception of cochlear implant users: A questionnaire, and its implications for a music training program,” *Int. J. Audiol.*, vol. 49, no. 2, pp. 116–128, Feb. 2010.

[12] Y.-Y. Kong, R. Cruz, J. A. Jones, and Z. Fang-Gang, “Music perception with temporal cues in acoustic and electric hearing,” *Ear Hear.*, vol. 25, no. 2, pp. 173–185, Apr. 2004.

[13] J. Huang, D. Gamble, K. Sarnlertsophon, X. Wang, and S. Hsiao, “Feeling music: Integration of auditory and tactile inputs in musical meter perception,” *PLoS ONE*, vol. 7, no. 10, p. e48496, Oct. 2012.

[14] J. Kreiman, D. VanLancker-Sidtis, and B. R. Gerratt, “Perception of voice quality,” in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. Malden, MA, USA: John Wiley & Sons, 2005, pp. 338–362.

[15] M. Cleary and D. B. Pisoni, “Talker discrimination by prelingually deaf children with cochlear implants: Preliminary results,” *Ann. Otol. Rhinol. Laryngol. Suppl.*, vol. 189, pp. 113–118, May 2002.

[16] M. Cleary, D. B. Pisoni, and K. I. Kirk, “Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear implants.,” *J. Speech Lang. Hear. Res.*, vol. 48, no. 1, pp. 204–223, Feb. 2005.

[17] T. Vongpaisal, S. E. Trehub, E. G. Schellenberg, P. van Lieshout, and B. C. Papsin, “Children with cochlear implants recognize their mother’s voice,” *Ear Hear.*, vol. 31, no. 4, pp. 555–566, Aug. 2010.

[18] Q.-J. Fu, S. Chinchilla, and J. J. Galvin, “The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users,” *J. Assoc. Res. Otolaryngol.*, vol. 5, no. 3, pp. 253–260, Sep. 2004.

[19] Q.-J. Fu, S. Chinchilla, G. Nogaki, and J. J. Galvin III, “Voice gender identification by cochlear implant users: The role of spectral and temporal resolution,” *J. Acoust. Soc. Am.*, vol. 118, no. 3, pp. 1711–1718, Sep. 2005.

[20] J. Gonzalez and J. C. Oliver, “Gender and speaker identification as a function of the number of channels in spectrally reduced speech,” *J. Acoust. Soc. Am.*, vol. 118, no. 1, pp. 461–470, Jul. 2005.

[21] J. W. Cray, R. L. Allen, A. Stuart, S. Hudson, E. Layman, and G. D. Givens, “An investigation of telephone use among cochlear implant recipients,” *Am. J. Audiol.*, vol. 13, no. 2, pp. 200–212, Dec. 2004.

[22] M. F. Dorman, H. Dove, J. Parkin, S. Zacharchuk, and K. Dankowski, “Telephone use by patients fitted with the Ineraid cochlear implant,” *Ear Hear.*, vol. 12, no. 5, pp. 368–369, Oct. 1991.

[23] S.-C. Peng, J. B. Tomblin, and C. W. Turner, “Production and perception of speech intonation in pediatric cochlear implant recipients and individuals with normal hearing,” *Ear Hear.*, vol. 29, no. 3, pp. 336–351, Jun. 2008.

[24] J. M. Lenden and P. Flipsen, “Prosody and voice characteristics of children with cochlear implants,” *J. Commun. Disord.*, vol. 40, no. 1, pp. 66–81, Jan. 2007.

[25] “Cochlear implant frequently asked questions,” *American Speech-Language-Hearing Association*, 2015. [Online]. Available: http://www.asha.org/public/hearing/Cochlear-Implant-Frequently-Asked-Questions/. [Accessed: 28-Jun-2015].

[26] “Cochlear implants: Patient health information,” *American Academy of Otolaryngology-Head and Neck Surgery*, 2015. [Online]. Available: http://www.entnet.org/content/cochlearimplants. [Accessed: 28-Jun-2015].

[27] “Cochlear implants: Benefits and risks of cochlear implants,” *U.S. Food and Drug Administration*, 06-Jun-2014. [Online]. Available: http://www.fda.gov/MedicalDevices/ProductsandMedicalProcedures/ImplantsandProsthetics/CochlearImplants/ucm062843.htm#d. [Accessed: 08-Jul-2015].

[28] P. Bach-y-Rita and S. W. Kercel, “Sensory substitution and the human–machine interface,” *Trends Cogn. Sci.*, vol. 7, no. 12, pp. 541–546, Dec. 2003.

[29] C. Lenay, O. Gapenne, S. Hanneton, C. Marque, and C. Genouëlle, “Sensory substitution: Limits and perspectives,” in *Touching for Knowing: Cognitive Psychology of Haptic Manual Perception*, Y. Hatwell, A. Streri, and E. Gentaz, Eds. Amsterdam, Netherlands: John Benjamins Publishing Company, 2003, pp. 275–292.

[30] K. A. Kaczmarek, J. G. Webster, P. Bach-y-Rita, and W. J. Tompkins, “Electrotactile and vibrotactile displays for sensory substitution systems,” *IEEE Trans. Biomed. Eng.*, vol. 38, no. 1, pp. 1–16, Jan. 1991.

[31] M. Auvray, S. Hanneton, and J. K. O’Regan, “Learning to perceive with a visuo-auditory substitution system: Localisation and object recognition with ‘the vOICe,’” *Perception*, vol. 36, no. 3, pp. 416–430, 2007.

[32] P. Bach-y-Rita, “Tactile sensory substitution studies,” *Ann. N. Y. Acad. Sci.*, vol. 1013, pp. 83–91, May 2004.

[33] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. A. Siegelbaum, and A. J. Hudspeth, *Principles of Neural Science*, 5th ed. New York, NY, USA: McGraw Hill Medical, 2013.

[34] D. Eagleman, “Can we create new senses for humans?,” presented at the TED2015, Vancouver, Canada, 16-Mar-2015.

[35] S. K. Nagel, C. Carl, T. Kringe, R. Märtin, and P. König, “Beyond sensory substitution—learning the sixth sense,” *J. Neural Eng.*, vol. 2, no. 4, pp. R13–R26, Dec. 2005.

[36] C. R. Fetsch, G. C. DeAngelis, and D. E. Angelaki, “Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons,” *Nat. Rev. Neurosci.*, vol. 14, no. 6, pp. 429–442, Jun. 2013.

[37] L. D. Rosenblum, “Primacy of multimodal speech perception,” in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. Malden, MA, USA: John Wiley & Sons, 2005, pp. 51–78.

[38] H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature*, vol. 264, pp. 246–248, 1976.

[39] T. M. Hopyan-Misakyan, K. A. Gordon, M. Dennis, and B. C. Papsin, “Recognition of affective speech prosody and facial affect in deaf children with unilateral right cochlear implants,” *Child Neuropsychol.*, vol. 15, no. 2, pp. 136–146, Mar. 2009.

[40] S. Levänen, V. Jousmäki, and R. Hari, “Vibration-induced auditory-cortex activation in a congenitally deaf adult,” *Curr. Biol.*, vol. 8, no. 15, pp. 869–872, Jul. 1998.

[41] E. T. Auer, L. E. Bernstein, W. Sungkarat, and M. Singh, “Vibrotactile activation of the auditory cortices in deaf versus hearing adults,” *NeuroReport*, vol. 18, no. 7, pp. 645–648, May 2007.

[42] J. P. Rauschecker, “Compensatory plasticity and sensory substitution in the cerebral cortex,” *Trends Neurosci.*, vol. 18, no. 1, pp. 36–43, Jan. 1995.

[43] D. R. Van Lancker and G. J. Canter, “Impairment of voice and face recognition in patients with hemispheric damage,” *Brain Cogn.*, vol. 1, no. 2, pp. 185–195, Apr. 1982.

[44] J. Kreiman, “Listening to voices,” in *Talker Variability in Speech Processing*, K. Johnson and J. W. Mullennix, Eds. San Diego, CA, USA: Academic Press, 1997, pp. 85–108.

[45] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, “Voice-selective areas in human auditory cortex,” *Nature*, vol. 403, no. 6767, pp. 309–312, Jan. 2000.

[46] M. Schürmann, G. Caetano, Y. Hlushchuk, V. Jousmäki, and R. Hari, “Touch activates human auditory cortex,” *NeuroImage*, vol. 30, no. 4, pp. 1325–1331, May 2006.

[47] T. Ro, A. Farnè, R. M. Johnson, V. Wedeen, Z. Chu, Z. J. Wang, J. V. Hunter, and M. S. Beauchamp, “Feeling sounds after a thalamic lesion,” *Ann. Neurol.*, vol. 62, no. 5, pp. 433–441, Nov. 2007.

[48] M. S. Beauchamp and T. Ro, “Neural substrates of sound–touch synesthesia after a thalamic lesion,” *J. Neurosci.*, vol. 28, no. 50, pp. 13696–13702, Dec. 2008.

[49] M. J. Naumer and J. J. F. van den Bosch, “Touching sounds: Thalamocortical plasticity and the neural basis of multisensory integration,” *J. Neurophysiol.*, vol. 102, no. 1, pp. 7–8, Jul. 2009.

[50] C. L. De Filippo and B. L. Scott, “A method for training and evaluating the reception of ongoing speech,” *J. Acoust. Soc. Am.*, vol. 63, no. 4, pp. 1186–1192, Apr. 1978.

[51] D. W. Sparks, L. A. Ardell, M. Bourgeois, B. Wiedmer, and P. K. Kuhl, “Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of connected discourse,” *J. Acoust. Soc. Am.*, vol. 65, no. 3, pp. 810–815, Mar. 1979.

[52] C. Wada, S. Ino, and T. Ifukube, “Proposal and evaluation of the sweeping display of speech spectrum for a tactile vocoder used by the profoundly hearing impaired,” *Electron. Commun. Jpn. Part III Fundam. Electron. Sci.*, vol. 79, no. 1, pp. 56–66, Jan. 1996.

[53] K. L. Galvin, G. Mavrias, A. Moore, R. S. Cowan, P. J. Blamey, and G. M. Clark, “A comparison of Tactaid II+ and Tactaid 7 use by adults with a profound hearing impairment,” *Ear Hear.*, vol. 20, no. 6, pp. 471–482, Dec. 1999.

[54] M. Rothenberg and R. D. Molitor, “Encoding voice fundamental frequency into vibrotactile frequency,” *J. Acoust. Soc. Am.*, vol. 66, no. 4, pp. 1029–1038, Oct. 1979.

[55] A. Boothroyd, “Wearable tactile sensory aid providing information on voice pitch and intonation patterns,” US4581491, 08-Apr-1986.

[56] D. Franklin, J. Franklin, and P. Hughes, “Method and apparatus for sound responsive tactile stimulation of deaf individuals,” US5035242, 30-Jul-1991.

[57] D. Eagleman, “Plenary talks: A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired,” in *2014 IEEE Haptics Symposium (HAPTICS)*, Houston, Texas, US, 2014, p. xvii.

[58] S. D. Novich and D. M. Eagleman, “A vibrotactile sensory substitution device for the deaf and profoundly hearing impaired,” in *2014 IEEE Haptics Symposium (HAPTICS)*, Houston, Texas, US, 2014, p. D79.

[59] M. J. Osberger, A. M. Robbins, R. T. Miyamoto, S. W. Berry, W. A. Myres, K. S. Kessler, and M. L. Pope, “Speech perception abilities of children with cochlear implants, tactile aids, or hearing aids,” *Am. J. Otol. Suppl.*, vol. 12, pp. 105–115, May 1991.

[60] J. M. Pickett and W. McFarland, “Auditory implants and tactile aids for the profoundly deaf,” *J. Speech Hear. Res.*, vol. 28, no. 1, pp. 134–150, Mar. 1985.

[61] H. Z. Tan, R. Gray, J. J. Young, and R. Traylor, “A haptic back display for attentional and directional cueing,” *Electron. J. Haptics Res.*, vol. 3, no. 1, pp. 1–20, Jun. 2003.

[62] R. I. M. Dunbar, “Neocortex size as a constraint on group size in primates,” *J. Hum. Evol.*, vol. 22, no. 6, pp. 469–493, Jun. 1992.

[63] M. Sambur, “Selection of acoustic features for speaker identification,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 23, no. 2, pp. 176–182, Apr. 1975.

[64] P. A. Luce and C. T. McLennan, “Spoken word recognition: The challenge of variation,” in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. Malden, MA, USA: John Wiley & Sons, 2005, pp. 591–609.

[65] L. C. Nygaard, “Perceptual integration of linguistic and nonlinguistic properties of speech,” in *The Handbook of Speech Perception*, D. Pisoni and R. Remez, Eds. Malden, MA, USA: John Wiley & Sons, 2005, pp. 390–413.

[66] W. D. Voiers, “Perceptual bases of speaker identity,” *J. Acoust. Soc. Am.*, vol. 36, no. 6, pp. 1065–1073, Jun. 1964.

[67] M. P. Gelfer, “A multidimensional scaling study of normal voice quality in females,” *J. Acoust. Soc. Am.*, vol. 81, no. S1, pp. S69–S69, May 1987.

[68] M. P. Gelfer, “A multidimensional scaling study of voice quality in females,” *Phonetica*, vol. 50, no. 1, pp. 15–27, 1993.

[69] J. B. Kruskal and M. Wish, *Multidimensional Scaling*. Beverly Hills, CA, USA: SAGE, 1978.

[70] F. Zheng, G. Zhang, and Z. Song, “Comparison of different implementations of MFCC,” *J. Comput. Sci. Technol.*, vol. 16, no. 6, pp. 582–589, Nov. 2001.

[71] S. Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 29, no. 2, pp. 254–272, Apr. 1981.

[72] J. N. Gowdy and Z. Tufekci, “Mel-scaled discrete wavelet coefficients for speech recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, 2000, vol. 3, pp. 1351–1354.

[73] L. Rudasi and S. A. Zahorian, “Text-independent talker identification with neural networks,” in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada, 1991, vol. 1, pp. 389–392.

[74] D. Van Lancker, J. Kreiman, and K. Emmorey, “Familiar voice recognition: Patterns and parameters: I. Recognition of backward voices,” *J. Phon.*, vol. 13, no. 1, pp. 19–38, Jan. 1985.

[75] D. A. Reynolds, “Overview of automatic speaker recognition,” presented at the The Center For Language and Speech Processing Summer Workshop, Baltimore, Maryland, 2008.

[76] S. S. Chen and P. S. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion,” in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, Virginia, US, 1998, vol. 8, pp. 127–132.

[77] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Trans. Speech Audio Process.*, vol. 3, no. 1, pp. 72–83, Jan. 1995.

[78] D. A. Reynolds, “Speaker identification and verification using Gaussian mixture speaker models,” *Speech Commun.*, vol. 17, no. 1–2, pp. 91–108, Aug. 1995.

[79] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digit. Signal Process.*, vol. 10, no. 1, pp. 19–41, Jan. 2000.

[80] M. A. El-Gamal, M. F. Abu El-Yazeed, and M. M. H. El Ayadi, “Dimensionality reduction for text-independent speaker identification using Gaussian mixture model,” in *Proceedings of IEEE 46th Midwest Symposium on Circuits and Systems*, Cairo, Egypt, 2003, vol. 2, pp. 625–628.

[81] L. R. Rabiner, “A tutorial on hidden Markov models and selected applications in speech recognition,” *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[82] P. Somervuo, “Experiments with linear and nonlinear feature transformations in HMM based phone recognition,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, 2003, vol. 1, pp. I–52 – I–55.

[83] S. M. Kamruzzaman, A. N. M. Rezaul Karim, M. Saiful Islam, and M. Emdadul Haque, “Speaker identification using MFCC-domain support vector machine,” *Int. J. Electr. Power Eng.*, vol. 1, no. 3, pp. 274–278, 2007.

[84] W. M. Campbell, J. P. Campbell, D. A. Reynolds, E. Singer, and P. A. Torres-Carrasquillo, “Support vector machines for speaker and language recognition,” *Comput. Speech Lang.*, vol. 20, no. 2, pp. 210–229, Apr. 2006.

[85] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, T. Leino, and A. Laukkanen, “I-vectors meet imitators: On vulnerability of speaker verification systems against

voice mimicry," in *Proceedings of Conference of the International Speech Communication Association*, Lyon, France, 2013.

[86] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," *Odyssey*, 2010.

[87] R. S. Johansson and A. B. Vallbo, "Tactile sensibility in the human hand: Relative and absolute densities of four types of mechanoreceptive units in glabrous skin," *J. Physiol.*, vol. 286, no. 1, pp. 283–300, Jan. 1979.

[88] R. S. Johansson and Å. B. Vallbo, "Spatial properties of the population of mechanoreceptive units in the glabrous skin of the human hand," *Brain Res.*, vol. 184, no. 2, pp. 353–366, Feb. 1980.

[89] R. S. Johansson and Å. B. Vallbo, "Tactile sensory coding in the glabrous skin of the human hand," *Trends Neurosci.*, vol. 6, pp. 27–32, 1983.

[90] G. A. Gescheider, S. J. Bolanowski, J. V. Pope, and R. T. Verrillo, "A four-channel analysis of the tactile sensitivity of the fingertip: Frequency selectivity, spatial summation, and temporal summation," *Somatosens. Mot. Res.*, vol. 19, no. 2, pp. 114–124, Jan. 2002.

[91] C. E. Sherrick, R. W. Cholewiak, and A. A. Collins, "The localization of low- and high-frequency vibrotactile stimuli," *J. Acoust. Soc. Am.*, vol. 88, no. 1, pp. 169–179, Jul. 1990.

[92] S. Weinstein, "Intensive and extensive aspects of tactile sensitivity as a function of body part, sex, and laterality," in *The Skin Senses*, 1st ed., D. Kenshalo, Ed. Springfield, IL, USA: Charles C Thomas, 1968, pp. 195–222.

[93] K.-U. Kyung, M. Ahn, D.-S. Kwon, and M. . Srinivasan, "Perceptual and biomechanical frequency response of human skin: implication for design of tactile displays," in *Proceedings of the First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems*, Pisa, Italy, 2005, pp. 96–101.

[94] S. Levänen and D. Hamdorf, "Feeling vibrations: enhanced tactile sensitivity in congenitally deaf humans," *Neurosci. Lett.*, vol. 301, no. 1, pp. 75–77, Mar. 2001.

[95] M. Hollins, R. Faldowski, S. Rao, and F. Young, "Perceptual dimensions of tactile surface texture: A multidimensional scaling analysis," *Percept. Psychophys.*, vol. 54, no. 6, pp. 697–705, Nov. 1993.

[96] R. D. Melara and D. J. A. Day, "Primacy of dimensions in vibrotactile perception: An evaluation of early holistic models," *Percept. Psychophys.*, vol. 52, no. 1, pp. 1–17, Jan. 1992.

[97] G. Park and S. Choi, “Perceptual space of amplitude-modulated vibrotactile stimuli,” in *Proceedings of IEEE World Haptics Conference*, 2011, pp. 59–64.

[98] L. A. Jones, J. Kunkel, and E. Piateski, “Vibrotactile pattern recognition on the arm and back.,” *Perception*, vol. 38, no. 1, pp. 52–68, 2009.

[99] L. M. Brown, S. A. Brewster, and H. C. Purchase, “Multidimensional tactons for non-visual information presentation in mobile devices,” in *Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, Espoo, Finland, 2006, pp. 231–238.

[100] R. W. Cholewiak, A. A. Collins, and J. C. Brill, “Spatial factors in vibrotactile pattern perception,” in *Proceedings of Eurohaptics Conference*, Birmingham, UK, 2001.

[101] P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proc. Inst. Phon. Sci.*, vol. 17, no. 1193, pp. 97–110, 1993.

[102] H. Traunmüller and A. Eriksson, “The frequency range of the voice fundamental in the speech of male and female adults,” 1993.

[103] T. McDaniel, S. Bala, J. Rosenthal, R. Tadayon, A. Tadayon, and S. Panchanathan, “Affective haptics for enhancing access to social interactions for individuals who are blind,” in *Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access*, C. Stephanidis and M. Antona, Eds. Springer International Publishing, 2014, pp. 419–429.

[104] “Haptic-Construction-Kit,” 2013. [Online]. Available: https://github.com/Haptic-Construction-Kit. [Accessed: 30-Apr-2014].

[105] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, “The DARPA speech recognition research database: Specifications and status,” in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[106] “AngelSim,” *Emily Shannon Fu Foundation*, 2014. [Online]. Available: http://www.tigerspeech.com/angelsim/angelsim_about.html. [Accessed: 30-Apr-2014].

[107] D. D. Greenwood, “A cochlear frequency-position function for several species—29 years later,” *J. Acoust. Soc. Am.*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.

[108] D. R. Ketten, M. W. Skinner, G. Wang, M. W. Vannier, G. A. Gates, and J. G. Neely, “In vivo measures of cochlear length and insertion depth of nucleus cochlear implant electrode arrays,” *Ann. Otol. Rhinol. Laryngol. Suppl.*, vol. 107, no. 175, pp. 1–16, Nov. 1998.

[109] M. W. Skinner, D. R. Ketten, L. K. Holden, G. W. Harding, P. G. Smith, G. A. Gates, J. G. Neely, G. R. Kletzker, B. Brunsden, and B. Blocker, "CT-derived estimation of cochlear morphology and electrode array position in relation to word recognition in Nucleus-22 recipients," *J. Assoc. Res. Otolaryngol.*, vol. 3, no. 3, pp. 332–350, Sep. 2002.

[110] W. Donaldson, "Measuring recognition memory," *J. Exp. Psychol. Gen.*, vol. 121, no. 3, pp. 275–277, 1992.

[111] G. A. Gescheider, *Psychophysics: The Fundamentals*, 3rd ed. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1997.

[112] S. E. Vollset, "Confidence intervals for a binomial proportion," *Stat. Med.*, vol. 12, no. 9, pp. 809–824, May 1993.

[113] A. M. Pires, "Confidence intervals for a binomial proportion: comparison of methods and software evaluation," *unpublished*.

[114] D. C. Montgomery, *Design and Analysis of Experiments*, 8th ed. Hoboken, NJ, USA: John Wiley & Sons, 2012.

[115] M. S. Bartlett, "The use of transformations," *Biometrics*, vol. 3, no. 1, pp. 39–52, Mar. 1947.

[116] C. Forbes, M. Evans, N. Hastings, and B. Peacock, "Bernoulli distribution," in *Statistical Distributions*, Hoboken, NJ, USA: John Wiley & Sons, 2010, pp. 53–54.

[117] M. Brookes, "VOICEBOX: Speech processing toolbox for MATLAB," 2014. [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html. [Accessed: 29-May-2014].

[118] P. D. Bricker and S. Pruzansky, "Speaker recognition," in *Contemporary Issues in Experimental Phonetics*, N. Lass, Ed. New York, NY, USA: Academic Press, 1976, pp. 295–326.

APPENDIX A

BOX-COX TESTS FOR RESPONSE TIME DATA

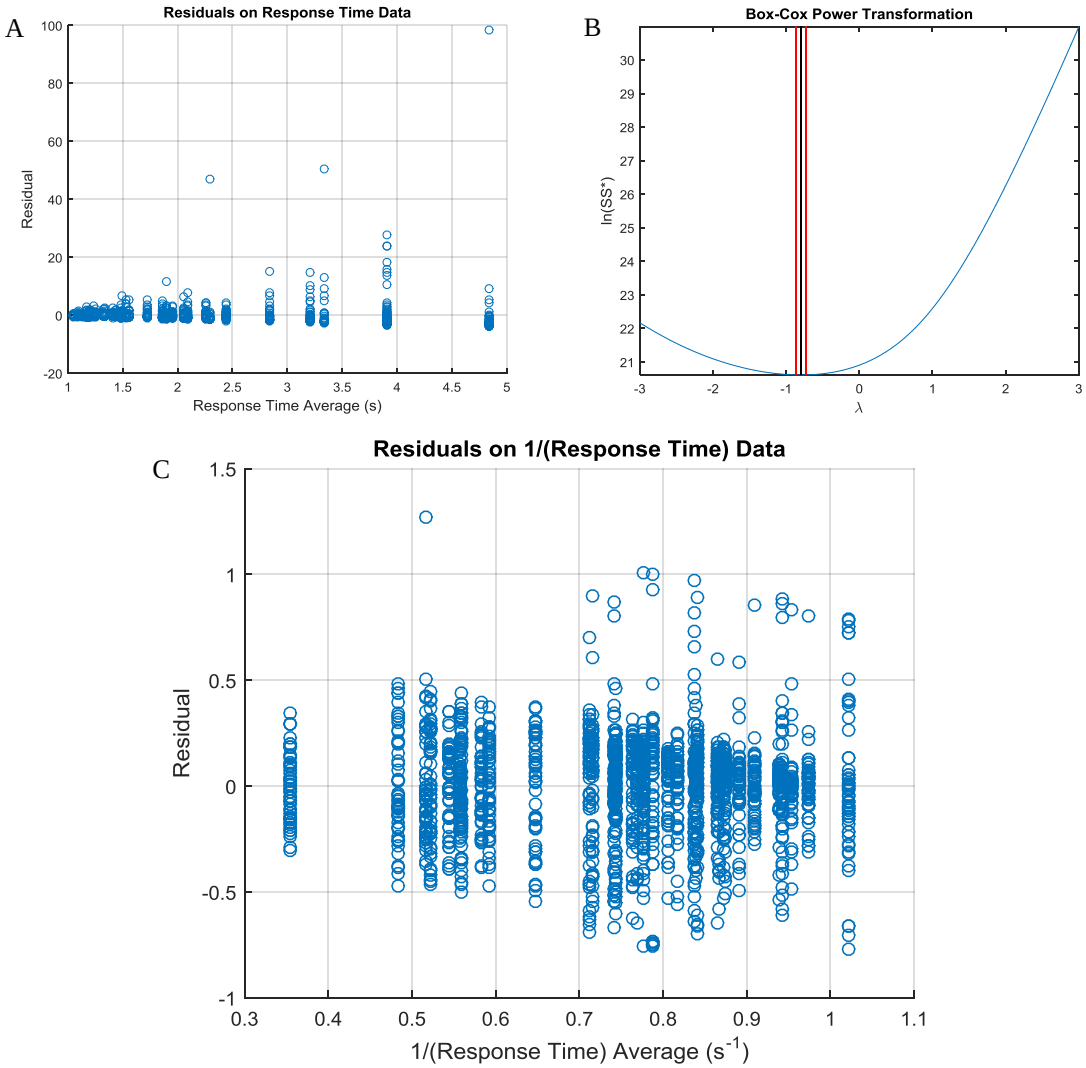

Fig. A1. Effects on residuals in the Box-Cox procedure for response times in Experiment 1. (A) Residual plots for the original response time data, showing the distinctly non-normal properties. (B) Adjusted sum of squares for Box-Cox power transforms. The best-suited power is closest to -1. (C) Residuals on the response times utilizing an inverse transformation, demonstrating the residuals are now approximately uniform and normal.

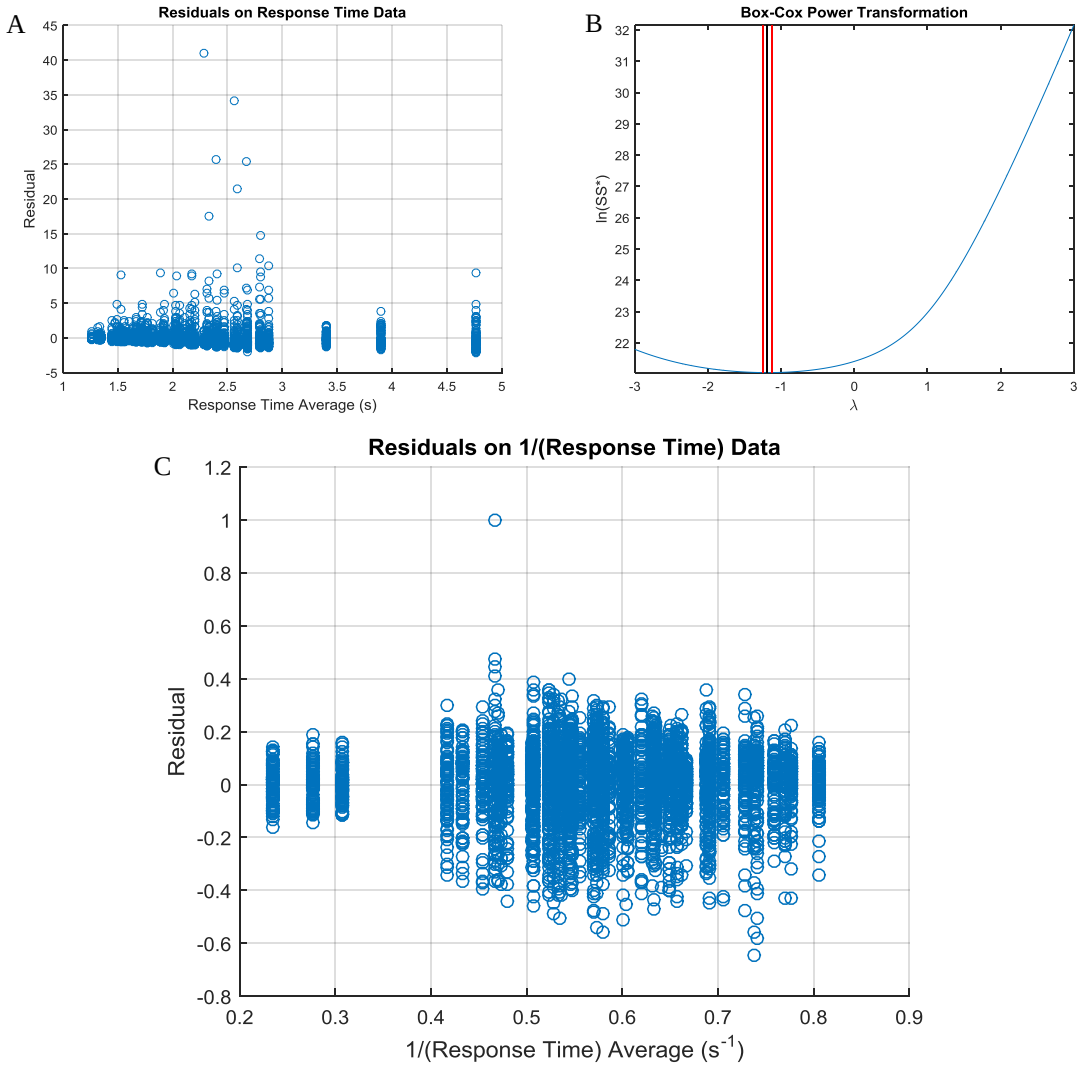

Fig.A2. Effects on residuals in the Box-Cox procedure for response times in Experiment 2. (A) Residual plots for the original response time data, showing the distinctly non-normal properties. (B) Adjusted sum of squares for Box-Cox power transforms. The best-suited power is closest to -1. (C) Residuals on the response times utilizing an inverse transformation, demonstrating the residuals are now approximately uniform and normal.

APPENDIX B

ANOVA TABLES FOR EXPERIMENT 2

TABLE B1
ANOVA FOR ACCURACY DATA IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Stage 1: Between Stimuli					
A	17.6528	5	3.53056	2.698792	0.073735
B(A)	15.6984	12	1.3082	5.16739	1.13 E-08
C	13.8184	2	6.9092	9.117846	0.001133
A*C	16.968	10	1.6968	2.239212	0.051437
B(A)*C	18.1864	24	0.757767	2.993178	1.33 E-06
Error	1080	4266	0.253165		
Discarded		0			
Total		4319			
Stage 2.1: Within All Audio Stimuli					
D	2.252	1	2.252	3.957125	0.069961
A*D	1.4644	5	0.29288	0.514637	0.760547
B(A)*D	6.8292	12	0.5691	2.21949	0.008949
C*D	0.2212	1	0.2212	1.331995	0.270918
A*C*D	1.9868	5	0.39736	2.392774	0.100127
B(A)*C*D	1.9928	12	0.166067	0.64766	0.802457
Error	720	2808	0.25641		
Discarded		35			
Total		2879			
Stage 2.2: Within All Haptic Stimuli					
E	0.0208	1	0.0208	0.104418	0.752153
A*E	1.3596	5	0.27192	1.36506	0.303845
B(A)*E	2.3904	12	0.1992	0.77688	0.675063
C*E	0.076	1	0.076	0.388482	0.544763
A*C*E	1.0744	5	0.21488	1.098381	0.410409
B(A)*C*E	2.3476	12	0.195633	0.76297	0.689464
Error	720	2808	0.25641		
Discarded		35			
Total		2879			
Stage 3: Within Combined Stimuli					
D*E	0.0002	1	0.0002	0.000717	0.979078
A*D*E	2.1328	5	0.42656	1.529254	0.252714
B(A)*D*E	3.3472	12	0.278933	1.059947	0.390708
Error	360	1368	0.263158		
Discarded		53			
Total		1439			

TABLE B2
ANOVA FOR RESPONSE TIME DATA IN EXPERIMENT 2

Source	Sum Sq.	d.f.	Mean Sq.	F	Prob>F
Stage 1: Between Stimuli					
A	11.1858	5	2.23716	0.736698	0.610173
B(A)	36.4409	12	3.036742	158.7902	0
C	3.0998	2	1.5499	5.477162	0.010978
A*C	3.4456	10	0.34456	1.217634	0.329068
B(A)*C	6.7914	24	0.282975	14.79666	5.37 E-58
Error	81.5649	4265	0.019124		
Discarded		0			
Total		4318			
Stage 2.1: Within All Audio Stimuli					
D	0.0007	1	0.000703	0.111478	0.74423
A*D	0.0989	5	0.01978	3.135535	0.048644
B(A)*D	0.0757	12	0.006308	0.315775	0.986881
C*D	0.0235	1	0.0235	1.135266	0.307624
A*C*D	0.1175	5	0.0235	1.135266	0.393729
B(A)*C*D	0.2484	12	0.0207	1.036177	0.411889
Error	56.0962	2808	0.019977		
Discarded		35			
Total		2879			
Stage 2.2: Within All Haptic Stimuli					
E	0.0194	1	0.0194	2.895522	0.114568
A*E	0.0694	5	0.01388	2.071642	0.13984
B(A)*E	0.0804	12	0.0067	0.36325	0.975952
C*E	0.0077	1	0.0077	0.578947	0.461416
A*C*E	0.1563	5	0.03126	2.350376	0.104563
B(A)*C*E	0.1596	12	0.0133	0.721078	0.732081
Error	51.774	2807	0.018445		
Discarded		35			
Total		2878			
Stage 3: Within Combined Stimuli					
D*E	0.00016	1	0.000162	0.028181	0.86948
A*D*E	0.2239	5	0.04478	7.810465	0.001765
B(A)*D*E	0.0688	12	0.005733	0.295972	0.990133
Error	26.4998	1368	0.019371		
Discarded		53			
Total		1439			

APPENDIX C

CONFIDENCE INTERVALS FOR EXPERIMENT 3

TABLE C1
CLASSIFIER ACCURACY AT MAXIMUM DURATION FOR VARIABLE NUMBER OF DIMENSIONS (SPEAKERS = 10)

Number of Dimensions	Estimate	95% CI
1	0.1550	0.1339 - 0.1787
2	0.3020	0.2743 - 0.3312
3	0.3730	0.3436 - 0.4034
4	0.5180	0.4870 - 0.5488
5	0.6390	0.6088 - 0.6682
6	0.6660	0.6362 - 0.6945
7	0.7280	0.6996 - 0.7547
8	0.7610	0.7336 - 0.7864
9	0.7870	0.7606 - 0.8113
10	0.8390	0.8149 - 0.8605
11	0.8310	0.8065 - 0.8530
12	0.8160	0.7908 - 0.8388

TABLE C2
CLASSIFIER ACCURACY AT MAXIMUM DURATION FOR VARIABLE NUMBER OF SPEAKERS (DIMENSIONS = 3)

Number of Speakers	Estimate	95% CI
2	0.8790	0.8573 - 0.8978
3	0.8140	0.7887 - 0.8369
5	0.6590	0.6291 - 0.6877
7	0.5470	0.5160 - 0.5776
10	0.3730	0.3436 - 0.4034
15	0.3220	0.2938 - 0.3516
20	0.2290	0.2040 - 0.2561

APPENDIX D

APPROVAL FOR HUMAN STUDY

ASU Knowledge Enterprise Development

APPROVAL: EXPEDITED REVIEW

Sethuraman Panchanathan
Knowledge Enterprise Development, Office of (OKED)
480/965-4407
panch@asu.edu

Dear Sethuraman Panchanathan:

On 11/3/2014 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Vibrotactile Patterns to Convey Auditory Cues
Investigator:	Sethuraman Panchanathan
IRB ID:	STUDY00001779
Category of review:	(4) Noninvasive procedures, (7)(a) Behavioral research
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• HRP-502a - TEMPLATE CONSENT SOCIAL BEHAVIORAL (7).pdf, Category: Consent Form; • HRP-503a - TEMPLATE PROTOCOLSOCIAL BEHAVIORAL.docx, Category: IRB Protocol; • Experimental Procedure.docx, Category: IRB Protocol; • Subject Information Form.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Study Participation Receipt Form.pdf, Category: Other (to reflect anything not captured above); • Email.pdf, Category: Recruitment Materials; • Flyer.pdf, Category: Recruitment Materials;

The IRB approved the protocol from 11/5/2014 to 11/2/2015 inclusive. Three weeks before 11/2/2015 you are to submit a continuation request for continuing approval or closure.

If continuing review approval is not granted before the expiration date of 11/2/2015 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: Shantanu Bala
Visar Berisha
Troy McDaniel
Austin Butts
Shantanu Bala
Sethuraman Panchanathan
Stephen Helms Tillery

APPENDIX E

APPROVAL OF STUDY REVISION

ASU Knowledge Enterprise Development

APPROVAL: MODIFICATION

Sethuraman Panchanathan
Knowledge Enterprise Development, Office of (OKED)
480/965-4407
panch@asu.edu

Dear Sethuraman Panchanathan:

On 2/16/2015 the ASU IRB reviewed the following protocol:

Type of Review:	Modification
Title:	Vibrotactile Patterns to Convey Auditory Cues
Investigator:	Sethuraman Panchanathan
IRB ID:	STUDY00001779
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• Subject Information Form.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Study Questions (3).pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • Flyer.pdf, Category: Recruitment Materials; • Study Participation Receipt Form.pdf, Category: Other (to reflect anything not captured above); • HRP-503a - TEMPLATE PROTOCOLSOCIAL BEHAVIORAL.docx, Category: IRB Protocol; • HRP-502a - TEMPLATE CONSENT SOCIAL BEHAVIORAL.pdf, Category: Consent Form; • Experimental Procedure.docx, Category: IRB Protocol; • Email.pdf, Category: Recruitment Materials;

The IRB approved the modification.

When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc:

APPENDIX F

DEFENSE PRESENTATION

[Consult Attached Files]

Biographical Sketch

Austin received his Bachelor of Science in Biomedical Engineering at Texas A&M University in 2013, focusing on instrumentation and graduating with Summa Cum Laude honors. He is finalizing his work for a Master of Science degree in Biomedical Engineering at Arizona State University, emphasizing a specialization in neural engineering. As part of his research in sensory substitution aids, he was awarded an ASU Graduate and Professional Student Association (GPSA) Jumpstart Grant for the Spring 2015 semester. He has also performed additional research in cochlear implants, designing a program interface and signal processing algorithms in sound quality simulations, and presented his findings at the Conference on Implantable Auditory Prostheses (CIAP) 2015. He plans to pursue a career in neural implants and hearing devices.