VOCAL CONNECTION

Rethinking the Voice as a Medium for Personal, Interpersonal, and Interspecies Understanding

Rébecca Kleinberger
Bachelor of Engineering
Arts et Métiers ParisTech, 2010

Master of Research in Virtual Environment, Imaging and Visualisation
University College London, 2012

Master of Mechanical Engineering
Arts et Métiers ParisTech, 2013

Master in Media Arts and Sciences
Massachusetts Institute of Technology, 2014

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** at the **Massachusetts Institute of Technology**
May 2020
© 2020 Massachusetts Institute of Technology. All rights reserved.

Author: RÉBECCA KLEINBERGER
Program in Media Arts and Sciences
16 February 2020

Certified by: TOD MACHOVER
Muriel R. Cooper Professor of Music and Media
Program in Media Arts and Sciences
Thesis Supervisor

Accepted by: TOD MACHOVER
Academic Head
Program in Media Arts and Sciences

VOCAL CONNECTION

Rethinking the Voice as a Medium for Personal, Interpersonal, and Interspecies Understanding

Rébecca Kleinberger

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning on February 16, 2020, in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** at the **Massachusetts Institute of Technology**

Abstract

Voices are ubiquitous and familiar, so much so that it is easy to forget how fundamentally important vocal signals really are to how we relate to others and to ourselves. Vocal experiences can take many forms (audible, tangible, silent, internal, external, neurological, remote, etc.) and offer great potential for bridging diverse fields. I am proposing a new approach for looking at the voice holistically, in its experiential nature, based on its propensity to connect. This dissertation introduces and examines methods for the creation of interactive voice-based experiences that foster novel and profound connections.

I present three projects to support and illustrate this approach by establishing connections at three levels: individual, interpersonal, and extending beyond human languages. The **Memory Music Box** establishes a sense of connection across space and time, and is specially designed to encourage conversation and to enhance a sense of connectedness for older adults. With the **Mumble Melody** initiative, I extract musicality from everyday speech as a way to access inner voice processes and help people who stutter gain increased fluency. Finally, with the **Sonic Enrichment at the Zoo** project, I present ways to improve connections within and between species—including between humans and animals—by exploring sonic and vocal enrichment interventions at the San Diego Zoo.

Each of these projects represents a different angle from which to consider the potential of the voice for creating new forms of connection. Such is the vision of this work. I consider the notion of connectedness broadly, including the raising of personal self-awareness, the creation of strong interpersonal bonds, and the potential to create new forms of empathetic understanding with other species. Although this research focuses on the voice, it extends beyond this realm. The broader themes examined through this work have implications in the fields of neurology, cognitive sciences, assistive technologies, human-computer interactions, communication sciences, and rapport-building. Indeed, since the voice is a versatile projection of ourselves into the world, it offers a unique perspective for the study and enhancement of cognition, learning, personal development, and wellbeing.

Thesis Supervisor: TOD MACHOVER
Muriel R. Cooper Professor of Music and Media
Program in Media Arts and Sciences

VOCAL CONNECTION

Rethinking the Voice as a Medium for Personal, Interpersonal, and Interspecies Understanding

Rébecca Kleinberger

The following person served as a reader for this thesis:

Thesis Reader: JANET BAKER

Research Affiliate

Media Laboratory

Massachusetts Institute of Technology

VOCAL CONNECTION

Rethinking the Voice as a Medium for Personal, Interpersonal, and Interspecies Understanding

Rébecca Kleinberger

The following person served as a reader for this thesis:

Thesis Reader: SATRAJIT GHOSH

Principal Research Scientist,

McGovern Institute, McGovern Institute for Brain Research, MIT

Assistant Professor of Otolaryngology–Head and Neck Surgery,

Harvard Medical School

Aknowledgements

For their help and inspiration, I am very thankful to:

My advisor: Tod Machover,

The members of my committee: Janet M. Baker and Satrajit S.Ghosh,

Other professors who have inspired me during my eight years at the Media Lab: Joseph Paradiso, Ethan Zuckerman, Roz Picard, Neil Gershenfeld, Pattie Maes, Hiroshi Ishii, Louis D. Braida, John J Rosowski,

Opera of the Future, Past and Present: Ben Bloomberg, Charles Holbrow, Nikhil Singh, Alexandra Rieger, Sizi Chen, Priscilla Capistrano, Nicole L'Huillier, Karsten Schuhl, Manaswi Mishra, Hanne Lienhard, Aaron Montoya-Moraga, Janice Wang, David Nunez, David Su, Bryn Bliska, Andy Cavatorta, Kelly Donovan, Hane Lee, Sarah Platte, Eyal Shahar, Adam Boulanger, Tim Savas,

All my UROPs: George Stefanakis, Janelle Sands, Chantine Akiyama, Galen Chuang, Sebastien Franjou, Sarah Sime, Jack Moore, Lydia Yu, Anne Harrington,

My extended lab family: Deepak Jagdish, Chia Evers, Paula Aguilera, Jonathan William, Xiao Xiao, Alisha Panjwani, Katia Vega, Alisha Panjwani, Cindy Hsin-Liu Kao, Tony Shu, Sunny Jolly, Udayan Umapathi, Harpreet Sareen, Javier Hernandes, Santiago Alfaro, Amna Cavalic-Carreiro, Brian Mayton, Gershon Dublon, Steven Keating, Daniel Novy, Nan-Wei Gong, Nan Zhao, Spencer Russell, Clément Duhart, Donald Derek, Kristy Johnson, June Kinoshita, Sujoy Kumar, Dávid Lakatos, Mirei Rioux, Mark Feldmeier, Jon Ferguson, Michal Firstenberg, Elliott Hedman, Jacqueline Kory, Natan Linder, Pragun Goyal, Elliott Hedman, Max Little, Daniel McDuff, Laia Mogas-Soldevila, Jifei Ou, Will Patrick, Sandra Richter, Daniel Smilkov, Yoav Sterman, Carlos Gonzalez Uribe, Thom Howe, Juliana Cherston, Artem Dementyev, Magaret Evans, Adam Horowitz, Joy Buolamwini, Sarah Taylor, Cameron Taylor, Jessica Artiles, Morris Vanegas, Charles Fracchia, Javier

Pietro, Darthur Petron, Edwina Portocarrero, Santiago Alfaro, Sunanda Sharman, Rachel Smith, Susan Williams, Felix Kraemer, Nicolas Hogan,

Linda Peterson, Keira Horowitz, Lily Zhang, Mahy El-Kouedi, Stacie Slotnick, Ryan McCarthy, Rickey Ishida, Cornelle King, Kevin Davis, Jessica Tsymbal, Bill Lombardi, Tom Lutz, John Difrancesco, Janine Liberty, Ellen Hoffman, Hellen Curley,

Other friends and moral supports at and beyond MIT: Michael Erkkinen & Anna, Gabriel Miller, Racheal Meager, Anya Burkart, Adi Hollander, Claudio Baroni, Arman & Donna Rezaee, Jessica & Alex Wallar, Ben Houge, Jutta Friedrichs, Thaddeus Bromstrup

But also: The National Stuttering Association Boston Chapter
Jacques Picard, Jacques Bojin, Jacques Paccard, le GP58 et la 1i209,
Cristina Amati, Marie Weber and Benjamin Lize,
All the community from La Maison Française in New House,
The bus drivers of Tech Shuttles and Cambridge Saferides West,

My Family: Babeth & Marion, les parents, Mamie, Mia the tiger, Poopok the dinosaur and Witzy the wolf,

Akito van Troyer and Petit, to whom, I dedicate this work.

Contents

List of Figures

Prelude

When in the womb, well before the fetus can open her eyes at six months of gestation[1], before she develops the sense of taste at week 11[2] or smell at week 20[3]; even before she can hear sounds, starting around week 16[4]; her first functional sensory perception is through the development of her somesthetic system which is functional starting in week eight, including the senses of touch, proprioception, and haptic perception[5]. The fetus is first responsive to touch stimuli. The first sensations she perceives from the outside world are tactile sensations. And the first tactile sensations reaching her are vibrations from her mother's heartbeat and voice [6]. The vocal connection with the mother is one of the first of all experiences.

[1] Peter G Hepper et al. Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, 1993

[2] Beverly J Cowart. Development of taste perception in humans: sensitivity and preference throughout the life span. *Psychological bulletin*, 1981

[3] Jean-Pierre Lecanuet and Benoist Schaal. Sensory performances in the human foetus: A brief summary of research. *Intellectica*, 2002

[4] Peter G Hepper and B Sara Shahidullah. Development of fetal hearing. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 71(2):F81–F87, 1994

[5] Jean-Pierre Lecanuet and Benoist Schaal. Sensory performances in the human foetus: A brief summary of research. *Intellectica*, 2002

[6] Peter G Hepper et al. Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, 1993; and William P Fifer and Christine M Moon. The role of mother's voice in the organization of brain function in the newborn. *Acta Paediatrica*, 83(s397):86–93, 1994

1 - Introduction

The voice is the center of our sonic universe. We are our voice, both to the ears of others and in our own perception of ourselves. Whether it is raised, found, given or heard, the voice is *much more* than words. While being speechless is being unable to find one's words, being voiceless is being denied social existence.

1.1 - What is the voice?

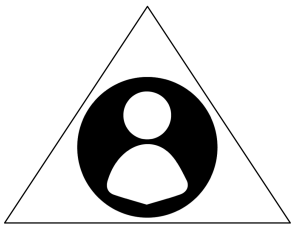
One can look at the voice from a very simplistic standpoint. In our body, the bio-mechanics of the voice starts with airflow from the lungs coordinated by the action of the diaphragm. The airflow builds up pressure in the larynx which activates self-sustained vibrations of the vocal folds creating a quasi-sinusoidal audible vibration. This vibration is filtered by the shape of the nasal and mouth cavities to create a vocal sound. This being said, I have explained *how the voice works*, but not really *what* the voice *is*. Is the voice found in the sounds produced by this mechanism? Is it the mechanism itself? Is it the organ that hosts the mechanism? Is it the sound as perceived by others or by the speaker? Or is it the physical phenomenon or the sound wave itself? Is it the moment of sound production? And what about the neural control of the production mechanism, or the perception mechanism? To tackle the question of *what*, I look at the how, the who, and the why of the voice.

- The HOW is the investigation, the field chosen to guide my inquiry.
- The WHO is the manifestation itself, the nature of the phenomenon.
- The WHY is its purpose, its most essential nature.

The HOW

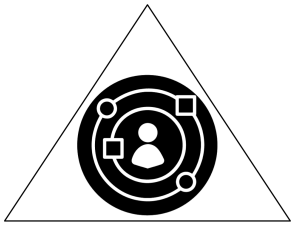
The voice has intrigued scientists, researchers, and artists from various different fields, including clinical medicine, acoustics, sociology, biology, physiology, neurology, communication and information theories, computer science, Digital Signal Processing, Natural Language Processing, linguistics, music, and more. Each of these fields has yielded huge amounts of valuable results. Surgeons and biologists are now able to understand the

HOLISTIC



pieces at play in the vocal apparatus and fix tissues, muscular disorders, and nodules that affect the voice. Neurologists are finding neural control centers and pathways of the voice in the brain. Evolutionary biologists have uncovered the evolution of the voice since our most ancient ancestors. Computer scientists can now analyze voices in terms of words or emotions and synthesize existing or new disembodied voices. The quantity of information in the voice has been analyzed from bitrate to semantic meanings. Languages and accents have their experts too, who discover new connections in verbal and vocal concepts of various populations around the globe. Sociologists explore what our words reveal about our societies. Psychologists analyze what our vocal patterns and Freudian slips reveal about our subconscious minds. Musicians, poets, and artists have elevated the voice to a spiritual level, capable of moving and touching crowds in the depths of their humanity. In this work, I consider the voice in its full breadth of meaning. I adopt a holistic approach to synthesize as many facets of the voice as possible.

EXPERIENTIAL

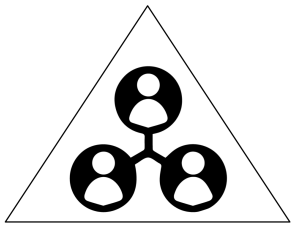


The WHO

The voice as a phenomenon arises from an entire ecosystem of cascading factors. The voice is more than a hundred muscles in activation, millions of neurons in control, hundreds of Hz in frequency range, billions of years in evolution, spread over 6,500 languages and infinite tonalities in the subtle expression of emotions. In this work, I look at the different factors and I choose to consider the voice in its experiential nature. In this dissertation, I define experience as a situation that has the potential to extract us from the default state of mind. We have an experience every time we remember that there is something to experience, something to be present for. Experiencing is about existing in the present. Manoury calls "aesthetic quantum" the "primordial conditions for someone to be in a favorable mental disposition to appreciate the aesthetic dimension of what they perceive"[7].

[7] Philippe Manoury. *La musique du temps réel.* Editions MF, 2012

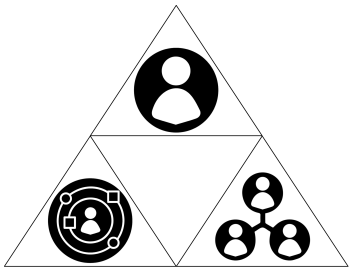
CONNECTED



The WHY

Beyond the direct studies of the voice, the voice itself is the tool that allows our species to function and strive, from its early beginnings up to the present. Why? I believe it is because the voice is absolutely central to our ability to connect with ourselves and with others—to transform otherness into togetherness. Reaching a better understanding of the potential of the voice to create connections opens the door to beneficial personal and global insights. This is also my reason to study the voice. Because it calls for the creation of novel types of experiences curated to establish or deepen the connections between various agents.

With the multiplicities of possible angles, modalities, and reasons to look at the voice, it can be challenging to embrace the voice comprehensively. This work is an attempt to do just that. I am proposing another way to look at the voice, holistically, in its experiential nature, based on its tentacular propensity to connect.



1.2 - Paradigm triangle

The multifaceted character of the voice is simultaneously the biggest challenge, most thrilling motivation, and most important foundation of this work.

The object—voice—is ubiquitous and familiar. However, beyond this simple archetype, there are other conceptual, sometimes paradoxical, worldviews through which one can understand the voice. For instance, the human voice can be seen both as what distinguishes us from other primates, and as having deep commonality with how other mammals and even birds communicate. The voice is both the universal ground that allowed humans to create common languages, and also the unique signature that lets us identify a specific person on the phone in less than a second.

In tackling the question of the nature of the voice, we are navigating this question in a multidisciplinary way, and outside established field boundaries. We must, therefore, establish our own set of concepts, thought patterns, theories, and postulates. Defining the voice, and the context within which one studies it, therefore appears as a paradigm question[8]. Instead of proposing one unique archetype of voices, this work articulates three different worldviews framed around the voice, each emerging from our unifying approach to looking at the voice holistically, in its experiential nature, and in its propensity to connect. Each proposed paradigm acts as a flashlight that illuminates and reveals anthologized hidden qualities and essences of the voice.

Our first paradigm frames voice beyond the words. In this context, we are focusing on studying the voice aside from verbal content, defined as all parts of the voice signal that can accurately be transcribed into words. In the last few decades, research in the field of computational linguistics and speech recognition has achieved phenomenal results in extracting semantic meaning from speech[9]. This verbal content has since enabled a wide range of applications, from healthcare and assistive systems[10] to helicopter piloting[11]. Though words are an important vocal component, they are far from reflecting the entirety of the message. In this work, I

[8] Thomas S Kuhn. The structure of scientific revolutions. *Chicago and London*, 1962

[9] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 1989

[10] Scott Durling and Jo Lumsden. Speech recognition use in healthcare applications. In *International conference on advances in mobile computing and multimedia*, 2008

[11] Martin Cooke et al. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 2001

argue that what remains when we remove the words—the music of the voice—is at least as meaningful, if not more so, for building rapport and connectedness. Indeed, from someone else's voice, we can potentially gain an understanding not only of the message but also of the current mental state of the speaker and, more deeply, help the speaker understand themselves as their own person. Recently, researchers have also started investigating emotion recognition through voice[12]. Those approaches can be interesting for their potential to help people and machines classify and recognize specific affective modes, but are limited in their scope for understanding the full extent of the vocal manifestation. This is due both to the diversity of personal histories[13], and to our inherent limitations in understanding and defining emotion[14]. The non-verbal parts of the voice cannot be reduced to the concept of emotion as defined today by the research community. The voice also contains less apparent information about our changing identities and the subtleties of human interactions. When interacting, we are sharing thoughts, but we are also sharing parts of who we are as of a moment in time. Besides words and emotions, the voice also contains clues about where we come from, who we are and where we might be heading. In this dissertation work, I focus on the potential of the voice to access deeper levels of connection and understanding.

[12] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 2001

[13] Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2): 203, 2002

[14] Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003

Our second paradigm frames the experiential nature of voices and proposes to shape a vocal phenomenology. In this regard, I question to what extent our experience of the voice deeply informs our experience of the world, information, and abstract thoughts. The voice can be seen as one of the first of all sensory experiences. Tactile feelings start developing at only a few weeks of gestation, and one of the first stimulus encountered is the tactile sensation from the vibrations of the mother's voice [15]. After birth, our relationship with our own voices starts developing as a dual mechanism. For each motor command sent to the body, the brain sends two messages, one to the muscles and the other informing part of the brain of the impending motion, as a prediction. This allows us to maintain a certain awareness of our physical trajectory without having to sense our body. A vocal action is ultimately a motor action[16]. This same duality of control mechanism thus also applies to the voice through a feedback and a feedforward mechanism[17]. The feedforward control signal is a predictive signal and a feedback control signal contains the perceptual (auditory, vibratory) signal that we gather from our ears and other senses. For the voice, this feedforward signal has also been associated with the experience of the inner voice[18] and is even understood as a possible missing link between thoughts and actions. **In this context, we can understand our experience of complex thoughts as not only linked with language but also as shaped by our more visceral experience of the voice.** In addition to providing a connection between thoughts and actions, voices also underlie

[15] Robert M Bradley and Charlotte M Mistretta. Fetal sensory receptors. *Physiological Reviews*, 1975; and Lise Eliot. *What's going on in there?: how the brain and mind develop in the first five years of life*. Bantam, 2000

[16] Jason A Tourville, Kevin J Reilly, and Frank H Guenther. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, 2008

[17] Joseph S Perkell. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of neurolinguistics*, 2012; and Frank H Guenther. *Neural control of speech*. Mit Press, 2016b

[18] Marcela Perrone-Bertolotti et al. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, 2014

our experience of text, data, and information. Indeed every piece of textual information read silently passes through the inner voice (see background section 2.3 for more information) One can wonder what this means in terms of the omnipresence of texts in our environment, or go a step further and query whether the inner voice may exist independently from language, or at least from human language. Do other vocal mammals experience an inner voice? Do birds rehearse their songs silently? Far from denying that non-verbal animals have complex inner lives, this theory might suggest that any species capable of producing vocal sounds would potentially experience an inner life or at least an inner sonic life.

Finally, our third paradigm frames the voice as a modern manifestation of social grooming. Here we contemplate how the use of the voice, timing, choice to whom we talk, and subtlety in how this voice is addressed, are markers of social dynamics. I concur with Dr. Scott that "dialogue is the new social grooming"[19]. Indeed, primates regularly use allogrooming (cleaning or maintaining another's body or appearance) to reinforce social structures, family links, and maternal behaviors; build new companionships; resolve conflicts and reconcile members of their communities [20]. I explore the theory according to which those behaviors are the direct origin of our vocal messages and also remain one of the most important parts of both the message transmitted and the consequences of our current vocal interactions. In other words, the content of most of our discussions might be seen as an excuse to have the discussion, rather than having the exclusive goal of transmitting information. In this paradigm, I elevate the notion of phatic communication and argue that at different levels, most of our vocal interactions are phatic interactions. Evolutionarily speaking, we could indeed argue that virtually none of the conversations taking place during a regular day contains any essential information for our immediate survival, so what really motivates those vocal interactions might have to be found somewhere else. In summary, it is not what I say that matters, but the fact that I say it, how, when, and to whom.

[19] Sophie Scott. Theres a lot more to conversation than words. What really happens when we talk, Aeon Videos, 2015

[20] John Sparks. *Allogrooming in primates*. Aldine Chicago, 1967

Based on these three paradigms, I propose a novel approach to consider the voice in research, experience design, and even in everyday life. An approach that respects its inherent multiplicity and its ramifications on many aspects of our outer and inner lives. Each proposed paradigm acts as a step to elevate the discussion to a new level. Our approach, based on the work of numerous scientists, aims at increasing our ability to leverage the central connecting propensities of the voice.

1.3 - Problem Space

This dissertation frames the voice according to our three new paradigms as three novel ways to look at the voice: (1) the experience of the voice, by virtue of its evolution through history and throughout our lives, is an instantaneous projection of our holistic fluid identity; (2) our experience of the voice deeply informs our experience of the world, of information, and of abstract thoughts; (3) the voice originates and remains anchored in social grooming.

In addition, the projects presented in this work span and extend over three contexts: the **Interpersonal context** as a medium for meaningful interactions with others; the **Personal context** as a platform to better self-understanding; finally the **Interspecies context** as a common ground to create more understanding across species. Those three angles and three contexts frame and structure our problem space of Vocal Connection as illustrated in figure 1.

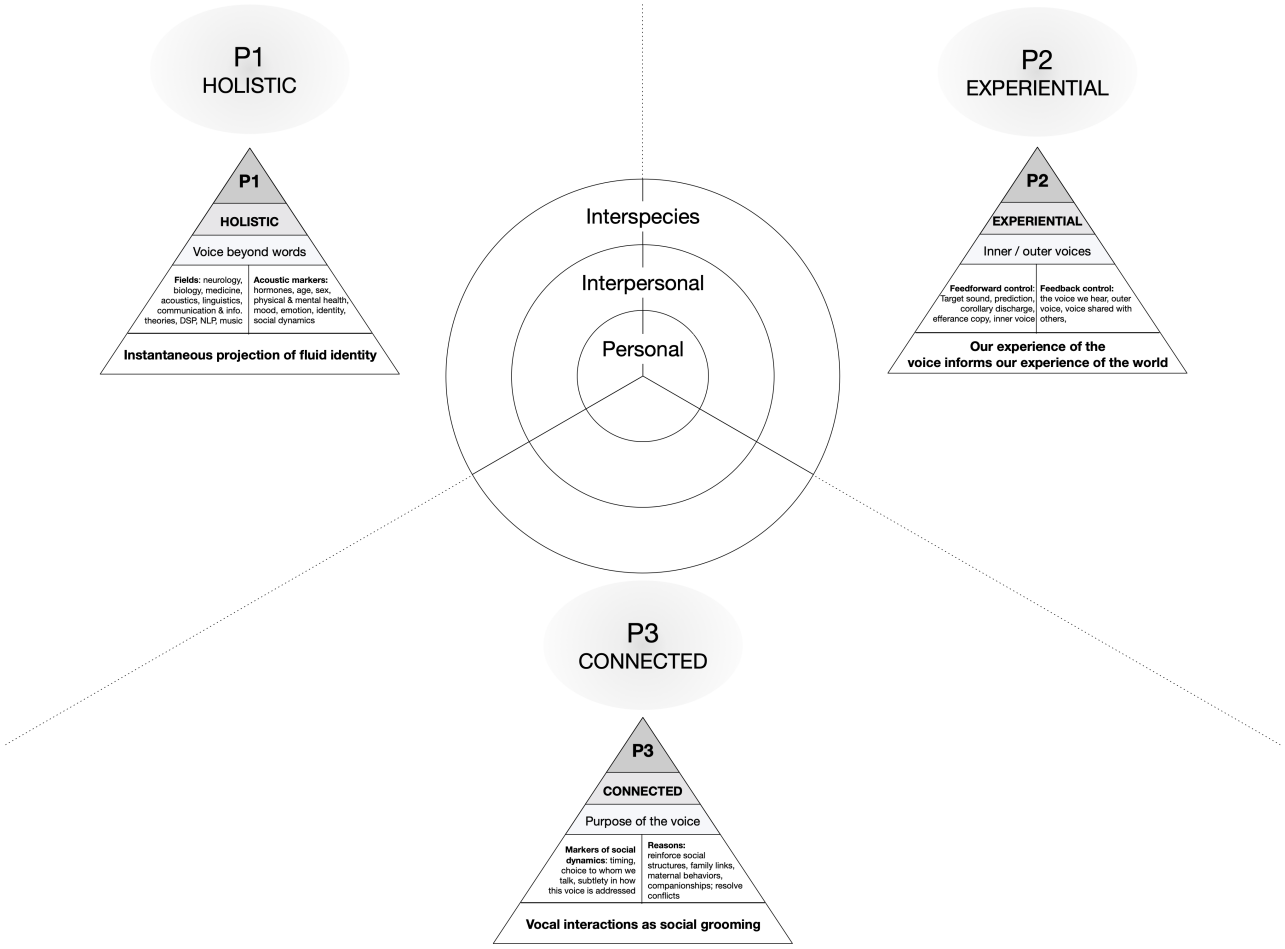


Figure 1: Problem Space of this dissertation framed by our three Paradigms (Holistic, Experiential, Connected) and structured by our three contexts (Interpersonal, Personal, Interspecies)

1.4 - Contributions

This work proposes a new approach for thinking about experience and interface design that uses the power of the voice to create connections between individuals, within different parts of the self, and across species. The contributions from this dissertation include:

- A cross-referenced reading of diverse fields related to voices. Different aspects of the voice have been studied in many distinct domains, from clinical medicine, biology, and surgery to psychology, physics, linguistics, communication theory, learning sciences, etc. However, there are few bridges between those different domains. In this dissertation, I propose methods and examples of how to use knowledge from those different fields to make them accessible in meaningful ways in order to create experiences of connection.
- A history of projects and interventions that embrace the characteristics of the Vocal Connection. This includes scientific visualization; technical learning about the voice in medical, therapeutic, or physical domains; and prior designs of artistic and perceptual experiences.
- A collection of Design Studies that explores different aspects of the meaning of voice and connection in various contexts.
- Three novel systems demonstrating the potential of Vocal Connection in different contexts: interpersonal, personal and interspecies. Including design, implementation, deployment, and evaluation of each system.
- A brief outline of future directions, unexplored areas, and example applications for continued research in the domain of Vocal Connection.

1.5 - Organisation

In the following chapter: *Background*, I present foundational knowledge from various fields that motivate and support this dissertation. Chapter 3: *Design Studies* covers preliminary explorations toward the design of experiences of Vocal Connection. In Chapters 4 to 6, I present three projects that offer different angles from which to consider the potential of the voice for creating new forms of connection: **Memory Music Box**; **Mumble Melody**; and **Sonic and Vocal Enrichment at the Zoo**. In Chapter 7: **Conclusion & Future Directions**, I discuss insights generated from the projects and proposes possible directions for future research inspired by the theoretical implications of this work.

This dissertation can be considered a continuation and expansion of my master's thesis on the topic of Self-Reflective Vocal Experiences[21]. Some of the material presented in the "background" and "design studies" sections are developed from and inspired by this previous work.

[21] Rébecca Kleinberger. Singing about singing: using the voice as a tool for self-reflection, 2014

2 - Background

There is a long history of research work on the voice. This chapter puts my work in context and presents the different scientific and inspirational domains that impact our understanding of Vocal Connection. After a succinct account of the vocal production and perception mechanisms, I present three sections to contextualize the three paradigms introduced in the introduction.

The first paradigm highlights the meaningfulness of the voice beyond words, and advocates for the proposition that there is something in the voice that words or emotions alone cannot capture. The voice is an instantaneous projection of our fluid identity, both personal and social. The voice tells us where we come from and where we are heading. Every vocal manifestation captures and projects our identity and the history of how it connects us on different levels. In the second section of this chapter, I trace a journey of **voices across time** to reveal different dimensions of this connection.

The second paradigm proposes that our experience of voices, both interior and exterior, informs our experience of the world. This frames the voice as a permeable membrane between the personal inner life and the social space we share with others. The way they interface influences our perception and processing of inner experiences (thoughts, mental models, subconscious, mental states) and outer information (text, data, discussions). As inner and outer voices developed as a dual mechanism, they maintain a special parallel correspondence. Based on this idea, the third section of this chapter proposes an exploration of **voices across the mind** and present a phenomenology of inner voices based on their potential for connection.

The third paradigm presents the voice in its fundamental social nature, to build and regulate companionships originating in social grooming. In this context, those behaviors are the direct origin of our vocal messages, and also remain one of the most important parts of both the message transmitted and the consequences of our current vocal interactions. With this in mind, the fourth section of this chapter presents some aspects of **voices across people** both in regards to issues of loneliness and disconnection, and also in terms of technological and societal creations surrounding the voice.

2.1 - Vocal Production and Perception Mechanisms

This section presents a basic examination of vocal production and perception mechanisms, and a basic taxonomy of some of the brain areas involved in speech, voice, and language. This introduction of relevant technical terms lays the foundations for a better understanding of the voice throughout this dissertation.

2.1.1 - Vocal Production Biomechanics in Humans

The voice production mechanism is composed of three subsystems[22]:

- **The air pressure subsystem**: The ability to produce vocal sounds starts with airflow from the lungs, which is coordinated by the action of the diaphragm and abdominal and chest muscles. Its role is to push air out of the lungs to provide and regulate air pressure, causing the vocal folds to vibrate.
- **The vibratory subsystem**: This action takes place within the larynx (also called the voice box), a complex system of muscles and cartilage that supports and moves the vocal folds. Located between the thyroid and arytenoid cartilages, these two muscles form an elastic window that opens and shuts across the trachea. The folds are open when breathing. When speaking, the air from the lungs pushes against the vocal folds and creates the self-sustained vibration of the tissue to produce sound. The vibratory subsystem produces voiced speaking sounds, singing sounds and gives meaning.
- **The resonating subsystem**: Using the muscles of the vocal tract—throat (pharynx), oral cavity and nasal cavities—the sounds from the vocal folds produce phones, other sounds and and a person's recognizable voice.

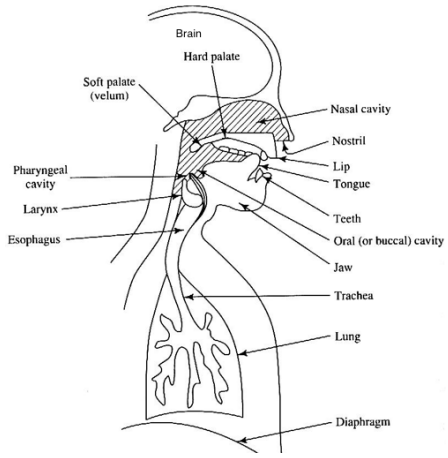


Figure 2: A schematic diagram of the human speech production mechanism

The voice box is essential for voice production but also plays a role in breathing, by opening the glottis and bringing both vocal folds apart, and in swallowing, by coordinating closing the glottis by bringing both vocal folds to the midline to prevent choking. Within the larynx, the thyroid, cricoid, and arytenoid cartilages house the vocal folds and the muscles responsible for movement and constriction. In the larynx, left and right vocal folds are composed of three distinct cell layers. The cover is composed of the epithelium (mucosa), basal lamina (or basement membrane zone), and the superficial layer of the lamina propria. The transition is composed of the intermediate and deep layers of the lamina propria. The body is composed of the thyroarytenoid muscle. Each vocal fold is about 11–15mm long in adult women and 17–21mm in men. A review of the physics of voice production can be found in Zhanga's work on the mechanics of human voice production and control[23]. Laryngeal muscle activation stiffens, deforms, or

[22] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000

[23] Zhaoyan Zhang. Mechanics of human voice production and control. *The journal of the acoustical society of america*, 2016

repositions the vocal folds, thus controlling the geometry and mechanical properties of the vocal folds and glottal configuration, and allowing fine control for pitch, loudness, and voice quality.

In a narrow sense, voice refers to the sound produced by the larynx. We distinguish voiced sounds resulting from vocal fold vibrations and unvoiced sounds that involve vocal muscles but no vibrations from the vocal folds. Examples of unvoiced sounds include whispering, fricatives, and pulsations. During whispering, air passes between the arytenoid cartilages to create audible turbulence during speech[24], while supralaryngeal articulation remains identical to normal speech. Fricatives result from airflow passing through narrow passages in the vocal tracts, and plosives come from the sudden release of a complete closure in the vocal cavities.

[24] John Laver. *Principles of phonetics*. Cambridge University Press, 1994

2.1.2 - Vocal Perception Mechanism in Humans

Sound perception in humans is a complex psychoacoustic process. We present a basic overview of the functioning of the ear and the specificity of self-generated sound perception. The ear is composed of three parts: the outer, middle, and inner ears[25].

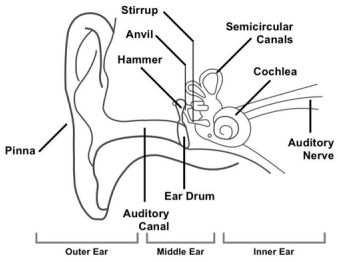


Figure 3: Diagram showing the structure of the human ear, detailing the parts of the outer, middle, and inner ear.

[25] A James Hudspeth. How the ear's works work. *Nature*, 1989

- **The outer ear** consists of the lobe and ear canal that protect the middle and inner ears as well as guide, prefilter, and encode the incoming sounds with location-specific features.
- **The middle ear** is composed of the eardrum, a thin membrane that vibrates with entering sounds and transforms air vibrations into fluid vibrations. The motion of the eardrum is transferred across the middle ear via three small bones called ossicles (hammer, anvil, and stirrup). These ossicles are supported by muscles that allow for free motion but can tighten up and stop the bones' motion in case of overly loud sounds.
- **The inner ear** consists of several tubes winding in various ways within the skull. Most of them, called semicircular canals, are used for orientation and balance. The tube involved in the hearing process, called the cochlea, is wound tightly and converts sound waves into neural signals. This conversion is made by hair cells on the cochlea which decode the frequency spectrum of the incoming sounds through inner and outer hair cells. The auditory nerves transmit this electrical signal to the brain.

For self-generated vocal sounds, we hear a combination of the signal transmitted via air and the one transmitted via cranial bone conduction[26]. The characteristics of this self-perceived voice are a boost of low frequencies and attenuation of high frequencies. The precise transfer function has been studied, enabling the design of systems to simulate the sound of the voice as heard by the speaker[27]. Figure 4 shows the transfer function that applies

[26] Georg V Békésy. The structure of the middle ear and the hearing of one's own voice by bone conduction. *The Journal of the Acoustical Society of America*, 21(3): 217–232, 1949

[27] Jonathan Berger and Song Hui Chon. Simulating the sound of one ' s own singing voice. 2003

to self-produced sounds as heard by the speaker. Other internal vibrations can, in certain cases, reach the ear canal, such as noises produced by chewing, swallowing, or even walking. To optimize *useful hearing*, those sounds, as well as our own vocalization sounds, are reduced by different mechanisms such as dampening by muscles or by the elasticity of the cervical vertebra, the bones forming the top of the vertebral column[28].

In addition, hearing is not a purely mechanical phenomenon of wave propagation, but is also a sensory and perceptual event: loudness perception is logarithmically filtered, and more frequently heard sounds have reduced perceptive effect due to long- and short-term habituation[29].

2.1.3 - Voice and the Brain

Several brain areas play a critical role in voice, speech, and language[30] among which:

- **The left inferior frontal gyrus** classically referred to as Broca's area, is involved in speech production, articulation, and word access in both written and spoken language. Damage to Broca's area is characterized by slurred and unclear words[31].
- **The Brodmann area 22** located in the left superior temporal lobe and classically referred to as Wernicke's area, is located in the posterior superior temporal lobe and is connected to Broca's area through a neural pathway called Arcuate fasciculus. It is associated with speech comprehension and language processing, both written and spoken..
- **The angular gyrus**[32] is located in the posterior part of the inferior parietal lobule and allows for the association of perceived words with different images, sensations, and ideas.
- **The motor cortex** controls 100+ muscles involved in the vocal production mechanism, including the laryngeal muscles, tongue, and lips.
- **The cerebellum**, located at the back of the brain, is thought to have considerable influence over language processing[33].
- **The auditory ventral stream** (AVS) is the pathway responsible for sound recognition, connecting several areas of the auditory cortex with the middle temporal gyrus and temporal pole.
- **The auditory dorsal stream** connects the auditory cortex with the parietal lobe and is responsible for sound localization
- **The basal ganglia** is involved in the selection and initiation of cortical patterns of activation for planned behaviors and thoughts. It has been associated with speech acquisition, optimisation in language learning, and motor planning and control

Sections 2.3: Voices Across the Mind, and 5.3.2: Background of the Mumble Melody Project cover in more detail the neural basis of the voice.

[28] Georg V Békésy. The structure of the middle ear and the hearing of one's own voice by bone conduction. *The Journal of the Acoustical Society of America*, 21(3): 217–232, 1949

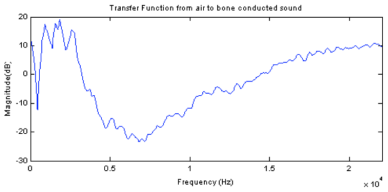


Figure 4: Transfer function of air to bone conducted sound from Berger et al.

[29] HSPaJT Fruhstorfer and year=1970 publisher=Elsevier others, journal=Electroencephalography and clinical Neurophysiology. Short-term habituation of the auditory evoked response in man; and Frederic G Worden. Auditory habituation. In *Physiological Substrates*. Elsevier, 1973

[30] Frank H Guenther. *Neural control of speech*. Mit Press, 2016b

[31] Jay P Mohr et al. Broca aphasia: pathologic and clinical. *Neurology*, 1978

[32] Mohamed L Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 2013

[33] Peter Mariën et al. Consensus paper: language and the cerebellum: an ongoing enigma. *The Cerebellum*, 2014

2.2 - Voices Across Time

In this section, I retrace the path of the voice through time at four different scales. Each of these different time scales offers perspective to consider the voice as a bridge between species, to our ancestors, and to our contemporaries:

- **From the Big Bang to Homo sapiens**: I offer a basic roadmap covering the natural history of vocal evolution across species.
- **From antiquity to today**: I account for specific aspects of human vocal history and the changes in societal perception and understanding of the voice throughout history.
- **From birth to death**: I propose a human-scale representation of the vocal evolution throughout life as a marker of fluid identity.
- **From vocal intent to acoustic signals**: I present a focused picture of a moment in the voice and how the stages of vocal production create a highly connected web in the brain.

2.2.1 - Genesis and Evolution of Human and Animal Voices

Natural History and Clues to the Evolution of the Voice

The voice, in its resonances, reflects the evolution of species and the specificities of our world and its history. Most sounds produced by humans and animals are made possible by the specific chemical composition of our atmosphere: 22 percent oxygen, 0.1 percent helium and 78 percent nitrogen. The air density allows for the propagation of the vocal sound waves we are familiar with[34]. On other planets, the difference in air density would affect the resonant frequencies of our voice, the same way a vocal track full of helium amplifies high-pitch components of the voice relative to low-pitch components, drastically changing the overall vocal timbre. Our ability to produce sounds is also tightly connected with fundamental biomechanics. Each of our organs requires energy to function. This energy is synthesized from proteins by mitochondria; the more energy an organ consumes, the more mitochondria its cells will contain[35]. And the vocal cords contain seven times more mitochondria than the muscles in our biceps.[36]

Considering this environment, how did the voice come to be? Which of our ancestors had a voice? And which of our primate ancestors could vocalize? When did the human voice as we know it appeared? Understanding the evolution of the voice in conjunction with the evolution of species is a major challenge[37]. Evolutionary biologists have a few clues to tackle those questions.

[34] ME Delany. Sound propagation in the atmosphere: a historical review. *Acta Acustica united with Acustica*, 1977

[35] Diane Sweeney and Brad Williamson. *Biology: Exploring Life: Laboratory Manual*. Pearson Education, Incorporated, 2006

[36] Jack Jiang, Emily Lin, and David G Hanson. Vocal fold physiology. *Otolaryngologic Clinics of North America*, 2000

[37] Asif A Ghazanfar and Drew Rendall. Evolution of human vocal production. *Current Biology*, 2008

First, evolutionary biologists have been looking at the fossilized auditory systems of extinct species. Dinosaurs had a very elementary form of hearing[38] and the slow development of additional and more complex ossicles can help us envision how small ancestral mammals perceived sounds. The evolution of mammalian auditory ossicles[39] marked an important evolutionary event in which bones in the jaw of reptiles were co-opted to form part of the hearing apparatus in mammals. Those fossilized skulls help us connect the dots from basic dinosaur hearing to more complex mammalian hearing.

Second, researchers have analyzed the size of frontal cortex areas to assess the part of the brain devoted to the voice. In 1912, Broadman assessed the part of the brain used for vocal sounds at 3.5 percent for the cat, 17 percent for the chimpanzee, and 30 percent for the human[40]. For extinct species, researchers infer brain constitution from ridges on the skull. Such work let us presume that Homo erectus could talk two million years ago[41]. Since the discovery of the Broca area—necessary for the production of language—in 1850 followed by the Wernicke area, used for the understanding of language, additional understanding of our ancestors' voices was made possible. The general fact that humans have the biggest brain/body mass ratio of all species[42] also enters into the equation, as this is one of the main differences between us and our closest ancestors and cousins. The brain differences include in particular the neocortex, which is considered the locus of language. From the skulls of extinct species and the brains of contemporary species, we can understand how the part of the brain devoted to voices has increased to its current size in humans.

Third, bipedalism also altered the structure of our skeleton and the angle formed between the skull and the body. The foramen magnum is the large hole on the underside of the skull where the spinal cord exits. In humans, the foramen magnum is positioned centrally, allowing the human body to be oriented vertically and creating an angle of 90 degrees between the skull and the spine. In chimpanzees and other apes, the foramen magnum is positioned towards the back of the skull, with the spinal cord exiting at an angle of 45 degrees[43]. This angle is also determined by the hyoid bone connected to the jaw. This angle not only allowed humans to become true bipeds, but also let our voice boxes to become bigger. Lucy, who lived 3.2 million years ago, was also a biped, but her head was not vertically aligned with her body, indicating that the vocal sounds that she could produce were more similar to primate vocalizations than to the voices of Homo sapiens[44]. The evolution of bipedalism alongside vocal capabilities sheds light on how our voices relate to the voices of our extinct primate ancestors.

[38] Paul C Sereno and Fernando E Novas. The complete skull and skeleton of an early dinosaur. *Science*, 1992

[39] Edgar F Allin. Evolution of the mammalian middle ear. *Journal of Morphology*, 1975

[40] K Brodmann et al. Neue ergebnisse über die vergleichende histologische lokalisation der grosshirnrinde mit besonderer berücksichtigung des stirnhirns. *Anatomischer Anzeiger*, 1912

[41] Kathleen Gibson. Tools, language and intelligence: Evolutionary implications. *Man*, 1991; and Marge E Landsberg. *The genesis of language: a different judgement of evidence*, volume 3. Walter de Gruyter, 2011

[42] Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 2012

[43] Martin Pickford. Orientation of the foramen magnum in late miocene to extant african apes and hominids. *Anthropologie*, 2005

[44] William H Kimbel and Lucas K Delezene. "lucy" redux: A review of research on australopithecus afarensis. *American journal of physical anthropology*, 2009

Fourth, the physiological changes in the voice boxes of living species can help us understand the branching evolution of the voice by looking at their laryngeal shape and composition. Humans are one of the rare known species to have a descended larynx, alongside the red deer, the hammer-headed bat, the wolf, and the koala[45]. Our voice box is located lower in the neck, around the fifth cervical vertebra. This particularly low position can be an issue, as the food canals can interfere with the air canal, making us incapable of breathing and drinking at the same time. Baby humans are born with the larynx at the second vertebra, which protects them from suffocation caused by swallowing the wrong way[46]. Between the second and eighteenth months, the larynx descends to its final position, around the fifth vertebra. This low position allows our voice box to resonate with a much wider range of frequencies, roughly from 50 to 1kHz (C6 note for a soprano voice). In comparison, the chimpanzee's small voice box does not allow them to produce most of the known vowels[47]. Interestingly, there are signs of laryngeal descent in the infant chimpanzee, but to a much lesser degree and not accompanied by a descent of the hyoid itself[48]. Our closest primate cousins present very similar—although diverging—evolution to ours. Laryngeal descent evolved independently at multiple places and times in the mammalian evolutionary life, demonstrating a complex example of convergent evolution.

[45] W Tecumseh Fitch and David Reby. The descended larynx is not uniquely human. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2001

[46] Robert Pracy. The infant larynx. *The Journal of Laryngology & Otology*, 97(10): 933–947, 1983

[47] Philip Lieberman. Can chimpanzees swallow or talk? a reply to falk. *American Anthropologist*, 1982

[48] Takeshi Nishimura, Akichika Mikami, Juri Suzuki, and Tetsuro Matsuzawa. Descent of the larynx in chimpanzee infants. *Proceedings of the National Academy of Sciences*, 2003

Fifth, genetic studies can also help us look at the genetic genesis of the human voice in particular: the gene FOXP2 (forkhead box P2) located on human chromosome 7q31 and seemingly unique to humans, is often associated with the human voice[49]. Genetic mutations in the gene or its absence result in severe articulatory disorders as well as linguistic and grammatical impairment. So far, only humans are known to carry this gene; it is one of the rare genetic differences between humans and chimpanzees (1.2 percent).

[49] Wolfgang Enard, Molly Przeworski, Simon E Fisher, Cecilia SL Lai, Victor Wiebe, Takashi Kitano, Anthony P Monaco, and Svante Pääbo. Molecular evolution of foxp2, a gene involved in speech and language. *Nature*, 2002

Sonic Expression & the Evolution of the Larynx from Fish to Humans

The variations in vocal production mechanisms among different species give us a map to reflect on how humans and animals experience voices. If our experience of the voice informs our experience of the world, then variations in vocal experiences between species may shed some light on how other species experience the world.

If one considers the voice as a way for living creatures to express themselves with sounds, laryngeal sounds are neither the only nor the first mechanism for auditory expression in the animal kingdom. Cicadas and other arthropods possess a percussion instrument equivalent to a voice that they play by changing the shape of their body[50] and hitting a ribbed

[50] James A Simmons, Ernest Glen Wever, and Joseph M Pylka. Periodical cicada: sound production and hearing. *Science*, 1971

membrane on the torso. Amphibians[51] use a wind instrument. Humans and many other mammals possess a vocal instrument that is a combination of wind and strings. Alike a wind instrument, the source of the sound comes from a column of air. However, this column of air activates the vibrations of the vocal folds similarly to strings.

This more complex wind/string instrument started from a more humble origin. The initial origin of the larynx can be found in the fish gills that are considered the first respiratory organs. Gills then evolved into the lungs of the Polypterus, a prehistoric fish that possessed both gills and lungs, and in the first aero-digestive organs found in the Ceratodus, discovered in Australia in 1931[52]. This original larynx is very similar to the one found in human embryos at gestational week 12. The amphibian's larynx presents a sphincteric capacity in addition to aero-digestive, allowing the animals to also produce sounds. Reptiles possess a more complex cartilage system, allowing for stronger muscles to develop. Crocodiles, however, breathe and eat through similar mechanisms. Swallowing happens in one continuous action, and the angle between their mouth and larynx is 0 degrees[53]. For mammals and birds, the larynx has a marked effect on the mechanism, as the thoracic motions are required for breathing.

The development of the larynx then branched out, resulting in very different structures in different species. In birds, the larynx is called the syrinx and is located at the base of the animal's trachea[54]. The syrinx produces sounds without vocal folds. For marine mammals, the air opening is located on top of their heads, allowing them to eat underwater without suffocating.

The human larynx also keeps clues to our closest ancestor through the presence of laryngeal pseudo valves, also called "false vocal cords." Necessary for salamanders to resist water pressure that would otherwise drown them, these false vocal chords are also used by chimpanzees to assist prehension and provide additional strength for holdin onto branches[55] For humans, they seem to only be useful for weightlifters and babies, who can hang from one hand and hold up their bodyweight easily during the first three months of their lives. When singers apply excessive pressure to their vocal cords, it can make the false vocal folds vibrate, as in the case of Louis Armstrong. In addition, the tight relationship between the upper limbs and the voice is rich in other ways, and a parallel could be made between the fine dexterity of the hands and the dexterity of vocal control. From the motor homunculus, one can see that the brain area controlling the motor command of the fingers is just adjacent to the brain area controlling the motor commands for the facial muscles, the lips, tongue, and larynx. In terms of vocal control acquisition, one central assumption is that because apes

[51] Jakob Christensen-Dalsgaard. Amphibian bioacoustics. In *Handbook of signal processing in acoustics*. Springer, 2008; and Robert Dudley and A Stanley Rand. Sound production and vocal sac inflation in the túngara frog, physalaemus pustulosus (leptodactylidae). *Copeia*, 1991

[52] MM Babiker. Development of dependence on aerial respiration in polypterus senegalus (cuvier). *Hydrobiologia*, 1984

[53] VE Negus. The mechanism of swallowing, 1942

[54] Stephen Nowicki and Peter Marler. How do birds sing? *Music Perception: An Interdisciplinary Journal*, 1988

[55] Lucie Bailly. *Interaction entre cordes vocales et bandes ventriculaires en phonation: exploration in-vivo, modélisation physique, validation in-vitro.* PhD thesis, 2009

have voluntary fine control of their hands but not their voices, language must have started from manual gestures and was only later transmitted to the vocal system[56]. This might help explain some of the hand gestures associated with speaking in people who "talk with their hands."

[56] Francisco Aboitiz. A brain for speech. evolutionary continuity in primate and human auditory-vocal processing. *Frontiers in neuroscience*, 2018

By examining this succinct exposition of vocal evolution from the Big Bang to Homo sapiens, we contend that all mammals and many other species share important similarities in their abilities to produce and perceive sound, supporting a global ecology of the voice shared amongst species. The way this evolution took place gives us some suggestions as to the main evolutionary advantages offered by vocal exchanges. Our third paradigm argues that the main purpose of vocal exchanges is to create togetherness.

In the following section, we focus on the variations of understanding of vocal manifestations throughout human history.

2.2.2 - Voices Throughout History

In addition to its evolution in natural history, the voice marked its presence throughout history and in the evolution of human civilizations. By retracing this journey in time we relate to our ancestors and connect to their understandings of vocal manifestations. We illustrate this evolution through two lenses that were primordial at the time: medical and biblical. First, we look at the evolution of the uderstanding of the voice. As a biomechanical manifestation from the body, the voice has naturally intrigued doctors and physicians. An understanding of both normal or abnormal functioning of the voice has helped us reach an ever-more accurate representation of the vocal apparatus. This evolving understanding reflects on the perception of vocal manifestations in time across civilizations. We then present a short account of the mystical understanding of the voice through time. Biblical narratives have accompanied humanity for two millennia. They had a strong influence on our ancestors' world perception and representation. If their voices have now disappeared, their understanding of vocal manifestations may still inform their everyday vocal experiences.

Medical Understanding

Because it is hidden from sight, the vocal apparatus is particularly hard to study. Its understanding has presented challenges and misunderstandings across millennia and aspects of its functioning still remain enigmatic today.

Ancient civilizations had often reached a certain understanding of the functioning of many of the body's organs through experiments and dissections on animals or corpses. Around 200 A.D., Greek physician and

philosopher Galen, also the father of the theory of humoral imbalance, clarified the anatomy of the trachea and was one of the first to demonstrate that the larynx generates the voice[57]. In one experiment, Galen used bellows to inflate the lungs of a dead animal. Original Greek-language texts and Byzantine medical writers reveal early knowledge of diagnostic and therapeutic techniques for laryngeal disease and cancer of the larynx[58]. However, in order to grasp the actual functioning of the human vocal apparatus, one cannot use animals, as their larynxes are quite different, and complexity of the voice makes it difficult to study in dead bodies.

The fact that the voice is so complex and recruits so many other organs (lungs, diaphragm, facial muscles, etc) has made it hard to study, but has also made it an asset/tool/flashlight to understand other conditions. Respiratory and heart anomalies often modify the sound of the voice because they change the acoustic characteristic of the trachea. Auscultation (from the Latin verb auscultare "to listen") is the action of listening to the internal sounds of the body. The term was introduced by René Laennec in 1837[59], but the act of listening to the voice and other body sounds for diagnostic purposes originates as early as Ancient Egypt[60]. The invention of Laryngoscopy by singing teacher Manuel García in 1854[61] allowed the vocal apparatus to be observed live for the first time.

Mystical Understanding

The place of the voice in biblical narrative demonstrates and reinforces the spiritual, mystical, and mysterious views humans had about the voice hundreds or thousands of years ago. The treatment of vocal manifestation in the bible can first be seen in who can or cannot talk. Biblical stories include instances of beastly talk, including the dialogues between Eve and the snake (Genesis 3 NIV) and the interaction between Balaam and his donkey (Numbers 22:21-38 KJV). Some have conjectured that all of the animals in the story of the Garden of Eden were able to talk[62]. In Jewish tradition, the Golem, a giant created by men from clay, cannot speak, as only God can give the gift of the voice[63]. If the voice is a gift from God, it can also be withdrawn as punishment, as in the Tower of Babel—an origin myth to explain the multiplicity of human languages. After destroying the Tower of Babel[64], God "confounds [human] speech so that they can no longer understand each other, and scatters them around the world" (Genesis 11:1–9). Voices and languages are here portrayed as highly sacred, as only God has the power to gift or take them back. In this context, the voice also appears in its proportion to connect, and its absence to create disconnection, scattering, and chaos. Biblical narratives also provide insights as to the ancient understanding of vocal production as shown in Psalm 137 5-6: "If I forget you, O Jerusalem / let my right

[57] Charles Mayo Goss. Galen on anatomical procedures (de anatomicis administrationibus). *Translation of the surviving books with introduction and notes by Charles Singer. Oxford University Press, New York, The Anatomical Record*, 1958

[58] Dimitrios Assimakopoulos et al. Highlights in the evolution of diagnosis and treatment of laryngeal cancer. *The Laryngoscope*, 2003

[59] René Théophile Hyacinthe Laennec. *Traité de l'auscultation médiate, et des maladies des poumons et du coeur*. Société Typographique Belge, 1837

[60] Carlos G Musso. Imhotep: the dean among the ancient Egyptian physicians. an example of a complete physician. *Humane Medicine Health Care*, 2005

[61] Anthony Jahn and Andrew Blitzer. A short history of laryngoscopy. *Logopedics Phoniatrics Vocology*, 1996

[62] George Savran. Beastly speech: intertextuality, balaam's ass and the garden of eden. *Journal for the Study of the Old Testament*, 19(64):33–55, 1994

[63] Lewis Glinert. Golem! the making of a modern myth. In *Symposium: A Quarterly Journal in Modern Literatures*. Taylor & Francis, 2001

[64] Theodore Hiebert. The tower of babel and the origin of the world's cultures. *Journal of Biblical Literature*, 2007

hand forget its skill! / Let my tongue stick to the roof of my mouth." This passage has been interpreted as illustrating left middle cerebral artery strokes, causing motor aphasia and right hemiparesis[65]. On the theme of speech impediment, God tells Moses to speak for him despite his stutter[66], illustrated by Shoenberg in Moses and Aron as "But my tongue is not flexible. Thought is easy; speech is laborious."[67]. Those instances show that the interest in and questioning toward vocal afflictions and differences have been ongoing for thousands of years.

The evolution of the voice box, the history of our understanding of the voice are ongoing and lead us toward an unknown future. At the individual level, our voice appear unique but throughout our lives it follows a similar journey as people who came before and those who will come after us.

[65] Luiz Antonio de Lima Resende, Silke Anna Theresa Weber, Marcelo Fernando Zeugner Bertotti, and Svetlana Agapejev. Stroke in ancient times: a reinterpretation of psalms 137: 5, 6. *Arquivos de neuro-psiquiatria*, 2008

[66] Marc Shell. Moses' tongue. *Common Knowledge*, 2006

[67] Arnold Schoenberg. *Moses und Aron: opera*, volume 8004. E. Eulenburg, 1984

2.2.3 - Transformation of the Voice Throughout Life

The voice can be seen as one of the earliest sensory experiences as, before birth, at just a few weeks gestational age, the first sensations perceived by the fetus are tactile sensations from the vibrations of their mother's voice[68]. After birth, the first screams also correspond to the first time the lungs are inflated and are a sign of a healthy newborn. After only one week, babies naturally orient their heads toward voices talking to them. The descent of the larynx occurs after only a few months[69]. Metaphorically, in those few months, the human baby relives a key element of the evolution of our entire species. Some have used this as an argument for the highly controversial Recapitulation Theory[70], also called the Meckel–Serres law, that argues for parallelism between embryonal development and the evolution of species and summarised by Ernst Haeckel as: "ontogeny recapitulates phylogeny"[71]. Without going this far, this parallel shows an interesting correspondence between different time scales of vocal evolution.

[68] Peter G Hepper et al. Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, 1993; and William P Fifer and Christine M Moon. The role of mother's voice in the organization of brain function in the newborn. *Acta Paediatrica*, 83(s397):86–93, 1994

[69] Robert Pracy. The infant larynx. *The Journal of Laryngology & Otology*, 97(10): 933–947, 1983

[70] Waldo Shumway. The recapitulation theory. *The Quarterly Review of Biology*, 7(1): 93–99, 1932

[71] M Elizabeth Barnes. Ernst haeckel's biogenetic law (1866). *Embryo Project Encyclopedia*, 2014

Throughout life, the voice carries signs of our changing identity. Amongst other markers, it reveals our hormonal and sexual identity in adults. One can identify quite accurately the gender and age range of a speaker only from hearing their voice[72] and even elephants can recognize those cues from human voices[73]. When babies are born, there are no audible differences between the screams and babbles of baby girls and boys. The major changes in the voice appear during puberty because of hormonal changes triggered by the pituitary gland and the hypothalamus. The larynx is a target organ for thyroid hormones, estrogens, progesterone, androgynes, and testosterone. The thyroid hormone acts on phonation, and patients with thyroid disorders often suffer from phonation disturbances[74]. Sexual hormones have several effects on the voice and are not only responsible for the acoustic differences between women and men, but also for individual

[72] Edward D Mysak. Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, 1959; and Hadi Harb and Liming Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 2005

[73] Karen McComb et al. Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. *Proceedings of the National Academy of Sciences*, 2014

[74] Lejla Junuzović-Žunić et al. Voice characteristics in patients with thyroid disorders. *The Eurasian journal of medicine*, 2019

differences, and temporal variations within one individual.

Estrogen, progesterone, and androgens secreted at puberty cause the development of the third and last cell layers on the vocal folds. From a homogenous tissue covering the folds three distinct functional layers form: a central muscle, a layer of stiff collagen wrapped in stretchy elastin fibers, and an outer layer of mucous membrane. This layer specialisation changes the harmonic characteristics of the individual and creates the typical timbre of an adult voice.

Testosterone changes the mucus structure, giving boys richer, lower harmonics[75]. In addition, the size and thickness of the vocal folds, as well as the thoracic cage, lung tidal volume, and body shape, are also affected by testosterone, creating a lower voice. Prepubescent castration prevents testosterone from affecting the larynx, as well as hardening the bone-joints, inducing unusually long bones, limbs, and ribs, and thus exceptional lung capacities in addition to extended prepubescent vocal range and rich high harmonics[76].

[75] Meredydd LL Harries, Judith M Walker, David M Williams, S Hawkins, and IA Hughes. Changes in the male voice at puberty. *Archives of disease in childhood*, 77(5):445–447, 1997

[76] James S Jenkins. The voice of the castrato. *The Lancet*, 1998

For women, the voice drops by a third of an octave at puberty. The thyroid modifications also induce the formation of the cricoid cartilage. The levels of sexual hormones affect mucus and muscles as well as the brain causing changes in the voice acoustics. Estrogens also create a more flexible voice in women with thicker and more lubricated vocal folds[77]. The same effects are seen in genital mucus. Progesterone leads to a phenomenon of desquamation (shedding of the outermost membrane or layer of tissue) and thickens the mucus, creating a dryer throat which is often balanced by the effects of estrogens. Premenstrual symptoms can also lead to gastric reflux due to the changes in esophageal muscle capacities. This reflux dries the vocal folds. Professional singers know these premenstrual vocal symptoms, which may reduce their lung capacities and erase some of their high harmonics[78].

[77] MB Schiff. Sex hormones and the female voice. *Journal of Voice-Official Journal of the Voice Foundation*, 13(3):424–424, 1999; and Ofer Amir and Tal Biron-Shental. The impact of hormonal fluctuations on female vocal folds. *Current opinion in otolaryngology & head and neck surgery*, 2004

[78] Filipa MB Lã, William L Ledger, Jane W Davidson, David M Howard, and Georgina L Jones. The effects of a third generation combined oral contraceptive pill on the classical singing voice. *Journal of Voice*, 21(6):754–761, 2007

After puberty, except phenomena linked with menstrual symptoms, the voice remains more or less the same for about 50 years. However, the way one uses their voice affects the health of the voice. Mental or physical health can alter the voice, from a simple cold changing the shape of the nasal cavity, or smoking causing swelling of the area near the vocal cords and decreasing lung capacities[79], to Parkinson's disease causing vocal turbulences[80]. Eventually, presbyphonia, or aging larynx symptoms, becomes apparent. This includes morphological changes in the coverage of mucosa, muscle, and cartilage. Due to aging and reduced hormone levels, the collagen in our folds stiffens and the surrounding elastin fibers atrophy and decay. This causes decreased flexibility and increased pitch in older voices.

[79] Julio Gonzalez et al. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 2004

[80] Geralyn M Schulz and Megan K Grant. Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease: a review of the literature. *Journal of communication disorders*, 2000

Reduced nerve endings also decrease the virtuosity and agility of controlling the voice in older adults. In addition, during menopause, the decline of the female hormones leads to a more masculine and drier voice. However, as fat cells can produce estrogens, bigger women often maintain more feminine and fluid voices than thin women. Some obese male tenor singers also have more estrogen than testosterone, leading to rounder, more feminine voices[81].

Besides normal hormonal changes, other biological, physical, and mental conditions can also influence the voice. Parkinson's disease affects the acoustics of the voice by creating unique patterns of nonlinearity and non-Gaussian turbulences[82]. Respiratory conditions and lung diseases also affect the voice and specific acoustic features can now be used as diagnostic markers[83]. The origin of chest pain can be associated with many types of physical or mental issues, such as cardiac problems, acute coronary syndromes, pulmonary embolisms, pneumonia, or panic disorder. In some cases, the analysis of the vocal sounds can help diagnose the origin of the pain[84]. The evaluation of voice dosimetry can help predict the probability of bronchial stenosis on patients who have normal voice sounds but weak breath sound[85]. Depression and other mental health conditions are now thought to affect timing elements in the vocal production mechanism[86]. Certain substances such as alcohol and drugs affect muscle control and response time which consequently affect the voice. For alcohol, vocal rhythm and formant shape are affected[87]. Marijuana and cocaine are known to alter time perception in rats[88]. This might suggest that such a drug might also affect speech timing and rhythm. In addition, cocaine abuse often induces vocal tics[89]. Smoking alters jitter, frequency, and tremor of the voice in younger adults[90]. The presence of nodules and polyps reduces pitch range[91]. Fatigue also marks the voice[92] by affecting the choice of word, increasing hesitation, and reducing the vocabulary of speakers. In all of those cases, the voice might be used as a diagnostic or detection tool.

Besides the influence of hormones and abnormal conditions, the specific sound of the voice is the result of many other variables. Other elements that make the voice of an individual unique include:

- **Anatomical elements**: shape and size of the vocal cords, vocal tract, nose, jaw, and lips, shape and size of other elements of facial morphology, physiology, health of the tissues, etc.
- **Vocal posture (learned & habits)**: accent, pacing, rhythm, muscle tension, tonus, glotization, breathiness, jitter, ventricularity, fry, etc.

However, our languages still lack the necessary vocabulary to truly characterize the uniqueness of a voice. The voice is not an object; it is alive and the manifestation of something deeper. It evolves throughout an individual's life and becomes enriched overtime. The voice help frame and enable thoughts. It is the link between individual and society.

[81] Jean Abidbol. Hormones and the voice. *The Singer's Guide to Complete Health*, 2013

[82] M. a. Little et al. Testing the assumptions of linear prediction analysis in normal vowels. 2006

[83] René Théophile Hyacinthe Laennec. *De l'auscultation médiate: ou, Traité du diagnostic des maladies des poumons et du coeur; fondé principalement sur ce nouveau moyen d'exploration*, volume 2. Culture et civilisation, 1819; Amon Cohen and AD Berstein. Acoustic transmission of the respiratory system using speech stimulation. *Biomedical Engineering, IEEE*, 1991; and Raymond LH Murphy Jr et al. Visual lung-sound characterization by time-expanded wave-form analysis. *New England Journal of Medicine*, 1977

[84] William Cayley. Diagnosing the cause of chest pain. *Am Fam Physician*, 2005

[85] FL Jones. Poor breath sounds with good voice sounds. a sign of bronchial stenosis. *CHEST Journal*, 93(2):312–313, 1988

[86] James C Mundt et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 2007; and William A Hargreaves et al. Voice quality in depression. *Journal of Abnormal Psychology*, 1965

[87] Florian Schiel et al. Rhythm and formant features for automatic alcohol detection. In *International Speech Communication Association*, 2010

[88] Ruth Ogden and Catharine Montgomery. High time. *Psychologist*, 25(8), 2012

[89] Francisco EC Cardoso et al. Cocaine-related movement disorders. *Journal of the Movement Disorder Society*, 1993

[90] Julio Gonzalez et al. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 2004

[91] Rahul K Shah et al. Relationship between voice quality and vocal nodule size. *Otolaryngology–Head and Neck Surgery*, 2008

[92] Harold P Greeley et al. Detecting fatigue from voice using speech recognition. In *Signal Processing and Information Technology*, 2006

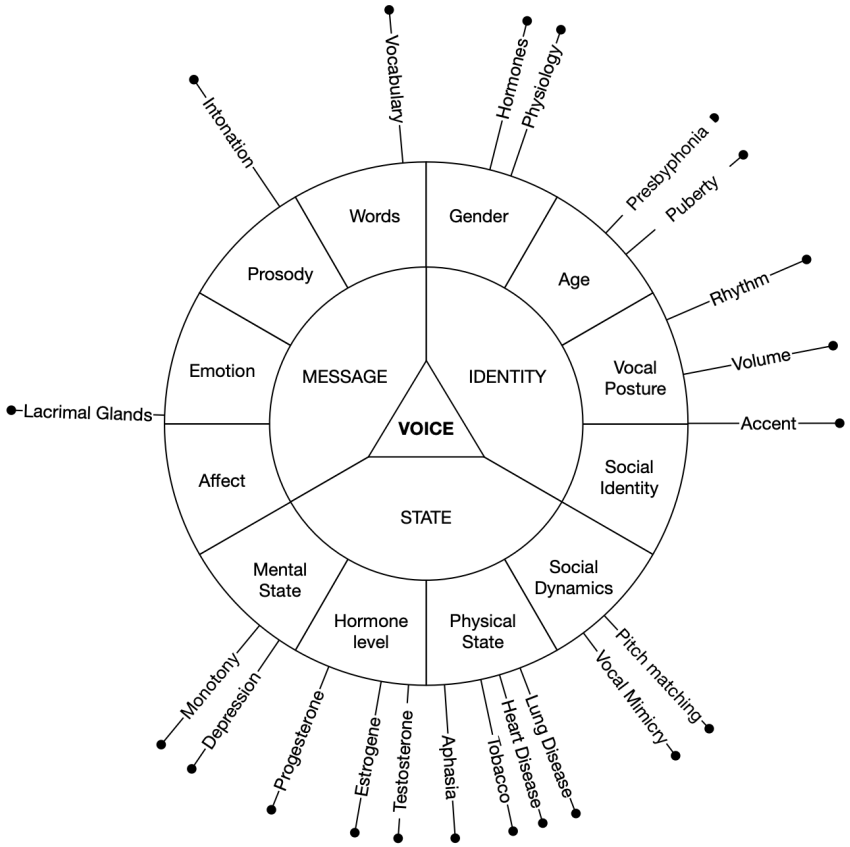


Figure 5: Every moment of the voice contains element of the speaker's message, identity and state

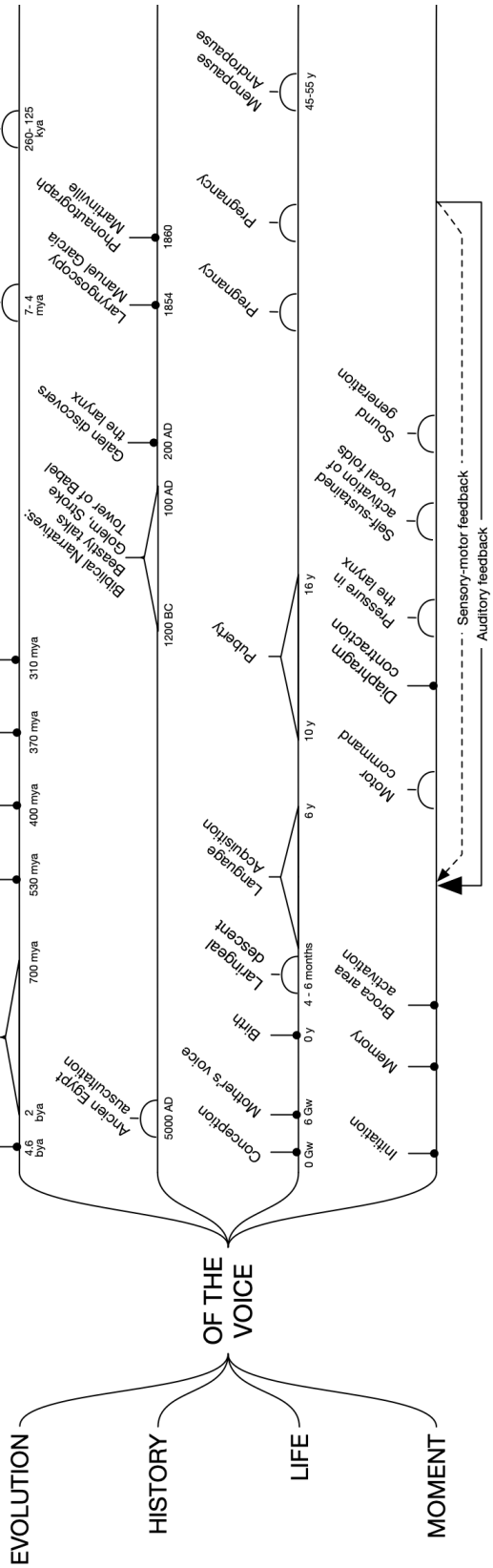
2.2.4 - A Moment of the Voice

Vocal production is a complex process that includes many stages in parallel or in series to transform conceptual ideas into acoustic signals. As described by Price[93] the different stages include: conceptualization of the intended message; word retrieval; selection of the appropriate morphological forms; sequencing of phonemes, syllables, and words; phonetic encoding of the articulatory plans; initiation and coordination of sequences of movements in the tongue, lips, and laryngeal muscles that vibrate the vocal tract; and the control of respiration for vowel phonation and prosody. In addition to this feed-forward sequence, auditory, and somatosensory processing of the spoken output is fed back to the motor system for online correction of laryngeal and articulatory movements. Dysfunction in the production mechanism leads to various speech disorders, such as aphasias or stuttering, and in neuropsychiatric disorders such as schizophrenia[94]. The neural substrates of these dysfunctions remain poorly understood. We will come back to these different stages in section 3 of the background chapter and in chapter 4: Mumble Melody

[93] Cathy J Price et al. A generative model of speech production in Broca's and Wernicke's areas. *Frontiers in psychology*, 2: 237, 2011

[94] Deanna M Barch. The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.*, 2005; and Juliana V Baldo et al. The role of inferior parietal and inferior frontal cortex in working memory. *Neuropsychology*, 2006

Figure 6: Evolution of the voice across time on four scales



2.3 - Voices Across the Mind

> *Well, then, thought and speech are the same; only the former, which is a silent inner conversation of the soul with itself, has been given the special name of thought. Is not that true?*
>
> **— Plato**
> (Sophist, 263e-264a)

The second paradigm introduced in this dissertation argues that our experience of the world is highly informed by our experience of voices, both interior, and exterior. The exterior, also called the outer voice, is the familiar phenomenon of producing vocal sounds. Interior voice refers to this silent manifestation of the voice experienced in silent thoughts, silent reading, or mind wandering. In this worldview, we appreciate the voice as a membrane, an interface between inner personal experiences, and the social space we share with others. The familiar outer voice is a key to understanding the more hidden and intimate manifestations of the inner voice. But the interiorized aspects of the inner voice also offer a new appreciation of the outer voice.

In this section, we present inner and outer voices as a dual mechanism and detail their special parallel correspondence. After presenting a succinct journey to the inner voice, we explore some insights into the evolution of inner voices through time, and finally propose a phenomenology of inner voices based on their potential for connection.

2.3.1 - Journey to the Inner Voice

In the same way that the visual cortex has a specific function for recognizing faces, the human brain processes vocal sounds differently than any other noise[95]. Human audition is specifically tuned to hear the voice of others, distinguish them, and subconsciously decode information from them. Research on the cocktail party effect—the ability to tune one's attention to a single speaker when many people are talking at the same time—highlights our ability to tune our attention to voice[96].

Beyond the general processing of voices, previous work indicates that our brain treats our own voices differently than the voices of others. Indeed, even in cases where many different voices are present in the sonic environment, our brains can usually distinguish between self-produced sounds and those that stem from the external world. Studies have highlighted that the auditory cortex is inhibited when processing self-generated vs

[95] Marianne Latinus and Pascal Belin. Human voice perception. *Current biology : CB*, 21, 2011

[96] Barry Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 1992

played-back speech sounds[97]. By comparing the auditory cortical response to self-produced speech versus prerecorded, it has been shown that the brain's response to self-produced speech is weaker[98]. The same result was obtained with vocalisation of single tone instead of speech[99]. These results suggest that during vocal production, there is an attenuation of the auditory cortex sensitivity and that the brain modulates its activity as a function of the expected acoustic sensation. This inhibition has been shown to be present even before the vocal production, during mental preparation for speech. Your brain actually prepares itself to not listen to your own voice.

[97] Nadia Müller et al. Listen to Yourself: The Medial Prefrontal Cortex Modulates Auditory Alpha Power During Speech Preparation. *Cerebral cortex*, 2014

[98] John F Houde et al. Modulation of the auditory cortex during speech: an meg study. *Journal of cognitive neuroscience*, 2002

[99] Mika H Martikainen et al. Suppressed responses to self-triggered sounds in the human auditory cortex. 2005

Further studies connect this finding to the phenomenon of corollary discharge[100]. Indeed, for each motor command, the brain doesn't send one signal stream but two. One signal is sent to the muscle and the second signal informing part of the brain of the impending motion. This second signal is a predictive signal called corollary discharge. It creates an efference copy that is used by the brain to predict the action in comparison with somatosensory feedback to confirm the target motor action. In some cases, such as tickling, the efference copy is also used to inhibit any response to the self-generated sensory signal which would interfere with the execution of the motor task[101].

[100] John F Houde et al. Modulation of the auditory cortex during speech: an meg study. *Journal of cognitive neuroscience*, 2002

[101] Sarah-Jayne Blakemore. Why can't you tickle yourself? In *The Anatomy of Laughter*. Routledge, 2017

The production of speech and vocal sounds involves more than a hundred muscles, including laryngeal, mandibular, lingual, palatal, and respiratory muscles. This motor control is one of the fastest and most precise motor behaviors routinely performed compared to any other human behaviour[102]. Similarly to almost all other intentional motor commands, it comes with its own efference copy of the phenomenon. During the vocal activity, this feedforward, predictive copy is constantly compared to the auditory and somatosensory feedback signal perceived by the speaker. This comparison is hypothesized to takes place in the basal ganglia[103], a group of subcortical nuclei responsible for motor control, motor learning, executive functions and behaviors, and emotions[104]. The efference copy is also used to inhibit the response from the self-generated auditory signal in the auditory cortex and thus the reduction of the auditory cortex activity. Feedforward and feedback mechanisms are tied together in most instances of vocal production.

[102] Ray D Kent. The uniqueness of speech among motor systems. *Clinical linguistics & phonetics*, 2004

[103] James R Booth, Lydia Wood, Dong Lu, James C Houk, and Tali Bitan. The role of the basal ganglia and cerebellum in language processing. *Brain research*, 2007

[104] José L Lanciego, Natasha Luquin, and José A Obeso. Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine*, 2012

Screams—of fear or surprise—are interesting counterexamples. Recent research suggests that instead of resulting (as previously thought) from a fight or flight response, the acoustic properties of the scream[105] actually act as trigger for the brain's fear center more effectively than almost any other sound. One possible element of explanation for this effect is that the motor action is so sudden and unexpected that the brain perceives a loud feedback auditory signal without a predictive copy, which leads to a very strong visceral reaction.

[105] Luc H Arnal, Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, and David Poeppel. Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, 2015

But the role of this feedforward signal might have evolved further than a simple prediction. This system is now thought to be a basis of the inner voice[106]. Such phenomena are most studied in the context of schizophrenia and auditory-verbal hallucinations (AVHs)[107]. It is thought that AVHs are caused by a breakdown in the efference copy pathway, leading the brain to think that the voice observed is not one's own but is part of the feedback signal[108]. Although AVHs are one of the most commonly studied instance of inner voice experience, they are only one extreme example of possible experiences of the inner voice. In more regular cases of inner voices, the efference signal is simply inhibited before an action takes place, leaving only the efference copy and leading to the perception of inner hearing[109]. And those inner voices are—literally—heard. They activate voice-selective areas in the auditory cortex[110] and often possess the auditory qualities of overt speech, such as recognizable identity, gender, pitch, loudness, and emotional tone[111].

Some researchers have estimated that approximately one-third of our conscious waking life consists of inner language[112] although it may present a high inter-individual variability. Inner voices seem to constitute an important part of our inner lives along with inner seeing, feeling, sensory awareness and non-symbolised thinking[113].

The inhibition of auditory neurons in the brain during phonation in humans was a key factor in unveiling the neural bases of the inner voice manifestation. However, inhibition during vocalization is also seen in other vertebrates, such as birds[114], bats[115], and even during sonic production in the cricket[116]. This might support the idea of looking for sonic inner lives in other species.

But in humans, how does the inner voice experience develop and what purpose does it serve? Lev Vygotsky suggested that inner speech develops from overt speech in early childhood. For Vygotsky, it first develops as a phase of expanded inner speech, in which internal dialogue retains many of the acoustic properties and turn-taking qualities of external dialogue, and then becomes condensed inner speech, in which the semantic and syntactic transformations of internalization are complete[117]. As to its purpose, a common theory proposes that we use inner speech as a rehearsal tool in working memory: internal verbal rehearsal can refresh the memory trace continuously, leading to better recall[118]. Other researchers have highlighted the possible role of the inner voice as a potential error monitor for external speech in regard to language acquisition, second language learning, or singing[119]

[106] Xing Tian et al. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, 2016

[107] Simon R Jones and Charles Fernyhough. Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and cognition*, 2007

[108] Marc Seal et al. Compelling imagery, unanticipated speech and deceptive memory: Neurocognitive models of auditory verbal hallucinations in schizophrenia. *Cognitive Neuropsychiatry*, 2004

[109] Xing Tian and David Poeppel. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, 2012

[110] Bo Yao et al. Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *Journal of Cognitive Neuroscience*, 2011

[111] Ruvanee P Vilhauer. Inner reading voices: An overlooked form of inner speech. *Psychosis*, 2016

[112] Bob Uttl et al. Sampling inner speech using text messaging. *Proceedings of the Canadian Society for Brain, Behavior, and Cognitive Science*, 2012

[113] Christopher L Heavey et al. The phenomena of inner experience. *Consciousness and cognition*, 2008

[114] James S McCasland and Masakazu Konishi. Interaction between auditory and motor activities in an avian song control nucleus. *Proceedings of the National Academy of Sciences*, 78(12):7815–7819, 1981

[115] Gerd Schuller. Vocalization influences auditory processing in collicular neurons of the cf-fm-bat, rhinolophus ferrumequinum. *Journal of comparative physiology*, 1979

[116] James FA Poulet and Berthold Hedwig. A corollary discharge maintains auditory sensitivity during sound production. *Nature*, 2002

[117] Charles Fernyhough and H Pashler. Inner speech. *The encyclopedia of the mind*, 2013

[118] Alan Baddeley. Working memory. *Science*, 1992

[119] Robert J Hartsuiker and Herman HJ Kolk. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, 2001

In this dissertation, I refer to the phenomena as "experiences of the inner voice." Others have focused on the concept of inner language[120], inner speech, or inner speaking[121]. However, I wish to focus on the purely acoustic properties as arising from our ability to produce sound, rather than the verbal and linguistic quality of the phenomena. This contributes to larger thinking about the sonic quality of voices being equally or maybe enven more meaningful to our cognitive abilities than speech and language.

[120] Hélène Loevenbruck. What the neurocognitive study of inner language reveals about our inner space. *Langage Intérieur/Espaces Intérieur, Inner Speech/Inner Space*, 2018

[121] Russell T Hurlburt et al. Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 2013

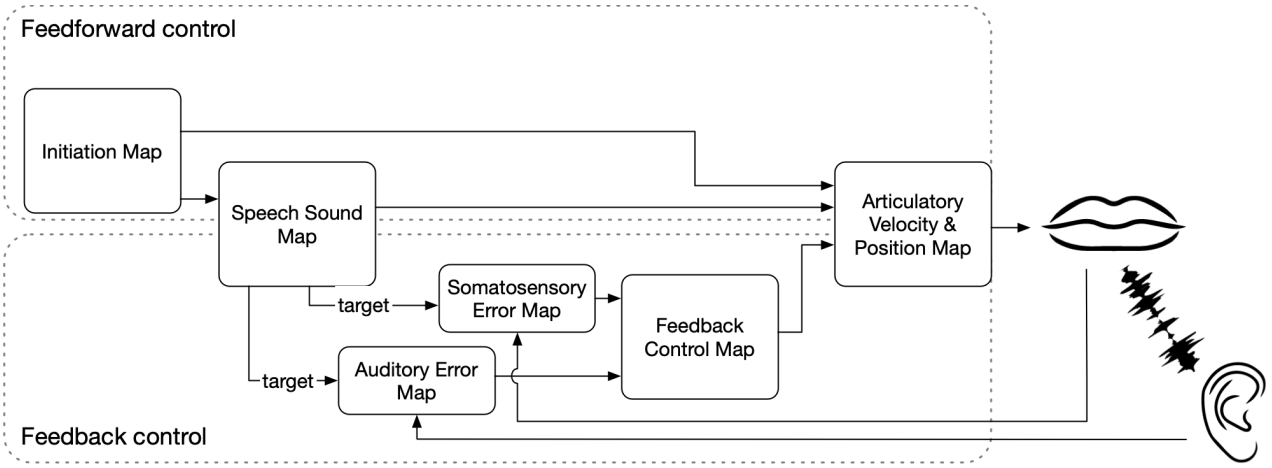


Figure 7: Feedforward/Feedback model of the voice, diagram adapted from (Guenther 2006)

2.3.2 - Evolution of the Inner Voice

In addition to providing a connection between thoughts and actions, voices also underlie our experience of text, data, and information. Most of us experience sonic manifestations of our inner voices when reading silently or even looking at a text without actually wanting to read it. Every piece of textual information absorbed silently passes through the inner voice at varying levels of verbalization that can be very abstract or sub-vocalized. In terms of the evolution of this phenomenon, some believe that the advent of silent reading drastically transformed readers' interior lives[122]. There are still debates around the historical origins of silent reading. Some literature researchers even argue that silent reading is a quite recent practice and might only have appeared around the fifteenth to seventeenth centuries.

[122] Paul Saenger. *Space between words: The origins of silent reading*. Stanford University Press, 1997

Though highly controversial[123], this theory is supported by several arguments. First, various texts report the transition from out-loud to silent reading around the fifteenth century, including the *Confessions from Saint Augustin*[124]. Such texts support the idea that, if not completely unknown in the ancient world, silent reading was at least so rare that whenever it was observed, it aroused astonishment. Secondly, *scriptio continua*, a style

[123] Bernard MW Knox. Silent reading in antiquity. *Greek, Roman, and Byzantine Studies*, 9(4):421–435, 1968

[124] Saint Augustine. *The confessions*. Clark, 1876

of writing without spaces or other marks between words or sentences, was the norm in Classical Greek and Latin writings[125]. It is thought that this typographical style almost required the reader to read out loud to decipher the meaning of the text. Over time, the performative reading system was replaced by faster silent reading[126] and punctuation marks and spaces between words became common[127].

Finally, another clue is sometimes used to support the idea that the understanding and perception of the inner voice have evolved over time the various reports of "hearing voices" throughout history. McCarthy-Jones retraces the evolution of societal views on auditory hallucinations and other experiences of inner voice[128], from common ghostly manifestations in cuneiform texts written by Babylonians (1050BC); to a seductive trait in antiquity; to Socrates's Daimonium; to biblical and Talmudic reports of hearing the voice of God. One hypothesis is that silent inner dialogues were less common in past civilisations and appeared as surprising and unnatural when they occurred.

More recently, the Hearing Voices Movement (HVM) is a peer-support group directed at creating ways for people who hear voices to exchange experiences and knowledge[129]. Their aims consist of raising awareness and promoting empowerment for people who hear voices, as well as challenging negative stereotypes associated with hearing voices. Recent work by Luhrmann shows that for people with psychotic disorders, voice-hearing experiences are also shaped by local culture—in the United States, the voices are harsh and threatening; in Africa and India, they are more benign and playful. The researchers suggest that this may be influenced by feelings of individuality or of belonging to a collective[130].

2.3.3 - A Phenomenology of the Inner Voice

There are different types of inner voices, and most of us experience some form of inner voice on a daily basis, in the context of self-awareness, past and future thinking, and emotional reflection. They manifest in various ways. Inspired by Perrone-Bertolotti[131] and Hélène Loevenbruck[132] I present a phenomenology of inner voices focused on their potentials for connection. To this end, I propose a categorization of inner voices according to three parameters: utterance, defining the type of delivery and sonic characteristics of the voices heard; controllability, describing the level of cognitive agency one has over the experience; and finally abstraction, detailing the complete or extended nature of the phenomenon.

[125] E Otha Wingo. *Latin punctuation in the classical age*, volume 133. Walter de Gruyter, 2011

[126] Richard A Lanham. *The economics of attention: Style and substance in the age of information*. University of Chicago Press, 2006

[127] Paul Saenger. *Space between words: The origins of silent reading*. Stanford University Press, 1997

[128] Simon McCarthy-Jones. *Hearing voices: The histories, causes and meanings of auditory verbal hallucinations*. Cambridge University Press, 2012

[129] Dirk Corstens, Eleanor Longden, Simon McCarthy-Jones, Rachel Waddingham, and Neil Thomas. Emerging perspectives from the hearing voices movement: implications for research and practice. *Schizophrenia bulletin*, 2014

[130] Tanya M Luhrmann, Ramachandran Padmavati, Hema Tharoor, and Akwasi Osei. Differences in voice-hearing experiences of people with psychosis in the usa, india and ghana: interview-based study. *The British Journal of Psychiatry*, 2015

[131] Marcela Perrone-Bertolotti et al. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, 2014

[132] Hélène Loevenbruck. What the neurocognitive study of inner language reveals about our inner space. *Langage Intérieur/Espaces Intérieur, Inner Speech/Inner Space*, 2018

UTTERANCE

Experiences of inner voice can take many shapes, and the voices themselves can range from manifestations very similar in acoustic quality to our own external voice, to the voices of other agents, and to more musical types of sounds.

Silent reading, inner reading voices, reminiscence of vocal memories, silent vocal thinking, internal monologue, internal dialogue, internal rehearsal of talk, etc., are all speech-based types of inner voice experiences. The utterance often corresponds to the subject's own spoken voice. But one interesting feature of inner speech is the flexibility in vocal characteristics that can be applied to them. There is evidence that silent reading and inner voice retain some of the properties of external, heard speech. One can, for instance, internalize and imagine the voice of someone else, such as their mother calling them for dinner. Another example is when silently reading a letter or email, one might hear the correspondent's voice. In silently reading "What's up, Doc?" or "I am your father," the specific individual voice of Bugs Bunny or Darth Vader might be heard internally. Silent reading is faster than oral reading, but silent reading speed can still be influenced by the imagined speaking speed of the character[133]. The flexibility in the type of utterance goes even further. I argue that non-speech types of inner voices can also be considered to be types of inner-voice experiences. Interiorizing a known song is a common phenomenon, where one can find themselves imagining not only the main vocal line sung by the singer, but also the different parts of the instrumentation. I believe that the internalization of non-vocal musical pieces fits the same process as speech. Famous pianist Glenn Gould was known for humming the melody while playing[134], which supports the idea of voice-based experiences even in the case of non-vocal music. With those examples, we can see that in experiences of the inner voice, the type of utterance can range from individual to vocal to musical.

When the delivery, or utterance, of our inner voice is modeled from the voice of someone else, it can be seen as a very intimate way to be connected with others: they are literally "in our heads." One powerful characteristic of dreams is that the dreamer is at the same time all the different dreamed characters[135]. This phenomenon demonstrates the existence of models of others and supports the theory of mind[136]. This phenomenon is also present during our waking hours, through our ability to multiply the number of voices and characters in our minds. Some projections are models of real people, others of fictional characters, and others completely imaginary. Our ability to internalize musical experiences involves an additional step in our connection to singers, musicians, or composers but might actually create a powerful visceral and deep connection, which doesn't rely on words but only on a shared internal model of acoustic aesthetic.

[133] Jessica D Alexander and Lynne C Nygaard. Reading voices and hearing text: Talker-specific auditory imagery in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 34(2): 446, 2008

[134] Paul Sanden. Hearing glenn gould's body: Corporeal liveness in recorded music. In *Liveness in Modern Music*. Routledge, 2013

[135] Paul Tholey. Consciousness and abilities of dream characters observed during lucid dreaming. *Perceptual and Motor Skills*, 1989; and Brigitte Holzinger. Conversation between stephen laberge and paul tholey, july, 1989. *Lucidity Letter*, 9(1), 1990

[136] Marvin Minsky. *Society of mind*. Simon and Schuster, 1988; and David Kahn and Allan Hobson. Theory of mind in dreaming: Awareness of feelings and thoughts of others in dreams. *Dreaming*, 2005

CONTROLLABILITY

The quality of inner voices ranges between highly controlled and uncontrollable. Auditory-vocal hallucinations are often reported to be uncontrollable. Most of us have experienced unstoppable earworms that can be considered a musical type of inner voice[137]. Reading can also be seen as unavoidable, given that it can not be "unlearned." Seeing a street sign or an advertisement often means that the verbal message is being transmitted from our eyes to our brain without our consent. In addition to reading, writing is also often accompanied by an internal voicing of the written thoughts that can be very hard to turn off. The "phonological mediation hypothesis," according to which spoken forms of words are retrieved before graphemic forms can be accessed, has been supported by studies on brain-lesioned patients[138]. This supports the idea that all experiences of written text are mediated by our experience of sounds and voices. And this holds whether the experience of text goes from the outside worlds toward our inner worlds, or vice versa. Indeed, according to the "recoding hypothesis," silent reading usually entails the conversion of the initial visual representation of a printed word into a speech-like form prior to accessing stored information about the meaning of the word. Readers usually translate the visual representation of a word into a speech representation in order to access the meaning of the word[139]. This gives substance to our second paradigm, according to which our experience of the voice informs our experience of the world. In addition, in its mediating nature, the voice also serves as a bridge between the world and our inner self. Making lists mentally and counting also recruit the inner voice, which also suggests that mathematical concepts might be voice-mediated.

However, the inner voice and inner speech are also often used in a willful manner, for instance when rehearsing for a talk or a conversation. It is involved in remembering personal past episodes, plays a role in autobiographical memories[140], and interacts with working memory to encode new material[141].

Another aspect in considering the controllability of the inner voice is the distinction proposed by Hurlburt between inner speaking and inner hearing, as reported by subjects undergoing Descriptive Experience Sampling. Both inner speaking and inner hearing are generated and perceived by the same person, but the agency shifts from a more conscious to a more subconscious form of control[142].

Another common form of the inner voice is verbal mind wandering, which consists of flowing, spontaneous, unconstrained, external-stimulus-independent verbal thoughts. Recognized for its beneficial outcomes in terms of prospection and creativity[143], mind wandering often emerges from a resting state and is often neither controlled nor parasitic.

From willful recall to verbal mind wandering, auditory hallucination or

[137] C Philip Beaman and Tim I Williams. Earworms (stuck song syndrome): Towards a natural history of intrusive thoughts. *British Journal of Psychology*, 2010

[138] Aleksandr Romanovich Luria. *Higher cortical functions in man*. Springer Science & Business Media, 2012

[139] Maryanne Martin. Speech recoding in silent reading. *Memory & Cognition*, 1978

[140] A Morin. Inner speech. encyclopedia of human behavior, w. hirstein, 2012

[141] Alan Baddeley. Working memory. *Science*, 1992

[142] Russell T Hurlburt et al. Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 2013

[143] Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015

parasitic auditory phenomena, the inner voice questions and challenges our sense of control over our own inner lives. However, the trained puppeteer pulling the string is none other than a deeper layer of one's own self. The voice not only connects our conscious self with the outside world of text, data, and information, it also creates a bridge between our conscious self and hidden parts of our subconscious.

ABSTRACTION

Inner speech plays a central role in human consciousness at the interplay of language and thought (Morin, 2005). Therefore it seems normal that the level of abstraction of the phenomenon might oscillate between the ethereal quality of thoughts and the concrete and finite nature of language. This abstraction (or condensation) seems to operate at different levels including articulation, phonology, lexicon, syntax, etc.

In 1881, Egger was the first to propose explanations as to why "la parole intérieure" is often accelerated compared to outer speech[144]. He argues that by not being limited by the speed of motor control and the necessity to breath between sentences, inner speech can go faster. In addition, he claims that, unlike talking to other people with whom we need to articulate our thoughts and use common ground expression and lingual rules, we do not need to follow linguistic and social rules when talking to ourselves. Therefore, the inner voice is not only physically condensed, but it can also be syntactically condensed. In his theory of inner speech, Lev Vygotsky developed this notion and proposes that inner speech is the developmental outcome of an internalization process emerging from social speech. Vygotsky came to his conclusion by observing children's overt self-directed speech (private speech) during cognitive tasks, and viewing it as a transitional stage in the transformation of interpersonal dialogues into intrapersonal ones[145] This internalization can be seen as a way to shape one's own individuality from interactions with others. In a way, our social connections inform the formation of the self and how we relate and connect to our inner selves.

[144] Victor Egger. *La parole intérieure: essai de psychologie descriptive*. Alcan, 1904

[145] Charles Fernyhough and H Pashler. Inner speech. *The encyclopedia of the mind*, 2013

Despite the frequent abstraction of the inner voice, it can also feel very concrete. The most concrete form of inner voice might be instances when it shadows the actual outer voice. Then, there is the case of inaudible whispering. Also used as part of sign language, mouthing is a type of silent speech consisting of moving the lips without producing sounds. Subvocalizing is often used during silent reading and tends to slow the reader by activating lips and guttural muscles[146].

[146] Maria L Slowiaczek and Charles Clifton Jr. Subvocalization and reading for meaning. *Journal of verbal learning and verbal behavior*, 1980

The scale from concrete to abstract inner voice can also be experienced in music. Classically trained musicians are often taught to not "sing the music in their head" but to rather internalize the sounds in a more intel-

lectualized manner[147]. However, one might internalize a song in various degrees of abstractions—from simply thinking about the song "Yesterday" by the Beatles in passing, for example, to actively mouthing the entirety of the song, in tempo. The mental ability to condense or extend acoustic phenomenon can be seen as a stairway from a concrete version of an auditory message, to a more personal, interiorized version that is only meant to be deciphered by ourselves. A bridge, a connection, a flexible missing link between thoughts and language.

[147] Stan Bennett. The process of musical creation: Interviews with eight composers. *Journal of research in music education*, 1976

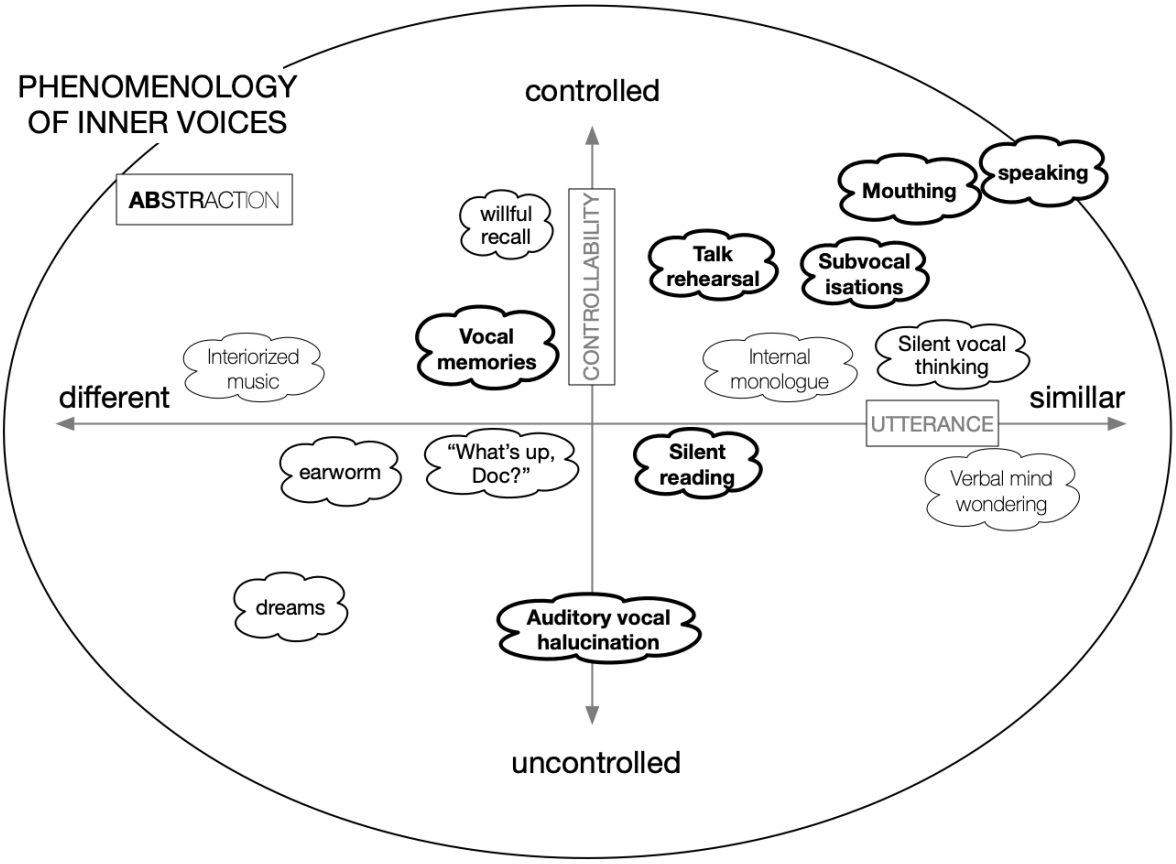


Figure 8: Phenomenology of the Inner Voice categorized according to three parameters: utterance, controllability, and abstraction,

2.3.4 - Studying the inner voice

Because of its invisibility to external observation, the inner voice has been resistant to empirical research. Direct methods for studying inner speech are often limited to self-report questionnaires, thought protocols, experience sampling, or the study of self-reported AVHs in the case of auditory hallucinations.

One interesting reason and entry point to study the inner voice is the

experience of atypical populations. Alderson-Day and Fernyhough propose seeing the inner voice as a lens to understand the experience of people with various neuroatypical conditions[148]. In regard to developmental disorders, they relay anecdotal findings indicating that descriptions of inner experience made by people with autistic spectrum disorder (ASD) who describe their inner experiences as "thinking in pictures" rather than inner speech. Their work also reflects on the controversy between the theory that impaired inner speech might be linked with ASD and other work that reports intact inner speech mechanism in people on the autism spectrum.

[148] Ben Alderson-Day et al. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 2015

Another population stands out as a potential key in exploring inner speech in an objective manner: The population of people who stutter (PWS). Stuttering is a neurodevelopmental disorder that affects $\sim$ 1 percent of the population worldwide. It is characterized by dysfluent speech with features of sound repetition, blockages, and prolongations[149]. According to Guenther's model[150], stuttering may be connected to discrepancies in the comparisons between the feedback and feedforward signals within the basal ganglia. The model suggests that altered auditory feedback (AAF), by acting on the neural bases of speech by disturbing the feedback control signal[151], might alleviate this problem. AAF consists of altering the outward voice to act on subconscious neurological mechanisms and render more fluent speech. This model also shows that the outer/inner voice mechanism can be used as a channel to affect the production mechanism of the voice. In chapter 5: Mumble Melody, I use those ideas to explore the idea of affecting mental processes by using this special auditory feedback channel.

[149] Stuttering Foundation. A nonprofit organization helping those who stutter, 2018. URL https://www.stutteringhelp.org/; and Jane E Prasse et al. Stuttering: an overview. *American family physician*, 2008

[150] Frank H Guenther. *Neural control of speech*. Mit Press, 2016b

[151] Ludo Max, Frank H Guenther, Vincent L Gracco, Satrajit S Ghosh, and Marie E Wallace. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary issues in communication science and disorders*, 2004

In summary, our silent inner lives remain connected to the physical manifestation and our experience of voices. Just like the spoken voice, the inner voice contains acoustic features[152], and we now have evidence of activations of temporal voice areas during silent reading[153]. One can wonder what this means in terms of the omnipresence of texts in our environment, or go a step further and query whether the inner voice exists independently from language, or at least from human language. Do other vocal creatures experience an inner voice? Do birds rehearse their songs silently? Far from denying that non-verbal animals have complex inner lives, this theory might suggest that any species capable of producing vocal sounds could potentially experience an inner life or at least an inner sonic life.

[152] Russell T Hurlburt et al. Exploring the ecological validity of thinking on demand: neural correlates of elicited vs. spontaneously occurring inner speech. *PLoS One*, 2016

[153] Marcela Perrone-Bertolotti, Jan Kujala, Juan R Vidal, Carlos M Hamame, Tomas Ossandon, Olivier Bertrand, Lorella Minotti, Philippe Kahane, Karim Jerbi, and Jean-Philippe Lachaux. How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. *The Journal of Neuroscience*, 32 (49):17554–17562, 2012

2.4 - Voices Across People

> *The aeroplane and the radio have brought us closer together. The very nature of these inventions cries out for the goodness in men — cries out for universal brotherhood — for the unity of us all,*
>
> **— Charlie Chaplin as the barber**
> (The Great Dictator, 1940)

The third paradigm frames the voice in its fundamental social nature, to build and regulate companionships originating in social grooming. Those behaviors appear to be the direct origin of our vocal messages, and also remain one of the most important parts of the message transmitted. The following section presents aspects of voices across people through the issue of disconnection, and in terms of inspirational prior art in Vocal Connection .

2.4.1 - Connection and Disconnections

The objective of this dissertation is to explore the creation of novel and deeper connections between various agents using the voice. I consider the notion of connectedness broadly, including the raising of personal self-awareness, the creation of strong interpersonal bonds, and the potential to create new forms of empathetic understanding with other species. Furthermore, I view these connections as primordial evolutionary motivations of the voice. This section explores the question of loneliness and the potential and risks of using technology in helping to create meaningful human connections.

Many have highlighted what seems to be an epidemic of loneliness in our societies[154]. Not to be confused with solitude, loneliness can be defined as a state of mind and the perception of being alone and isolated. Loneliness can affect many parts of a person's life, from health[155], to increased risks of suicidal thoughts[156], to chronically increasing our cortisol level which in turn increases risks of heart attack[157]. Loneliness might also result in epigenetic impacts to our offspring, as the emotional and physical impacts of loneliness can trigger cellular changes that alter gene expression and may be passed on to our children[158].

In addition to affecting our physical and mental health, loneliness affects our overall sense of wellbeing, life meaning, and personal achievement. Hospice workers recently issued a report on the most frequently expressed regrets of people on their deathbeds[159]. Two of the top five regrets touch on the frequency and quality of interpersonal connections: "I wish I'd had

[154] Sarvada Chandra Tiwari. Loneliness: A disease? *Indian journal of psychiatry*, 2013

[155] Julianne Holt-Lunstad et al. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on psychological science*, 2015

[156] Ariel Stravynski and Richard Boyer. Loneliness in relation to suicide ideation and parasuicide: A population-wide study. *Suicide and Life-Threatening Behavior*, 2001

[157] Nicole Vogelzangs et al. Urinary cortisol and six-year risk of all-cause and cardiovascular mortality. *The Journal of Clinical Endocrinology & Metabolism*, 2010

[158] Luc Goossens et al. The genetics of loneliness: linking evolutionary theory to genome-wide genetics, epigenetics, and social science. *Perspectives on Psychological Science*, 2015

[159] Bronnie Ware. *The top five regrets of the dying: A life transformed by the dearly departing.* Hay House, Inc, 2012

the courage to express my feelings," and "I wish I had stayed in touch with my friends."

Loneliness may be a common human emotion but it is also a complex and unique experience for each individual. Sarvada Chandra Tiwari proposes a categorization of loneliness into three types according to its causes[160]. **Situational loneliness** can occur when socio-economic and cultural milieu prevents people from creating meaningful connections; **developmental loneliness** can happen when a person is not able to balance the need for individualism and the need to be related to others; and **internal loneliness** can occur when low self-esteem heightens the feeling of loneliness regardless of the actual social level of intimacy with others.

[160] Sarvada Chandra Tiwari. Loneliness: A disease? *Indian journal of psychiatry*, 2013

Today, various debates rage over the effect of technology and the rise of social media as either a catalyzer for loneliness and lack of meaningful connection, or as a possible tool against it. More recent work from experts argues for the latter. In her book *Alone Together* [161], Sherry Turkle argues that relentless online digital connection is not real social intimacy and that social media use contributes to the decline in social intimacy. Interventions aiming at targeting the problem of loneliness can either touch on the contextual parts of the human connections or try to influence the depth and meaningfulness of the interaction. Indeed, it is a hard problem to increase the meaningfulness of interactions, it can seem easier to act on peripheral characteristics of the connection by creating pleasant triggers for connection, increasing its frequency, affecting the dynamic of the connection between two agents.

[161] Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Hachette UK, 2017

However, we believe that it is possible to create curated interventions that affect the depth of the interaction and allow agents to reach a greater level of intimacy and understanding. We know for instance that some subtle contextual elements can negatively affect the depth of human connection. For instance, the simple presence of a cellphone on a table lowers the quality of in-person conversation[162]. In her book *Reclaiming Conversation* [163], Sherry Turkle also mentioned the seven-minute rule inspired by one of her students. Many people seem to have lost the ability to concentrate for that long. However it may take some time for a conversation to reveal meaningful. And thus, the idea of setting a rule for oneself: in any face-to-face conversation, she would wait seven minutes before she gave up.

[162] Shalini Misra, Lulu Cheng, Jamie Genevie, and Miao Yuan. The iphone effect: the quality of in-person social interactions in the presence of mobile devices. *Environment and Behavior*, 48(2):275–298, 2016

[163] Sherry Turkle. *Reclaiming conversation: The power of talk in a digital age*. Penguin, 2016

2.4.2 - Grooming Talking

In the context of the loneliness epidemic, the voice can appear as a natural and ancestral way to transform otherness into togetherness. I believe that this is the most important reason for the voice. In this regard, I elevate

the concept of phatic interaction and argue that, at different levels, most of our vocal interactions are phatic interactions. This theory is based on past research as well as natural vocal phenomena. Subconscious vocal mimicry is the natural phenomenon of adapting one's vocal delivery to a conversational partner. We often adapt our vocal parameters, such as speed, volume, accent, breathiness, etc. to the people we are talking to as a way to empathize and appear more likable[164]. In the case of motherese—or infant/animal directed speech—adults adapt and exaggerate their prosody to appeal and capture the attention of non-verbal babies and animals. In those cases, again, it is the act of talking and the joint attention that matters more than what is being said. Ethnographer Malinowski established the theory that in so-called *primitive languages* the voice is used more as a mode of action than an instrument for thoughts compared to what he considered more "intellectual languages"[165]. A century later, ethnographers and language researchers have moved away from the simplistic opposition of savage versus intellectual languages. Additionally, some languages may have gone to greater lengths to obscure the importance of the connection established over the intellectual reach of spoken words. My work is also inspired by Scott, who considers laughter as a form of social grooming[166]. Here, I extend on this idea to include more general vocal behaviors.

[164] Rachel M Miller, Kauyumari Sanchez, Lawrence D Rosenblum, James W Dias, and Neal Dykmans. Talker-specific accent: Can speech alignment reveal idiolectic influences during the perception of accented speech? *The Journal of the Acoustical Society of America*, 127(3):1958–1958, 2010

[165] Bronislaw Malinowski. The problem of meaning in primitive languages. *Language and literacy in social practice: A reader*, pages 1–10, 1994

[166] Disa A Sauter, Frank Eisner, Paul Ekman, and Sophie K Scott. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107 (6):2408–2412, 2010

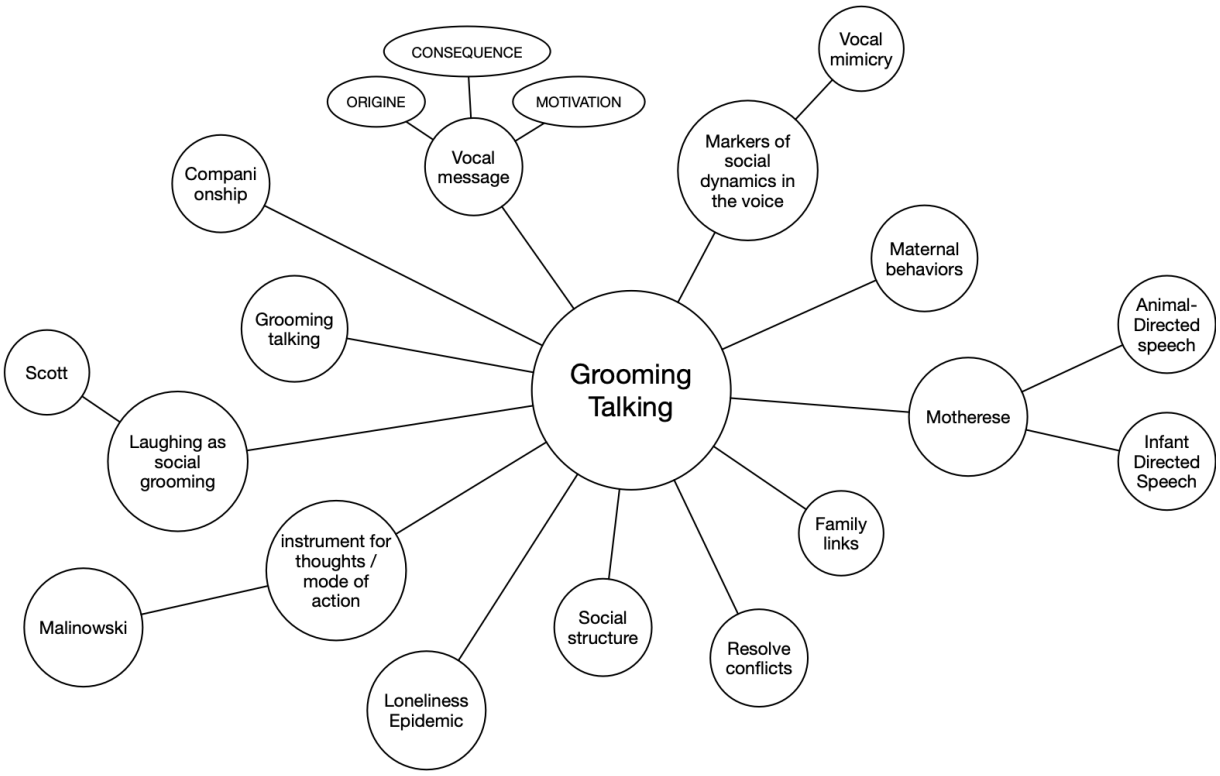


Figure 9: Grooming Talking: behavioral connections as primordial motivations of the voice

2.4.3 - Foundational Previous Work on Vocal Connection

Some prior works in technology, design, and art are particularly relevant as inspiration and illustrations of leveraging the potential of the voice to create meaningful connections. Starting with technologies that enabled the voice to transcend space—such as the tin can telephone in the 19th century and then the telephone in 1876[167]—and time—with audio recording techniques such as the phonautograph, phonograph, and microphone[168]. Claude Shannon's work on information theory provided fundamental contributions to natural language processing and computational linguistics. In his article from 1951, "Prediction and Entropy of Printed English[169]," Shannon gives a statistical foundation to language analysis by identifying upper and lower bounds of entropy on the statistic of English. The discipline of voice computing also has a rich history starting with Wolfgang von Kempelen's Acoustic-Mechanical speech machine in 1784[170] and Edison's Dictation Machine in 1879[171]. Such inventions provided new perspectives on the voice at an individual and experiential level. Today, processing and synthesis of voices have reached new horizons. In 1952, the first documented speech recognizer, Audrey, was invented by Bell Laboratories. Fully analog, Audrey could only recognize digits with pauses in between[172]. Dragon Systems pioneered Dragon Dictate (1990) and Dragon NaturallySpeaking (1997) the world's first commercial speech recognition software for dictation and transcription (also including Voice Search and Command/Control.) These personal computer-based systems featured Hidden Markov-Models, a powerful stochatsitc processing methodology probabilistically modelling multiple sources of information[173]. New voice models have also enabled the advent of natural language processing, speaker identification, emotion recognition from speech, diarisation or speech synthesis. Some specific uses for such technologies directly target the sense of connectedness, such as the company VocaliD[174], which enables the creation of unique personalized synthetic voices for children and adults with voice disabilities.

My work is also inspired by embodied techniques and historical phenomena around vocal connection. Anne Sullivan used the Tadoma method, developed by Sophie Alcorn at the Perkins School for the Blind, to teach Helen Keller to read voices[175]. To understand spoken words, Keller learned to place her hand over the face, mouth, and neck of her conversational partner to feel the vibration and airflow from their voice. This example is an insightful illustration of vocal connection that plays on the holistic voice as composed of tactile elements as well as sounds. Ventriloquism, also known as the ability to "throw" one's voice, dates back to Ancient Greece and was originally used in religious practices[176]. Multiplying one's own voice and distancing it from one's own body creates some mysticism and a feeling of estrangement and awe even today. Similarly, the voices of

[167] Herbert Newton Casson. *The history of the telephone*. AC McClurg & Company, 1910

[168] Vesa Välimäki et al. Digital audio antiquing-signal processing methods for imitating the sound quality of historical recordings. *Journal of the Audio Engineering Society*, 2008

[169] Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 1951

[170] Homer Dudley and Thomas H Tarnoczy. The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, 1950

[171] Theresa M Collins et al. *Thomas Edison and Modern America*. Boston: Palgrave Macmillan, 2002

[172] M Manjutha, J Gracy, P Subashini, and M Krishnaveni. Automated speech recognition system–a literature review. *Computational Methods, communication techniques and informatics*, 2017

[173] James Baker. The dragon system–an overview. *IEEE Transactions on Acoustics, speech, and signal Processing*, 1975

[174] Camil Jreige, Rupal Patel, and H Timothy Bunnell. Vocalid: personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 259–260. ACM, 2009

[175] Helen Keller and Annie Sullivan. *The story of my life*. Doubleday, 1904

[176] Valentine Vox. *I can see your lips moving: The history and art of ventriloquism*. Kaye & Ward, 1981

castrati created a sense of amazement arising from the apparent disconnect between a childlike/feminine voice emerging from an adult male. While it is, fortunately, illegal to mutilate young men to produce castrati, voice actors may create a similar effect.

Some important instances of vocal connection were developed by artists through music, performance art, movies, and architecture. The use of silence and voices in the work of Charlie Chaplin provides an example of the power of looking at the complex relationship between human sounds, connection, and words. Because of his incredible success during the silent film area, Chaplin was one of the rare artists to have successfully resisted the talkies. Indeed he continued producing acclaimed silent films, such as *City Lights* in 1931, even among the new world of sound. Modern Times from 1936 is another striking example. In a film blaming technology for the Great Depression, he makes the choice to use a recent important technology of the time—sync sound—but to subvert its intended use. Indeed, he doesn't use dialogue, but only the voice itself, through the famous "Nonsense Song," sung in gibberish. For this scene, Chaplin chooses voice over words to pass along an important message. He does not need dialogue, only voice, to make the audience understand the story. After this movie, his famous character of the Tramp was never seen on screen again. However, four years later, the silent star finally spoke out, in *The Great Dictator*, which utilized actual dialogue for the first time in Chaplin's career.

In terms of immersive art, Vocal Connection has been an important concern in architecture and acoustic design for millennia. Atriums in Ancient Greece were designed to suppress low-frequency background noise, while passing on the high frequencies of performers' voices. Researchers are today rediscovering the techniques used at the time to obtain such acoustic results, from the geometry and row alignment[177] to the limestone in the seats, which created a sophisticated acoustic filter[178]. Byzantine churches were also designed specifically to make sounds both intelligible and immersive[179]. The field of music and performative art also led to inspiring instances of vocal connection. Alvin Lucier's piece "I am sitting in the room," written in 1969, marks a milestone in terms of vocal consideration beyond words and in terms of experience. The piece features Lucier recording himself reading a text, and playing the recording back into the room, re-recording it and re-playing it repeatedly until the words become unintelligible. Then his voice becomes a pure vessel to reveal the acoustic properties of the room. The script read in the piece is the following:

> I am sitting in a room different from the one you are in now. I am recording the sound of my speaking voice and I am going to play it back into the room again and again until the resonant frequencies of the room reinforce themselves so that any semblance of my speech, with perhaps the exception

[177] Nico F Declercq and Cindy SA Dekeyser. Acoustic diffraction effects at the hellenistic amphitheater of epidaurus: Seat rows responsible for the marvelous acoustics. *The Journal of the Acoustical Society of America*, 2007a

[178] Nico F Declercq and Cindy SA Dekeyser. The acoustics of the hellenistic theatre of epidaurus: the important role of the seat rows. *Canadian Acoustics*, 2007b

[179] Arthur M. Noxon. *Understanding Church Acoustics*. Acoustic Sciences Corporation, 2001; and Peter K. McGregor. Revealing the acoustic mysteries of Byzantine churches, 2019. URL https://faithandform.com/

> of rhythm, is destroyed. What you will hear, then, are the natural resonant frequencies of the room articulated by speech. I regard this activity not so much as a demonstration of a physical fact, but more as a way to smooth out any irregularities my speech might have.

This sound art piece is also very meaningful in our work in light of Lucier's severe stutter, which shines a different light on the concept of smoothing irregularities from speech. Other inspirational musical and performance pieces include John Chowning's *Phone*. The 1981 piece was the first to use FM synthesis to generate voice-like sounds[180]. This allowed for a more complete view of the voice beyond language and more centered on timbral qualities. Laurie Anderson's *Handphone Table* uses bone conduction to convey sound from a vibrating table directly to the visitor's ears. Visitors hear the table only while seated, elbows making contact with particular points of the tabletop and hands covering the ears. This is another instance of looking at the voice holistically and experientially to create connection. More recently, Tod Machover's *Philadelphia Voices* takes this concept to a new level. The piece uses a symphony of voices to create a dialogue between individual voices and their place in the community. This compositioon also uses the individual voices of the performers and contributors to reflect on the power of the communal voice of the people to shape the future of democracy.

[180] David B Schwarz. *X: An Analytical Approach to John Chowning's Phone*. PhD thesis, Citeseer, 2010

Figure 10: Laurie Anderson's *Handphone Table* Photo credit:calisphere.org

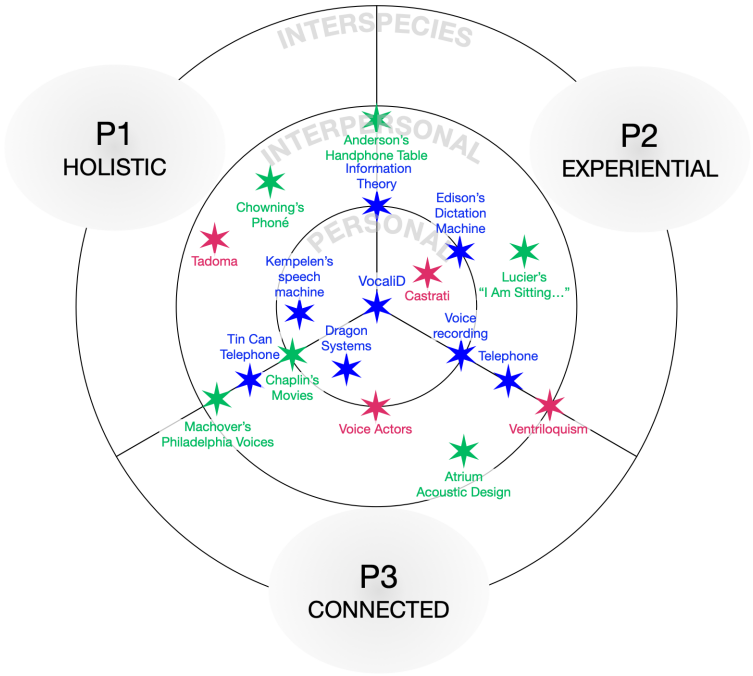


Figure 11: Inspirational Prior Art organised through the paradigms of the Vocal Connection Framework

In summary, figure 12 compiles the foundational knowledge presented in this chapter into a coherent problem space. From here, the preliminary projects presented in chapter 4, as well as the three main projects of this dissertation work, will be mapped and discussed using the framing of this problem space.

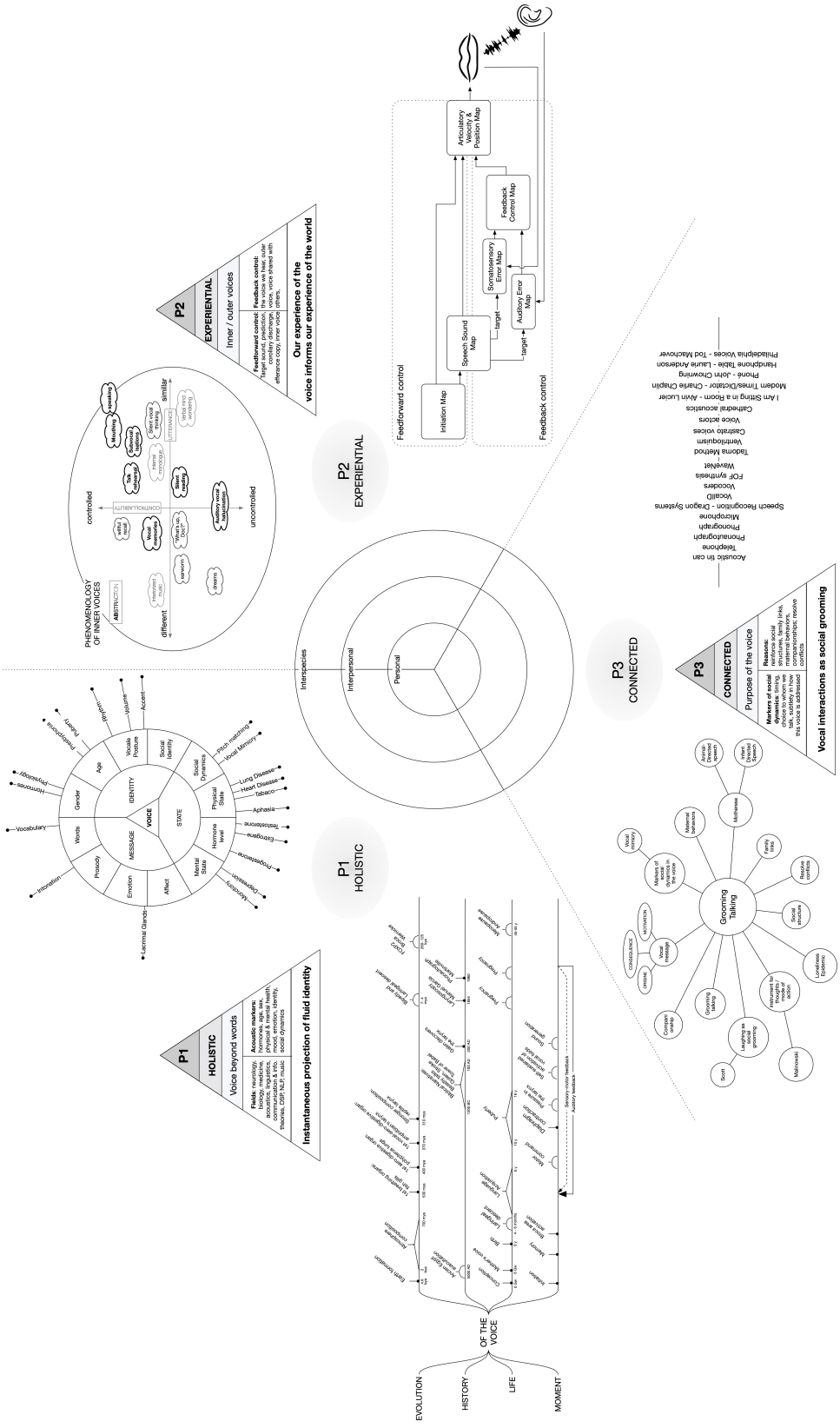


Figure 12: Mapping of the Problem Space

3 - Design Studies

The previous chapter presents theoretical explorations that led to the development of the framing of our problem space. The crystallization of those ideas and approaches was also supported by a series of early practical explorations, including preliminary design studies, ideations, development, and deployment of devices, software, and experiences. This chapter presents a summary of five preliminary explorations in the field of Vocal Connection. The *ORB* was developed as a design and art project to explore the physicality of one's voice and become aware of the tactile vibration it produces in the body. The *MiniDuck* project is an augmented book that plays on the musicality of everyday speech. *SIDR* is a deep learning-based real-time system for speaker identification to support group emotional intelligence. *Nebula* is a voice-controlled interactive web interface that merges the voices of an entire city into one. Finally, *Fleur Pulmonaire* is a robotic tool that reflects on the breathing of two people synchronously and shows how they might interconnect.

For each project, I succinctly present design rationales, application scenarios, and preliminary findings. I then frame those early explorations in terms of specificity of the connection established through the experience and the paradigm employed to consider the voice. Those explorations participate in the genesis of this dissertation work and guide the journey toward vocal connection.

3.1 - The ORB

Project

Tod Machover's Vocal Vibration installation explores the relationships between human physiology and the resonant vibrations of the voice. Built as part of the installation, the ORB (Oral Resonant Ball) is a small hand-held ceramic object that translates the voice of the user into tactile sensations[181]. The original version developed as part of my master's thesis required the use of two networked computers. This version was used for the original Vocal Vibrations exhibition presented in Paris in 2014[182]. In 2015,

Figure 13: The ORB (image credit: Bold Design)

[181] Rébecca Kleinberger. Singing about singing: using the voice as a tool for self-reflection, 2014

[182] Charles Holbrow, Elena Naomi Jessop, and Rébecca Kleinberger. Vocal vibrations: A multisensory experience of the voice. In *NIME*, 2014

Figure 14: Insole application using the same technology as in the ORB

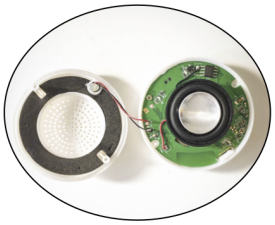


Figure 15: Anchor project: when turned on, the user feels the device vibrate along with any surrounding voices, and no vibration if the voice comes from inside their head

Figure 16: A portable monitoring device that senses the voice through a throat microphone and gives tactile feedback through a vibrating bracelet

Figure 17: Partition benches, pop-up installation at the MFA from the Hot Milks foundation

[183] Rébecca Kleinberger. Vocal musical expression with a tactile resonating device and its psychophysiological effects

the exhibition was brought to Cambridge, Massachusetts alongside the development of a portable version of the ORB. The new version contains all the electronics in the small enclosed polyurethane base. This simple plug-and-go version allowed me to offer simple demonstrations to hundreds of visitors during Lab tours, exhibitions, workshops, etc.

Lessons and Findings

The ORB provides a vocal experience of connection between audition and touch. It was originally built as a very simple way to provide everyone with a novel, surprising, often soothing experience of their own voice[183]. The original experience of the ORB had the user listen to and vocalize along with a six-minute musical composition by Tod Machover and experience their own voice through vibrations. Over the years, more than 200 people have played with the ORB during tours, lab demos, and exhibits in Mexico, Lyon, and Amsterdam. People who have experienced their voices through the ORB have expressed an overwhelmingly positive response to the object, even after timeframes of just a few seconds, and even without music. Women seem to react more strongly than men, and general feedback ranges from "very calming" to "very engaging." User feedback also provided us with other use-cases currently in various stages of ideation or deployment.

One derived application imagined by Adi Hollander is an insole form factor for deaf users to help enhance lipreading (figure 14). This application lies at the frontier between a personal context, as it changes the way an individual processes voice, and an interpersonal context, as it eventually aims at facilitating conversations. This idea also focuses on the establishment of a connection.

With help from Adam Haar Horowitz and Ishaan Grover, we developed Anchor, a pocket-sized device to help people with schizophrenia who experience auditory hallucinations determine whether a voice comes from inside their heads or not (figure 15); not knowing if a voice is real often causes distress, as people rarely dare to ask others if they can hear the voice too.

I have also worked on versions to help professional singers improve their physical awareness of the voice, and bracelet form factors to monitor people's use of their voice throughout the day (figure 16).

Finally, through the Hot Milks Foundation, I took part in building a pop-up installation commissioned by the Boston Museum of Fine Art. The Partition installation was composed of a pair of benches that vibrate alongside audio recordings or real-time voice input from a microphone at the other end of the room. Visitors were invited to share secrets into a

microphone that could be sensed, but not heard, by people sitting on the benches (figure 17).

Those ideas were helpful in extending the field of possibilities in terms of connectedness with one's own voice and the voices of others. This project marked an important step in my work in broadening the definition of the voice from a mere sound to a multisensory embedded experience. It was also an early experiment in using a physical object as a way to connect a person with their own voice. To create this bond, we take a step back in order to exteriorize an internal phenomenon. By creating initial estrangement, we aim to build a stronger connection with this familiar phenomenon. Work on the ORB is ongoing toward research and potential commercialisation.

Figure 18 details the contexts and paradigms tackled by the ORB and derived projects. This early exploration helped me shed light on the potential of vocal connection technology in the personal context while slightly touching on the interpersonal aspect when it comes to conversational support. Indeed, the technology touches on the interpersonal context as it encourages dialogue (Partition), enhances understanding (Insole), reveals one's role in the shared vocal space (Monitoring Bracelet) or supports serenity in a social context (Anchor).

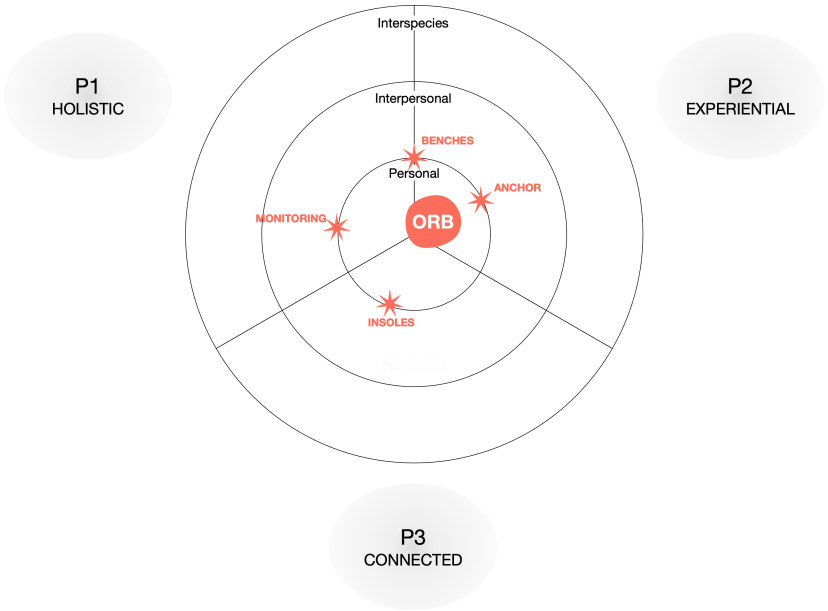


Figure 18: Mapping of the ORB and derived formfactors and applications of the technology onto the Vocal Connection space

Figure 19: MiniDuck book

3.2 - MiniDuck

Project

MiniDuck is a multi-modal augmented book developed in 2013 in collaboration with Akito van Troyer. It was exhibited at the MIT Media Lab as part of the Other Festival. The project has the appearance of a normal book but is sonically interactive. Each uniquely designed page tells a story about sounds and music, and contains text and graphical elements inspired by visual poetry and the Oulipo poets. We used different graphical techniques to encourage readers to touch and connect with the drawings, enriching their intellectual experiences with touch and music. When a user touches the text on the page, the book plays music corresponding to the vocal musicality of that specific section of the text. The music is pre-composed from recordings of someone reading the text. We then extracted the pitch contour and harmonized the score. Then we chose a timbre associated to the semantic of the page. The resulting MIDI score was associated with the page. The system recognizes which page is currently open, detects which word or group of words the finger is pointing at, and plays music accordingly. By letting their finger slide and run on the page, the reader controls the tempo of the melody. MiniDuck translates the musicality of the reading of a book from a textual medium to a musical medium and aims to make people aware of the omnipresence of music in their everyday life. While the reader indulges in reading the book, the specially processed audio accompanies the reader through the reading and the semantic understanding of the texts. More details are available on the project website[184].

[184] Kleinberger Rebecca Van Troyer Akito. Miniduck webpage. URL http://web.media.mit.edu/~akito/Projects/MiniDuck/

Lessons and Findings

Our goal was to translate the internal musicality of reading a book from a textual medium to a musical medium, and make people aware of the omnipresence of music in their everyday lives. This project was an important stage in the author's design journey, as it is the source of the later implementation of SpeechCompanions that translates the musicality of speech in real-time which could have several applications. In addition to being applicable to musical composition, we are also currently assessing its potential to influence users' moods and choice of words, and to improve fluency for people who stutter. The MiniDuck project mixed three media into one experience: sound, sight and touch. The multiplicity of media doesn't aim at creating a whole new experience but rather emphasizing what is already present in the traditional experience of reading. The other application of this project was to unveil the existing latent human experience of voices. Silent versus out-loud reading are fundamentally different processes in the brain[185], and are an interesting doorway to studying the experience of the inner voice. Similarly to the ORB, MiniDuck is designed

[185] Paul Saenger. *Space between words: The origins of silent reading*. Stanford University Press, 1997

to enhance the connection of different senses—in this case, vision, audition, and touch—as well as connecting silent and out- loud reading experiences.

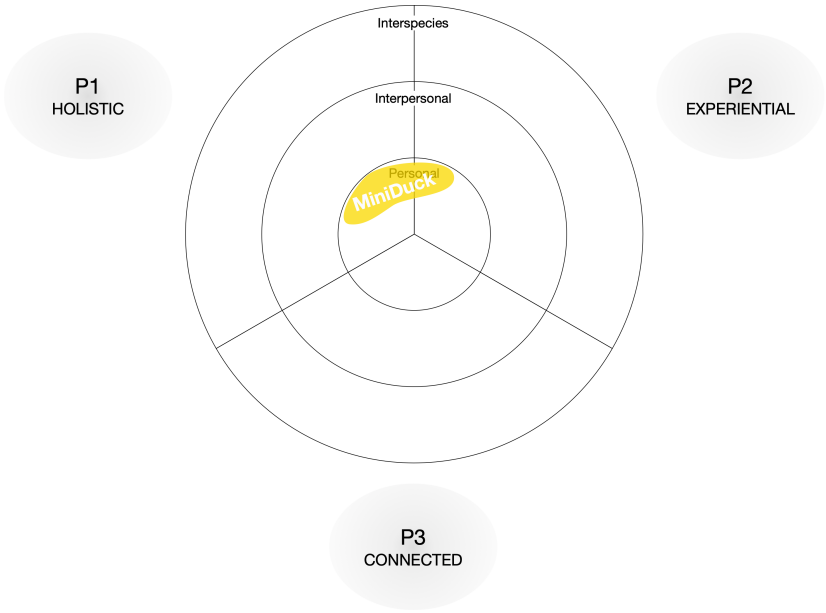


Figure 20: Mapping of the MiniDuck project onto the Vocal Connection space: The MiniDuck provides a personal experience to reflect on the holistic aspects of voices.

3.3 - SIDR

Project

SIDR is a real-time, deep-learning based speaker identification system developed in collaboration with Dr. Clement Duhart. By providing a real-time visualization of the use of the shared vocal space, this systems aims at bringing more awareness to turntaking during conversations and meetings. We consider each of our individual voices as a flashlight to illuminate how we project ourselves in society and how much sonic space we give ourselves or others. Thus, turn-taking computation through speaker-recognition systems has the potential to help us understand social situations or work-meeting dynamics. The SIDR system uses the same Convolutional Neural Network pipeline as presented by Mayton in [186] and is resilient to noise and adapts to room acoustics, different languages, and overlapping dialogues. While existing systems require several microphones for each speaker or coupling video and sound recordings for accurate recognition of a speaker, SIDR only requires a medium-quality or computer-embedded microphone.

Figure 21: SIDR project

[186] Brian Mayton et al. The networked sensory landscape: Capturing and experiencing ecological change across scales. *PRESENCE: Teleoperators and Virtual Environments*, 2017

Lessons and Findings

We originally developed this system as a tool to help us understand turn-taking and group intelligence but then realized the potential of creating interventions as real-time feedback for people's use of the shared vocal space. Could such a system, if implemented within companies, help increase participation of minorities and women to reduce biases or at least make participation inequalities more visible? The system designed for this project is at the base of some of our current work on classifying and understanding animal vocalization and helped us gain insight on how to design intervention systems to connect groups and individuals. The resulting experience creates connection based on the unique vocal signature of our voice when observed holistically.

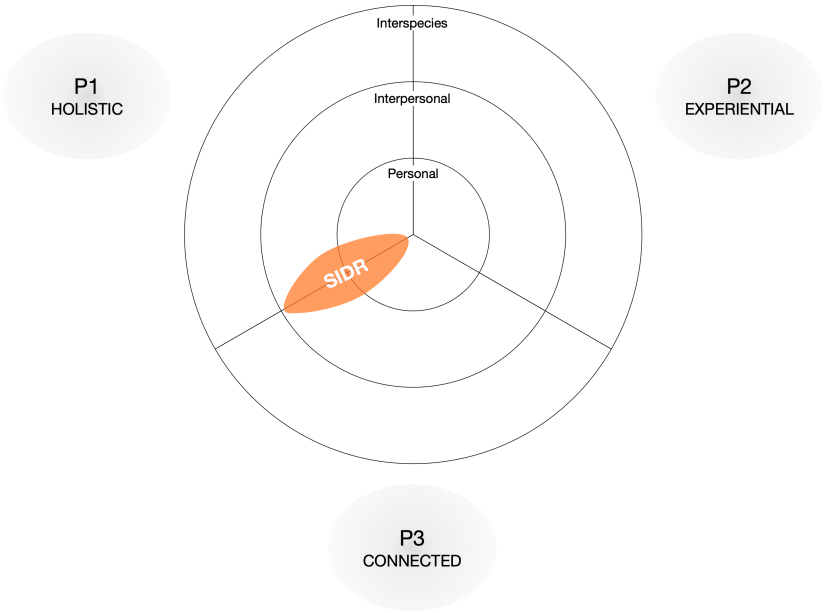


Figure 22: Mapping of the SIDR project onto the Vocal Connection space.

3.4 - Nebula

Figure 23: Nebula interface

[187] Akito van Troyer. Constellation: A tool for creative dialog between audience and composer

Project

Nebula is an interactive web interface I developed as part of Tod Machover's *Philadelphia Voices* project. Nebula is a voice-controlled interactive software app that allows users to conduct a choir of diverse vocal sounds by using only their own voices as input. The system is based on the Constellation project by Akito van Troyer, which takes sonic materials and organizes them visually to let anyone compose creative soundscapes[187]. Nebula uses hundreds of vocal samples that are represented as individual stars and organized by perceptual and spectral audio features. The samples

are triggered and activated when the user sings or produces any sound with the voice. The voice is analyzed in real-time, and this analysis is then used to trigger and mix a cascade of sounds with similar features. The voice becomes a conductor's baton that creates a dialogue without words between the individual and the community. And once a participant uses Nebula, their own voice, first used as a controller, is then transformed into a new sample adding an additional star to the experience for all subsequent participants. The result—a final cosmos of voices—provided material used amongst other sonic and text material for the *Philadelphia Voices City Symphony* composition.

Lessons and Findings

About 60 unique users tried Nebula either online or as part of an on-site workshop run by the *Philadelphia Voices* team. Through observation of the workshop participants (mainly children), we had quite positive responses as the interactivity, both visual and sonic, kept them engaged for an extended period of time. The overwhelmingly positive public reception of Machover's final piece both in Philadelphia and New York highlighted the power of losing/merging one's voice within the voices of the community. The objective of this project was to explore the potential of connecting the individual and the collective.

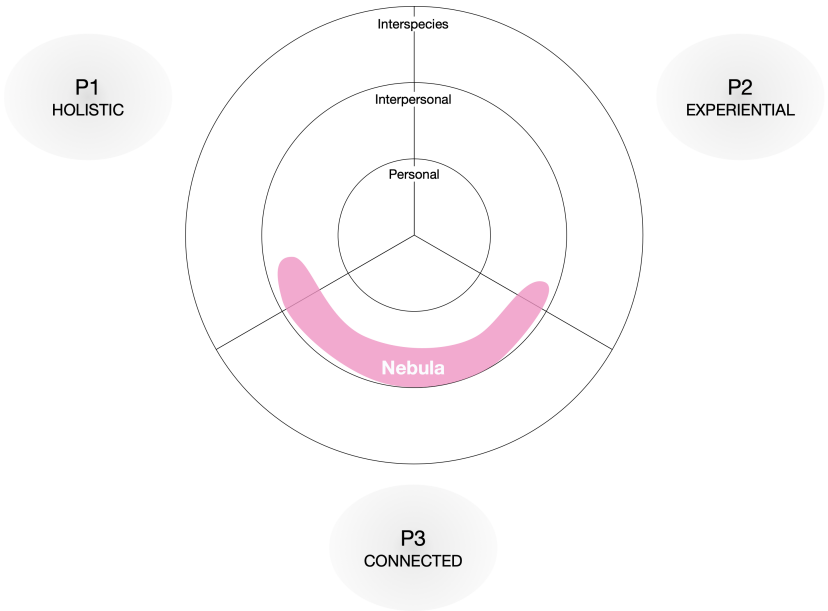


Figure 24: Mapping of the Nebula project onto the Vocal Connection problem space

3.5 - Mapping of the Design Journey

Those four preliminary projects provided inspirations and insights for the three main projects that constitute the core of this dissertation work.

Presented in the next chapter, the *Memory Music Box* project is a low-threshold platform for older adults to encourage conversations to take place. We designed it as an interactive portal within a familiar form factor—a jewelry music box—to reinforce the sense of connectedness across generations while sustaining personal memories. This project is the result of design explorations to encourage conversations to take place. The voice is an essential connecting tool, but in some situations, the road to conversations is paved with obstacles. Anchored in the framing of this dissertation, the Memory Music Box tackles both the personal and interpersonal contexts by reinforcing intergenerational links and sustaining memories. The device is highly experiential and aims at creating a seamless connection between the user and the correspondent. The MiniDuck project was instrumental in the development of the Memory Music Box in leveraging the potential of an ordinary everyday object in the design of interactive vocal portals. The ORB helped us realize the power of creating simple experiences of the voice as if viewing it with a new lens. We also used this approach in the design of the Memory Music Box.

Underlying our interpersonal experience of vocal connection is the relationship we have with our own voice. Beyond the work on vocal interactions between people, this dissertation also tackles the more intimate context of the relationship with one's own voice. As introduced in the background section, stuttering can be seen as both a key to explore the neural basis of the inner voice, and as a field where our research approach could potentially make a meaningful impact on people's lives. In chapter 5, I present *Mumble Melody*. For this project, we extract musicality from everyday speech as a way to access inner voice processes and help people who stutter gain increased fluency. The MiniDuck project marked my first experimentations in extracting musicality from speech. We then used the same underlying process, in a more substantial way and in real-time, in the development of the Mumble Melody system. The real-time digital signal processing used within the Mumble Melody project also derives from the Nebula project, where I explored multiple ways to treat vocal speech signals as musical mediums.

Moreover, we - humans - are not the only vocal creatures. This evolutionary tool is shared by tens of thousands of species around the globe. I believe that learning to listen to their voices as well is an important part of this work in light of our three paradigms. First, because many vocal acoustic

markers are also present in some animal voices, such as hormone levels[188], upbringing[189], stress level[190], thermal comfort[191] and other markers. Understanding animal voices in a holistic manner can bring insights to our understanding of human voices beyond words. Second, looking at animal voices in light of our own experience of voices and inner voices may open new doors to better interspecies understanding. And finally, animal voices are often used to support social structural behaviors[192] which may shed light on the origin of the use of the human voice as a form of social grooming. In this context, the last project presented in chapter 6 of this dissertation is a series of explorations and interventions conducted at the San Diego Zoo under the general umbrella of *Sonic and Vocal Enrichment at the Zoo*. This initiative is also highly informed by the preliminary projects presented in this chapter. The SIDR project uses the same underlying principle and the same platform as we later used in our Zoo initiative, for the recognition of Panda vocalization. The Nebula project also inspired us to think deeply about the uniqueness of zoo soundscapes and the importance of species or humans coevolving and cohabitating within the same city or habitat.

In summary, figure 25 illustrates the mapping of influence from our four preliminary projects into the design and deployment of the three main projects presented in the following chapters.

[188] Megan A Owen et al. Dynamics of male–female multimodal signaling behavior across the estrous cycle in giant pandas (ailuropoda melanoleuca). *Ethology*, 2013; and B D Charlton et al. Vocal cues to male androgen levels in giant pandas. *Biology Letters*, 2010

[189] Ellen C Garland, Jason Gedamke, Melinda L Rekdahl, Michael J Noad, Claire Garrigue, and Nick Gales. Humpback whale song on the southern ocean feeding grounds: implications for cultural transmission. *PloS one*, 2013; Volker B Deecke, John KB Ford, and Paul Spong. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. *Animal behaviour*, 2000; and Robert E Lemon. How birds develop song dialects. *The Condor*, 1975

[190] Vasileios Exadaktylos, Mitchell Silva, Daniel Berckmans, and H Glotin. Automatic identification and interpretation of animal sounds, application to livestock production optimisation. *Soundscape Semiotics-Localization and Categorization*, pages 65–81, 2014

[191] Daniella Jorge de Moura, Irenilza de Alencar Nääs, Elaine Cangussu de Souza Alves, Thayla Morandi Ridolfi de Carvalho, Marcos Martinez do Vale, and Karla Andrea Oliveira de Lima. Noise analysis to evaluate chick thermal comfort. *Scientia Agricola*, 2008a

[192] Donald H Owings, Eugene S Morton, et al. *Animal vocal communication: a new approach*. Cambridge University Press, 1998

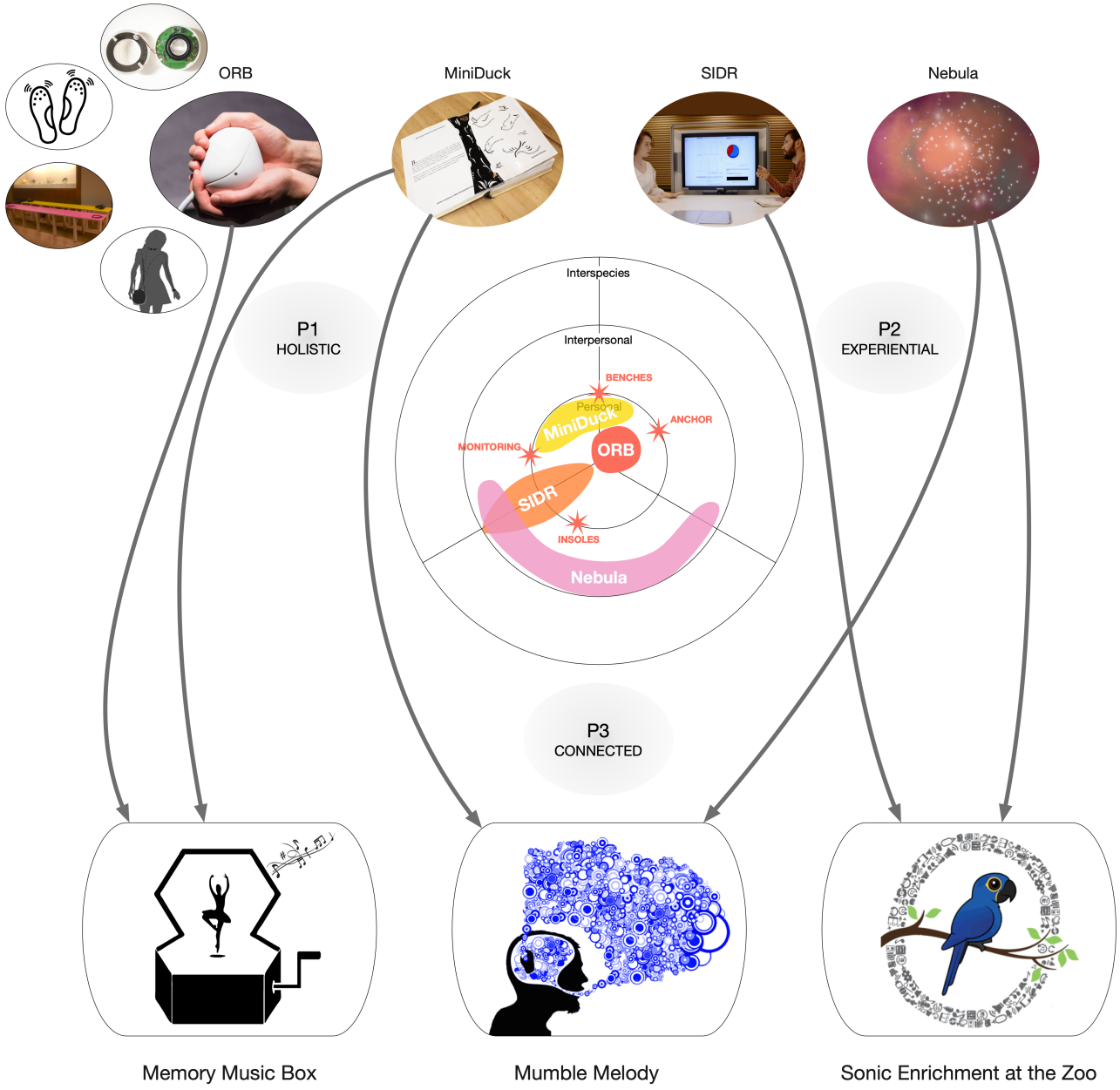


Figure 25: Graphics describing how the preliminary projects inform the design of the three main projects: Memory Music Box, Mumble Melody and Sonic Enrichment at the Zoo

4 - Memory Music Box

The voice evolves throughout our lives. As babies, our brain pathways are not yet trained and our voice motor commands mechanisms are not yet skilled enough to control the voice[193]. At this age, parents and society invest in individual infants, knowing that it is through vocal interaction and benevolent "motherese" that learning will occur[194]. However, during the latest stages in life, our voices, control mechanisms, and even memories can fade and come back at stages similar to our infant beginnings. But what happens after the words disappear? In Okinawan culture, the Kajimaya () tradition tells us that when an elderly person reaches 97 years old, they are considered to have gone back to childhood and are entitled to be considered and treated as such. However, most western cultures are not well equipped to acknowledge and manage senior citizens with late-stage dementia or Alzheimer's disease[195]. Once the words have gone missing, once the memories of one's life are no more, one is often left unable to communicate, share life experiences, or express even the most mundane thoughts. However, we believe that even in such apparently desolate and isolated stages, what remains of the voice—even without the words—can carry meaning and allow us to maintain invaluable connections between individuals. While it may seem daunting and awkward for loved ones to carry on conversations with elderly relatives who can no longer understand what is being said, no one should doubt the importance of such vocal connections, which benefit both the person with cognitive decline and the speaker.

[193] Morris Michael Lewis. *Infant speech: A study of the beginnings of language*. Routledge, 2013

[194] Deborah G Kemler Nelson et al. How the prosodic cues in motherese might assist language learning. *Journal of child Language*, 1989

[195] Gill Livingston et al. Dementia prevention, intervention, and care. *The Lancet*, 2017

Such vocal experiences offer a remarkable example of the three paradigms proposed in this dissertation work. First, by recognising the existence and power of the voice beyond words and emotion; second, by believing that for elderly people with late-stage dementia, the voices of others and their engagement with those voices is a determinant part of how they perceive the world around them and whether they consider themselves to belong to it; and third, by considering that behaviors that reinforce social structures (in this case, by generating wellbeing and creating a bridge between the parties) are still the most critical parts of the vocal message.

Although face-to-face interactions might be best, they are sometimes not practically possible. Indeed, visiting relies on some level of independent mobility and is often impossible for less able-bodied persons or for those who live farther away. In such cases, the telephone is an invaluable tool. It can help maintain emotionally intimate relationships or simply create a feeling of togetherness through the voice, even for those with cognitive decline. Participating in frequent telephone conversations has been shown to not only reduce the loneliness perceived by older adults but also to affect perceived health indirectly and significantly[196]. However, there can be limitations in the use of the telephone as a connecting tool. Some mental and physical health disabilities can affect telephone usage, such as simple deterioration in hearing that may discourage telephone communication. For people with cognitive decline, even using a telephone by themselves might become too complicated, requiring them to rely on caretakers to take the initiative and dial the correspondent, which removes agency from the older adult[197]. Finally, those with more profound cognitive loss, in more advanced stages of dementia or Alzheimer's, can lose their understanding of the concept of telephones and even forget their loved ones, leaving little hope that they can actively seek those meaningful vocal connections.

[196] Bronwyn S Fees et al. A model of loneliness in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 1999

[197] Thomas A Ala et al. Using the telephone to call for help and caregiver awareness in alzheimer disease. *Alzheimer Disease & Associated Disorders*, 2005

For these reasons, we believe that it is possible to develop tools for older adults to maintain relationships through active, remote vocal communication that are better adapted to their abilities and needs. In this work, we introduce the concept of Cognitively Sustainable Design as a way to conceive devices and interactions that will maintain their relevance for older adults even as they become less cognitively able.

In this context, we designed the Memory Music Box; a platform to increase connectedness. Countless research projects have been implemented for the support of elders through monitoring, tracking, and memory augmentation. Despite advances in the Information and Communication Technologies field (ICT) aimed toward providing new opportunities for connection, challenges in accessibility increase the gap between elders and their loved ones. We approach this challenge by embedding a familiar form factor with innovative applications while performing design evaluations with our key target group to incorporate multi-iteration learnings. These findings culminated in a novel design that facilitates elders in crossing technology and communication barriers. Based on these findings, we discuss how future inclusive technologies for older adults' need to balance ease of use, subtlety, and Cognitively Sustainable Design.

4.1 - Introduction

4.1.1 - Motivations

Older adults often experience a lack of connectedness which leads to social isolation[198]. In recent years, the HCI community has shown interest in designing interventions for individuals facing symptoms related to aging, dementia, or general memory loss. We hope to contribute to this conversation through the Memory Music Box; revealing how technologies that often alienate older generations can be reimagined to help them connect. The Memory Music Box is an interactive system that the elder controls with the natural gesture of opening and closing the box. The device transforms an intuitive everyday object into a portal for video calling and a personalized memory slideshow, to create a sense of connectedness and reminiscence.

[198] David J. Weeks. A review of loneliness concepts, with particular reference to old age. *International Journal of Geriatric Psychiatry*, 1994; and Graeme Hawthorne. Measuring social isolation in older adults: Development and initial validation of the friendship scale. *Social Indicators Research*, 2006a

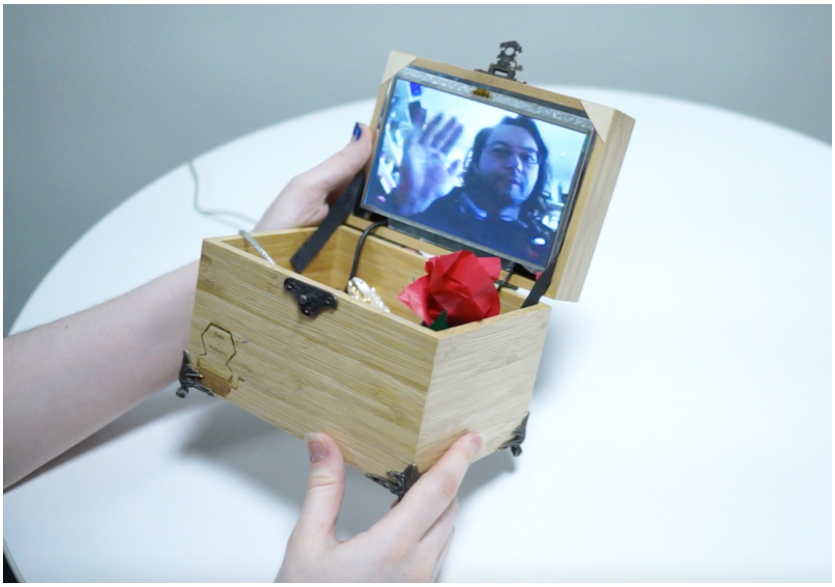


Figure 26: Accessible video call with correspondent from box to establish direct vocal connection

Although more communication-support technologies exist than ever before, most social media platforms and ICT systems are not designed with elder-accessibility in mind, and and present barriers to older adults who are not technologically literate. Even simple platforms pose challenges, as components like the touch-screens used in most interfaces require a certain amount of hand moisture that is often lacking in older skin[199]. Even devices designed for autonomous use by older people usually have a figurative "expiry date" as the platforms become increasingly difficult to use as elders age or develop pathologies. To remain relevant and future proof, we employ a concept we have coined—Cognitively Sustainable Design—consisting in creating something so intuitive and accessible that one can continue using it despite an increase in age, disability, or cognitive impairment.

[199] Tobias Kalisch and other. Cognitive and tactile factors affecting human haptic performance in later life. 2012

Although we as researchers support lifelong learning, we believe that learning should never be a barrier for the simple privilege of connecting with loved ones. The Memory Music Box system is designed to increase social connectedness without technological fluency requirements. If an elder can open a standard music box, they can now connect to their family.

Based on focus groups, prior research, and interviews, we have identified numerous barriers to connectedness in older adults. The two most prevalent are technological accessibility and emotional barriers. Older adults may feel such concern about intruding upon their loved ones' schedules that they may stop attempting to establish contact. Our device addresses both factors through an intuitive design that allows for conversations to take place. Based on early feedback, we selected the form-factor of an old-fashioned music box, a neutral yet familiar object in which many older adults keep valuable and sentimental items such as photographs, lockets, medals, and trinkets from the past. Brereton and colleagues observe the potential of habituated objects to increase adoption by the older population[200]. Our work follows aspects of this approach by using the affordances of a very familiar object to return control to the user, with the goal of increasing their sustained engagement with the device. Within the box is an embedded, interactive, connected system, controlled by the simple, natural gestures of opening and closing the box. Secondly, our design uses serendipity to support connectedness. As opening the box delivers a subtle phone notification to the grandchild-user, elders can feel confident that they aren't being bothersome when they reach out. This makes connecting less of a chore than a whimsical and incidental part of life. This idea relates to Wright's work on the role of aesthetics in experience-centered design[201].

[200] Margot Brereton et al. The Messaging Kettle : Prototyping Connection over a Distance between Adult Children and Older Parents. *Proc. CHI*, 2015

[201] Peter Wright et al. Aesthetics and experience-centered design. *ACM Transactions on Computer-Human Interaction*, 2008

A wide range of individuals can benefit from the Memory Music Box. However, the device was initially designed to help individuals of two specific and often intersecting demographics. The first includes elders who have difficulty using existing technologies to connect with their loved ones. We hope that replacing the need to navigate apps on a phone or tablet with the simple gesture of opening a box will reduce the technological barrier to feeling connected. Secondly, elders experiencing memory loss and cognitive decline could benefit from the aesthetically pleasing, familiar design, unconsciously associating it with a pleasant experience. This would encourage them to serendipitously open the box.

Although grandparents and grandchildren were included in our study, the relationship pairing is mostly in place to provide context in regard to life stage and technological comfort levels than a specific familial relationship. We also believe the Memory Music Box device can be potentially beneficial

for other relationship pairings or for remote pen-pals. To simplify, we call grandparent-users (GpU) the person using the box and grandchild-users (GcU) the correspondent on the other side.

With this project, we hope to shift the conversation of elder wellbeing to include isolation reduction using music, images, and vocal interactions. We present a holistic approach to targeted design in our collaboration with our user groups. Both older adults and the younger generation aided us in shaping new ICT solutions to improve connectedness. Firstly, we present background information on the social problem of elder isolation and technological barriers in systems designed for the elderly, and make a case for the inclusion of music. As noted in the subsequent section, numerous devices exist, but few truly tackle the grandparent-user's need for connection and agency. Of those devices, the ones that do are laden with technological complexities. As one grandparent (and new iPad owner) commented in our focus group, "[My family] said I could see the grandkids on here but to be honest, I mostly use it as a paperweight. I can't figure where it turns on or if it needs a plug!" On this basis, we engaged in the research challenge of innovating solutions that can connect grandparent- and grandchild-users. We present how we involved these key actors in the design of the Memory Music Box via thoughtful affordances and a simple, engaging user experience design. We also present how we gathered feedback and iterated our design based on an online survey and focus group session. Our findings are presented in an analysis synthesizing the results. Finally, we discuss the implications of our findings for further designs, research limitations, and innovations for the future.

4.1.2 - Acknowledgements

Though this project was primarily led by the author, the conception and development of the Memory Music Box system was conducted in collaboration with multiple stakeholders. The project and evaluation were done in collaboration with Alexandra Rieger, Janelle Sands, and Janet Baker. This project also received help from Akito van Troyer, Sara Sime, and George Stefanakis in the preparation of the design of focus group interactions. All of the research aspects of the project mentioned in this chapter are the personal contribution of the author. The project was initially sparked following three meetings with Dr. Maya Geddes, MD, a behavioral neurologist then at the Brigham and Women's Center for Brain/Mind Medicine (CBMM), a clinic specializing in aging. Through observing patients and meeting with clinical staff, the need for accessible interfacing arose, as clinicians noted that issues of connectedness continue to be pervasive in aging communities. Throughout our design phases, we consulted with two specialists working at CBMM, Dr. Geddes and neurologist Dr. Michael

Erkkinen, MD. In parallel, we worked closely with our elderly collaborators as well as younger researchers who struggle with maintaining long-distance connections with their grandparents. The Memory Music Box emerged from these conversations and presents an approach to the challenges outlined in this chapter: attenuating isolation by creating connections through improving technological accessibility and reducing emotional barriers. This project was made possible by the staff and the residents of the Hebrew Senior Life Center in Brookline, Massachusetts, and we wish to thank them for their excitement and collaboration. We also wish to thank the authors' respective grandparents for their inspiration. This research protocol was approved by MIT research ethics board under protocol 1806425177.

More details on this project can be found in the publication that resulted from it, "Supporting Elder Connectedness through Cognitively Sustainable Design Interactions with the Memory Music Box.[202]"

[202] Rebecca Kleinberger, Alexandra Rieger, Janelle Sands, and Janet Baker. Supporting elder connectedness through cognitively sustainable design interactions with the memory music box. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 2019c

4.2 - Context

To situate our work, we consider previous research conducted in the areas of HCI, assistive technology, eHealth, neuroscience, and social psychology.

4.2.1 - Loneliness

Risks of Loneliness

Loneliness and a lack of meaningful connections are now regarded as possible causal factors for a lack of overall cognitive performance and faster cognitive decline, contributing to the onset and acceleration of dementia and memory loss symptoms[203]. A 2006 survey reported that as many as 84 percent of adults over 65 suffer from loneliness in the US[204] Jaremka et al. found higher rates of memory difficulties among breast cancer survivors who experienced high degrees of loneliness[205]. Older individuals with increased feelings of loneliness are also more likely to develop dementia[206]. In contrast, strong and socially meaningful connections were found to be associated with less cognitive decline in older adults[207]. In light of the findings in the aforementioned studies, it is imperative to focus our efforts on designing technologies that support elder connectedness.

[203] Robert S Wilson et al. Loneliness and risk of alzheimer disease. *Archives of general psychiatry*, 2007

[204] William Lauder et al. A comparison of health behaviours in lonely and non-lonely populations. *Psychology, Health & Medicine*, 2006

[205] Lisa M Jaremka et al. Cognitive problems among breast cancer survivors: loneliness enhances risk. *Psycho-Oncology*, 2014

[206] Tjalling Jan Holwerda et al. Feelings of loneliness, but not social isolation, predict dementia onset. *J Neurol Neurosurg Psychiatry*, 2012

[207] Catherine Helmer et al. Marital status and risk of alzheimer's disease a french population-based cohort study. *Neurology*, 1999; and Shari S Bassuk et al. Social disengagement and incident cognitive decline in community-dwelling elderly persons. *Annals of internal medicine*, 1999

The Social Media Gap

Current technologies facilitating connection are ubiquitous and highly used. Social media is among the most prevalent, and young adults participate in numerous networks to stay connected. Although it has changed the way younger generations connect, older generations are often left out. The gap between younger users and elders grows exponentially as most social networking systems are not designed with elders in mind[208]. There

[208] Joseph Wherton et al. Designing technologies for social connection with older people. *Aging and the Digital Life Course*, 2015

are many barriers to using these communication technologies for older populations as they have a higher adoption threshold for new technology. We see the Memory Music Box as a bridge between the two worlds. The "grandchild" portal allows tech-savvy loved ones to easily share images and music from their social media platforms so elders can stay in the loop.

Long Distance Solutions

In *Social Isolation and Loneliness in Old Age: Review and Model Refinement*[209], Wenger et al. define loneliness as the subjective state of negative feelings associated with perceived social isolation, a lower level of contact than desired or the absence of a specific desired companion. In their *Model of Loneliness in Older Adults*[210] Fees et al. showed feelings of loneliness decrease one's evaluation of physical wellbeing and that the frequency of telephone contact affects loneliness more than the frequency of in-person contact with others. This indicates that increasing the frequency of video calling could have beneficial effects on perceived feelings of loneliness.

[209] G Clare Wenger et al. Social isolation and loneliness in old age: review and model refinement. *Ageing & Society*, 1996

[210] Bronwyn S Fees et al. A model of loneliness in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 1999

4.2.2 - Technologies Designed for Elders

One Way vs Two Way Connections

Existing technological devices designed in the context of aging facilitate caregivers in monitoring and tracking patients via off-the-shelf devices such as Keruve[211], an Alzheimer's GPS Tracking Device, SmartSole[212], a GPS unit that can be hidden in a shoe, or PocketFinder, a personal GPS to track your elderly parents on foot or via car. Research ventures suggest the use of monitoring devices as well, especially for individuals in rural areas[213]. While loved ones can receive real-time data regarding an elder's whereabouts, we feel that it is important to design for more than a one-way connection. Although useful for patient safety, these designs fail to answer the needs of older users. In the wake of aging, we focus on improving the quality of life through the quality of experience. With its engaging familiar appearance and technological capability, the Memory Music Box is designed to provide meaningful experiences, a sense of reassuring connections.

[211] Keruve. Keruve alzheimer's gps tracking device, 2008. URL http://www.keruve.com/

[212] SmartSole. Smartsole:hidden, wearable monitoring and recovery solution for wandering. URL http://gpssmartsole.com/gpssmartsole/

[213] Hung-Huan Liu et al. Mobile guiding and tracking services in public transit system for people with mental illness. In *TENCON IEEE*, 2009

Technology and Aging

The other branch of technology designed for elders features tools for mild cognitive decline, such as the COGNOW day navigator[214], a cognitive prosthetic device to help users in remembering facts and people; or the Multi-Agent Personal Memory Assistant system, a distributed memory assistant platform concept[215]. Although momentarily useful, studies reveal[216] that these systems are time-bound—unable to provide relevant support as individuals continue to age. Conversely, the MMB is designed to age with

[214] FJ Meiland et al. Cogknow: development of an ict device to support people with mild dementia. *Journal on Information Technology in Healthcare*, 2007

[215] Ângelo Costa et al. Multi-agent personal memory assistant. *Trends in practical applications of agents and multiagent systems*, 2010

[216] Kellie Morrissey et al. The value of experience-centred design approaches in dementia research contexts. In *CHI 2017*

the elder via an experience-centered approach. We focus less on assistance and more on creating a sustained sense of wellbeing by integrating accessible connectedness into everyday life. For the last few decades, electronic healthcare (eHealth) has focused on designing technology-mediated connections between patients, relatives, and healthcare providers[217]. The challenges addressed by researchers in this domain overlap with ours, as they also have to ensure that their systems are accessible to older users. The difference lies in the objective: while their system focuses on medication management, our system aims to improve connectedness for its own sake, not as a way to monitor medication intake.

[217] World Health Organization. *World report on ageing and health*. World Health Organization, 2015; and Anne Marie Kanstrup et al. Designing connections for hearing rehabilitation. In *DIS*, 2017

Some previous aspirational projects tackle the question of connectedness. For example, the "Family Portrait" project[218] presents an interesting perspective in connecting two households, although the system does not support music or video calls. The "Experiences To Go" interface[219] designed in 2013 appears to have laid some of the groundwork in establishing that video conferencing platforms, although vital to long-distance connections, are not sufficient as freestanding interfaces. While their proposed touchscreens pose a difficulty for older hands, the focus on agency concurrently amplifies our design concepts. The MMB incorporates video aspects while removing the need for touch-activated controls. While myriad platforms have sought to support elder connectedness, studies reveal that one of the most difficult relationships to maintain (especially long-distance) is the grandparent-grandchild relationship. The paper on "Connecting Grandparents and Grandchildren"[220] further provides detailed insights on how grandparent-grandchild dynamics continue to adapt as life expectancy and communication technologies change. This article identifies that older grandchildren encounter a greater amount of communication difficulty with grandparents, a gap that prior technologies do not fully address. Although the rise in social media has fueled global connectedness for younger generations, it has alienated elders. The 2013 study "Enriching In-Person Encounters Through Social Media: A Study on Family Connectedness for the Elderly"[221] approached this via the researchers' Tlatoque system. The interface transfers social media feeds from desktops to a touchscreen digital picture frame device. While elders could see updates through the interface, they could not easily communicate through the platform without prior knowledge or tutorials. Furthermore, using the system required several non-intuitive hand motions to commence operation. Our design builds upon Tlatoque by including "newsfeed"/slideshow elements while removing contact and outreach barriers.

[218] Elizabeth D Mynatt et al. Digital family portraits: supporting peace of mind for extended family members. In *CHI*. ACM, 2001

[219] Kori Inkpen et al. Experiences2go: sharing kids' activities outside the home with remote family members. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013

[220] Karyn Moffatt, , et al. Connecting grandparents and grandchildren. In *Connecting Families*. 2013

[221] Raymundo Cornejo et al. Enriching in-person encounters through social media: A study on family connectedness for the elderly. *International Journal of Human-Computer Studies*, 2013

Sometimes, barriers can occur at the point of conversation, "Hello, is Grandma there?"[222] focuses on important aspects of intergenerational relationships and how stories or added media formats help to bridge the

[222] Hayes Raffle et al. Hello, is grandma there? let's read! storyvisit: family video chat and connected e-books. In *CHI*. ACM, 2011

gap and inspire talking points. This further influences our platform design, which allows for music and image sharing via the portal. While aspects of MMB's low-barrier design are explored above, our work also draws from larger surveys of this design space. "Desiring to be in touch in a changing communications landscape: attitudes of older adults"[223] provides a broad overview of the landscape of connection challenges faced by older adults. The study sought to "draw a distinction between using technology to mediate intimacy and using it simply for the expression of emotion." The focus groups included "a 30-minute discussion, loosely structured around a number of prompts on topics such as the types of communication media that were normally used, triggers for making contact, and whether there were people who the participants would like more, or less, contact with." The study revealed that grandchildren (both children and adolescents) were viewed by elders as "too busy for contact." In light of these findings, researchers organized design sessions alongside their research team which included designers and researchers. The implications which arose from the focus groups provided guidelines for more nuanced and improved connectedness, which our design incorporates. The main guidelines identified in the study include: (1) "the importance of communication being personalized"; (2) "technologies that allow for a more focused, intense means of communication"; and (3) "contact should feel non-intrusive." In regards to personalization, the MMB focuses on this from the outer shell of the box to the inside. Outside, the box is fully personalizable, a canvas awaiting a personal imprint. Inside, the custom slideshow further develops intimacy. We feel the MMB also answers the need for technologies that support "intense, focused communication." Without distracting widgets or complex features, the interaction is as close to an in-person experience as we could achieve, short of including a life-sized hologram. Thirdly, the MMB is nonintrusive. It is careful to respect the privacy of both parties, as the screen and camera are only active if it is open and fully in use. While the Family Window (FW)[224] project eliminates formal barriers to initiating long-distance conversations, the continuity of being on camera could be viewed as intrusive. Our work extends this to the older population by creating a system without FW's extensive and possibly complex control mechanisms, while also reducing traditional outreach barriers in a nonintrusive manner.

[223] Siân E Lindley, , et al. Desiring to be in touch in a changing communications landscape: attitudes of older adults. In *CHI*. ACM, 2009

[224] Tejinder K Judge et al. The family window: the design and evaluation of a domestic media space. In *CHI*. ACM, 2010

4.2.3 - A Case For Inclusion of Music

Memory

Based upon research revealing the benefits of music to older adults, we feel it would be a mistake to exclude it when designing for elders. Music can be used as an anchor point to access personal memories in a beneficial manner and is frequently used in reminiscence therapies[225]. Music engages diverse regions of the brain, and particular memories not only remain preserved, but

[225] Susan Bluck and Linda J Levine. Reminiscence as autobiographical memory: A catalyst for reminiscence theory development. *Ageing & Society*, 1998

can be recalled upon hearing a familiar melody. According to Janata's fMRI study[226], there is a specific "hub" in the brain that links autobiographical memories to specific songs. This hub is located in the medial prefrontal cortex (MPFC), a region slower to atrophy during the progression of regular aging, Alzheimer's disease, or dementia. As people age and encounter later-stage memory loss, music-associated memories in this surviving "hub" are recalled when a particular song is played. As music is like a passcode to the MPFC "memory drive," we felt it was imperative to design our project around it. In doing so, we assessed how music could be used as an anchor point to access personal memories in a beneficial manner.

[226] Petr Janata. The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex*, 2009

Music and Connectedness

The MMB is not only a music box in form, but features a musical playlist, remotely curated by the correspondent, accompanying a photo gallery that elders can hear every time they open the box. According to Ayelet Dassa and Dorit Amir's 2014 study[227], familiar songs increased conversations and connectedness in older individuals. When technologies for elders do incorporate music the experience is often static, such as the project Studio Meineck[228]. Although this project also supports reminiscence, it does not provide users an outlet for increased music-triggered conversation, while the Memory Music Box allows elders to converse with loved ones.

[227] Ayelet Dassa et al. The role of singing familiar songs in encouraging conversation among people with middle to late stage alzheimer's disease. *Journal of music therapy*, 2014

[228] Fares Kayali et al. Elements of play for cognitive, physical and social health in older adults. In *Human Factors in Computing and Informatics*. Springer, 2013

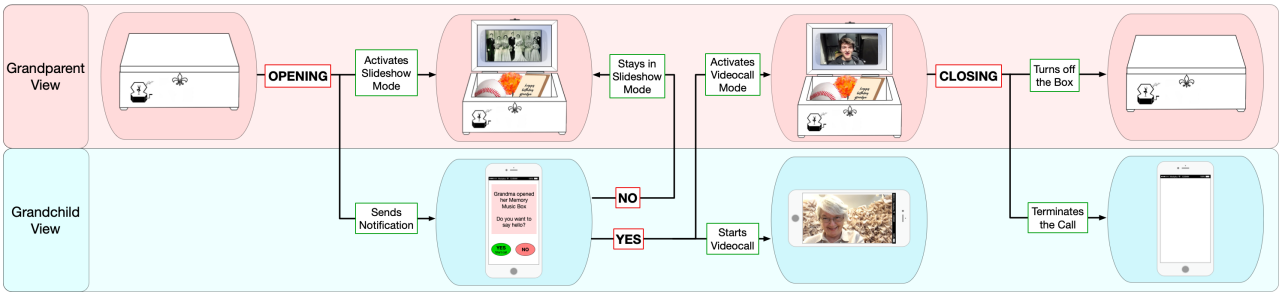


Figure 27: User experience diagram for the Memory Music Box

4.3 - Design

4.3.1 - Interaction Design

Figure 27 represents an overview of the system through a user experience diagram showing in parallel the grandparent-user and grandchild-user views and interactions. The user actions are shown in red boxes and only consist of opening and closing the box for the grandparent-user, and pressing the "yes" or "no" button on a phone for the grandchild-user. The interaction design is detailed in the following subsections.

Grandparent-User

In focus groups, elders appreciated that the only user input needed is the opening of the box. Upon opening, the box automatically enters a mode called "slideshow" where a carefully orchestrated series of photos and music will be displayed. The content of the slideshow is curated remotely by the grandchild-user through an online interface.

(a) Slideshow interaction (b) Videocall interaction

Figure 28: Grandparent-user interactions

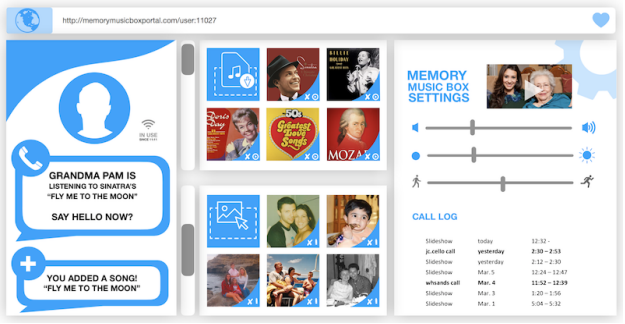
While entering the slideshow mode (Figure 28.a), the box also sends a discreet phone or email notification to the grandchild-user to inform them that the box is open. If the grandchild is available for a video call, they simply touch the "say hello" icon for the box to switch to call-mode. The grandparent is then automatically connected to the grandchild via video-call (Figure 28.b). The embedded speaker, microphone, and camera in the box allow for a seamless video call. The call is stopped when the grandparent-user closes the box or the grandchild-user ends the call. If the grandchild-correspondent is not connected or not available for a call, the box remains in slideshow mode where the familiar music plays and updated photos of family and friends are displayed on the screen. The grandchild can then make a call at a later time using a regular phone. This interaction is designed such that every opening of the box leads to a feeling of connectedness and comfort, whether the correspondent is available or not. In addition, this design also cover the unfortunate situation of grandparent-user calling repeatedly, or calling back after only a few minutes, having forgotten the previous interaction. In this case, the grandchild-user can make the choice to not answer, knowing that the box will still provide a sense of connectedness.

In addition to its technological interface, the Memory Music Box offers a space to hold important items for the user. From childhood mementos to important written reminders, the box can create a portal of reconnection through real objects, allowing the user to explore its contents in both a tangible and digital way.

Grandchild-User

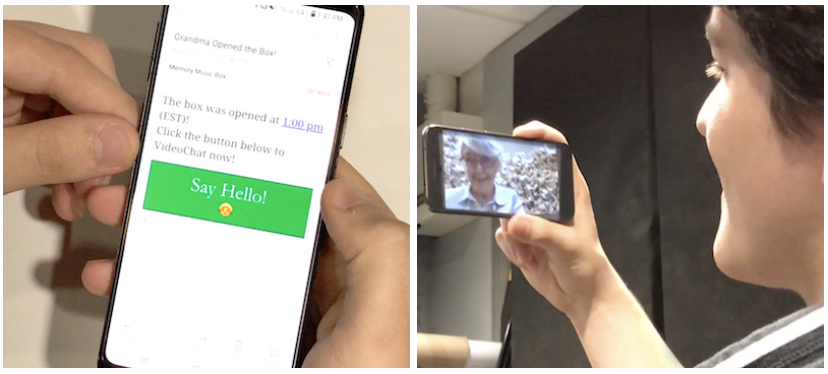
The interaction requires the grandchildren to be relatively tech-savvy, as they have to be able to use an email client and be familiar with photo-sharing platforms such as Facebook or Instagram. This covers a high proportion of young adults but might exclude young children and young adults with disabilities. Future iterations could explore this further. The grandchild-user has two ways to interact with the device. Asynchronously, they are in charge of curating the content of the slideshow using a password-protected online platform (Figure 29). Their personal account on the platform is tied to the unique ID of the Android system within their grandparent's box for privacy concerns. The correspondent can edit music and images at any time, and changes are remotely transferred to reflect in the box. The online correspondent interface also allows them to control the Wi-Fi connection, battery life, and volume of the device.

Figure 29: Grandchild-user video call interactions



The second way for the grandchild-user to interact with the box is through video call. We designed the box interaction to ensure that calls can only be initiated by the grandparent-user, to provide them with agency and control over the box interaction.

Figure 30: Grandchild-user video call interactions



(a) Notification reception (b) Videocall interaction

Each time the box is opened, the grandchild-user receives a subtle email or text notification on their phone informing them that the box is open

(Figure 30.a). They can decide to ignore the notification if they are unable to answer, or they can press a button that automatically connects them to the grandparent through the calling feature (Figure 30.b).

4.3.2 - System Design

First Prototype

We developed an early-stage prototype (diagram Figure 31) using an embedded Linux computer (Raspberry Pi), in a box specifically built to fit the electronics. The Raspberry Pi was connected to a USB microphone, a 5" touchscreen, a Pi camera, and a mini speaker connected through an 1/8-inch cable. The speaker was attached to the bottom part of the box to use the resonance of the container and increase the overall volume. All the hardware was hidden behind a one-sided mirror frame. A pressure sensor was used to detect if the box was opened. This original system used Python to sense the interaction and launch the two modes. The video call mode was implemented using the Google Hangout API and the slideshow mode was made using Pyglet and Pygame libraries.

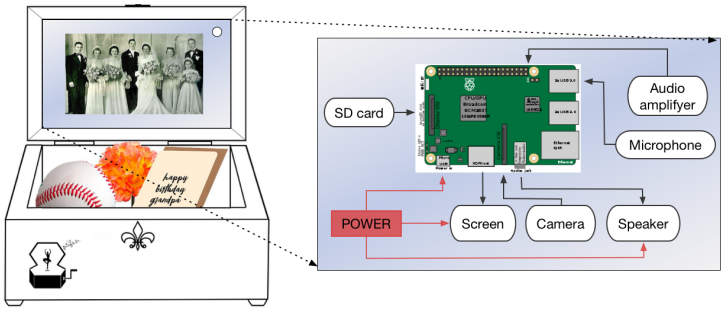


Figure 31: First prototype system diagram

The first prototype featured selected design elements and aesthetic detailing ensuring that the box could be recognized as a piece of craftsmanship. The designs give the impression that the box was hand-made especially for the user. This perpetuates an atmosphere of appreciation for the object. We used bamboo for the box and antique brass for the decorated feet, hinges, and latch. The top of the box is laser etched with a design personalized to the taste of the recipient. This design was used as a tangible object to discuss and review user interactions with a team of designers and physicians. The prototype offered good flexibility in terms of development and choice of hardware. The modularity allowed us to go through a series of implementations to quickly iterate the interface and experience design based upon feedback.

Second Prototype

Once we selected specific interaction components, we decided to implement the system on an Android-based platform. Indeed, using an inexpensive Android device and turning its processor into a Raspberry-like board gave us the same flexibility, comparable performance, memory, storage and all the I/O necessary for a fraction of the cost. The activation-trigger pressure sensing function was replicated by using the front camera as a light sensor to detect when the box is in use. Our current prototype is entirely embedded within an Android device, placed horizontally behind a mirrored frame in the lid of a hand-crafted, old-style music box (diagram Figure 32). The device is mounted to ensure that the front-facing camera and microphone are in an optimal position. As all “buttons” are inaccessible behind the frame, unintended interactions cannot occur. The front-facing camera, speakers, microphone, and screen are employed during slideshow and call modes, while position sensors allow the box to respond to being opened or closed. More compact, elegant, and tamper-proof, this design direction renders our system more easily deployable for the future as an affordable kit. We envision that this "off-the-shelf" kit could simply contain an empty, neutral box (to be decorated in DIY-fashion), the mounting frame, and a personal PIN and URL to download the application to any available phone.

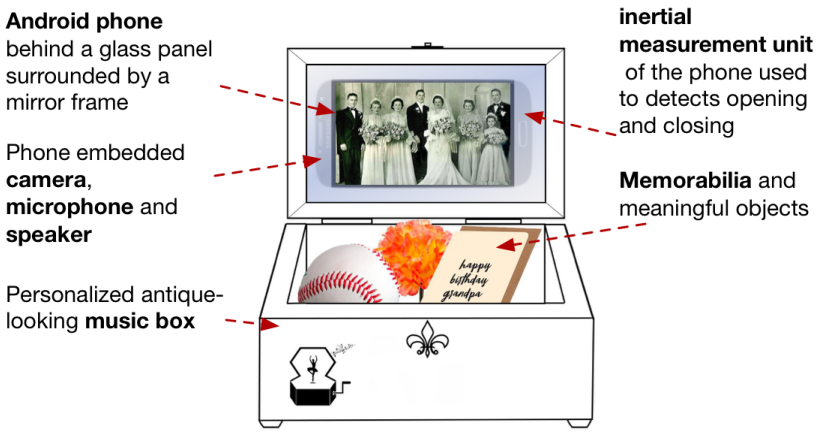


Figure 32: Second prototype system diagram

The software was built using Java and relies on the Skype for Business API for call services. It seamlessly transitions between an active mode when the box is opened and a sleeping mode when the box is closed, while the system periodically reads light sensors from the front-facing camera to determine whether the box has reopened, it portrays a dark screen with no audio. Upon opening the box, the application immediately joins a unique video call in the background, but the camera and microphone remain off while the user enjoys a fullscreen slideshow of photos, accompanied by music. The application automatically transitions between slideshow and

call modes depending on whether the other person is present in the unique video meeting. The video call is self-contained within the application and is protected by all the privacy and security protocols supported within the Skype platform.

This design and set-up prioritize simplicity to cater to grandparent-users at different levels of physical and cognitive abilities. Considering grandchild-users' tech-savviness and attempting to maximize connectivity, this prototype features a grandchild-user notification system in which grandchild-users are alerted when the box opens with an email to their personal account. They can follow an active video call link within that email to join the call on the "Skype for Business" application on their own phone or computer. With its scalability for multiple boxes, its simplicity and reliability for grandparent-users when encountering weak internet connections, this prototype offers promising improvements in deployability and connectivity. A future version could allow to expend the connection to more than one grandchild-corespondent.

4.4 - Evaluation

Our two target groups were grandparent-users (GpU) and grandchild-users (GcU). While GpU focus groups were conducted in person, GcU surveys were completed through a secure online survey. Both focus groups facilitated our participatory approach to the design refinement of the Memory Music Box. Throughout the conception and development processes, we iterated our design choices to maximize the adoption potential of the device for older adults. The next section presents a formal assessment of our design in comparison to existing frameworks through the *lens of elder technology* adoption.

4.4.1 - Technology Adoption Assessment for Older Adults

In order to evaluate the Memory Music Box system, we implemented Lee's adoption factors as presented in[229]. The ten identified adoption factors that predict older adults' adoption of technology are: experience, usability, emotion, social support, value, confidence, affordability, accessibility, technical support, and independence. These factors are also aligned with the work of Suh et al., featuring six constructs that predict user burden with computer systems[230]. The presented categories are revealed to stymie initial adoption of technological systems (1 - difficulty of use, 2 - physical, 3 - time and social, 4 - mental and emotional, 5 - privacy, and 6 - financial). For each of the ten adoption factors, we reflected on our design choices to understand the adoption potential for older adults.

[229] Chaiwoo Lee et al. Perspective: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. *Journal of Product Innovation Management*, 2015

[230] Hyewon Suh et al. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *CHI*. ACM, 2016

- **EMOTION:** This factor seems the most clearly broached, as it relates to our concept of Cognitively Sustainable Design. By creating a device to encourage emotional attachment, the interactive system can remain relevant for the user even as cognitive and physical abilities decline. Users have a higher probability of projecting emotional qualities onto our system as it is personalized[231], although this is not guaranteed. Some might argue that a stuffed toy or robotic form factor could also reach this goal.
- **USABILITY:** The user interaction is kept very simple without superfluous features as to not overwhelm the user, even if they have physical and cognitive limitations or low technology familiarity. If the box's function is not clearly explained or understood, an elder might expect their grandchild to answer their calls continually, ignoring the slideshow. This could cause disappointment or confusion. We have explored addressing this in future iterations by adding a short reminder message before the slideshow plays.
- **VALUE:** The Box is designed as a personalizable artifact to be gifted by the GcU. Thus it is likely to carry a higher affective value in addition to providing lifestyle benefits. The outside appearance can be embellished through collaborative crafting sessions with grandparents and correspondents, further increasing its value.
- **CONFIDENCE:** Unlike many current devices, there is no need to fear unknowingly pushing a wrong button as there are no buttons on our system. One could even introduce the box to a user by saying that there are no right or wrong ways to use it. In addition, the device is designed to empower the GpU as the one having the agency to initiate calls.
- **EXPERIENCE:** When trying to interact with a laptop, it is not the opening of the laptop that stops elderly adults but what happens afterward. While the system inside the MMB might feel foreign to the GpU, its outside appearance invites interaction and functions analogously to other familiar systems, like a regular jewelry box. No further knowledge is required for the user to benefit from it, and once the box is opened is runs entirely by itself.
- **SOCIAL SUPPORT:** Contrary to the very individual experience of talking on the phone, the box can also be shared with others physically present around the user. Participants expressed interest in sharing their memory boxes with friends with the goal of learning more about one another. However, the screen might be too small for several people to effectively interact with the object; it is primarily designed as an individual experience.
- **AFFORDABILITY:** The design choices of the prototype allowed us to reduce costs as it only requires three key elements: a free downloadable application, a low-cost Android phone and an empty box with the right inset dimensions. If the device were to be deployed, we would aim for

[231] Ruth Mugge, Jan PL Schoormans, and Hendrik NJ Schifferstein. Emotional bonding with personalised products. *Journal of Engineering Design*, 20(5):467–476, 2009

an open-source platform with a tutorial for the grandchild to build the box and download the software platform.

- ACCESSIBILITY: The device is designed to be gifted to the GpU to reduce acquisition barriers. There is no need for the user to have prior knowledge of the existence of the device. We hope to make the designs publicly available, along with a simple tutorial targeted to GcU population. Using open-source code, an old phone, and an existing box, they can easily create the experience. Accessibility could become an issue in case of severe physical limitation if the user is unable to use their hands.
- TECHNICAL SUPPORT: In the current use scenario, technical support would be assured by the grandchild user. From the online platform, GcUs have access to information such as Wi-Fi information, audio volume, luminosity, battery level, and the time and length of each interaction. This allows GcUs to not only identify possible technical problems but also to be proactive when noticing that the box is not being used. This system caters to any relatively tech-savvy grandchild users who are comfortable with computers and internet use.
- INDEPENDENCE: Apart from keeping the battery of the box sufficiently charged, grandparent users can tend to the use and care of the box independently as long as there are no physical limitations impeding dexterity. In designing the box, we aimed to increase the feeling of agency and engagement. Indeed, contrary to existing digital frames that continuously display images, our device requires an active control action from the user. This design choice was guided by the decision to keep the user's agency central to the experience.

To summarise, the affordances of the Memory Music Box target most of the factors that are shown to decrease adoption barriers. With our iterated design based on the ten adoption factors, we led a focus group and conducted a survey to gather feedback from our two target populations (GpU and GcU).

4.4.2 - Grandparent-user - Focus group

Focus Group Methodology

To learn about target user-group applications of the Memory Music Box, we collaborated with the Hebrew Senior Life Center in Brookline, Massachusetts, an independent senior living facility. The format of a focus group was chosen because, before deploying, we wanted to get feedback about potential users' first impressions and finalize the prototype in co-designing with our target population. Elders at the center were told in advance about the workshop-format focus group and registered to attend. The co-ed group included ten individuals recruited through flyers and advertisements by

the center. Our participant group included eight women and two men, all between the ages of 70 and 95. The participants were divided into three groups of three to four participants, and the group discussions were each guided by two researchers. In each group, participants had the opportunity to interact with the Memory Music Box during the session and were invited to discuss three topics of interaction—connection, memories, and usability—for a total of an hour and a half, during which participants were presented with a selection of drinks and refreshments. We opened with a musical prelude and welcome introduction before moving into the main part of the session. The focus group yielded positive feedback from participants following the user-interaction experiences, as well as general insights and suggestions to increase accessibility.

While the current study has been conducted with healthy older adult populations per our initial IRB clause, future iterations will be sparked by a broader array of individuals with diverse abilities with the goal of co-designing increasingly accessible MMB's. In the future, we plan on involving diverse populations in testing the device, including children with disabilities and older adults with early-stage Alzheimer's disease.

The workshop had 10 participants and featured four main parts: introduction, questionnaire, interaction, and assessment. As participants arrived they were greeted and presented with consent forms One helper from the center and four researchers were present to answer any questions. At the start of the session, elders were shown a video of the Memory Music Box's functions and features. The narrated video revealed a grandparent and grandchild using the box to communicate in a narrative everyday setting. Following the introduction of the concept, researchers shared their enthusiasm for learning from participant interactions and experiences with the Memory Music Box, then participants were asked to separate into two smaller groups. The focus group segment of the workshop was arranged with a head researcher and an assisting undergraduate researcher present at each table. Researchers led the focus groups with interactive Memory Music Box prototypes and the Memory Music Box questionnaire. The questionnaire was developed based upon accessible design research[232], assessments charts of devices for elderly individuals[233], and previous work on assessing connectedness[234].

Focus Group Questionnaire

The questionnaire featured two distinct portions: pre-experience and post-experience. Before interacting with the Memory Music Box, we sought to learn about the participants' current state of connectedness. To assess this, the first portion of the questionnaire posed detailed questions about

[232] Anne Marie Piper et al. Exploring the accessibility and appeal of surface computing for older adult health care support. In *CHI*. ACM, 2010

[233] Marcela D Rodríguez et al. Home-based communication system for older adults and their remote family. *Computers in Human Behavior*, 2009; and Chaiwoo Lee et al. Perspective: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. *Journal of Product Innovation Management*, 2015

[234] Daniel T van Bel et al. Social connectedness: concept and measurement. *Intelligent Environments*, 2009; and Graeme Hawthorne. Measuring social isolation in older adults: development and initial validation of the friendship scale. *Social Indicators Research*, 2006b

current correspondences with grandchildren/loved ones and overall satisfaction rates. Although one listener indicated that they did not have friends/relatives outside the center, all focus group participants stated that they had specific friends or relatives with whom they enjoyed connecting. Although all participants had some form of contact with their correspondents, 80 percent of the participants expressed dissatisfaction with the frequency of connection. None of the participants possessed the technological experience necessary to video-conference with their loved ones via Skype or any alternative platforms and therefore were unable to see their relatives in real-time with the exception of participant number 6. This participant's family installed Facebook live notifications on her phone so she could see her loved ones at the push of a button, but could not interact, nor could they see her. To learn more about the rate of dissatisfaction and lack of connection, we posed the question "What makes it difficult to connect with your friends/family?" The responses revealed outstanding concerns about the busyness of their correspondents (as participants hesitated to reach out in fear of interrupting full schedules) and technological challenges. All the participants stated that current technologies did not support their connectedness with their correspondents, with one participant specifying that although some technologies could be helpful, there was a lot of room for improvement. The post-experience sessions included questions about the subject's overall impression of the device, projected frequency of use and expected reception by their grandchild-correspondent. During the sessions, researchers at their respective tables asked each participant the intended questions and took note of the answers on prepared charts. Charts included rating scale systems so participants' responses could be synthesized into quantifiable information alongside qualitative quotations. Following the sessions, results were analyzed to ascertain general views of potential grandparent users.

Focus Group Interaction

Following the first part of the questionnaire, participants interacted with prototypes of the Memory Music Box. Participants were able to pass the box around, video-conference with another researcher, and experience a slideshow of music and images. Each participant took the time to explore the box's features and demo the experience firsthand. Following the interactive session, researchers asked participants about their thoughts. Some of the responses included: "I think this is wonderful," "This would be great to have," "It reminds me of Skype but much more accessible," as well as accessibility questions such as: "Wonderful, I'd love to have one to share with others. Could the screen size be bigger/brighter and the volume higher?" or "I love the easy-to-handle system. Maybe it could be lighter to open for arthritic hands."

Focus Group Design Feedback

The focus group not only helped to confirm many components of our design but also supported specific changes in the latest prototype and online interface. For instance, participants mentioned that they would prefer the system to be preset on the highest possible microphone and speaker volume. The positioning angle of the lid was satisfactory for most participants but not all. Further data analysis will help to inform whether creating a click-based adjustable angle positioning system could improve usability. One participant mentioned that adding a large handle on the lid could help low-dexterity elders to better enjoy the interaction. One focus group participant said that she "would love this on the nightstand." In a future version, we would like to include an automatic blue tone reduction on the display if the system detects that the box is being opened at night to prevent circadian rhythm disruption. Congruently, the box could automatically select more relaxing music from the playlist.

Focus Group Findings

A notable finding arose from the focus group: elders saw immense potential for two distinct kinds of connectedness leveraged through the Memory Music Box; 1) Interfamilial Connectedness and 2) Interpersonal Connectedness. Interfamilial Connectedness was currently unsatisfactory for 80 percent of the participants, due mostly to respect for loved ones' busy schedules and a lack of accessible technologies. Participants expressed that they felt their connections would improve with the box, as it is an accessible device that sends subtle and non-intrusive alerts to their point-of-contact, allowing outreach to be unobtrusive and leading to frequent connections. The second form, Interpersonal Connectedness, relates to sharing one's personally curated Memory Music Box slideshow with peers and friends inside the senior living center. Reminiscence Therapy[235] research reveals that both music and pictures can support autobiographical memories, making it easier to recall and share life events in conversation. Many of the participants noted that they would enjoy sharing their memory boxes to learn more about one another, deepening conversations to include life experiences and current family updates.

[235] Kai-Jo Chiang et al. The effects of reminiscence therapy on psychological well-being, depression, and loneliness among the institutionalized aged. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 2010

Feedback for Future Work

Although our novel design has many beneficial aspects, our team looks forward to further developing the MMB. The design thus far has been hewn through a variety of stakeholder input, constructive feedback, and evaluations at myriad intervals in the prototyping process. In awareness of the efforts that have allowed us to develop the MMB to this extent, we also acknowledge the potential for growth and improvement. Overall,

we received overwhelmingly positive feedback from the focus group. In parallel, we received a few negative but highly relevant comments from participants highlighted within this section. When the focus group event began, a few potential participants disclosed that they did not wish to join the study. When asked why the individuals commented that they were either "not interested in technology" or lacked technological fluency. None of these individuals saw the MMB but seemed to shy away from the mere terminology before the chance of interaction. There were no tests conducted where an elder simply opened the MMB without knowing what would be inside. Instead, all participants were primed and aware of the functionality prior to interaction. Perhaps an emphasis on the music box element could spark a more open-minded and exploratory user-feedback session. This inspires us to consider reformatting future study experiences so that the word “technology” will not be an additional barrier. Further comments are as follow:

"This would be great for people with accessibility challenges. If someone has only coarse motor skills left, could they even open the box?" This concern was raised by a focus group participant who was a retired accessible designer. She noted that although she felt the form-factor would be accessible for numerous conditions, there are some which would require the box to have a more easily graspable handle. Fortunately, the box is fully customizable and "grandchildren' users or local caregivers could help to affix a handle, tether or similar assistive modality as needed. In spite of this, we are considering future user testing to identify a system of adaptive modular grips that can be snapped into place.

"What if the "grandchild" person I want to talk to is also technologically illiterate?" Although our current design aims to address the generation gap, future iterations could support box-to-box contact features. Initial plans for this system include similar low-threshold "plug-and-play" usage alongside non-intrusive notifications via a gentle "glow" when a MMB is opened.

"Could the screen size be bigger/brighter and the volume higher?" While the grandchild is able to adjust brightness and volume settings, screen size is limited by the box itself. In some cases, a larger screen size could be beneficial although it would increase the size of the MMB, which is currently no bigger than a traditional music box. However, having learned of this point of interest, we anticipate exploring sizing options in future prototype testing.

"I frequently see friends and family in person, they all live nearby!", one of the participants told us. For this participant, perhaps the box is only a gateway to exploring technology, as it is not needed for connectedness. As creators of the MMB, we acknowledge that there will be cases where

the MMB may not be required; we were glad to hear of this participant's frequent interactions with loved ones, while simultaneously acknowledging that many elders are separated from their relatives. It is our hope that the Memory Music Box can act as a bridge: bringing connections and good memories into places of isolation.

4.4.3 - Grandchild-User - Online Survey

To gather feedback from potential GcUs, we conducted an online survey regarding current interactions with the grandparent; general feedback on the device; and feedback about the online interface. For this portion of the survey, we opted for an online survey rather than an in-person study, due to time constraints. We feel this decision is justified as our device specifically targets young, busy individuals for the "grandchild" category. The online component ensured that they could be included. Thirty-one adults between 18 and 54 years of age, recruited through a student mailing list, participated in the survey.

Current Interactions with Grandparents

The survey revealed that participants think about their grandparents far more often than they interact with them (see Figure 33). Eight-seven percent of participants stated that they only visited their grandparents a few times a year or less. Seventy-four percent of participants stated that they only called their grandparent on the phone a few times a year or less; this number was 87 percent for video calls. When asked who initiates the calls, 45 percent responded that they always initiate phone calls, 31 percent answered that it is 50/50, and 60 percent of GcUs responded that they always initiate video calls.

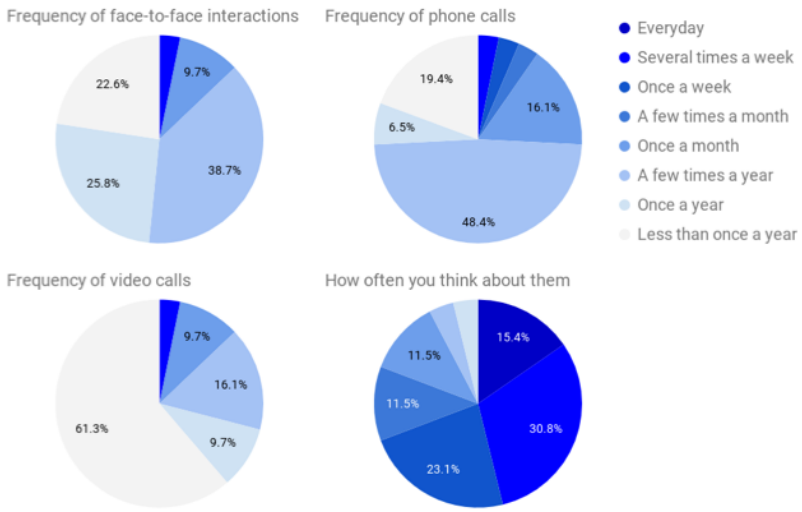


Figure 33: Frequencies of current interactions

Across all answers, communication tends to last from five minutes to an hour, and the content of the conversations generally stays at a superficial level (52 percent) despite being very caring (52 percent) (see Figure 34). Most people revealed not knowing what music their grandparent enjoys (74.19 percent) and even more have the impression that their grandparent does not know about their musical taste (96.77 percent). Only about a third of participants reported that their grandparent often discusses their past interests and memories (29.03 percent) and only a quarter of grandchildren reported frequent sharing of their own interests and hobbies (22.58 percent).

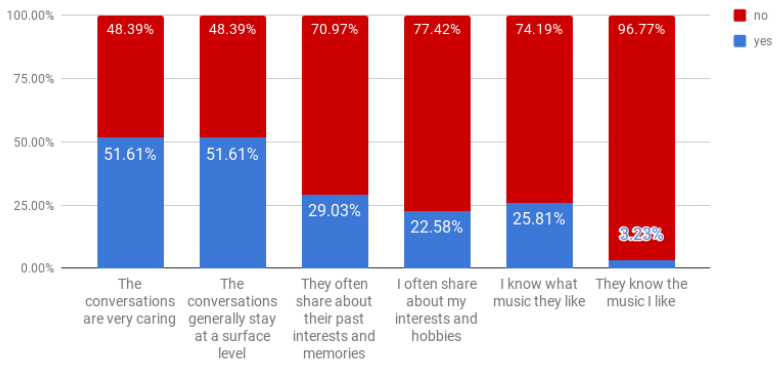


Figure 34: Content of current interactions with grandparent

When asked if any factors limit their grandparent's ability to connect, 40 percent answered that memory loss is an obstacle for connection, 36 percent mentioned physical difficulties with using a device, 24 percent reported that their grandparents did not own a phone or computer, and 80 percent stated that their grandparents experienced general difficulties with using technologies (see Figure 35).

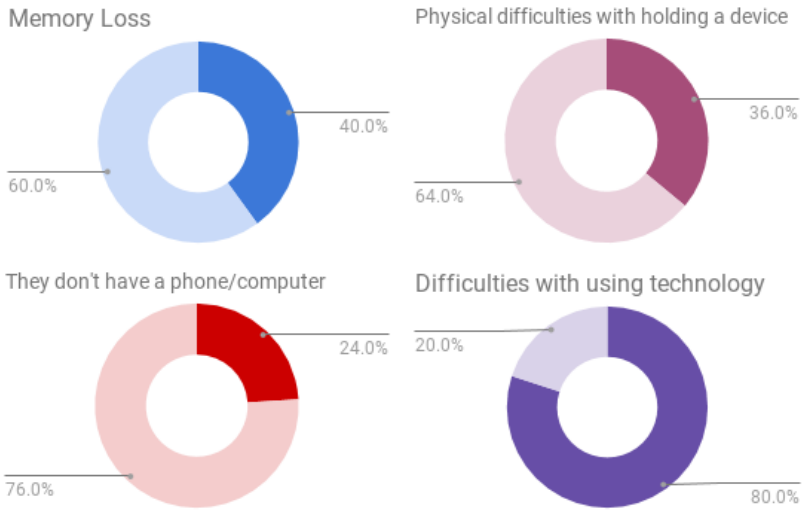


Figure 35: Factors limiting the connection, lighter color means "no" and darker color means "yes"

Globally, only 32.3 percent of grandchild participants reported feeling happy with current interactions with their grandparents and only 38.7 percent of grandchild participants felt that their grandparents were satisfied with their current interactions.

General feedback on the device

After asking about current information, the participants were introduced to the Memory Music Box device through an introductory text and a short video, then asked to provide feedback (see Figure 36).

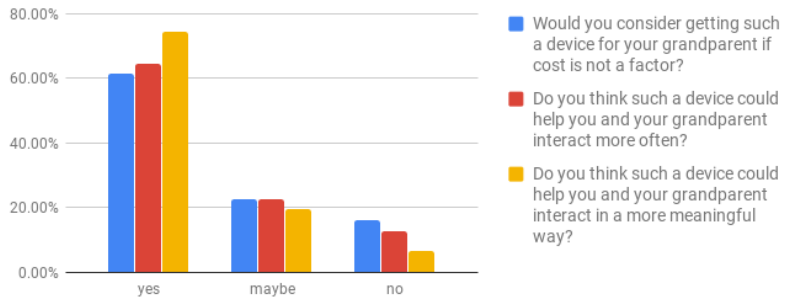


Figure 36: General thoughts on the Memory Music Box device

Most found the device interesting and desirable for their own interactions with their grandparents and we gathered feedback such as "It seems great!! and really easy for them to use," "Very cool!", "looks really awesome!", "I love it," "easy to use form-factor for older generations, auto-connected which is convenient," "it is very cute and it could make a lot of people very happy," "It would definitely be helpful for facilitating communication between me and my grandparents," or "My grandpa would love something like this!" While many believed that the device would improve the frequency of contact (See Figure 14), a small percentage questioned whether the box format would be beneficial. One respondent stated that they "think it is interesting. It requires a good internet connection. If it doesn't work a few times my grandmother might get frustrated." The design of the box proved not to be ideal for everyone, with participants saying "Seems ok. Does not feel that different from a tablet." The conceptual idea of incorporating reminiscence therapy into technology catering to the elderly in order to spark memory gain was quite popular, as 74 percent agreed doing so could produce more meaningful conversations.

Feedback on the Grandchild Interface

When asking participants how frequently they would update the media content on the box, 3.23 percent said they would do so every day, 9.68

percent said several times a week, 45.16 percent said "several times a month," 19.35 percent said "about once a month," and 23 percent answered "a few times a year" (see Figure 37).

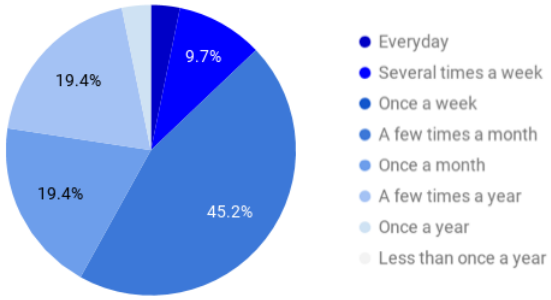


Figure 37: Projected frequency of content updates

We also asked participants how user-friendly they think the box would be for their grandparent. Most participants thought the ergonomics of the device would be user-friendly for their grandparent, as well as thinking the online interface would be user-friendly for themselves (see Figure 38).

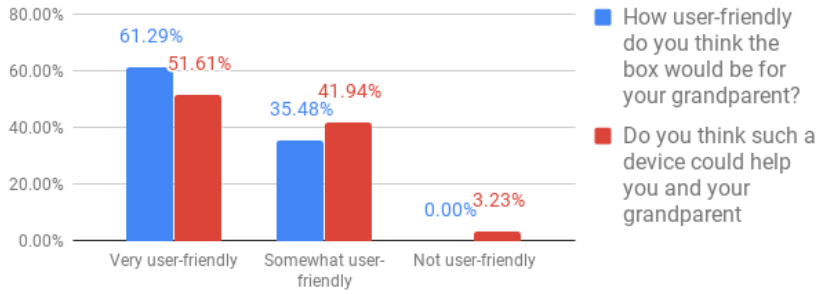


Figure 38: On user-friendliness

Finally, we asked grandchild participants how often they think their grandparent might open the box if they had such a device. About half of all participants thought that their grandparent would open the box every day and 87 percent think that their grandparent would open the box at least once a week.

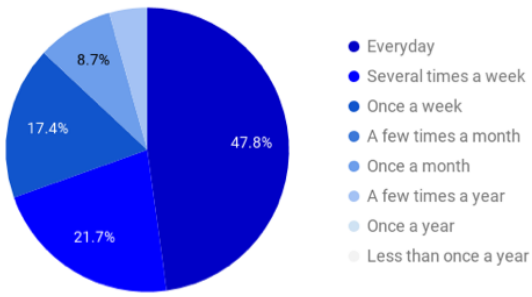


Figure 39: Projected frequency of grandparent interactions

4.4.4 - Insights

Several grandchild participants worried that such a device could create a heavy commitment and they would feel guilty not answering notifications or forgetting to update photos and music often enough. Indeed, we received comments such as "I would feel guilty not picking up if I were busy, knowing they were looking at pictures of me," or "It's a very sweet interaction, but I'm afraid if the grandchild doesn't answer, the grandparent will be sad. But maybe it's ok because if they see new photos or hear new music, they would know that I thought about them." However, in regard to feedback obtained from the focus group with older adults, grandparents are generally highly aware of their grandchildren's busy schedules, and tend to avoid initiating phone or video calls because of it. Moreover, grandparents from the focus group expressed that they were happy to spend a lot of time looking at a limited set of photographs. Even minor improvement on this end, such as an extra photograph or two a year, would be welcomed.

Another major insight from the survey is how dissatisfied grandchildren are with their current interactions with grandparents. Only a third indicated satisfaction with their current interactions. In addition, we observed an important disparity between how often grandchildren think about their grandparents and how rarely they actually call/video call or have face-to-face interactions with them. We compared those frequencies with the reported projected frequencies at which grandchildren would update the photos and music on the box. More than three quarters of participants reported that they would update media content at least once a month (77 percent), which is three times as many as reported calling their grandparent at least once a month (25 percent), and about six times as many as reported video-calling (13 percent) or having face-to-face interactions (13 percent) with their grandparent at least once a month. This insight reinforces the idea that cognitive sustainable devices, such as the Memory Music Box, might help bridge the gap between how seldom grandchildren interact with their grandparents and how often they think about one another.

When looking at the content of current interactions, more than half of the grandchild participants indicated that discussions often remained at a superficial level, and some seemingly important details about each other's lives, such as musical tastes, are not often discussed. As few as 4 percent of grandparents seem to know what music their grandchild listens to and only about a quarter of grandchildren know their grandparent's musical taste. More generally, only 30 percent of grandparents share their pasts, interests, and memories, while 23 percent of grandchildren share their interests and hobbies. Both the focus group and the online study revealed a real desire for ways to trigger more meaningful and personal interactions.

One focus group participant explained, "For music, there is a generational divide. I mainly listen to classical music and country but they listen to very different things that I don't relate to. They call it 'Phish' I think. I would be interested in listening to what they like, maybe we could understand more of each other if they could share it, with the box it could be fantastic." Similarly, grandchildren seemed hopeful that well-curated triggers could improve not only the quantity but also the quality of interactions with their grandparents, as 74 percent said they believe such a device could help them and their grandparents interact in a more meaningful way. Here, music acts as a language that can unexpectedly express unknown identities and further explain ones that are known, which in turn deepens social connections. In future work, these factors need to be tested longitudinally for a longer period of time. Another interesting next step would be to run an evaluation with people with memory loss and dementia and see how our thinking plays out in terms of sustained use.

Some study participants expressed concerns regarding their grandparents using the box: "Love it, but I hope they can actually use it without problems." Others were worried that lack of infrastructure on their grandparents' end would affect the usability: "it is a nice concept and would make it easier for my grandparents to call me. However, I would need to set up wifi for them at their home (they don't have wifi)"; "Very interesting concept. I would like to know how robust it is in different circumstances (i.e., have you considered an alternative to a no-WIFI environment where Skype calls are not feasible)." Based on these results, we feel the creation of a non-Wi-Fi iteration is an important avenue for future work. While most participants felt the Memory Music Box would be an ideal device for their grandparents, a few participants noted that their grandparents were already technologically fluent and wouldn't require the ease of access provided by the MMB: "It's cool. My grandparents are very competent with technology (they are on Facebook, know how to use a smartphone, etc). They have some pride surrounding the issue. It's possible that they might consider it slightly condescending." Both concerns noted in the feedback (no Wi-Fi and technology proficiency) reveal extreme ends of their respective spectrums, shining a light on the specificity of the population that might benefit from the box.

A major objective in our qualitative study was to examine how grandparents experience the empowering potential of being an actor in personalized ICT solutions. While elders have responded with overwhelmingly positive feedback, we remain aware that our design could seem inadvertently limiting or patronizing if an extremely tech-fluent grandparent desires to have further control of the system. In spite of this, the Memory Music Box can still be a pleasant artifact—far more interactive than a digital picture frame

(which even tech-savvy individuals enjoy) and furthermore could help to increase confidence with more complex technologies. However, at its core, our device is meant for users who feel helpless with existing ICT options and are seeking easier alternatives, while keeping a level of independence in the process.

4.5 - Contributions of the Memory Music Box project

When approaching the challenge of creating a design for the aging population, we acknowledge the intersectional individualities that accompany this age group. Unlike younger target populations, elders frequently experience rapid changes in their physical and cognitive capabilities as well as their living situations. For example, an elder aging in place with their family may be moved to a skilled nursing facility where healthcare takes precedence over connectedness. Although creating a one-size-fits-all design proves unlikely in these situations, our development of the terminology Cognitively Sustainable Design presents an answer. In spite of dramatic life changes, we envision the Memory Music Box will remain a familiar object in an elder's domestic space. While further longitudinal evaluations with the MMB will enable us to hone design elements, we wish to bring the novel concept of Cognitively Sustainable Design to the HCI conversation.

We envision HCI as a space through which to address gaps in opportunity for older generations. Moving beyond monitoring, the Music Memory Box focuses exclusively on empowering agency and facilitating meaningful connections. The design is uniquely poised to appeal to a diverse user base due to low threshold usability features as well as a design that encourages continued use through later years. Our novel approach ensures that accessibility is forged through intuitive design, implementing technology within a familiar platform already common within the domestic space. The Memory Music Box presents a platform for serendipitous communication between older and younger generations while also mediating curated images and music, encouraging reminiscence and connectedness.

More broadly, our work fits in the rhetoric around target population empowerment in HCI[236]. We believe that our findings are a call to a broader discussion about the use of HCI to empower elders who are not familiar with technology, especially through the lens of facilitating connectedness. On one hand, any solution that does not tackle this problem frontally could add more layers of complexity between the two individuals trying to connect, which could obscure the root problem by simply addressing symptoms. Indeed, many technologies designed "for" older adults are aimed more at providing reassurance to caregivers. These interventions often omit the perspective of the older adult and fail to probe the long-term impact

[236] Amanda Lazar et al. Designing for the third hand: Empowering older adults with cognitive impairment through creating and sharing. In *DIS*, 2016

of the technology on their mutual relationship. On the other hand, as the shift in technology has been so important these last few decades, many born after the 1960s have acquired sufficient technological literacy that could allow them to keep up more efficiently with future advances. Our elders today could be the last generation of humans to feel completely disconnected from technology (at least in most developed countries and urban environments); as researchers, we believe we have a duty to support them nonetheless.

4.6 - Conclusion for the Memory Music Box project

There are many approaches to supporting elders with new innovations. While there are numerous areas needing technological intervention, research in the fields of aging and empathic design reveals that loneliness and isolation are major contributors to a rapid decline in cognitive function. In spite of this, many innovations do not support the connectedness with which so many elders struggle. In the context of Cognitively Sustainable Design, the Memory Music Box is proposed as a novel contribution to the discussion around familial isolation and builds on previous work in this space within the interface research community. Our proposed design speaks to the broader design-research and positive-aging communities about the importance of empowerment, autonomy, and self-expression through the personalization and application of design objects by older individual users, especially those experiencing cognitive impairment. Our findings confirm that the current state of connectedness for older individuals, especially those in care facilities, is waning during this time of technological revolution. Based upon our surveys, focus groups and user tests analyzed throughout this work, we propose that the Memory Music Box has the potential to helps fill an existing gap in the wellbeing of elderly individuals by facilitating connectedness.

Ideally, there should not be a major learning curve in order to achieve the simple task of talking to a family member. Although there are many systems with a higher learning threshold, we envision our device as a low-barrier gateway for individuals to become more comfortable with technology. Positive interactions can instill elders with agency and in some cases the confidence to expand their circle of technologies to include more complex interfaces.

Whether or not one expands their technological horizons, the box provides immediate access. We can foresee a future where the correspondent can select different interaction levels in the box so elders can expand their technological capabilities within their comfort-zones. We envision the Memory Music Box as something with which older individuals can

age. Crafted within the principles of Cognitively Sustainable Design, the box will remain a familiar and intuitive object with which elders can continue interacting, even as memory loss increases and mobility decreases.

Although the MMB fills an important gap, we can envision other populations who could similarly benefit from this device. Specific minor changes in the form factor and usage flow could allow individuals with developmental disabilities, or reduced cognitive or motor function, and young children to engage with the MMB as well.

Although the purpose of our work is currently geared to elders, there is great opportunity in the exchange of information, legacy, and wisdom for the younger individual: a kind of guidance and mentorship that youth in all of its haste is seldom privy to. From our data with the "young correspondents/grandchild" group, grandchildren are seeking a deeper connection with their elders. We refer to those advanced in years as sages and elders because of their wisdom hewn from decades of living: knowledge garnered from life experiences and major events impacting our world.

Non-western cultures often look to the oldest members of their communities for guidance. We pose the question, what has happened to our elders? Why are they so often far away in care facilities, segregated from the general population without the chance for information transfer? Our elders need us as we need our elders. Sadly, due to the connection fallout, it is almost as though we have established two different planets: one planet that uses technology and one that doesn't. Often we cast the one without technology in a pejorative light, considering the people there to be old-fashioned or obsolete. However this is not a result of elders falling behind—this is an issue of equity, the weight of which we feel upon our shoulders. The Memory Music Box would afford a low-cost, equal-access platform to facilitate meaningful connections across these divides.

4.7 - Vocal Connection and The Memory Music Box

Connection is at the center of the Memory Music Box project. Bridges are established through time, space, and generations, and across different cognitive abilities. The media of the connection starts with the physical object. The tangibility and aesthetic of the object are designed to inspire a sense of familiarity and closeness. Beyond the object, the images and music are also used as media for connection. The connection to the correspondent is established remotely through the video calls and asynchronously through the sharing of experience and memories. Finally, the connection with friends is reinforced and given support through the shareability of the object. The vocal aspects are present in the video calls enabled by the objects but also

in the establishment of a common language through the music and memories shared. All those techniques and strategies are deployed to support the sharing of vocal experiences even when the words and meanings have gone missing.

Ultimately, this project is the result of design exploration to encourage conversations to take place. In terms of our three paradigms, the Memory Music Box tackles both the personal and interpersonal contexts. The box reinforces the links across generations while sustaining memories through he personalized slideshow. The device is highly experiential and aims at creating a seamless connection as a platform between the user and the correspondent. Figure 40 frames the Music Box project in the context of the thesis.

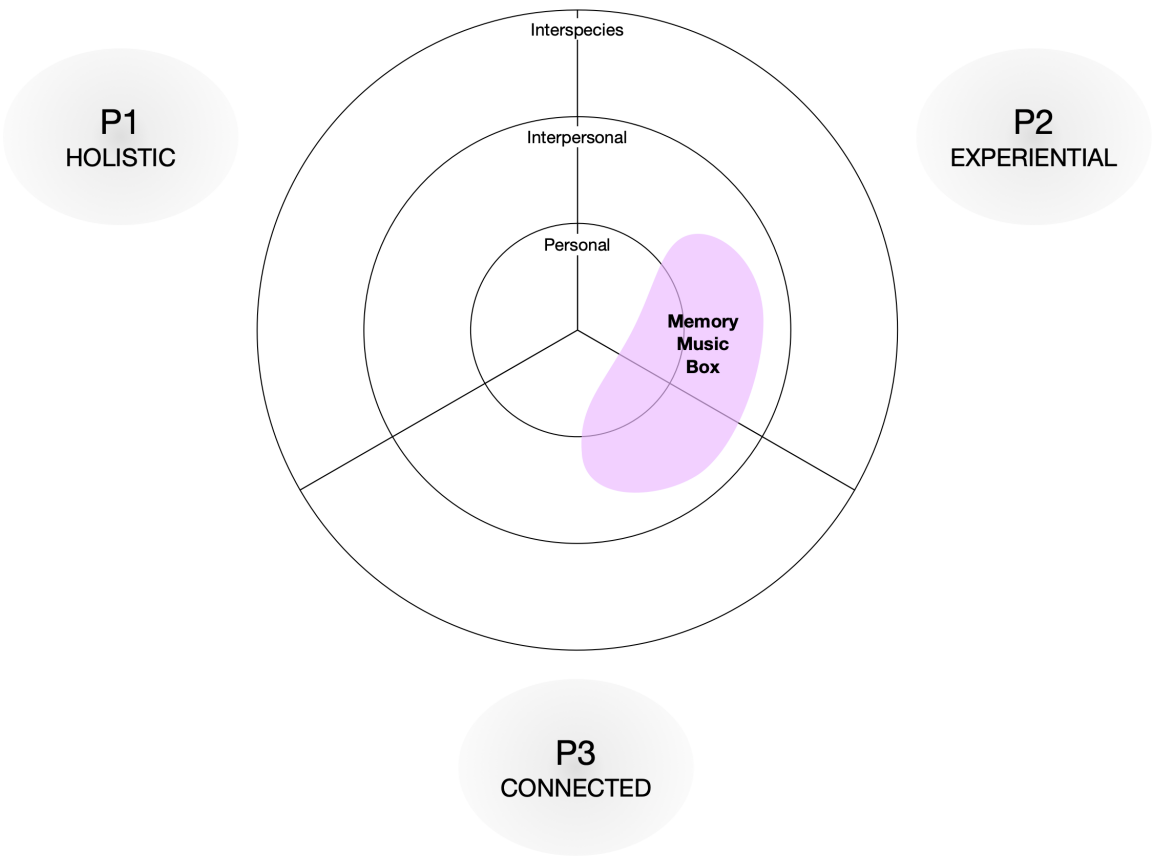


Figure 40: Mapping of the Memory Music Box project onto the Vocal Connection space

5 - Mumble Melody

This chapter presents the Mumble Melody initiative. This initiative aims at shaping a better understanding of the relationship between inner and outer voices. This work adopts the inner voice model presented in the Background chapter, and fits within the framing of our three paradigms. After introducing some of the complex interconnections between voice, music and the brain, I present two studies that explore the potential effects of modulating vocal auditory feedback using music. In the first study, we explore the effect of those feedback on mood and valence. In the second study we explore the effects of a series of different feedback modes on fluency for people who stutter (PWS). The proposed explorations can be seen as ways to examine some aspects of our second paradigm, according to which our experience of the words is strongly informed by our experience of voices.

5.1 - Introduction

5.1.1 - Voice Music in the Brain

This work begins with the constatation that, despite the existence of neural overlaps[237], our brain processes voices differently than other sounds, such as noise or music[238]. This effect is even stronger with our own voice, as it results from intentional motor control. However, the objective acoustic differences between music, noise, and voice are less clear, as evidenced by multitudes of voice-based auditory illusions. Vocal auditory illusions are vocal experiences or exercises in which what is presented to the ears perceptually transforms into another percept. Diana Deutsch's work has been influential in uncovering and studying several of those illusions. Instances of voice-based auditory illusions include:

- **Verbal transformation effect**: An auditory illusion involving radical changes in what is heard in a clear recording of a word or phrase repeated many times on a loop of audiotape[239]
- **Phonemic restoration effect**: A perceptual phenomenon where under certain conditions, sounds actually missing from a speech signal can be restored by the brain and may seem to be heard[240]

[237] Isabelle Peretz, Dominique Vuvan, Marie-Élaine Lagrois, and Jorge L Armony. Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015

[238] Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 2015

[239] Richard M Warren and Richard L Gregory. An auditory analogue of the visual reversible figure. *The American journal of psychology*, 1958

[240] Richard M Warren. Perceptual restoration of missing speech sounds. *Science*, 1970

- **Semantic satiation**: A psychological phenomenon in which repetition causes a word or phrase to temporarily lose meaning for the listener, who then perceives the speech as repeated meaningless sounds[241]
- **Speech-to-song illusion**: When spoken language comes to be perceived as musical in quality. The illusion can be elicited by continuously repeating a spoken phrase[242]: a repeated segment of speech will be perceived by the brain as music while the single segment will be heard as speech.
- **McGurk effect**: A perceptual phenomenon that demonstrates an interaction between hearing and vision in speech perception[243]
- **Yanny or Laurel**: For the same noisy speech utterance, different people reported hearing either "Laurel" or "Yanny" [244]
- **Auditory pareidolia**: Hearing indistinct voices in random noise [245]
- **"Rubber voice" illusion**: A stranger's voice, when presented as the auditory concomitant of a participant's own speech, is perceived as a modified version of their own voice[246]

These illusions speak for highly efficient "shortcuts" used by the brain when processing vocal sounds. Much of human perception is the result of the brain filling in gaps in the data from our senses. Research on voice and body ownership[247] even suggests that one can create the illusion of someone else's voice being yours, meaning that vocal agency can be self-attributed even in the absence of prior intention, feed-forward prediction, or priming, and have a preceding effect.

In our work, we are particularly interested in making use of the strangeness provided in such illusions for two reasons. First, because such playful examples bring awareness to the brain processes involved in speech and voice processing. Those instances of auditory illusion also raise the question as to whether one can learn to perceive something differently. Second, those brain pathways involved in speech production and perception are instrumental in the development and functioning of the inner voice and—we believe—our inner self. In affecting the voice-related neural pathways, can we affect deeper parts of the mind?

5.1.2 - Speech Companions

The work presented in this chapter is based on altering vocal auditory feedback. There are infinite ways to affect/transform/perturb the spoken voice. Many past studies have been based on changing the fundamental frequency or formants of the spoken voice. In other cases, researchers added delays on the voice or overlaid someone else's voice on top of the subject's voice. For our work, we created a series of "speech companions" that use

[241] MF Bassett and CJ Warne. On the lapse of verbal meaning with repetition. *The American Journal of Psychology*, 1919

[242] Diana Deutsch, Trevor Henthorn, and Rachael Lapidis. Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 2011

[243] Kaisa Tiippana. What is the mcgurk effect? *Frontiers in Psychology*, 2014

[244] Daniel Pressnitzer, Jackson Graves, Claire Chambers, Vincent De Gardelle, and Paul Egré. Auditory perception: Laurel and yanny together at last. *Current Biology*, 2018

[245] Michael Nees. Hearing ghost voices relies on pseudoscience and fallibility of human perception. 2015

[246] Zane Z Zheng, Ewen N MacDonald, Kevin G Munhall, and Ingrid S Johnsrude. Perceiving a stranger's voice as being one's own: A 'rubber voice'illusion? *PloS one*, 2011

[247] Domna Banakou and Mel Slater. Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking. *Proceedings of the National Academy of Sciences*, 2014

musically modulated auditory feedback (MMAF) as a way to transform the voice. The idea is to loosely use musification and timbre modulations in order to leverage the existing brain pathways associated with music perception instead of, or in addition to, the vocal brain pathways. In the first study, the use of music allows us to test the influence of different musical emotional modalities on mood and valance. In the second study and the final proposed protocol, both targeting adults who stutter, the use of music is purposeful for different reasons. First, it is thought that the brain areas involved in music perception and production are more global and are not affected by stuttering, as attested to by the fact that many people who stutter (PWS) are fully fluent when singing. The idea is to leverage those functional areas and repurpose them toward speaking fluency. Second, the use of music is also meaningful purely for aesthetic reasons. PWS often carry stigmas associated with their own voices, and this project also aims to increase people's appreciation and liking of their voices. Third, musical feedback often contains rules and parameters that are understandable by everyone without training. Such multiple parameters could also enable the user to control and modify parameters to personify their feedback.

Speech companions originally arose from the idea of creating a real-time musical accompaniment of the spoken voice. Here the concept of "music" is considered rather loosely as "organized sound" and ranges from discreetly returning the voice on the closest note on a major scale, to changing the vocal timbre to a more breathy, whispery quality, or to synthesising guitar or piano chords following the voice's natural rhythm and melodies. Because there is musicality in everyday speech, our speech companions aim less at imposing external musicality than at extracting and using the existing musical aesthetic of the voice and transforming it to increase awareness of it.

5.1.3 - Organization

In this chapter, we are particularly interested in asking whether we can use auditory feedback perturbation as a mode of inner voice control/training to change brain pathways during speech. Can we train our brains to switch between different ways of perceiving our own voices? Can we use the outward voice as a way to access the inner voice, our control over it and our relationship with it, and ultimately reach even deeper parts of our subconscious?

In this chapter we propose three preliminary explorations toward answering those questions:

- First, a study on people who do not stutter to establish whether the use of altered feedback based on music can affect their inner state of mood and valence.

- Second, a preliminary study on 24 adults who stutter to evaluate the effects on their fluency of novel auditory feedback based on music. The improvement offered by musical modes over control feedback perturbation can be seen as a sign suggesting that affecting the aesthetic (musical) part of the signal effectively alters the brain pathways involved with the inner voice.
- Third, a protocol to a longitudinal study that would aim at testing the longer-term effects of musically modulated auditory feedback on fluency as well as the potential of using the feedback as a training to learn—even in the absence of perturbation—to acquire and maybe ultimately master the new ability to mentally switch perception modes between speech or music, thus intentionally mentally switching perception mode on the voice.

5.1.4 - Acknowledgements

Though this project was primarily led by the author, the conception and development of the Mumble Melody initiative was conducted in collaboration with multiple stakeholders and with support and close guidance from Tod Machover, Satrajit Ghosh and Janet Baker.

For the initial study on the effects of speech companions on valence described in the following section, the development and evaluation of the systems were conducted in collaboration with George Stefanakis and Sebastian Franjou. All participants gave written informed consent to the study, in person at the research location, under the protocols approved by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (protocol number: 1802248976R001). More details on this project can be found in the publication that resulted from it: *Speech Companions: Evaluating the Effects of Musically Modulated Auditory Feedback on the Voice*[248]

Regarding the work on the effects of Musically Modulated Auditory Feedback on the fluency of adults who stutter, the ideation process, research work, study design and result interpretation were done in close collaboration with Dr. Michael Erkkinen, MD, neurologist at Brigham and Woman's hospital. George Stefanakis collaborated in developing the system, conducting the study and running data evaluation. Participants gave informed consent, and all protocols were approved under MIT Committee on the Use of Humans as Experimental Subjects (protocol number: 1906848293). More details on this project can be found in the preliminary extended abstract publication that resulted from it: *Fluency Effects of Novel Acoustic Vocal Transformations in People Who Stutter: An Exploratory Behavioral Study*[249]

[248] Rébecca Kleinberger, Stefanakis George, and Sebastian Franjou. Speech companions: Evaluating the effects of musically modulated auditory feedback on the voice. 2019b

[249] Rebecca Kleinberger, George Stefanakis, Satrajit Ghosh, Tod Machover, and Michael Erkkinen. Fluency effects of novel acoustic vocal transformations in people who stutter: An exploratory behavioral study, 2019d

5.2 - Effects of Speech Companions on mental state

> *There is music in everyday speech and often some kind of poetry in the way people talk.*
>
> — **Tony Schwartz**
> (Introduction to a track on Millions of Musicians (1954))

Changing the way one hears one's own voice, for instance by adding delay or shifting the pitch in real time, can alter vocal qualities such as speed, pitch contour, or articulation. But can altered vocal feedback also affect deeper parts of the mind such as mood and valence? We created new types of auditory feedback called Speech Companions that generate live musical accompaniment for the spoken voice. Our system generates harmonized chorus effects layered on top of the speaker's voice that change chord at each pseudo-beat detected in the spoken voice. The harmonization variations follow predetermined chord progressions. For the purpose of this study, we generated two versions: one following a major chord progression and the other one following a minor chord progression. We conducted an evaluation of the effects of the feedback on the speakers and present initial findings assessing how different musical modulations might potentially affect the emotions and mental state of the speaker, as well as semantic content of speech and musical vocal parameters.

5.2.1 - Introduction

This work seeks to assess how different musical feedback modulations might affect the general mental state of the speaker, semantic content of speech, emotions in vocal tonalities, and vocal parameters of musicality. Modulated Auditory Feedback uses digital signal processing to transform the way someone hears their own voice. Modulated Auditory Feedback has documented effects on how someone speaks in terms of speed, articulation, and fluency. For example, adding a short delay to the voice can lead to prolongation of vowels, repetition of consonants, increased intensity of utterance, and other articulatory changes [250]. A short delay (20-150ms) can help people who stutter become more fluent [251], but a longer delay (higher than 200ms) can lead to jammed speech [252].

In recent years, the research community has investigated the possible effects of altering vocal auditory feedback for the regulation of emotions [253]. In these studies, modulated feedback is used covertly to make the voice sound more calm, sad, happy, or fearful by manipulating formants and

[250] Aubrey J Yates. Delayed auditory feedback. *Psychological bulletin*, 1963; and Grant Fairbanks and Newman Guttman. Effects of delayed auditory feedback upon articulation. *Journal of Speech & Hearing Research*, 1958

[251] Joseph Kalinowski et al. Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language & Communication Disorders*, 1996

[252] Grant Fairbanks. Selective vocal effects of delayed auditory feedback. *Journal of Speech & Hearing Disorders*, 1955

[253] Jean Costa et al. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *CHI*. ACM, 2018; and Jean-Julien Aucouturier et al. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 2016

overall pitch, and by adding filters. The researchers then established the effects on the subjects, as measured through self-reported emotions and levels of stress. Our approach, Musically Modulated Auditory Feedback, consists of producing aesthetic musical manipulation of the voice instead of covert intonation and testing the effects on the speaker's fine-tuned ability to shape their voice and speech. We conducted a study to assess whether specific Musically Modulated Auditory Feedback can induce particular effects and modulate emotional content from the voice, in addition to affecting vocal parameters. Our objectives are twofold: first, we are interested in studying the potential regulatory effects of music when woven into voice. Second, we wish to bring more awareness to the intrinsic musicality present in everyday speech and explore possible research applications based on perceiving the spoken voice as an inherent musical signal. These applications range from infant-directed speech and language acquisition to speech pathology and aphasia reeducation. Such research could also be useful for music composers or lead to new tools for phonologists to characterize human speech. We present the background supporting our inquiries in terms of neurology, music and emotion, and self-perception theory. Then we present the study design and detail the data analysis and results.

5.2.2 - Background

5.2.2.1 - Musicality of everyday speech

Speech is one of humanity's richest and most ubiquitous forms of communication. Its richness lies in the combination of linguistic and nonlinguistic information. Musicality is a crucial nonlinguistic component of speech, incorporating the tempo and rhythm of the speaker along with the pitch variation and unique texture of vocal sound. In casual everyday speech, individuals possess a unique musicality, rhythm, and melodic style. In 1954, urban folklorist and sound archivist Tony Schwartz proposed the idea that "there is music in everyday speech, and often a kind of poetry in the way people talk" [254]. In our work, we aim to increase awareness of the beauty and diversity of musicality in our everyday experience of voices. Vocal, non-verbal behaviors such as prosody, tone, loudness, breathiness, accent, pitch envelope, and tempo are all parameters that are most often unconsciously controlled when speaking, but they implicitly convey a great deal of information. For instance, pitch intervals can reveal changes in mood[255] or hormone levels[256], and tempo information can be a marker of depression[257]. Prosody and especially pitch accentuation can also be used to modify semantic content[258]. Our system creates different types of musical layers on top of the spoken voice by extracting existing musicality from speech and aims to bring more awareness to this intrinsic musicality.

[254] Tony Schwartz and Richard Kostelanetz. Interview with tony schwartz, american hörspielmacher. *Perspectives of New Music*, 1996

[255] Steven K. Blau. Musicality of speech changes with mood. *Physics Today*, 2010

[256] David R Feinberg et al. Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and behavior*, 2006

[257] Ying Yang et al. Detecting depression severity from vocal prosody. *Transactions on Affective Computing*, 2013

[258] Vivek Kumar Rangarajan Sridhar et al. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*. International Speech Communication Association Campinos, Brazil, 2008

5.2.2.2 - Music and Emotion

The influence of music on emotion is not a novel concept. As early as 350 B.C., Aristotle characterized different musical modes by the emotions they evoked[259], and throughout the classical age of music, the "feel" of a piece was often married to more objective qualities like tempo and chord. In terms of valence, minor-keyed pieces and melodies are traditionally associated with sad, nostalgic, or morose atmospheres, including Chopin's *Funeral March* and Mozart's*Requiem*. On the other hand, major keyed-pieces are classically associated with joyful, strong, and uplifting atmospheres, such as Mendelssohn's textit{Wedding March' and Rossini's *William Tell Overture*. Whether some innate qualities of the major and minor tonalities informed theory and popular opinion, or vice versa, is a philosophical inquiry which is not to be dwelled upon, but the popular social perception of the major and minor chords, for hundreds of years in the Western tradition, has been that the former is classically joyful, while the latter is often considered sorrowful[260]. Of course, there are exceptions; many pieces of music exist which do not follow this categorization. Furthermore, the perceptions of individual pieces can vary widely from person to person. Work by Hevner et al. in 1935 elucidates the various affective qualities of the major and minor musical modes [261] proposing that major is dynamic, more natural and fundamental than minor, and "expresses varying degrees of joy and excitement" and that "[the major] sounds bright, clear, sweet, hopeful, strong, and happy," while the minor "express gloom, despair, sorrow, [and] grief." Many theories and studies have supported this notion of musical modes having intrinsic emotional connotations implicit within them, and several support the idea that music can indeed evoke strong emotional responses in listeners [262].

[259] Stephen Halliwell et al. *Aristotle's poetics*. University of Chicago Press, 1998

[260] Robert G Crowder. Perception of the major/minor distinction: I. historical and theoretical foundations. *Psychomusicology: A Journal of Research in Music Cognition*, 1984

[261] Kate Hevner. The affective character of the major and minor modes in music. *The American Journal of Psychology*, 1935

[262] John A Sloboda. Music structure and emotional response: Some empirical findings. *Psychology of music*, 1991; and Klaus R Scherer. Expression of emotion in voice and music. *Journal of voice*, 1995

Although the findings that minor chords have a negative valence effect have been presented in much prior works on music emotion, as of the time of this writing, we haven't found any prior work assessing effects of the use of minor vs major keys when interactively woven into spoken voice. In this work, we are proposing a step toward assessing the unconscious effects of auditory musical transformation of speech.

5.2.2.3 - Self-Perception Theory

In his work on self-perception, Daryl Bem postulates that we come to know our own attitudes, emotions, and other internal states partially by inferring them from observation of our own overt behaviors. He argues that internal cues are "weak, and ambiguous" and that we often have to rely on external cues to understand our own behaviors the same way an outside observer would[263].

[263] Daryl J. Bem. Self Perception Theory, 1972

This theory suggests that it is partly by monitoring the way we overtly express our emotions that we infer our emotional state and attitudes. Multiple studies support this theory, by showing that forcing the outside symptoms of emotion can reinforce said emotion in the subject [264]. Similar results have been obtained for vocal expression of emotion: subjects asked to imitate vocal patterns associated with specific emotions (such as laughter) reported their emotions being affected accordingly [265]. The previously mentioned studies involve active cooperation from the subject, but further studies have found similar effects in cases where subjects didn't have to consciously adjust their behavior or weren't even aware of anything being modified. Subjects who heard their voices processed in real-time to make it sound as if they were happy, sad, or afraid experienced changes in tension and self-reported positivity usually associated with the experience of the corresponding emotion. This suggests an influence of the perception of the subject's own voice alone on their emotions, although they never consciously noticed any modification of their voices [266]. Similarly, participants whose voices were modified to sound calmer and fed back to them in real time during relationship conflicts reported feeling less anxious than those having unmodified feedback [267]. These studies suggest that emotions can be regulated by feeding back a modified version of a speaker's voice in real time even if the modification is not consciously detected.

In our work, we explore this field by modifying the subjects' fed-back voices to match purely musical expressive features. Links between prosodic and musical emotional features have been suggested, such as the use of the interval of a minor third for affects of negative valence for both speech and music [268]. By mapping the fed-back voice to musical attributes considered happy or sad. we hypothesize similar emotional responses to those induced by previous non-musical manipulation.

5.2.2.4 - Neural Basis

A large body of work conducted on neural control of speech has been accumulated in Frank Guenther's book of the same name. A key idea presented in the book is that of neural auditory feedback control, which is operated by means of a feedback/feedforward mechanism. In this scheme, it is suggested that fluent speech is dependent on fluid, logical, sensory feedback streaming back to the speaker. It is for this reason, Guenther asserts, that delayed auditory feedback results in a range of dysfluent behavior, up to and including complete cessation of speech [269]. The importance of auditory feedback in speech production has been further demonstrated by studies on the effect of modified real-time and delayed feedback on

[264] Paula M Niedenthal. Embodying emotion. *science*, 2007

[265] Elaine Hatfield and Christopher Hsee. The impact of vocal feedback on emotional experience and expression. 1995

[266] Jean-Julien Aucouturier et al. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 2016

[267] Jean Costa et al. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *CHI*. ACM, 2018

[268] Meagan E Curtis and Jamshed J Bharucha. The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, 2010

[269] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5

speech and sustained vowel sounds. It was found that modification of the fundamental frequency (F0) of the feedback voice produces a compensatory opposing shift in the pitch of the resultant sound for both sustained vowels and speech [270], due to brain overcompensation. Formant shifts in feedback have also been found to produce compensatory changes in the spectral characteristics of the voice [271], even when participants were consciously informed of the modifications and instructed not to compensate [272]. Thus it appears that auditory feedback plays a crucial role in speech production, to the point where it sometimes cannot be ignored even if the speaker is consciously trying to inhibit its effects. The neurological basis of our study is to interfere with the encephalic speech-feedback mechanism by overlaying the stream of one's own raw voice, using musical modulations. The goal is to monitor the alterations in the resulting feed-forward mechanism of new speech being produced. We also seek to analyze the semantic nature of speech produced when the backward-fed vocal audio is substantially altered in either major or minor chord progressions.

[270] Theresa A Burnett et al. Voice f0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, 1998; and Theresa A Burnett et al. Voice f0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice*, 1997

[271] John F Houde and Michael I Jordan. Sensorimotor adaptation of speech i: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, 45(2), 2002

[272] Kevin G Munhall et al. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, 2009

5.2.2.5 - Measure of Musical Parameters in Speech

It can be difficult to assess and characterize the musicality of speech. The question is so challenging that quite often, researchers assess the "level of musicality" by asking experts with extensive music training to subjectively rate vocal sound samples. In *Music, Language and the Brain* [273], Aniruddh D. Patel distinguishes musical and linguistic sound systems in the way they carry pitch, timbre, rhythm, and melody. One way to assess pitch is through the analysis of the mean pitch (Pm) of a vocal sample. Pm provides information about the fundamental frequency (F0) of a subject's voice. Males with lower voices, for instaance, will have smaller F0, thus lower Pm. The level of melodicity can be very roughly accessed through the measurement of the pitch standard deviation (Psd) of a vocal sample. Psd gives cues about the pitch envelope in speech: the lower the Psd in a given phrase, the more monotone and concentrated around the main pitch the voice will be. Contrary to a lot of musical systems and instruments that present a fixed timbre, speech is also fundamentally a system of organized timbral contrast, as timbral variation in vocal sound is the basis of phoneme production. In addition, on account of the shape of formants, subtle vocal timbral variation is an important characteristic of phonation. Timbre in speech can be measured through different parameters, such as jitter, breathiness, or harmonic-to-noise ratio (HNR expressed in dB). HNR is a more global way to see timbre as it indicates the energy concentration of the sound around the main pitch. HNR represents the degree of acoustic periodicity. An HNR of 0 dB means that there is equal energy in the harmonics and in the noise. And an HNR of 20dB means that 99 percent

[273] Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010

percent of the energy of the signal is in the periodic part. Singing voices have higher HNR than spoken voices [274]. Pm, Psd, and HNR are used in this study as the measurement of the variation of musical parameters of speech.

[274] Elizabeth L Stegemööller et al. Music training and vocal production of speech and song. *Music Perception: An Interdisciplinary Journal*, 2008

5.2.3 - Speech Companions

We created new types of auditory feedback called Speech Companions that generate real-time musical accompaniment to the spoken voice. The Speech Companions used for this study are based on a type of active harmonizer. A harmonizer is a pitch shifter that combines the shifted pitch version with the original sound to create a harmony with two or more notes. Our system combines the original vocal signal with two extra layers creating a musical chord. A constant harmony chord being played in a sustained manner can create a very dull effect. In live or studio music production, harmonizers are often controlled manually by a keyboard that changes chords to make it more reactive. For our study, we wanted the feedback to react to the inherent rhythm of speech. Our system triggers a new chord, from a predetermined set, at each pseudo-beat of speech.

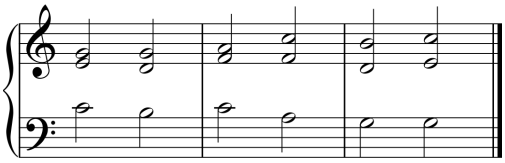
Pseudo-beats are triggered at near-regular intervals determined by minimum delay and natural attacks in the voice. Sound attacks correspond to onsets or peaks in the intensity of the sound signal. After each chord change, the system counts down the chosen delay in milliseconds and then waits for the next speech onset to generate the next pseudo-beat that controls the next chord change. When chords change at regular intervals, the feedback seems very static and creates a ticking clock effect that can feel stressful and alter the natural speech rhythm. By using the pseudo-beat method, we ease the chord variation into the organic speech tempo to respect the built-in musicality of speech. The system was implemented using Max MSP 8[275] for pseudo-beats detection and with MHarmonizerMB[276] for Reaper64 to create the harmonization.

[275] Cycling74. Max MSP, 1919. URL https://cycling74.com/

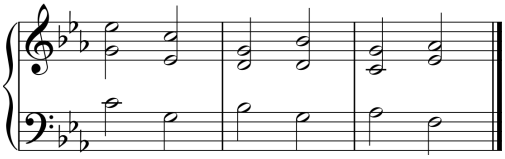
[276] Melda Production. Mharmonizermb, 2019. URL https://www.meldaproduction.com/MHarmonizerMB

The system randomly draws a chord to harmonize from a predetermined chord progression—either major or minor. The chord progressions were chosen to unambiguously convey the key and mode regardless of which order the chords were played in, as they were to be fed to the subjects in random order. The key of C was chosen, and the chords are in the modes of C Ionian (major) and Aeolian (natural minor) (see Figure 41). Although commonly used by classical composers, the harmonic minor was avoided as the augmented second interval can sound jarring or exotic to Western listeners. This interval is usually avoided by following proper voice leading rules, but this wasn't possible due to the random order of the chords. The aeolian or natural minor mode, commonly found in popular

music, was chosen instead to bypass this problem. The chords are voiced in the mid-range so that the harmonized feedback would not sound too distant in pitch from the normal voices of most subjects. The range and spread of the chords were kept comparable (see Figure 41). The major chords are triads, while the minor chords are sometimes enriched to convey more tension and sadness.



(a) Major chord progression



(b) Minor chord progression

Figure 41: Chord progression used for for the major (a) and minor (b) mode of our study

Figure 42 illustrates the use of the pseudo-beat to trigger changes in the MIDI track that always lasts longer than a minimum delay and is ultimately triggered by speech onsets and phone attacks from the raw voice. The result generates harmony changes in the processed voice (middle track) that exhibit different spectrum peaks than the raw voice. In this case, each chord lasts a minimum of 3000ms but can extend longer if no attack is detected. The volume was kept the same for all participants and was loud enough to mask the actual voice. We hypothesize that such feedback might affect the valence of the speaker as well as the musicality of their speech.

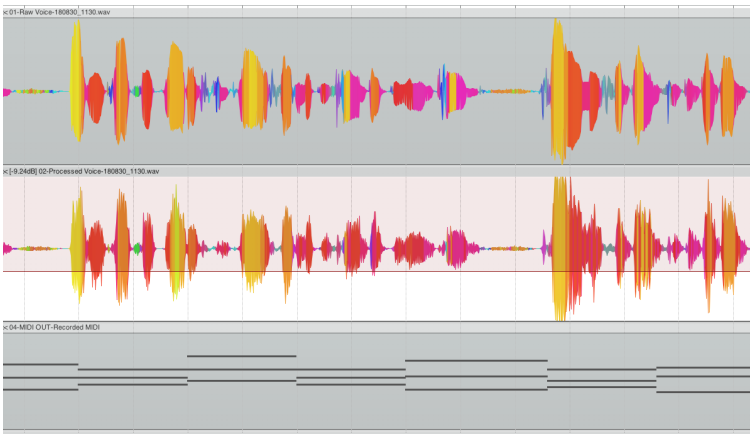


Figure 42: Illustration of the Speech Companion in use: attacks in the raw voice (top track) trigger the MIDI chords (bottom line) that control harmony changes in the processed voice (middle track)

5.2.4 - Study Design

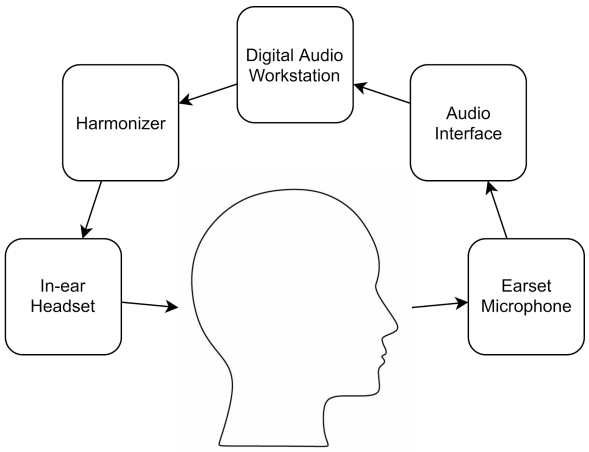
5.2.4.1 - Participants

The institutional review board approved this study. The sample comprised 20 adults (11 women and nine men). There were two groups: one group of 10 adults received the "major scale" condition, and the other group of 10 adults received the "minor scale" condition. No compensation was offered to the participants. The study was organized over five days, in which we measured respectively one, three, six, five, and five participants. The settings were identical throughout the five days in terms of environment, microphone settings, audio loudness, and lighting.

5.2.4.2 - Study Setup

The study was conducted in a soundproofed room to reduce background noise. We used a Countryman E6 directional ear-set microphone and a Babyface RME Pro audio interface connected to a computer to record the voice, and a pair of Bose SoundSport headphones to provide audio feedback. The SoundSports are very open (i.e., let outside sound in), which allowed the interactions between the subject and the researcher to remain natural. The researcher giving the instructions also wore a pair of SoundSport to monitor the quality of the feedback heard by the subject. The loudness of the feedback was set just loud enough to effectively cover the speaker's voice without sounding unnaturally loud (around 60dB).

Figure 43: System Flow



5.2.4.3 - Method

The study was composed of two phases (baseline and musical feedback), each containing the same three tasks (reading, mood assessment, and

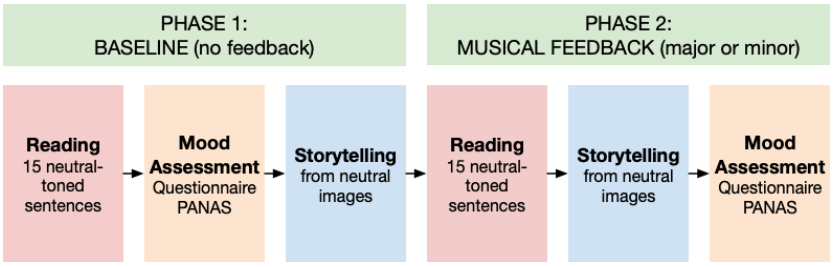


Figure 44: Order of the Study

storytelling). Subjects were initially fitted with in-ear headphones and a microphone. During phase 1, subjects did not hear any feedback through the headphones but still had to wear them to get accustomed to them in preparation for phase 2.

- Task 1 was a reading exercise to normalize the subject's mood to neutral at the start of the study. To this end, we used an adapted version of the Velten mood induction process (Velten MIP) method [277]. As we wanted to induce a neutral mood in all participants, we asked them to read a series of 15 trivial and factual statements that carry no emotional load, extracted from the 50 sentences used in the Velten MIP version used by Isen and Gorgoglione [278]. This reading task aims at initiating the same neutral common ground for each subject.
- Task 2 consisted of filling out a short mood questionnaire to measure self-reported affect. We used the Positive and Negative Affect Schedule (PANAS) methodology [279]. The PANAS was chosen for its robustness, replicability, and widespread use, to allow for easy comparison with other works. To limit demand effects, it was issued in digital form where only one question was visible at a time. This prevented the subject from seeing the whole questionnaire and influencing global results by correlating their answers to several questions.
- In task 3, subjects were shown four images from the IAPS image database and asked to construct a narrative loosely based on the images. IAPS is a database of images for experimental investigations of affect [280]. Each of the chosen images was in the valence range 4.5–5.5, signifying emotional neutrality. A scaled score of 1 on the valence portion of the IAPS image scale means unhappy, while 9 means happy. Images with neutral-evoked emotions could go either towards the more joyous, or the more depressed, semantically and tonally.

For the entirety of phase 1, no audio feedback was played through the headphones, though audio from the microphone was being actively recorded. This initial "neutral" portion of the protocol was used as a baseline to evaluate the effects of the musical modes.

[277] Emmett Velten Jr. A laboratory task for induction of mood states. *Behaviour research and therapy*, 1968

[278] Alice M Isen and Joyce M Gorgoglione. Some specific effects of four affect-induction procedures. *Personality and Social Psychology Bulletin*, 1983

[279] David Watson et al. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 1988

[280] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1997

The study then entered phase 2, where the subject had to repeat the same three tasks while hearing the Musically Modulated Feedback. For each subject, either the minor or major chord harmonizer was tested and each subject listened to their voice modulated at a volume sufficiently high so as to mask their own speaking voice.

- For this phase, task 1 subjects red 15 new neutral sentences
- For task 2, the subjects were given four different images from the IAPS image database, from which to generate a new story
- For task 3, the subject was asked to fill out a new randomized PANAS

In phase 2, we switched tasks 2 and 3 to give the subject more time to get used to the feedback before measuring their self-reported mood, in order to get a better sense of the change of mood induced by the study.

The musical modulations were then turned off and we asked the subjects to share their best guess about the purpose of the study, to determine whether they were aware that their mood and tone were being investigated. Indeed, past research has shown that the results of studies on affect might be skewed or unintentionally affected if subjects are aware that their mood is being monitored [281]. At the end of the experiment, we verified that all the participants had remained unaware that the study was about affect; we informed them of the actual objective through a short debriefing session and asked them not to divulge it to other potential participants.

[281] Rainer Westermann, Kordelia Spies, Günter Stahl, and Friedrich W Hesse. Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of social psychology*, 1996

5.2.5 - Data Analysis

The collected data were processed into three categories: the PANAS results were processed into numerical data, the stories generated (two per subjects) were analyzed in two ways: as text to assess semantic content, and as audio samples to assess changes in vocal affect and musicality.

5.2.5.1 - Self-Reported Affect

The PANAS was completed by the subject twice: once as part of the baseline evaluation, and once after the musical-feedback task. The questionnaire gives us scores for positive (PA) and negative affect (NA), which are subtracted to obtain a valence score V normalized between -1 and 1. To limit the variations due to differences in initial mood between subjects, we analyzed the variation in valence induced by the experience by subtracting the valence prior and post-study. This allowed us to only take into account mood changes from baseline induced during the study. Changes in valence were then compared between minor and major scale groups.

5.2.5.2 - Semantic Content

To analyze the semantic content of the speech, the audio recordings of the constructed narrative based on the pictures from IAPS in tasks 1.3 and 2.3 were transcribed to text using Dragon NaturallySpeaking [282], and the text outputs were then reviewed manually and corrected to assure accurate transcription of speech. These text transcriptions were processed using the SentiWordNet database which scores words based on their positivity and negativity [283]. For each subject, we compared the difference in average positive, negative, and total scores from the SentiWordNet analysis between the baseline story and the story invented by the subject while hearing musical feedback.

[282] Dragon. Naturallyspeaking, 2019. URL https://www.nuance.com/dragon.html

[283] Stefano Baccianella et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2010

5.2.5.3 - Emotion Analysis from Vocal Intonations

Emotional vocal qualities were analyzed using the speech emotion recognition software OpenVokaturi [284]. OpenVokaturi is a software development kit (SDK) developed by Vokaturi to provide an analysis of the basic emotions from the speaker's vocal intonations. It is worth noting that the SDK is presented as having an accuracy on the classification of only 66.5 percent, which limits the validity of the results [285]. Vokaturi provides percent likelihoods for neutrality, happiness, sadness, anger, and fear. Each speech audio sample was analyzed using the OpenVokaturi pre-trained model. Scores for positive and negative affect were constructed by way of a weighted sum (Positive Affect = Happiness; Negative Affect = (Anger + Fear = Sadness) / 3), in a similar fashion to the PANAS's way of summing different positive and negative reported emotions to construct positive and negative affect [286]. We then took the differences between the scores for speech segments produced under the musically modified feedback and those produced under normal feedback conditions. To mitigate the effects due to subject particularities, we considered the relative change in affect between the baseline phase and the musical feedback phase rather than absolute affects.

[284] Vokaturi. Vokaturi. URL https://developers.vokaturi.com/getting-started/overview

[285] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2012

[286] David Watson et al. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 1988

5.2.5.4 - Vocal Parameters

We used Praat [287] for the analysis of vocal and musical parameters of speech. For the speech samples of the narrative generated by subjects in phase 1 and 2, we extract mean pitch (Pm), pitch standard deviation (Psd) and harmonic-to-noise ratio (HNR) of the voiced sections of speech. A vocal sound is said to be "voiced" when it originates from the vocal chord and not only from the air leaving the lips (e.g., all vowels and diphthongs are voiced, consonants can be either voiced or unvoiced). The analysis parameters were set in Praat as follows: the pitch was computed by autocorrelation

[287] Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002

between 44 and 400Hz with an octave jump cost of 3.5 on voice sections defined with a silence threshold of 0.05, a voicing threshold of 0.25, and a voice/unvoice cost of 0.15. Detected pitch was also visually validated by researchers. For this section, we hypothesize that whatever the mode (major or minor), speech from phase 2 might have different Pm, Psd, and HNR than speech from phase 1.

5.2.6 - Results

We report findings on data comparing changes in valence between the major scale and the minor scale group as well as changes of vocal parameters (Pm, Psd, and HNR) induced for both groups by the experience. All t-tests were preceded by an F-test to determine whether the samples should be assumed to have equal or unequal variance and the relevant paired t-test was then run accordingly. The significance level of all tests was set to $p = 0.05$.

5.2.6.1 - Results from Self-Reported Affect

We hypothesized that the minor mode would induce a more negative mood and that the major mode would induce a more positive mood in the subjects. This was evaluated in three different ways. The first was self-reported mood by means of a digital version of the PANAS questionnaire. We observed trends concurring with our hypothesis as the average in valence change was higher for the major scale group (3.3 percent) than for the minor scale group (1.2 percent) However, a two-tailed t-test didn't show statistical significance. It is interesting to note that both group's general mood seemed to slightly increase after the study (with major mode increasing more than minor mode) which might be due to the surprise and novelty effect..

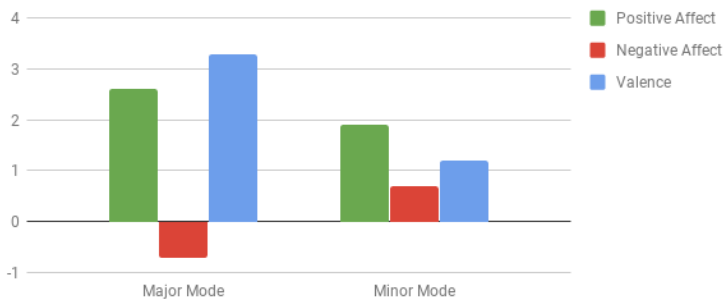


Figure 45: Difference in self-reported positive and negative affect in percentage

5.2.6.2 - Results from Semantic Content

The semantic score analyses conducted on major and minor chord progressions centered around positive, negative, and valence word scores, which give holistic, normalized, numerical attributes of the degree to which the words spoken by a subject leaned more towards positive or negative speech. The valence score was calculated as the sum of the positive and negative scores. We used the Natural Language Toolkit (NLTK) [288] to obtain these scores, and the text was obtained from the subjects' image narratives, from phases 1 and 2.

[288] Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002

We computed the differences in semantic scores from phase 1 to 2 of the study and then compared these across major and minor modes. We used a two-tailed t-test on the valence results as well as on the positive and negative results, and while our results didn't show statistical significance, they still present the expected trends. We specifically observed that subjects from the minor group had a negative score difference (difference in a holistic evaluation of negative words from phase 1 to phase 2), on average almost six times higher than those with the major mode; one-tail two-sample $t(18) = -1.0$, $p = 0.33 > 0.05$. Still, we cannot reject the null hypothesis with respect to semantic results.

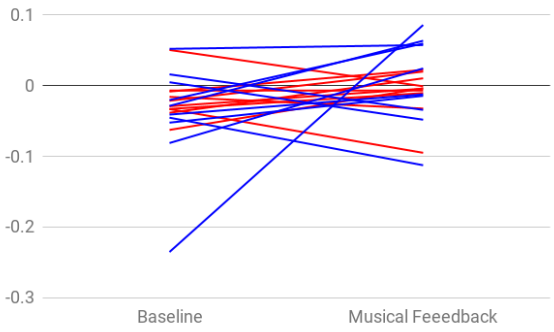


Figure 46: Evolution of semantic score valence (normalized between -1 and 1) between baseline and musical feedback for all participants. The blue lines represent the subjects from the minor group and red lines represent subjects from the major group

5.3.6.3 - Results from Emotion Analysis from Vocal Intonations

The third portion of the analysis was a comparison of the major and minor groups in terms of emotions extracted from the voice. To accomplish this, we used the speech emotion recognition software Vokaturi. As in the previous analyses, the speech used was obtained from the subjects' image narratives, from phases 1 and 2. We grouped the normalized Vokaturi data into three areas: positive affect, negative affect, and valence.

In accordance with our hypothesis, the negative affect score was found to be significantly greater for subjects subjected to the minor mode compared to those subjected to the major mode. The two-tailed t-test, $t(18) = -2.68$,

$p = 0.015 < 0.05$, agrees with this finding and thus we can reject the null hypothesis here. We also localized this difference to vocal parameters indicating sadness and anger, which implies significantly that the minor mode heightens these emotions in the speaker.

Furthermore, compared to subjects from the minor mode, subjects from the major mode saw their valence, or the difference between positive and negative affect scores, increase on average by over five times; two-sample $t(18) = 2.76$, $p = 0.013 < 0.05$. This serves to show that those who listened to the major mode feedback were much more vocally positive than negative, as compared to those with the minor mode. Although not significant, observed trends also suggest that the major mode increases happiness and positive affect in speakers. The significance of these results should also take into account the relatively low accuracy of the OpenVokaturi tool.

5.2.6.4 - Results from Vocal Musical Parameters

When analyzing vocal musical parameters, we hypothesized that regardless of key (major or minor), speech from phase 2 might have different Pm, Psd, and HNR than speech from phase 1.

A paired-samples two-tailed t-test was conducted to compare Pm between baseline and musical feedback conditions. There was no significant difference in the Pm between baseline (M=154.6Hz, SD=45.6Hz) and musical feedback (M=156.6Hz, SD=47.2Hz) conditions; $t(19)=-0.80$, $p = 0.430 > 0.5$. This indicates that fundamental frequencies didn't change much in speakers with or without feedback.

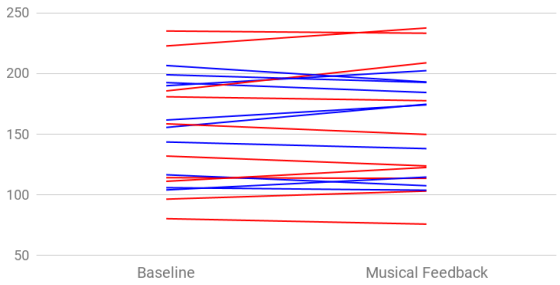
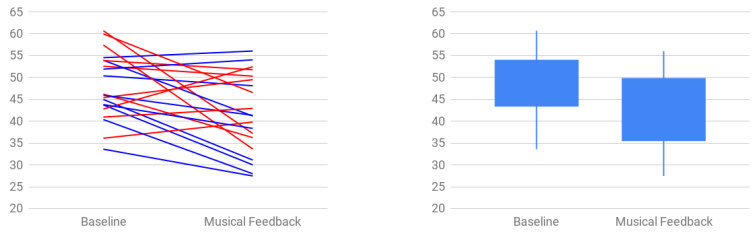


Figure 47: Evolution of mean pitch (in Hz) between baseline and musical feedback for all participants (blue lines for subjects in the minor group and red lines for subjects in the major group

However, significant differences were observed when running a paired-samples two-tailed t-test to compare Psd between baseline (M=47.9Hz, SD=7.5Hz) and musical feedback (M=41.8Hz, SD=9.0Hz) conditions; $t(19)=3.024$, $p = 0.0069 < 0.05$. This result indicates that speakers became slightly more monotonous and pitch envelopes were less accentuated when

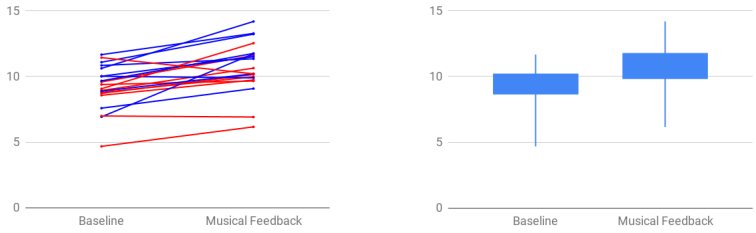
hearing musical feedback. We might have expected that musical feedback would make subjects more melodic but instead, it seems that as the melodic and harmonic matter was added to their speech, they became more conservative in terms of accent, pitch contours, and melody in their own produced speech.



(a) (b)

Figure 48: Pitch standard deviation evolution (in Hz) between the baseline and the musical modes (blue lines for minor group and red lines for major group) (a) and for the entire population (b)

Finally, significant differences were also obtained when running a paired-samples two-tailed t-test to compare HNR between baseline (M=9.2 dB, SD=1.7dB) and musical feedback (M=10.7dB, SD=1.9dB) conditions; $t(19)=-5.0$, $p = 0.000087 < 0.05$. This indicates that the spoken voice becomes more singing-like with a more precise and accentuated pitch.



(a) (b)

Figure 49: Harmonic-to-Noise ratio evolution (in dB) between the baseline and the musical modes for each participant (with blue lines representing the minor group and red lines representing the major group) (a) and for the entire group (b)

Those two results indicate that in terms of timbre, the spoken voice becomes more music-like, but in terms of pitch envelope, the speaker becomes more cautious and conservative. This could indicate that the subjects were distracted; further explorations should assess that potential element. This could also indicate that they were paying more attention to listening and integrating their own voice as music rather than language.

5.2.7 - Discussion and Future Work

When analyzing the data for possible valence and musical effects of musically modulated auditory feedback, we have observed some preliminary results suggesting a trend in the expected direction: self-reported valence became more positive for subjects hearing the major mode than for those

hearing the minor mode, though not to a statistically significant extent, on account of the small sample size. There were no significant results shown in the analysis of semantic content of speech, either, suggesting that, if present at all, cognitive mood change due to major or minor chords is marginal. However, our study showed significant changes in vocal emotionality and in vocal musicality with a higher harmonic-to-noise ratio and lower pitch standard deviation. This suggests that the feedback makes people's voices more song-like while reducing their pitch envelope, and changes their vocal (but not verbal) emotional content. Additional studies would be necessary to better understand these effects and the factors contributing to them.

This exploratory work presents several limitations both in the context and format of the study. Relatively small sample size and possible order effects are elements that have to be addressed in our future studies. The next stage of the work will also include a different type of baseline where the subject hears their voice amplified at the same loudness without any modulation. Another possible comparison could be with a mode where subjects hear music unrelated to their speech, though previous studies have indicated that this might create a high level of distraction. Indeed, listening to music has been shown to be detrimental to short term memory and cognitive tasks such as reading or writing words that the listener has just heard [289]. In our study, it seemed that the modulated feedback didn't seriously affect the ability of the subjects to speak and concentrate. Our subjects seemed sometimes slightly less cognitively and vocally fluent with the feedback but to a lesser degree than one would expect with background music at the same loudness. However, it might still be interesting in the future to assess the level of distraction induced by the system and see how distraction might be reduced when musical stimuli are responsive to user input compared to non-interactive stimuli such as background music.

[289] Pierre Salamé and Alan Baddeley. Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology Section A*

Further investigations are required to test factors such as novelty effects, social dynamic-related bias, or task-induced variability. The musically altered feedback was quite novel and unusual for many, and some subjects found it to be amusing or intriguing. Such reactions would tend to indicate an initial boost of positive affect that could then skew the results and mitigate the expected variations, especially in the minor direction.

In future explorations, we are interested in comparing the reaction to Speech Companions that adjust to the subject's natural voice range, and see if the system can be made to blend even more with the natural musicality of the voice, compared to imposing externally defined musicality onto it. Furthermore, in this study, the vocal modifications were made obvious, and the subjects were informed of the presence and general characteristics of the modifications. It would be of interest to determine whether a more

subtle modification (e.g., lower feedback volume) would have comparable, enhanced, or reduced effects, similar to the way pitch shift compensation has been studied for both uninformed [290] and informed [291] subjects. Finally, due to the number of people surveyed, we did not include a group with no feedback as a control. In future research, we intend to test additional subjects, some of whom will not hear any feedback, while others will only hear background music while they speak. These extensions would help to buttress the findings of this study.

[290] Theresa A Burnett et al. Voice f0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, 1998; and Theresa A Burnett et al. Voice f0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice*, 1997

[291] Kevin G Munhall et al. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, 2009

5.2.8 - Project Conclusion

In this study, we created a new type of audio manipulation to generate real-time changes in the perceived voice through Musically Mediated Auditory Feedback. Classification results significantly indicate that such feedback might alter voice quality and emotion valence detected from voice tonalities. Significant changes in vocal timbre and pitch variation were observed showing the potential to affect speech musicality at a subconscious level.

This early exploration proposed original ways to manipulate the voice in real time as a way to potentially affect internal mental and physical processes in speakers. By musically altering the way people hear their own voice, we also aim to raise questions about the existing underlying effects of musicality already present in the voice and its reinforcing potential in terms of enhanced emotional regulation, self-awareness, and musicality, in the context of everyday speech.

Speech is one of the richest and most ubiquitous modalities of communication used by human beings. Its richness lies in the combination of linguistic and nonlinguistic information it conveys. Musicality is a crucial nonlinguistic components of speech; it includes the tempo and rhythms of the speaker as well as the pitch variation and unique texture of the vocal sounds. Abstracting musicality from a speech in real-time presents several challenges, but explorations in the domain of musically modulated speech and feedback could open doors to explore real-life situations where the music of speech impacts speakers or listeners such as in the contexts of infant-directed speech, language acquisition, human-animal communication, speech pathology, aphasia re-education, or even music learning and musical composition.

In the following section, we explore the potential of musically modulated auditory feedback to affect fluency in people who stutter (PWS)

5.3 - Effects of Speech Companions on Fluency for PWS

> *I regard this activity not so much as a demonstration of a physical fact, but more as a way to smooth out any irregularities my speech might have*
>
> **— Alvin Lucier**
> ("I am sitting in a Room" ending with a reference to Lucier's stutter (1981))

As introduced in the background chapter, stuttering can be seen as a key to exploring the experience and neural pathways associated with the inner voice. But stuttering is more than that. Also called stammering, stuttering exists in every culture and in every language. Although still a mysterious condition, it is thought to be linked to neurological, genetic, and environmental factors.

When the flow of speech is interrupted by repetitions, prolongations, or blockages, it can severely impact communication. Often, people who stutter (PWS) develop secondary symptoms such as eye-blinking or other tic-like movements that can be very stressful and disruptive during interpersonal interactions. PWS can experience a fear of making phone calls or speaking in front of an audience. They may develop low self-esteem, social phobia, and depression[292]. The stigmas associated with stuttering have unfortunately been relayed by movies that mock stuttering especially, some targeted to young people. Such characters include Disney's Donald Duck and Porky Pig from the Looney Tunes (who anecdotally was originally voiced by Joe Dougherty, a PWS, but as he couldn't "control" his stutter, he was replaced by Mel Blanc) [293].

Transient dysfluencies in children are usually distinguished from persistent developmental stuttering (PDS), i.e., stuttering that began in childhood and persisted into adulthood. Indeed, according to the NIH, 5–10 percent of children stutter and 75 percent of them will outgrow it and recover by the age of 12. However, for the remaining 25 percent, stuttering will most often persist as a lifelong condition. Very rarely, someone can develop stuttering from brain trauma or medication (neurogenic stuttering). In most cases, stuttering cannot be treated, but people often learn to manage it[294]. However, one frequent consequence of this communication disorder is to isolate PWS who will even often choose career paths to limit vocal interactions to a minimum[295].

In the last decade, the views on stuttering from speech-language pathologists (SLPs) and patients have evolved. Throughout this project, we were

[292] Gordon W Blood and Ingrid M Blood. Bullying in adolescents who stutter: Communicative competence and self-esteem. *Contemporary Issues in Communication Science and Disorders*, 2004

[293] Marc Shell. Animals that talk. *Differences: A Journal of Feminist Cultural Studies*, 2004

[294] NIH NIDCD. Stuttering. URL https://www.nidcd.nih.gov/health/stuttering

[295] Geraldine Bricker-Katz, Michelle Lincoln, and Steven Cumming. Stuttering and work life: An interpretative phenomenological analysis. *Journal of fluency disorders*, 2013

amazed by the diversity of views PWS have on their own disfluencies. Although stigmas and shameful feeling still exist, there is a movement toward accepting one's stutter and considering it just another way of speaking, rather than something that needs to be treated. The advent of the National Stuttering Association, as well as self-help groups, have been helping with this evolution. SLPs are also slowly modifying their techniques from ways to control or hide disfluencies, to ways to accept them. We highly respect this movement and wish to support a more global societal change to accept stuttering not as a problem that should be fixed, but as a different way to speak that calls for a different way to listen. Stuttering is a condition that mainly affects tempo and speed of communication. Words may take a little bit more time to come. Learning to also take the time to listen without guessing the words to come can actually be a real exercise in humility.

Despite the important effect that stuttering can have on people, it can sometimes become an unexpected source of greatness. Marilyn Monroe recounts in interviews that to overcome severe stuttering, she was advised by her vocal coach to "always whisper/breath her way into sentences" to reduce disfluencies, thus creating her iconic voical style[296]. Alvin Lucier's stutter is often considered the inspiration of his famous piece *I am sitting in a room*, which ends with the sentence "I regard this activity not so much as a demonstration of a physical fact, but more as a way to smooth out any irregularities my speech might have"[297]. Rowan Atkinson learned various techniques such as over-articulation to reduce his stutter. He later on used those techniques to build the character of Mister Bean[298]. Lewis Carroll had a severe stutter, but he found himself vocally fluent when speaking with children[299]. In addition, stuttering—with its repetitions, unexpected tensions, and variations in rhythm—has also been an inspiration for music, with hundreds of songs containing stuttering such as *Barbara Ann* (The Beach Boys), *Bennie And The Jets* (Elton John), *Changes* (David Bowie), *My Generation* (The Who), etc.

[296] Jeffrey Meyers. *The genius and the goddess: Arthur Miller and Marilyn Monroe*. University of Illinois Press, 2012

[297] Alvin Lucier. I am sitting in a room. 2000

[298] Sehar Shoukat. Rowan atkinson to mr. bean: A story of weakness to success-case study. *SJSS)*, 2019

[299] Joseph S Attanasio. The dodo was lewis carroll, you see: Reflections and speculations. *Journal of fluency disorders*, 1987

For all these reasons, we try to be sensitive in this work in the vocabulary used, and talk about increasing fluency rather than "improving" fluency. In this context, the goal of this project is not to "fix" stuttering but to propose:

- A tool based on MMAF to increase fluency that is more efficient than existing techniques
- A tool to increase fluency that might be more pleasant to use than existing techniques
- A tool to help PWS gain appreciation for their own voice
- A tool that might potentially make itself obsolete after helping the user to gain control over the perception pathways of their voice

Our first in-lab study provides a first level of evaluation for the first

three objectives and the proposed longitudinal protocol aims at assessing the feasibility of the fourth objective.

A few questions need to be raised in regards to using such a tool as an assistive system or as a practice system. In the current stage of our research it is still unclear whether our system would be more relevant as an assistive device to increase fluency for a specific context or situation (for instance in stress-evoking situation, such as giving a talk, participating in an interview, or making a phone call) or whether our system would better be used as a practice tool for training at home or during sessions with a speech-language pathologist. This question will be approached in the discussion session and raised through the longitudinal study protocol.

5.3.1 - Introduction

We present the design and evaluation of a system that alters the auditory feedback of the voice in a musical way to help adults who stutter who may wish to improve their fluency and their relationships with their voices. Stuttering is a speech disorder associated with left inferior frontal structural anomalies [300] that affects approximately 1 percent of the population worldwide. This condition is characterized by involuntary repetitions, blockages, and prolongations of sound during speaking[301]. According to Guenther's model [302], stuttering would be caused by discrepancies in the comparisons between the feedback and feedforward signals within the basal ganglia. The model suggests that altered auditory feedback, by acting on the neural bases of speech by disturbing the feedback control signal[303], might alleviate this problem. Prior studies have indeed shown that fluency can be improved by playing back one's own voice fractions of second after it is spoken—this is known as delayed auditory feedback (DAF)[304]. The ideal delay to increase fluency is between 50 and 150ms [305], and a delay above 150ms is inversely known to highly disturb speech rate and fluency even for people who do not stutter[306]. Altering the pitch of one's own voice with frequency-shifted auditory feedback (FAF)—making it sound lower or higher—produces a similar, fluency-evoking effect.

Most previous research conducted on auditory feedback was conducted in the 1980s and almost exclusively on DAF and FAF feedback. Similarly, most available devices and systems only use DAF and FAF, even though technological advances now allow us to create much more interesting real-time modulation of the voice. Indeed, several devices and systems can now be found that use modulated auditory feedback to help stutterers with fluency, but they have flaws in terms of usability, adaptation, level of personalization, social acceptance, and price. Hardware such as SpeechEasy

[300] Anne L Foundas et al. Anomalous anatomy of speech–language areas in adults with persistent developmental stuttering. *Neurology*, 2001

[301] Stuttering Foundation. A nonprofit organization helping those who stutter, 2018. URL https://www.stutteringhelp.org/; and Jane E Prasse et al. Stuttering: an overview. *American family physician*, 2008

[302] Frank H Guenther. *Neural control of speech*. Mit Press, 2016b

[303] Ludo Max, Frank H Guenther, Vincent L Gracco, Satrajit S Ghosh, and Marie E Wallace. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary issues in communication science and disorders*, 2004

[304] Andrew Stuart et al. Effect of monaural and binaural altered auditory feedback on stuttering frequency. *The Journal of the Acoustical Society of America*, 1997; and Anne L Foundas et al. The speecheasy device in stuttering and nonstuttering adults: Fluency effects while speaking and reading. *Brain and language*, 2013

[305] Joseph Kalinowski et al. Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language & Communication Disorders*, 1996

[306] Jennifer Chesters, Ladan Baghai-Ravary, and Riikka Möttönen. The effects of delayed auditory and visual feedback on speech production. *The Journal of the Acoustical Society of America*, 2015

tackles the problem of social acceptability by proposing a very discreet system, but at a very high cost (up to $4000), and the device is not easily personalizable. Several phone applications are now on the market that offer FAF and DAF at a low price, such as Speech4Good ($4.99), DAF Assistant ($12.99), Fonate DAF ($1.99) or Easy AFF ($9.99) [307]. With such applications, the delay and frequency shift is controllable by the users and changeable at any time, but the inherent delay of the system can limit the fluency effect. In addition, such systems work better when used with wired headphones, which can cause problems with social acceptability. In addition to those drawbacks, both delayed and frequency-shifted auditory feedback have proven to be helpful in the short term but not in the long term [308]. This can be explained in small part by social acceptability and discomfort, as DAF and FAF are unpleasant and affect other vocal parameters, but also by neural plasticity, as the brain adapts to the new feedback after a few days or weeks.

[307] Judy Kuster. Some apps for Stuttering. URL https://www.mnsu.edu/comdis/kuster/appsforstuttering.html

[308] Ryan Pollard et al. Effects of the speecheasy on objective and perceived aspects of stuttering: A 6-month, phase i clinical trial in naturalistic environments. *Journal of Speech, Language, and Hearing Research*, 2009

In this project, we introduce a new type of system with two characteristics. First, instead of simply delaying or pitch-shifting the voice, our system creates musically modulated feedback using embedded, custom-made modules that create real-time musical accompaniments of the voice. Several clues indicate the potential of such an approach. First, most adults who stutter are fully fluent when they sing [309]. Secondly, although the theory is still controversial, it is believed that the brain typically predominantly processes speech and music in independent areas (Broca's and Wernicke's areas for production and perception of speech are located in the left hemisphere [310]; music is considered a more "whole-brain" phenomenon, but the perception of music involves mainly the right hemisphere [311]). In addition, it is known that persistent developmental stutterers, who comprise most of the population of adults who stutter, use compensatory behaviors [312] through overactivation in the right hemisphere. Preibish showed that during reading tasks, systematic activation of a single focus in the right frontal operculum (RFO) was negatively correlated with the severity of stuttering. Luc's work on silent and oral single words while reading also supports the hypothesis that stuttering adults show atypical lateralization of language processes [313]. Such previous work seems to indicate the potential for the right hemisphere to help alleviate the limitations of the left hemisphere during speech. In addition, we wish to assess whether musical feedback of the voice can help subjects gain awareness of and training in the timing and rhythms of speech. Indeed, researchers agree that the main dysfunction in PWS is the impaired ability of the basal ganglia to produce correct timing cues to initiate the next motor segment in speech [314].

[309] E Charles Healey et al. Factors contributing to the reduction of stuttering during singing. *Journal of Speech, Language, and Hearing Research*, 1976; and Gavin Andrews et al. Stuttering: Speech pattern characteristics under fluency-inducing conditions. *Journal of Speech, Language, and Hearing Research*, 1982

[310] Wilder Penfield and Lamar Roberts. *Speech and brain mechanisms*, volume 62. Princeton University Press, 2014

[311] Robert J Zatorre, Pascal Belin, and Virginia B Penhune. Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, 6(1):37–46, 2002

[312] Christine Preibisch et al. Evidence for compensation for stuttering by the right frontal operculum. 2003

[313] F Luc, Robert M Kroll, Shitij Kapur, and Sylvain Houle. A positron emission tomography study of silent and oral single word reading in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research*, 2000

Secondly, the device aims to be user-friendly and interactive by providing the user with musical parameters that can be changed. As it is

[314] Per A Alm. Stuttering and the basal ganglia circuits: a critical review of possible relations. *Journal of communication disorders*, 2004

also believed that the brain habituates quickly to changes in auditory feedback, which seems to be why existing devices don't work long-term, we hypothesize that allowing the user to change the parameters of the musical feedback will avoid habituation, as there are infinite ways to transform a signal into music. This personalization also aims at improving the general relationship that PWS may have with their voices.

In this study, we measure the effects on fluency of each mode, including a no-feedback baseline, three control modes (*Raw Voice*, *Delay*, and *Pitch-Shift*), and eight novel acoustic "musically modulated" feedback modes (*Pop*, *Retune*, *Bubble*, *Piano*, *Harmony*, *Reverb*, *Whisper*, and *DJ*). For this study, the modes were implemented on a desktop computer using a combination of custom made DSP systems and off-the-shelf audio software. Some of the mode are transformations classically used in the music industry and others are novel modulation developed for the sole purpose of this dissertation work as explained in the Methodology section. We also collected biographical data and per-subject perception of the modes on their fluency and likeness of their voice to test the following hypotheses:

- H1: MMAF increase fluency compared to baseline
- H2: Some of the MMAF modes have a stronger fluency-evoking effect than the control modes (*Raw Voice*, *Delay* and *Pitch-Shift*)
- H3: The subjects feel more prone to using some of the MMAF modes compared to control modes
- H4: Some of the modes increase the subject's likeness toward their voice

In addition, we collected biographical information as well as interview questions that were used to gather quantitative and qualitative insights into the personal experience of the relationship PWS have with their inner and outer voices.

5.3.2 - Background

Developmental stuttering is a motor-speech disorder characterized by dysfluent speech with hallmark features of sound repetition, blockages, and prolongations[315]. In the mature phenotype, these core features are usually accompanied by secondary behaviors of tension, adventitious movements, and others. The precise types of dysfluency and their severity vary across individuals and within individuals over time, often changing based on the speaking context. Fluency can be modulated by altering aspects of the speaking environment, and is often improved by singing, speaking in-time with others (i.e., choral speech), speaking with an externally-cued rhythm (e.g., metronomic speech), and several others.

[315] O Bloodstein and N Bernstein Ratner. A handbook on stuttering new york. *NY: Thomson-Delmar*, 2008

The neural mechanisms underlying developmental stuttering are not clearly established, although several models have been proposed[316]. An overview of basic principles of the neural organization of language may be helpful to frame the more detailed discussion of stuttering.

[316] Soo-Eun Chang, Emily O Garnett, Andrew Etchell, and Ho Ming Chow. Functional and neuroanatomical bases of developmental stuttering: current insights. *The Neuroscientist*, 2018

A basic organizing principle of the brain is the idea that sensory and perceptual information tends to be represented posteriorly and action/motor representations more anteriorly. This is also true with speech and language processing. The sounds of language—phonology—are processed in the posterior and superior portions of the temporal lobe bilaterally[317], whereas the motor sequences needed to produce sounds are processed anteriorly within a network within the left frontal lobe[318]. Phonological processing is important for translating basic sound elements into language symbols that are used to construct words and phrases. An influential model separates subsequent steps in cortical language processing into two broad functional pathways: a ventral stream that supports "sound-to-meaning" translations, and a dorsal stream support "sound-to-speech" translation[319]. The idea of a dual-stream pathway is broadly consistent with the organization of the visual system, whereby visual information can be used to construct meaning (the ventral "what" pathway) or to guide motor or cognitive actions (the dorsal "where" or "how" pathway).

[317] Gregory Hickok. The functional neuroanatomy of language. *Physics of life reviews*, 2009

[318] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5

[319] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 2007

The ventral language stream supports the brain's ability to use phonological information as a symbolic code whose specific combinations support the activation of specific concepts. The link between phonological word forms and their specific associated concepts (i.e., the "lexical-semantic interface") is believed to occur with the middle temporal gyrus bilaterally (although with a leftward bias), which sits adjacent to phonological processing areas dorsally (the posterior superior temporal areas) and semantic processing areas ventrally (the inferior temporal gyrus). Focal damage to this lexical-semantic interface causes poor verbal comprehension of words; individuals are able to demonstrate intact conceptual knowledge when the ideas are introduced through non-verbal means.

The dorsal language stream provides the link between phonology and motor expression[320]. The connections between certain sounds and the specific motor programs needed to produce them are acquired through trial-and-error feedback during development (and to a lesser extent throughout life). Specific articulatory gestures produce specific sounds; through repeated exposures, the brain learns to link these motor-sensory relationships in a process known as action-perception coupling[321]. Over time, these pathways become bidirectional and tightly co-activated. In adults, hearing certain speech sounds causes activation of both auditory/phonological and motor-articulatory areas. Hearing speech automatically activates the motor

[320] Gregory Hickok, John Houde, and Feng Rong. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, 2011

[321] Wolfgang Prinz. Perception and action planning. *European journal of cognitive psychology*, 1997

plans required to produce those speech sounds. To note, this principle of action-perception coupling is not limited to speech; it is part of a broader "mirroring" system that may be important for imitation learning, empathy, and a host of other important human faculties[322]. The cortical area important for performing this sensory-motor translation is an area within the Sylvian fissure at the junction of the temporal and parietal lobes known as "Spt" [323]. This area has specific connections with articulatory motor areas in the frontal lobe and phonological areas in the posterior superior temporal areas. Damage to this region causes speech abnormalities characterized by errors with repetition and numerous mispronunciation errors (known as phonological paraphasic errors), particularly for longer and more phonologically or phonetically complex words[324]. This is known, aptly, as "conduction aphasia"[325]. This area may also serve as a phonological buffer where phonological information is stored or kept "online." This buffer is important to allow for comprehension of grammatically complex sentences, particularly those with temporally distant subject-verb relationships. Spt damage is also associated with deficits in phonological working memory[326], which is important to support grammatical and articulatory processing.

Motor articulatory planning (e.g., "phonetic encoding") and production are primarily supported by left frontal processes[327]. Fluent speech production requires individuals to plan a specific sequence of articulatory movements at precise temporal intervals. Errors in the order or timing of movements lead to poor articulation. Stored motor programs containing the articulatory motor plans required to produce certain sounds may be located within left lateral premotor areas[328]. Timing signals are initiated by the dorsal medial frontal premotor cortex in an area known as the supplementary motor area (SMA)[329]. The neural connections between the SMA and the lateral premotor areas are known as the frontal aslant tract[330]. The motor plans generated by these premotor connections have direct connections with the so-called "primary" motor cortex whose cell bodies traverse along a descending pathway and directly synapse with cranial nerve and spinal nerves. The motor signals generated by the primary motor cortex are part of the motor effector system that actually produces the skeletal muscle movements supporting speech.

Words appear to be organized into hierarchical structures to denote dependent relationships between them through grammatical processing. This is accomplished by changing the order of words in a sentence (i.e., syntax) and by altering them (i.e., morphology), and both of these processes are supported by left prefrontal areas[331].

This core cortical system supporting speech and language—phonological decoding, the ventral and dorsal streams, phonetic encoding, and gram-

[322] James M Kilner, Karl J Friston, and Chris D Frith. Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 2007

[323] Soo-Eun Chang, Emily O Garnett, Andrew Etchell, and Ho Ming Chow. Functional and neuroanatomical bases of developmental stuttering: current insights. *The Neuroscientist*, 2018

[324] Maria Luisa Gorno-Tempini, Simona Marina Brambati, Valeria Ginex, Jennifer Ogar, Nina F Dronkers, Alessandra Marcone, Daniela Perani, Valentina Garibotto, Stefano F Cappa, and Bruce L Miller. The logopenic/phonological variant of primary progressive aphasia. *Neurology*, 2008

[325] D Frank Benson, William A Sheremata, Remi Bouchard, Joseph M Segarra, Donald Price, and Norman Geschwind. Conduction aphasia: a clinicopathological study. *Archives of Neurology*, 1973

[326] Bradley R Buchsbaum, Juliana Baldo, Kayoko Okada, Karen F Berman, Nina Dronkers, Mark D'Esposito, and Gregory Hickok. Conduction aphasia, sensory-motor integration, and phonological short-term memory–an aggregate analysis of lesion and fmri data. *Brain and language*, 2011

[327] Nina F Dronkers. A new brain region for coordinating speech articulation. *Nature*, 1996

[328] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5

[329] Ingo Hertrich et al. The role of the supplementary motor area for speech and language processing. *Neuroscience & Biobehavioral Reviews*, 2016

[330] Anthony Steven Dick et al. The frontal aslant tract (fat) and its role in speech, language and executive function. *cortex*, 2019

[331] Lorraine K Tyler et al. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*, 2011; and Matti Laine et al. Left hemisphere activation during processing of morphologically complex word forms in adults. *Neuroscience Letters*, 1999

matical processing—do not represent the complete picture. Subcortical structures such as the basal ganglia, cerebellum, and thalamus also contribute. The cerebellum (and its associated subcortical network) is important for the smoothness in motor production. The basal ganglia appear to play an important role in the selection of articulatory actions based on the speaking context[332]. The basal ganglia receive cortical inputs relaying the ongoing state of the speaker. These inputs may originate in auditory/phonological areas, somatosensory areas, motor areas, cognitive areas, and others, which contain information about the speech context. The integration of these signals by the basal ganglia (e.g., pattern detection) supports the selection and initiation (and termination) of specific motor programs. This type of motor selection often occurs "effortlessly," with contextual input driving sequential motor selections in an automated process not requiring top-down conscious control. Damage to the cortical inputs, the basal ganglia itself, or the cortical/motor outputs can lead to speech symptoms.

[332] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5

An influential schema in the neural control of speech is the DIVA/GODIVA model put forth and updated by Frank Guenther and colleagues[333]. An important principle in this model is the notion of both feedforward and feedback regulatory control of speech. In the feedforward model, motor plans are organized in the frontal lobes primarily via the premotor areas (for sequencing and timing) and are executed by motor effector systems, including the motor neurons, cranial and spinal nerves, and muscles. There are associated subcortical networks that contribute to this process, too. The executed actions produce overt sensory perturbations that are detected by the nervous system's sensory/perceptual systems. For the feedback component, frontal motor systems, in addition to sending a feedforward signal to execute the motor plan, also send a copy of the motor plan to the sensory cortices that are used to predict the expected sensory consequences of the proposed action. The sensory predictions are compared with the overt sensory feedback. If there are discrepancies in actual vs predicted sensory signal, that mismatch error is calculated and used to update the ongoing (and subsequent) motor plans. This is the notion of using feedback to guide action.

[333] Jason A Tourville and Frank H Guenther. The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, 2011

Neurological accounts of stuttering conceive of the condition as resulting from dysfunction within cortical-basal ganglia networks[334]. Broadly speaking, cortical processes represent different aspects of the speaking context (e.g., sensory conditions, articulatory conditions, cognitive conditions) or different aspects of the motor plan (e.g., motor sequences, initiation/timing signals), and the basal ganglia recognizes contextual patterns and matches them to motor programs and selects/initiates their cortical timing maps. Problems representing the speaking context due to cortical dysfunction, or in matching context to motor plans due to basal ganglia dysfunction or

[334] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5; and Catherine Theys et al. A crucial role for the cortico-striato-cortical loop in the pathogenesis of stroke-related neurogenic stuttering. *Human Brain Mapping*, 2013

dysregulation, or in motor planning itself from premotor damage, can all lead to symptoms of stuttering.

The notion of left hemisphere cortical dysfunction as a primary driver of stuttering comes from several observations. PWS show an atypical activation of the right hemisphere during speaking tasks. While fluent speakers primarily activate the left perisylvian areas, PWS have bilateral activation[335]. Supporting this functional activity data is the finding that stuttering severity correlates with reduced gray matter and white matter integrity in the left hemisphere (particularly between motor and sensory areas)[336]. Previously normal speakers who acquired stuttering after a stroke are most apt to have left-sided injury (and their subcortical connections)[337]. Apraxia of speech, a dysfluency related to motor-articulatory planning deficits, is associated with left frontal lobe damage[338].

Intrinsic basal ganglia disruption or its dysregulation can also lead to stuttering. Acquired stuttering can occur after basal ganglia damage. The hypothesized role of the basal ganglia is to recognize context and use it to guide the sequential selection and timing of motor plans. The basal ganglia's default setting is to "inhibit" the initiation/termination of action. When the appropriate context (i.e., set of inputs) is detected, the inhibitory setting is released, and the appropriate action selected via activation of the appropriate premotor cortices. The inability to perform this function would lead to prolongations or repetitions of the previous syllable/action, or blockages in motor function, which are the core symptoms of stuttering. These functions depend on the integrity of cortical representations (i.e., how well the brain represents context), their connection with the basal ganglia, the intrinsic machinery of the basal ganglia, and the motor programs available for selection[339].

The mechanisms of basal ganglia functioning are complex and beyond the scope of this review[340], but the neurotransmitter system dopamine plays an important role[341]. Diminished or excessive dopaminergic signaling causes dysregulation of the "gain" switch for action selection, which can lead to changes in the signal-to-noise ratios via regulation of multiple parallel pathways (e.g., direct, indirect, hyperdirect, etc.). Excessive dopaminergic signaling may increase the "noise" and link the detected context to multiple actions instead of one, causing delays in the initiation of the desired action (leading to core symptomatology). Diminished dopaminergic signaling may reduce the "signal," causing the inhibition of action selection. This idea of dopaminergic regulation of the basal ganglia function as a contributing factor in stuttering is supported by evidence that pharmacologic alteration of dopamine signaling with neuroleptics[342] and/or stimulants can impact speech fluency in PWS[343].

[335] Michel Belyk, Shelly Jo Kraft, and Steven Brown. Stuttering as a trait or state–an ale meta-analysis of neuroimaging studies. *European Journal of Neuroscience*, 2015; and AR Braun et al. Altered patterns of cerebral activity during speech and language production in developmental stuttering. an h2 (15) o positron emission tomography study. *Brain: a journal of neurology*, 120, 1997

[336] Shanqing Cai, Jason A Tourville, Deryk S Beal, Joseph S Perkell, Frank H Guenther, and Satrajit S Ghosh. Diffusion imaging of cerebral white matter in persons who stutter: evidence for network-level anomalies. *Frontiers in human neuroscience*, 2014

[337] Catherine Theys et al. A crucial role for the cortico-striato-cortical loop in the pathogenesis of stroke-related neurogenic stuttering. *Human Brain Mapping*, 2013

[338] Jonathan Graff-Radford et al. The neuroanatomy of pure apraxia of speech in stroke. *Brain and language*, 2014

[339] Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5

[340] Marjan Jahanshahi et al. A fronto–striato–subthalamic–pallidal network for goal-directed and habitual inhibition. *Nature Reviews Neuroscience*, 2015

[341] Per A Alm. Stuttering and the basal ganglia circuits: a critical review of possible relations. *Journal of communication disorders*, 2004

[342] Anne K Bothe et al. Stuttering treatment research 1970–2005: I. systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. *American Journal of Speech-Language Pathology*, 2006

[343] Per A Alm. Stuttering and the basal ganglia circuits: a critical review of possible relations. *Journal of communication disorders*, 2004

Fluent speech requires both the representation of context and the ability to use it to initiate sequential actions. Left hemisphere dysfunction, say due to reduced white matter integrity, may cause degradation of the sensory predictions of upcoming motor-speech actions (e.g., errors in the efference copy), producing a mismatch error between predicted and overt sensory consequences of speech. When there is a mismatch of the articulatory plan and the expected sensory consequences, the basal ganglia may not recognize the context and select the appropriate next articulatory action. There is evidence that in PWS, reduced activity in the auditory cortex during speech is associated with increased fluency[344], presumably representing a reduced contribution to auditory feedback processing as a contributor to the speech motor planning by the basal ganglia.

[344] Michel Belyk, Shelly Jo Kraft, and Steven Brown. Stuttering as a trait or state–an ale meta-analysis of neuroimaging studies. *European Journal of Neuroscience*, 2015

Guenther argues that alterations in auditory feedback (e.g., delayed auditory feedback, pitch shifts, etc) produce excessive mismatch errors, which has the effect of reducing the basal ganglia's reliance on this input as a contextual guide for ongoing motor selection. This allows the smaller articulation-sensation mismatch errors associated with stuttering to pass through undetected, thereby enhancing fluency.

The role of the right hemisphere is debated. It is not entirely clear whether this activity is the cause of dysfluency or instead a compensatory response. Stuttering severity appears to correlate with increased right hemisphere white matter tract integrity)[345], suggesting a compensatory role for these fibers. The notion of atypical laterality as a driver of stuttering is longstanding, and originally emerged from observations about left-handedness being overrepresented in stutterers. This cause vs compensation debate remains open, although there appears to be more support for the latter. Activities known to activate right-hemispheric structures, such as singing, are known to be fluency-evoking; there may be other mechanisms behind this. Even if the right hemisphere is compensatory, it is also unclear how effective the compensation is. There is some evidence that recovery may be more effective when accomplished via other mechanisms (e.g., reducing activity in the auditory cortex).

[345] Shanqing Cai, Jason A Tourville, Deryk S Beal, Joseph S Perkell, Frank H Guenther, and Satrajit S Ghosh. Diffusion imaging of cerebral white matter in persons who stutter: evidence for network-level anomalies. *Frontiers in human neuroscience*, 2014

This neuroanatomical background is important for understanding the rationale for the use of altered auditory feedback. Changes in the overt auditory perception of one's voice through the use of altered feedback is a well-known fluency-evoking factor[346]. In these paradigms, overt speech is captured, processed, and returned to the speaker, often using microphones and headphones as the means of direct interface. Well-studied examples include delays, frequency/pitch shifts, and masking effects. Delayed auditory feedback (DAF) and frequency altered feedback (FAF) are known

[346] Michelle Lincoln, Ann Packman, and Mark Onslow. Altered auditory feedback and the treatment of stuttering: A review. *Journal of fluency disorders*, 2006

to reduce stuttering rates during oral reading by 40–85 percent and can increase listeners' perception that the resulting speech sounds "natural." The effects are most robust for oral reading tasks, although they are also observed during monologue. Reductions in fluency are accompanied by only modest benefits in speed rate, suggesting a possible effect of slow rate as a driving factor for DAF [347].DAF, FAF, and their combined use are used to treat stuttering through several commercially available products, although there are currently no FDA-approved devices.

[347] Joy Armson and Michael Kiefte. The effect of speecheasy on stuttering frequency, speech rate, and speech naturalness. *Journal of Fluency Disorders*, 2008

DAF has fluency-evoking effects at delays as short as 50ms[348], although there is some suggestion that longer delays may be more fluency-evoking. Pitch shifts of greater than a quarter of an octave are most effective. There is evidence that commercially available products are effective for some patients, and the effect may be lasting[349]. There are many anecdotal reports suggesting that delays are unpalatable and distracting to the listener.

[348] Joseph Kalinowski et al. Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language & Communication Disorders*, 1996

[349] Andrew Stuart et al. Investigations of the impact of altered auditory feedback in-the-ear devices on the speech of people who stutter: initial fitting and 4-month follow-up. *International Journal of Language & Communication Disorders*, 2004

There are several potential mechanisms underlying the observed reduction in fluency associated with altered auditory feedback. First, musical perception is known to cause widespread, bilateral activity within multiple cortical networks[350]. Altering the spoken voice into a form that is musical may enhance the flow of activity through compensatory networks (right or left hemisphere). PWS almost universally show a large reduction in dysfluency when singing[351], although the mechanisms for this are debated. Another potential mechanism involves the generation of large auditory prediction errors. In these paradigms, the overt sound of one's voice becomes altered externally, leading to a mismatch when compared with one's internal expectations. Very large mismatch errors, as discussed above, may be ignored by the basal ganglia as a guide to action selection, thereby allowing the smaller, stuttering-associated prediction errors to pass through undetected. This latter mechanism is naturally subject to habituation: it adapts and updates its predictive models to account for the externally altered auditory feedback.

[350] T Braun Janzen and MH Thaut. Cerebral organization of music processing, 2018

[351] O Bloodstein and N Bernstein Ratner. A handbook on stuttering new york. *NY: Thomson-Delmar*, 2008

There are other reported fluency-evoking factors in addition to singing, such as whispering, talking in an accent, metronomic speech (i.e., speaking rhythmically to external pacing cues), and others[352]. The mechanisms for these effects are debated, although the observations provide unique opportunities to empirically test different types of feedback.

[352] O Bloodstein and N Bernstein Ratner. A handbook on stuttering new york. *NY: Thomson-Delmar*, 2008

Based on these principles, an ideal feedback system would generate sufficient prediction errors through external auditory manipulations, have a mechanism where the feedback parameters could be changed over time to counteract habituation effects, provide some degree of musicality, and not be overly distracting or bothersome to the listener.

The research protocol employed in this study does not directly answer questions about the neural activity associated with altered auditory feedback. It builds upon previous observations about how auditory perturbations affect stuttering by changing the speech envelope in novel ways.

5.3.3 - MMAF Modes Description

The main novelty of this work is the use of novel vocal feedback modulations, our musically modulated auditory feedback modes (MMAF modes) used in the context of stuttering to alter subjects' voices. Eleven different MMAF modes were tested for each subject, in random order. The digital signals were processed using Max MSP[353] and Reaper 64, a digital audio production application for computers[354].

[353] Cycling74. Max MSP, 1919. URL https://cycling74.com/

[354] Reaper. Digital audio workstation, 2019. URL reaper.fm

We describe each mode from three different angles: technically in terms of how they were implemented; perceptually in regards to how they sound to the user and affects them subjectively; and in terms of expected brain effect, describing how important vocal feedback qualities might be processed (F0: fundamental frequency, Vq: Vocal quality or breathiness, Filter: vocal track shape and how the sources are modified, Env: Environmental and room effects).

The modes are presented here from simple implementation end effects to more complex computation. We start with the three baseline modes used for comparison (*Raw Voice*, *Delay*, and *Pitch-Shift*). These modes are often considered state-of-the-art in auditory feedback used for fluency improvement. We then describe our eight novel modes of auditory feedback (*Reverb*, *Whisper*, *Harmony*, *Bubble*, *DJ*, *Piano*, *Pop* and *Retune*)

Raw Voice

In terms of implementation, the *Raw Voice* mode simply takes the subject's own voice and plays it back to them with minimal modification and latency. The delay is kept under 6ms (buffer size (= 256)/ sample rate (= 44100) = 0.0058). The direct audio feed-through was performed using g. Perceptually, the sound is slightly different from that heard by the subject in everyday conversation, more similar to hearing oneself on a video or recording machine. Indeed, not only is the sound filtered by the audio equipment, but it is also more similar to how others perceive the subject's voice than it is to their own perception of their voice. In terms of neurological response, the slightly foreign aspect could have an effect on the way the brain perceives the voice but overall the feedback should clearly be perceived as the speaker's voice. The slight feeling of estrangement might slightly

affect the brain perception.

Delay

This mode adds a 100ms latency to the *Raw Voice*. We used the delay object in MAX MSP then routed the signal to Reaper and fed it back to the subject. Perceptually, the sound is akin to an echo of one's own voice, or a delay on the phone or video-call conversation, that can be perceived as quite disorienting. The formant, voice qualities, and filter are respected and retained. Several past studies have explored the neurological potential of DAF in alleviating stutter and it has been used in research and in existing commercial products. Delays of 50 to 200ms have previously been reported as effective in enhancing fluency in stutterers[355]. Delay has also been shown to increase disfluencies in people who do not stutter[356].

[355] C Woodruff Starkweather. *Fluency and stuttering.* Prentice-Hall, Inc, 1987

[356] Grant Fairbanks and Newman Guttman. Effects of delayed auditory feedback upon articulation. *Journal of Speech & Hearing Research*, 1958

Pitch-Shift

The *Pitch-Shift* mode (also called frequency-shifted auditory feedback or FAF) was implemented using the Reaper plugin ReaPitch FX. To keep the delay minimal, we used the Simple Windowed Fast variant algorithm (20ms window, 10ms fade), with a -2 semitone shift applied. The delay was 30ms. The two-semitone negative shift was used to keep the speech intelligible and the voice as natural as possible. We choose to shift the voice lower, as it is often reported to be more pleasant for people to hear themselves with a lower, deeper voice than a higher voice. The chosen algorithm shifts all frequencies equally and thus doesn't respect formants. This creates a slightly unnatural voice, however, as the shift was of only two semitones it is not clearly perceivable. FAF has been shown to increase fluency for people who stutter[357]. Previous work has shown good results either a full octave up or down or half an octave up or down. A quarter-octave pitch shift reduces stuttering by about 35 percent [358] However, the fact that our algorithm doesn't respect formant might be perceived as less naturally connected to the speaker's voice. This mode is still clearly connected to the subject's voice and respects envelope and vocal qualities.

[357] Ulrich Natke, Juliane Grosser, and Karl Theodor Kalveram. Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons. *Journal of Fluency Disorders*, 2001

[358] Ulrich Natke and Karl Theodor Kalveram. Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *Journal of Speech, Language, and Hearing Research*, 2001

The following height modes are the novel interventions tested in comparison to the three control modes. Some of those modes (*Reverb*, *Harmony*, *Retune*, *DJ*) are classically used in other contexts such as musical post-production or vocal effects for singers. Others (*Whisper*, *Bubble*, *Piano*, *Pop*) have been designed specifically for this study.

Reverb

To implement the *Reverb* mode, we modeled a plate reverb in the style of Griesinger using Max MSP. We choose a medium room size, average decay

time, and high-frequency damping with a high diffusion rate. The resulting feedback has a 25ms delay. The reverberations added to the voice are akin to the experience of being in a very large empty room or a cathedral, creating slightly unexpected, though familiar, effect. The voice fundamental frequency, other formants, quality, envelope, and other vocal qualities are preserved. This mode only affects the perception of room acoustics. We choose to implement and create this mode because it offers a good balance between familiarity and strangeness. In addition, the effect is perceived as external to the speaker rather than as internal.

Whisper

The *Whisper* mode is obtained by source-filter convolution of the *Raw Voice* signal with pink noise filtered with a low pass filter obtained using the averaged spectral distribution of several whispered voice recordings of men and women. The final signal respects the consonants but imposes the pink noise timbre to the vowels and "voices" sounds. The balance threshold for convolution is chosen for a peak amplitude 10 times higher than the previous 40 samples. The resulting delay is 35 ms. The objective was to obtain an effect similar to someone whispering. The actual result is convincing, though slightly artificial. It creates an effect of a breathy voice and slightly muffled sound—akin to someone whispering close to one's ear. This experience is unexpected but still familiar. Vocal aspects such as envelope, consonants, and rhythm are respected which still renders this feedback mode as very connected to the voice. However, the vocal qualities are transformed into an extreme form of breathiness. The vowels are not voiced anymore. Our decision to implement and include this mode was motivated by two considerations: anecdotal knowledge about the fluency effect of whispering; and the use of easy onset as a fluency technique.

Harmony

The *Harmony* mode is obtained by layering three pitch-shifted versions of the original vocal signal on top of the *Raw Voice* in Max MSP. The delay obtained is 45ms. This created a chorus effect, as if several people were following the speaker's voice in a harmonized fashion. This mode combines the effect of pitch shifting and choir speech by blending an original version of the voice while respecting all of the original vocal parameters, with additional versions where F0 and the formats are transformed but the other parameters remain the same.

Bubble

For the *Bubble* mode, the raw vocal signal was processed using Max MSP to extract a continuous pitch contour of the spoken voice using the yin

algorithm for real-time pitch extraction. The pitch contour is then smoothed and subsequently used to control the sinusoidal oscillator generating a periodic waveform at the voice pitch. Because of the smoothing, this generated an envelope around the attacks creating a perceived delay of 70 ms. The feedback sounds akin to a sinusoidal oscillator following the pitch and amplitude of the voice. One could describe it as reminiscent of a "bubbly" or "underwater" feeling, as if the voice were very muffled. Contrary to all the previously described modes, this mode transforms the voice so much that we suspect the brain might not perceive it as a vocal sound. The mode respects fundamental frequency and amplitude but not the other formants, as it heavily filters the signal into a pure sine tone.

DJ

The *DJ* mode contained the most complex mix of vocal transformation and effects. For this mode, the raw signal was processed using the Reaper plugin VocalSynth2. The custom mode consisted of Distortion, Delay, Filter, and Chorus attributes and Biovox and Polyvox vocal modifiers, in addition to vocal key-correction. The mode generated a delay of 35ms. The feedback can be compared to the sound of a vocoder in the sense that it produces a guttural, deep sound. Furthermore, there is some constantly present background noise in this mode. This was one of the original modes used in the study. We were interested in evaluating how multiple effects and a more complex mix of modulation would affect the brain.

Piano

The *Piano* mode was implemented as a succession of keystrokes matching a subject's voice frequencies and amplitude. Each attack higher than a threshold triggers a MIDI note sent by Max MSP to Reaper. The note's pitch and amplitude are chosen using the amplitude and fundamental frequency of the voice measured using the yin algorithm. The mode measures the pitch 10ms after each attack and plays the closest MIDI note on the C major scale. This generated a delay of 10ms. The mode creates a succession of piano notes synchronized and in pitch-harmony with one's voice. The mode accentuates the rhythm of speech and sounds like a piano being played along to the ebb and flow of one's voice. No prior studies have analyzing the effects of such modulations on stutter. We suspect that even though it is connected to the voice, the sound would not be perceived neurologically as a voice but rather as an accompaniment to the voice, a background effect that accentuated the musicality and rhythm of speech.

Pop

The *Pop* mode is identical to the one used in the previous study described in section 5.2. It was implemented by detecting the pitch of the voice (using the yin algorithm) at each pseudo-beat, mapping the incoming vocal signal to the closest note on the C major scale, and then shifting the forthcoming speech to harmonize with that note, until the next pseudo-beat a few seconds later. We used the Reaper Plugin MHarmonizerMB to implement the mode with a 50ms delay. The *Pop* mode is very uplifting and jocular, so much so that it could draw attention away from the contents of the actual speech itself and more towards the musical aspect. It can be compared to using a sort of autotune on one's voice. This original mode was implemented for this study. There have not been previous studies investigating the effects of such modulations on fluency. This mode contains a signal very similar to the original voice preserving the voice-like quality (F0, formants, amplitude, quality, room amplitude) but also introduces additional layers of harmony (with modified F0, formants) as well as an element of rhythm through the use of pseudo-beats that highlight the natural rhythm of the voice.

Retune

Regarding the technical implementation of *Retune*, we used MAX MSP to implement the mode by matching the subject's fundamental frequency to the nearest note on the C major scale using the yin algorithm and a windowed pitch shifter. The mode had a delay of 45ms. This mode is a very subtle change to the voice, almost indistinguishable from one's own voice for the listener. Indeed, several subjects asked us if this mode was identical to *Raw Voice*. It does, however, have a nuanced aspect of pitch shift, and thus could potentially help subject to focus on the musicality in their voice. The effect is most obvious during pitch swipe as the subject would hear their voice quantized into semitonal steps. Though this mode doesn't technically preserve formants, the very slight shift in pitch doesn't significantly affect the vocal qualities nor the room acoustics. We expect that this mode would be perceived by the brain quite similarly to the *Raw Voice* mode while making the voice seem more musical.

5.3.4 - Materials and Methods

Participants

Twenty four adults who self-identify as stutterers consented to take part in the study. Participants were recruited through flyers and the mailing lists of the Boston chapter of the National Stuttering Association. The majority of participants were from the Cambridge/Boston area, and all were proficient English speakers. All participants gave written informed consent to the study in person at the research location, under the protocols approved

by the Massachusetts Institute of Technology Committee on the Use of Humans as Experimental Subjects (protocol number: 1802248976R001). To be considered an adult with persistent developmental stuttering for purposes of this study, each participant had to fulfill the following criteria: (1) being self-identified as an adult who stutters and (2) exhibiting a minimum of 2.5 percent of SLD/syllables during the baseline fluency evaluation. The final group used for data analysis consisted of 16 individuals (4 women, two left-handed, mean (SD) age = 31 ($\pm$ 7) years). Participants were only excluded if they exhibited less than a minimum of 2.5 percent of SLD/syllables during baseline fluency evaluation.

Environment Setup

The study was conducted in a soundproof room to reduce distraction and background noise, and help data analysis. Subjects were seated comfortably across a table from the researcher administering the study. An assistant researcher was also present in the room to take notes. The researcher faced a computer monitor (with the screen not in the subject's view) and two GoPro HERO 4 cameras were placed in the room to record the subject. One camera was directly facing the subject and the other recorded them on their left side. The cameras were small and inconspicuous so as to not distract the subjects, but subjects were informed of the recording and consented to it prior to the study. After reading and signing the consent form, and before starting the study per-se, each subject was fitted with a microphone and a pair of in-ear headphones. The microphone used was a Countryman E6 Directional Earset for speaking, placed on the left side at about 1.5cm from the corner of the subject's mouth. If the subject had a beard, the microphone was shaped to not be in direct contact with the skin, so as to avoid crackling sounds from friction.

The microphone output was fed to an audio mixer (RME Babyface) and routed to a Mac Mini processor using Max MSP and Reaper 64 before being returned to the subject's ears through a pair of Bose SoundSport headphones to provide audio feedback. The SoundSport headphones are very open (i.e., easily let outside sound in), which allowed the interactions between the subject and the researcher to remain natural. The researcher giving the instructions also wore a pair of SoundSport headphones to monitor the quality of the feedback heard by the subject. The loudness of the feedback was set just loud enough to effectively mask the subject's voice without being unnaturally loud. Subjects' raw voice input and processed voice speech samples were recorded internally in Reaper and also video-recorded with two cameras (GoPro HERO4).

Experimental Procedures

When the subjects arrived at the testing site, they were first offered a glass of water and sat with the experimenter for a few minutes to answer possible questions and to limit the novelty effect and stress of talking with an unfamiliar person. The experimental procedure then contained four parts: preliminary questionnaire, baseline, experimental conditions, and interviews. The system is only turned on during part 3 (experimental conditions), but the microphone and headphones were worn during all sections to habituate the subject to wearing them and avoid creating additional variables during the procedure. The four sections are summarized below:

Preliminary Questionnaire: After being given a short introduction to the purpose of the study and being fit with the microphone and earphones, subjects were asked to fill out a preliminary questionnaire regarding their history of stuttering, demographic/health information, and musical background. Some of the questions included in this section were presented to accrue information on the nature of a participant's stutter in ways that may not otherwise have been discussed; i.e. whether they stutter in their dreams and whether they have a sense of an impending stutter. The results of this questionnaire are expounded further in the data analysis section. Subjects spent approximately eight minutes answering the questionnaire

Fluency Baseline: Subsequently, the subjects were given an assessment for the purpose of determining their fluency baseline. The baseline ran for approximately 15 minutes and was adapted from the SSI4. It contained neutral open questions, image description tasks, reading tasks, word repetition tasks, diadochokinetic tasks, singing tasks, and a metronomic speech task. The baseline speech task was acquired before any of the other conditions with the device in place.

Main Core using MMAF Modes: The core of the study consisted of testing the effect of the feedback modes on fluency. Each subject started with *Raw Voice* mode followed by the 10 other modes in random order. Each mode testing section lasted from five to seven minutes and contained a shortened version of the baseline fluency assessment. It was composed of four tasks presented in random order: open question, reading task, vocal diadochokinesia, and word repetition. Most of the testing was composed of the questions and the reading tasks that represented 90 percent of the testing time. After each mode, subjects were presented with a short questionnaire about the mode to assess their self-perception of their fluency, distraction and potential use of the mode in real-life. Halfway through the modes (about 30 minutes), the subjects were given a five-minute break.

Interview: After the mode testing, the subjects were given a 10-minute

interview to assess their experience. Open questions included: "How did you find the experience?", "Which was your favorite mode and why?", "Is there any mode you would add/remove?", etc. For the interview, subjects were asked to continue wearing the microphone and headphones. The interview recordings were also used to assess their fluency after using the modes. The entire study ran for about 90 minutes, sometimes up to 120 minutes for subjects with severe dysfluencies.

5.3.5 - Data analysis

The data analysis consists of evaluating three primary data streams acquired through the course of the experiment: the preliminary questionnaire data, the mode-review questionnaire, and the audiovisual recordings of the study. The audio recordings of the study sessions were used to generate a script of the entire experiment. We used Dragon NaturallySpeaking to generate a first draft of the text. Two data coders manually corrected the text and time-stamped the different sections and subsections of the experiment using the video recordings.

For each participant and each task, we counted the number of stuttering-like-dysfluencies (SLD) characterized as sound/syllable repetitions, monosyllabic whole-word repetitions, sound prolongations, or blocks (i.e., inaudible or silent fixations or inability to initiate sounds) [359], similar to the coding used in the SSI-4[360]. The primary and secondary coders (FA and SA) received training using the fourth edition of the Barry Guitar /textitManual on Stuttering[361]. Each SLD was manually identified on the script using the video recordings by FA or SA.

To assess inter-rater reliability, two experimental sessions were independently encoded by both the principal data coder and by a second coder. The principal data coder measured the same recordings a second time to assess intra-rater reliability. For both inter- and intra-judge reliability, the Pearson product-moment correlation coefficient was computed. Reliability for judgments of SLDs was good for both inter-rater reliability (r = .955) and intra-rater reliability (r = .997).

Once the SLD was identified, we established a count of dysfluencies per syllable (SLD/syl), per minute (SLD/m), and per word (SLD/w) for each subsection. The syllable count was obtained using an online syllable counter and the time was measured using the time-stamps of each section. For each subject, we obtain a quantitative assessment of their SLD/syl for each tasks (simplified into "reading tasks" and "'speaking tasks") and for each mode (11 feedback modes + initial baseline testing + final interview with no feedback).

[359] Mark W Pellowski and Edward G Conture. Characteristics of speech disfluency and stuttering behaviors in 3-and 4-year-old children. *Journal of Speech, Language, and Hearing Research*, 2002

[360] G Riley. Ssi-4 stuttering severity instrument fourth edition, 2009

[361] Barry Guitar. *Stuttering: An integrated approach to its nature and treatment*. Lippincott Williams & Wilkins, 2013

To assess the effect of the modes on fluency we computed the percentage of fluency variation from baseline for each mode M as

$$DeltaF_{Mode} = 1 - \frac{[SLD/syl]_{Mode}}{[SLD/syl]_{Baseline}}$$

A $DeltaF_{Mode}$ of 0 percent means that the fluency remained unchanged. A positive $DeltaF_{Mode}$ means that fluency increased while using the mode (i.e. $DeltaF_{Mode}$=50% means that the number of dysfluencies per syllable was divided by two, and $DeltaF_{Mode}$=50% means that, when using the mode, the subject did not have any dysfluency). A negative $DeltaF_{Mode}$ means that, when using the mode, the subject had a reduced fluency compared to baseline.

5.3.6 - Results

Fluency modifications

Figure 50 shows the $DeltaF_{Mode}$ results for each mode and for each subject. The results seem to indicate an increased fluency for most modes compared to baseline that is consistent for all the subjects (average increase for all modes of 46 percent, for control modes of 34 percent, for all new modes of 51 percent and for the best mode of 67 percent over all subjects)

Our first hypothesis H1 was to test whether Musically Modulated Auditory Feedback increase fluency compared to baseline. To evaluate the significance of fluency increase between modes and baseline, as well as comparisons amongst modes, we ran a series of Wilcoxon sign-ranked tests on the paired SLD/syl values. We obtain a matrix of p values for all the pairwise comparisons shown in figure reffig:origpmatrix. To control for multiple comparisons, we used a corrected p threshold of 0.01 instead of 0.05.

DeltaF	Baseline	NF_final	Raw voice	Delay	Pitch Shift	Bubbles	Pop	Retune	DJ	Piano	Harmony	Reverb	Whisper
MM_01	0.00%	54.21%	28.32%	44.80%	48.99%	36.23%	88.71%	79.92%	38.33%	25.30%	68.87%	51.49%	79.72%
MM_02	0.00%	35.61%	11.02%	55.41%	59.96%	71.31%	68.49%	100.00%	71.23%	65.11%	89.53%	47.27%	82.64%
MM_05	0.00%	53.35%	6.61%	-18.87%	-38.45%	10.16%	6.42%	18.76%	35.09%	-4.39%	6.69%	43.84%	27.61%
MM_08	0.00%	55.84%	24.25%	13.40%	50.06%	66.68%	5.04%	56.78%	43.30%	-25.45%	73.04%	37.69%	52.08%
MM_10	0.00%	76.59%	77.76%	13.76%	13.50%	50.55%	50.38%	63.31%	100.00%	13.66%	100.00%	64.01%	78.76%
MM_11	0.00%	8.86%	58.46%	83.41%	22.15%	80.22%	32.58%	71.51%	16.57%	-4.42%	100.00%	53.61%	41.54%
MM_13	0.00%	19.60%	49.13%	84.49%	80.92%	71.99%	100.00%	85.35%	86.30%	30.88%	56.33%	92.28%	92.81%
MM_14	0.00%	41.08%	55.71%	57.05%	27.15%	66.75%	79.61%	29.85%	48.25%	16.49%	61.46%	52.60%	63.16%
MM_15	0.00%	10.38%	69.02%	26.01%	62.31%	89.65%	23.81%	44.87%	59.27%	-13.74%	56.31%	90.61%	79.64%
MM_16	0.00%	5.87%	22.65%	54.75%	8.90%	7.44%	42.52%	44.07%	-17.69%	57.78%	66.43%	63.10%	49.26%
MM_17	0.00%	40.84%	21.68%	35.40%	34.71%	11.14%	50.87%	28.82%	35.00%	18.60%	46.66%	47.50%	37.65%
MM_18	0.00%	4.32%	-7.03%	16.83%	3.78%	1.73%	25.57%	40.23%	27.52%	-38.76%	58.76%	86.21%	66.28%
MM_19	0.00%	10.30%	45.63%	48.50%	89.60%	34.09%	54.61%	68.88%	77.53%	50.41%	90.15%	67.91%	62.40%
MM_20	0.00%	18.99%	16.32%	-14.47%	-7.88%	20.89%	12.66%	14.53%	40.85%	26.74%	38.80%	65.44%	38.40%
MM_23	0.00%	55.22%	-7.95%	52.29%	37.02%	26.53%	62.38%	56.23%	61.72%	27.42%	79.07%	80.97%	76.76%
MM_24	0.00%	50.12%	15.83%	49.37%	70.56%	58.08%	72.74%	74.30%	74.23%	29.12%	79.70%	82.81%	83.58%

Figure 50: evolution of fluency (measured in SLD/syl) compared to baseline for each mode and for each subject. Dark green represents a major fluency increase (75 to 100%) while a red color represents a decrease of fluency compared to baseline

Wilcoxon signed rank test														
p Matrix	Baseline	NF_final	Raw voice	Delay	Pitch Shift	Bubbles	Pop	Retune	DJ	Piano	Harmony	Reverb	Whisper	
Baseline	X	4.38E-04	7.76E-04	0.0072	0.0097	4.38E-04	4.38E-04	4.38E-04	6.43E-04	0.0151	4.38E-04	4.38E-04	4.38E-04	
NF_final	X	X	0.5014	0.8361	0.9588	0.5014	0.0703	0.0703	0.1208	0.0627	0.0072	0.0084	0.0113	p < 0.01
Raw voice	X	X	X	0.438	0.5695	0.034	0.0437	0.02	0.0113	0.1961	0.0011	0.0013	6.43E-04	p < 0.05
Delay	X	X	X	X	0.7564	0.5695	0.0113	0.0229	0.1208	0.0703	0.0032	0.0027	0.0023	
Pitch Shift	X	X	X	X	X	0.4691	0.0557	0.0131	0.1477	0.1477	0.0045	0.0032	0.0016	
Bubbles	X	X	X	X	X	X	0.2343	0.098	0.1477	0.0229	0.0131	0.0097	0.0045	
Pop	X	X	X	X	X	X	X	0.4691	0.796	0.0027	0.1337	0.1477	0.0879	
Retune	X	X	X	X	X	X	X	X	0.8361	0.0016	0.1627	0.0879	0.0229	
DJ	X	X	X	X	X	X	X	X	X	0.0038	0.0946	0.0151	0.0386	
Piano	X	X	X	X	X	X	X	X	X	X	4.38E-04	6.43E-04	5.31E-04	
Harmony	X	X	X	X	X	X	X	X	X	X	X	0.8767	0.9588	
Reverb	X	X	X	X	X	X	X	X	X	X	X	X	0.5349	
Whisper	X	X	X	X	X	X	X	X	X	X	X	X	X	

Figure 51: p matrix showing the statistical significance of pairwise comparisons between modes. A light green shows a 50% confidence interval and dark green shows 99% confidence interval

The Wilcoxon signed rank test shows that the observed difference between all but one of the MMAF modes (mean SLD/syl = 2.8% for *Bubble* mode, 2.3% for *Pop*, 2.3% for *Retune*, 2.4% for *DJ*, 1.9% for *Harmony*, 1.8% for *Reverb*, and 1.7% for *Whisper*) and the baseline (meanSLD/syl = 4.7%) are significant, p<=0.0006 for each of those modes.

The *Piano* mode shows a slight increase in fluency (mean SLD/syl= 3.8%) compared to baseline (mean SLD/syl = 4.7%) but not significantly if using the corrected p value threshold, p=0.015>0.01 so we can not reject the null hypothesis for this mode.

However, for H1, we can reject the null hypothesis that, for all but one mode, the samples are from the same population, and we might assume that most of the MMAF modes (namely, *Bubble*, *Pop*, *Retune*, *DJ*, *Harmony*, *Reverb* and *Whisper*) caused a significant increase in fluency.

Our second hypothesis H2 was to assess whether some of the MMAF modes have a stronger fluency-evoking effect that the control modes (*Raw Voice*, *Delay* and *Pitch-Shift*)

Our three best candidates, ranked 1, 2 and 3 in average fluency increase, are *Harmony* (mean SLD/syl = 1.9%) , *Reverb* (mean SLD/syl = 1.8%), and *Whisper* (mean SLD/syl = 1.7%).

The Wilcoxon signed rank test shows that the observed difference between our three MMAF modes and the three control modes (meanSLD/syl = 3.4% for *Raw Voice*, 3.1% for *Delay*, and 3.2% for *Pitch-Shift*) are all significant, p<=0.004 for each of the pairwise comparisons.

Thus, we can reject the null hypothesis for H2 and and assume that our three best modes (*Reverb*, *Harmony* and *Whisper*) are significantly more fluency-inducing than the three control modes (*Raw Voice*, *Delay* and *Pitch-Shift*)

Personal preferences and usability

Quantitative insights: To assess personal preferences and usability of the system, we used the data from the mode-review questionnaire that participants filled immediately after each mode. The questionnaire contained four quantitative questions scored from 1 to 5 and one open comment field. The questions were related to the perceived fluency and distraction of the modes as well as the participant's appreciation of their voice with the mode and finally the projected use of the mode in real-life.

On average, the *Reverb* mode ranked best on perceived fluency; it ranked equal best with *Raw Voice* on projected usability and voice likeness and ranked second after *Raw Voice* on distraction. Although pairwise comparisons using sign-ranked tests didn't show significance, this is still an interesting trend. The *Piano* mode systematically ranked last on all questions.

Given our small sample size, clustering subjects by gender, age, education or musical background is difficult as both groups would be very small also with very different group sizes. However we believed that it could be interesting to look at differences in personal reactions by age given the recent trends in stuttering acceptance rather than hiding. We split the subjects into two distinct groups, on the median age of the test population (30 years), resulting in eight subjects with age less than 30, and eight with age above 30. We then performed a comparison over questionnaire responses to the *Reverb* mode, to observe response trends differentiating the older half of the population from the younger. Among the four questions in the questionnaire, two had the most interesting results: "Did you feel your fluency increased when using this mode?" and "Could you see yourself using this mode in your everyday life?". For the former, using a two-sample t-test with a one-tailed p-value of 0.032, it was observed that the younger population felt their fluency increased with the *Reverb* mode by 1.375 points more, on average, than the older population. As for the second question, while we did not obtain a sufficiently significant result ($p = 0.069$) to assert statistical relevance, it is still interesting to note that, on average, the younger population rated the usability of the Reverb mode in everyday life 1.125 points higher than the older population.

Qualitative insights: In addition to the numerical questionnaire data ascertained from the study, we also asked participants to submit their own additional feedback on the modes. Using these additional comments, we compiled a "canonical comment" for each mode, encapsulating all of the user feedback. Subjects often gave insightful comments, both about how the modes affected their speech and about the nature of the mode on its own.

For many modes, including *Whisper* and *Bubbles*, there was a stark division in comments regarding how distracting the mode was, with some subjects calling the modes extremely distracting, and others saying there was very little distraction. Other modes had comments relating more to the actual musical content of the mode itself. For example, one subject described the Harmony mode as akin to having a "companion/ghost" accompaniment to their voice, and the *Pitch Shift* as having an eerie sound. Other experiential comments were noticeable such as one subject liking the *Reverb* mode as it gave them the sense of being in a large space inside their head making them more free and relaxed.

Demographics and Personal Experience of PWS

The subject demographics were obtained via the preliminary questionnaire data on the 16 participants, and they provide insights into the population of the study. With respect to the gender of participants, 75 percent identify as male, while 25 percent identify as female—a statistic which is corroborated by data on the prevalence of stuttering among different genders. The subjects ranged in age from 21 to 45 years of age, and two were left-handed. All study participants had prior college education, and 31.2 percent were married or in a domestic partnership. With regard to their musical background, all participants listen to music, and 68.8 percent do so daily. Furthermore, all participants regularly sing aloud to themselves, with half of them reporting singing aloud to themselves daily, and 57.2 percent singing in front of others regularly. Regarding their stuttering history, more than half (73.3 percent) of participants were currently attending or had attended speech therapy.

5.3.7 - Discussion

This study re-demonstrates the well-described observation that altered auditory feedback can lead to increased fluency in people who stutter. All modes except one (the *Piano* mode) were associated with increased fluency when compared with baseline. Hearing one's voice through headphones, even without any superimposed digital acoustic transformations, leads to increased fluency. This may be because recorded vocal sounds delivered via headphone speakers (as opposed to travelling through the air from the mouth to the ear) is a form of altered acoustic feedback and is expected to change vocal perception. This supports the idea that many types of acoustic transformations, even those that do not fundamentally alter the speech sounds or timing, may be able to increase fluency.

The *Piano* mode provided altered auditory feedback but was not associated with a significant increase in fluency. In a third of the participants, fluency was even reduced. This may suggest that not all types of auditory

feedback are fluency-evoking and that the effect depends in part on specific transformations. The piano mode removes the nuances of the speech sounds and sends its output as a piano sound matched for pitch and rhythm only. This acoustic transformation has lost much of its connection to the original speech and was rated significantly as highly distracting. This might suggest that alterations in auditory feedback require an optimal level of overlapping sonic features with original speech action for maximal effectiveness.

The data suggest that not all effective forms are equally effective at increasing fluency. Three modes (*Reverb*, *Harmony*, and *Whisper*) demonstrated increased fluency beyond that of control auditory feedback modes (*Delay*, *Pitch Shift*, *Raw Voice*). The other modes were not significantly different from these control modes. This supports the notion that the fluency effects of AAF may be sensitive to specific combinations of transformations. Further exploration of this effect using different types of acoustic transformations along multiple combinatorial axes appears to be underexplored and is potential opportunity for further study. Our MMAF modes demonstrate this potential, although the study was not designed to find the optimal combination of parameters to maximize the fluency effect.

The study did not specifically isolate why certain modes were more effective at increasing fluency than others. Shared features of the three modes include perseveration of the natural rhythmic/temporal components of the voice, conservation of attack characteristics, and a sense of familiarity of sound (i.e., having previously heard one's voice like this). Regarding the latter point, most people have whispered, heard their voice reverberating in an empty room, or sung together with other voices (e.g., choral effect) at some point prior to the study. These perceptual changes, while certainly not the usual manner of speaking, are also not totally unfamiliar to the listener. The other modes— *Delays*, basic *Pitch Shift*, *DJ*, *Pop*, *Retune*, and *Piano*—are likely to be unfamiliar vocal self-perceptions. The degree of familiarity may follow a U-shaped curve in terms of its effectiveness. This requires further exploration. It is interesting to note that the *Whisper* and musical modes mirror the anecdotal reports of these ways of speaking being fluency-evoking in stuttering.

Although the study did not address this directly, there are likely to be individual differences in the fluency effects of specific combinations of acoustic transformations. Some subjects may respond well to certain modes and not others. Given the expected habituation effects, an individual's responsiveness to a given mode may diminish over time. The ability to offer continuous changes in the types of auditory feedback provided could offer a method to overcome these effects. Altering the musicality of the modes, by shifting combinations of pitches, delays, adding white noise or

reverberation, or changing vocal timbre, could provide a nearly limitless set of updates.

Speech samples recorded after exposure to the MMAF modes (e.g., post-testing) were associated with increased fluency relative to pre-testing controls. The reasons for this are uncertain. It is possible that the modes' fluency effects persisted because of adaptations in the speech system due to learning. If true, this raises the potential for MMAF to be used as a training method to induce changes in the motor-speech control systems. The current, dominant conception of AAF in stuttering is a method to be used when one wishes to speak more fluently. There is some evidence that the effect of simple delays or pitch shifts do not persist after discontinuing the use; this has been explored using our multi-dimensional MMAF modes. An alternative explanation for improved post-testing fluency is one of social interaction. Fluency is known to increase in settings where there is lower social anxiety; the time spent with the examiner may have changed the social dynamic in such a way that led to fluency increases. However, we aimed to partially account for this by having subjects arrive earlier to interact with and ask questions to the experimenter for 10 minutes before the study. Future studies should consider a non-MMAF mode baseline test embedded within the randomization structure to account for this uncertainty. Notably, the fluency increases observed in the study cannot be fully accounted for by social factors, as the post-testing fluency effects were significantly less than the three most effective modes.

Other limitations of the study include our controls. The *Pitch-Shift* mode, for example, utilized an interval that was not the most fluency-evoking in literature controls. This was an oversight in the study design and limits the certainty of the findings as a comparison against historical controls. We also utilized an online syllable counter to quantify the number of syllables used. While this was uniform across subjects, this introduces a variable of uncertainty to the results.

In terms of user experience, the MMAF modes did not differ significantly in how participants rated their subjective fluency, usability, likeness, or distractibility. Unsurprisingly, the trends demonstrated the untransformed *Raw Voice* mode to be the most usable, liked, and least distractible. *Reverb* trended highest on these subjective ratings, which was slightly higher in younger participants compared with those of more advanced age. These findings appear to demonstrate comparably with current state-of-the-art alterations in terms of their listener experience. To note, user commentary about the modes suggested a highly-variable user experience, offering further support for the aim of personalization based on fluency assessments and user tolerability.

There were a number of interesting findings within the questionnaires. Given the relationship of music and fluency, we asked about participants' relationship with music. The subjects in the study were overall highly involved in music, with almost 70 percent reporting daily listening and over half who sing or play a musical instrument. These rates of musical participation are higher than what is reported for US residents as a whole (citation). Half the participants reported singing out loud to themselves on a daily basis, and more sing to themselves covertly. Given these observations and the ability of singing to modulate fluency, it is tempting to consider that music and stuttering must be deeply related and intertwined.

While almost no participants reported stuttering in their covert speech, it is worthy of note that almost half stutter while dreaming. This shed light on the potential differences between internal perception of intentional covert speech production and dream state speech. That dream state speech is more like overt speech than covert speech may suggest that dream states are more akin to generated episodic memory or imagination states.

This study does not address the neurological mechanisms underlying MMAF-related fluency effects. This is an important area for future investigation. That most MMAF modes led to increased fluency suggests the normal mechanisms of motor-speech selection and timing are generated by AAF. This may include large auditory prediction errors (expected vs overt) such that the auditory inputs to motor selection are ignored or inhibited when using the speech context to guide the timing of motor programs. Problems in action selection are felt to be related to an unrecognizable context whereby the basal ganglia do not detect enough features to guide the motor program selection. In theory, problems could be due to small errors in a large number of contextual features, or a large error in a small number of features. The most effective MMAF modes, we suggest, are those that are somewhat familiar to the listener and that hold some critical features of speech constant (e.g., attack, timing). These type of MMAF produces a largely recognizable context with a few features containing large prediction errors. While speculative, this observation may provide a clue as to how the brain utilizes certain aspects of context and the deviation from it to facilitate fluency in stuttering.

5.3.8 - Conclusion

Persistent developmental stuttering is a complex motor-speech disturbance characterized by disfluent speech and other secondary behaviors. Stuttering phenotypes and severity vary across and within individuals over time and

are affected by the speaking context. In people who stutter, speech fluency can be improved by altering the auditory feedback associated with overt self-generated speech. This is accomplished by modulating the vocal acoustic signal and playing it back to the speaker in real-time. Most research to date has focused on simple delays and pitch shifts. New embedded systems, technologies, and software enable a re-evaluation and augmentation of the shifted feedback ideas. The current study explores alternative, novel modulations to the acoustic signal with the goal of improving fluency. In summary, ten out of eleven experimental modes yielded improvement in fluency compared with the pre-testing baseline, suggesting most types of headphone-based auditory feedback to be fluency-inducing. Compared with the "raw voice" mode, three feedback modes were associated with statistically significant fluency benefits ("whisper", "harmony", and "reverb"), suggesting a fluency benefit of these acoustic transformations beyond that of merely providing feedback. These modes were more fluency-evoking than simple pitch shifts or delays. Post-testing speech was associated with significant improvements in fluency compared to the pre-testing baseline, suggesting the procedure itself to yield persistent short-term fluency benefits even in the absence of ongoing acoustic feedback. This post-testing fluency benefit was robust but significantly less than the modes "harmony", "reverb", and "whisper".

Our study re-demonstrates the well-described fluency benefits of altered acoustic feedback and extends this finding to novel acoustic transformations with stronger effect sizes. While the temporal persistence of fluency in these modes remains uncertain (and requires longitudinal study), the identification of multiple fluency-evoking feedback modes may offer the potential to overcome the habituation effects and intolerable listening experiences that limit the effectiveness of existing feedback technologies.

5.4 - Mumble Melody in the context of Vocal Connection

This chapter presented the Mumble Melody initiative aiming at shaping a better understanding of the relationship between inner and outer voices within the framing of our second paradigm. Specifically for the work on stuttering, the objectvies of our system were to create a tool based on MMAF that increase fluency more efficiently and in a more palatable way than existing techniques. Beyond short term effects based on subconscious neural response, we are also very interested in exploring whether such a tool could potentially make itself obsolete after helping the user to gain control over the perception pathways of their voice. By asking subjects not only to simpley react to the feedback but somehow internalize mental feedback and learn to modulate their vocal intend accordingly. This hypothesis requiers longitudinal testing. A study of this kind could be an important step toward evaluating possible gain of agency over inner and outer voice pathway.

The Mumble Melody initiative covers several key areas of our Vocal Connection space. The initiative can be seen as a way to examine some aspects of our second paradigm, according to which our experience of the words is strongly informed by our experience of voices. Indeed, it highly touches on the experiential part of the voice, the membrane between inner and outer voices while using concepts from the holistic voice paradigm and from the grooming talking framework. The project arises from a personal view on the voice - the intimate relationship people have with their own voice - but also covers the interpersonal context, as it influences dialogues and someone's ability to hold fluid conversations. Figure 52 frames the Mumble Melody initiative in the context of the thesis. By creating a bridge between those different archetypes of the voice, this project bring insights on the potential to use the voice as a way to access the mind in novel ways using Musically Modulated Auditory Feedback.

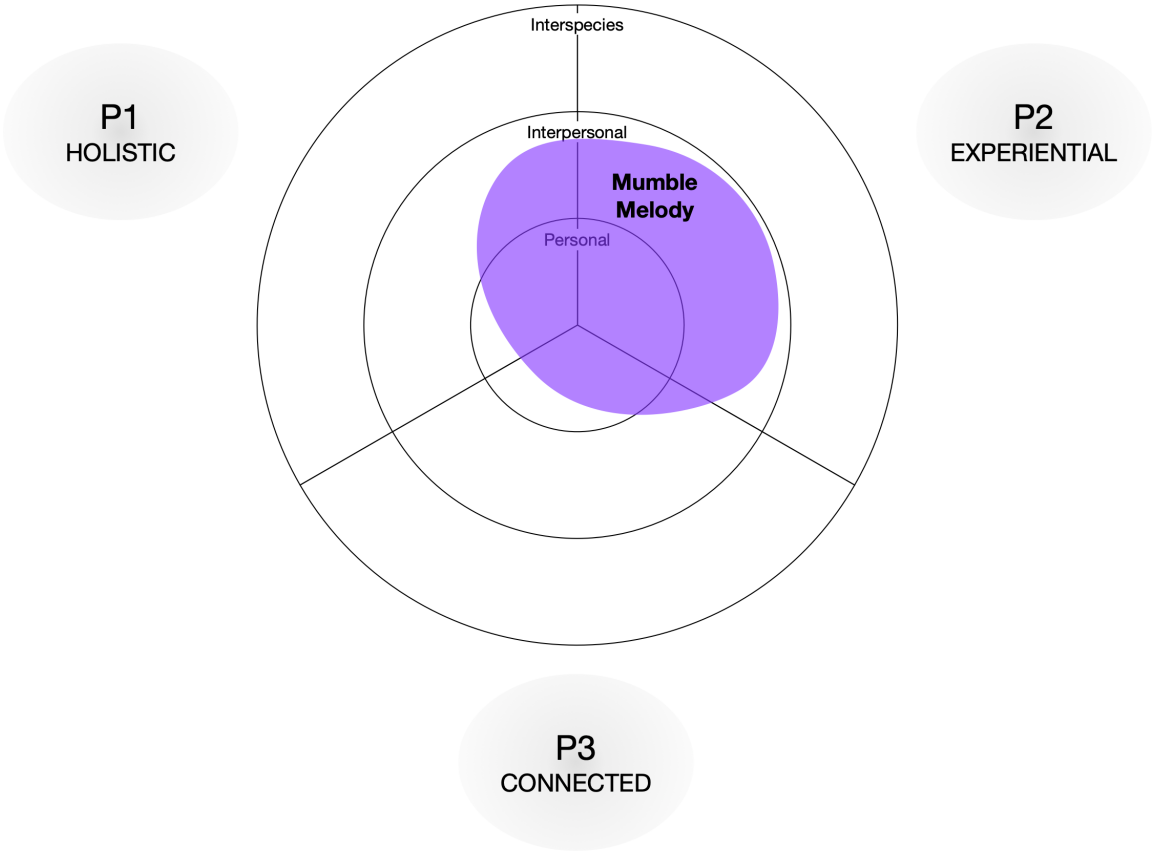


Figure 52: Mapping of the Mumble Melody initiative onto the Vocal Connection space

6 - Sonic and Vocal Enrichment for Animals in Managed Care

> *Some people talk to animals. Not many listen though. That's the problem*
>
> — A.A. Milne

Humans are not the only vocal creatures. This evolutionary tool is shared by tens of thousands of species around the globe. Animal voices may have an important part to play in understanding the origins and potential of vocal connections and learning to listen to those voices is an important part of this work in light of our three paradigms:

Our first paradigm frames the voice aside from verbal and emotional content. With nonhuman animals, there isn't a "distraction" of words when contemplating vocal interactions, although we do need to acknowledge the case of some social species that experts believe to have developed complex languages, such as bees, cetaceans, and elephants[362]. While the study of those instances is beyond the scope of this thesis, we consider them to further support the potential of the voice to be more than solely "verbal," as different aspects of sonority, different uses of the spectrum, different transmission media (water, ground), and even different organs of perception are shown to transmit complex and organized meaning.

[362] Irene M Pepperberg. Animal language studies: What happened? *Psychonomic bulletin & review*, 2017

Our second paradigm considers the extent to which our experience of voices informs our experience of the world. Instead of using the animal context to guide our understanding of the human voice, this point allows us to use our human experience as a flashlight to reveal elements of animal cognition and inner lives. This framing may shed light on the apparent enigma of understanding how animals think without words as we know them. If human inner lives arise not from language but from our ability to generate, perceive, and internalize sounds, then maybe we have commonalities in cognition with all vocal species. Is there a common ancestor in our evolution, a missing link, who was the first to be able to "think in sounds"?

Our third paradigm frames the voice as offering markers of social dynamics, looking first at its potential as a form of social grooming. Animal voices are known to be used to support social structural behaviors[363] and often respect conversational rules such as turn-taking[364]. The inspiration for considering the human voice as such comes from an effort to trace its origin, in light of our understanding of modern wildlife.

In addition to sound, some species have additional senses that are very developed, such as smell, and we acknowledge the possibility of multisensory inner lives. However, we still argue that some commonality in our way of thinking might be shared by other species with developed auditory abilities. This supports the idea that inner voices can exist independently from language, or at least from human language. Do other vocal mammals experience an inner voice? Do birds rehearse their songs silently? This theory might suggest that any species capable of producing vocal sounds would potentially experience a complex inner life, or at least an inner sonic life. This reasoning does not deny the intelligence of non-vocal species. On the contrary, we absolutely acknowledge the incredible intelligence of animals such as octopi and ants. We would, however, need a totally different model to understand their cognitions. They certainly think, but presumably do not share a common "sound-thinking" ancestor.

This conjecture is, of course, still at the stage of speculation. It would require a major endeavor to attempt to demonstrate it, which is outside the scope of this dissertation work. In this work, we have simply opened the door to such possibilities and asked questions. In this chapter, we offer early explorations of those questions and present initial observations and experimentations, works in progress in the more defined context of animals in managed care.

6.1 - Introduction

6.1.1 - Context

One does not have to look far to observe a powerful disconnect between humans and other species in our societies[365]. Zoos are at the pinnacle of this disconnect, demonstrating a fascination for other animals and the will to understand, educate, and see through one's own eyes; but also displaying the asymmetry between visitors in search of entertainment and captive animals suffering from a profound lack of meaningful interactions, not to mention their need for natural behaviors.

This being said, zoos have evolved to respect and protect species, as

[363] Donald H Owings, Eugene S Morton, et al. *Animal vocal communication: a new approach*. Cambridge University Press, 1998

[364] Laurence Henry et al. Social coordination in animal vocal interactions. is there any evidence of turn-taking? the starling as an animal model. *H., Casillas, M., Levinson, SC, eds.(2016). Turn-Taking in Human Communicative Interaction. Lausanne: Frontiers Media. doi: 10.3389*, 2016; and Cecilia P Chow et al. Vocal turn-taking in a non-human primate is learned during ontogeny. *Proceedings of the Royal Society B: Biological Sciences*, 2015

[365] Robert Costanza et al. Sustainability or collapse: what can we learn from integrating the history of humans and the rest of nature? *AMBIO: A Journal of the Human Environment*, 2007

well as educating the public about them, and caregivers are often very sensitive to the needs of the animals they care for. Enrichment is a way for zookeepers and caregivers to enhance the quality of life of captive animals, to enable them to express their natural behaviors, and to reduce stereotypic behaviors. Also called behavioral enrichment, it aims at improving the quality of captive animal care by identifying and providing "the environmental stimuli necessary for psychological and physiological well-being" [366]. One of the first evaluations of enrichment apparatus dates back to 1978 [367]. There exist different types of enrichment based on cognitive, social, food, environmental, and sensory stimuli. Concerning sensory-based enrichment, Wells [368] concludes that the greatest benefits for animal welfare are obtained through enrichments that target the dominant sense of the animals. Audition is a dominant sense for many species, and a multitude of studies have evaluated the welfare benefits of music—including natural sounds [369], classical music [370], country music[371], and radio broadcasts[372]—in enhancing the welfare of captive animals. However, most of the existing studies only evaluate the effects of static recordings and thus do not provide insight into how sonic interactiveness, i.e., customizing the sounds based on responses from animals, might improve the efficacy of such sonic stimuli. Kim-McCormack[373] shows the growing relevance of interactive digital applications for captive primates and insists on the importance of giving control to the animal.

We believe that a better understanding of animals' sonic world, as well as giving them options to shape their sonic surroundings, could yield substantial benefits. However, the conservation toolbox is lacking real-time tools that could help researchers and caregivers better understand animal experiences. Machine learning techniques, including deep learning, have recently been used to classify wildlife photographs and animal morphological characteristics [374], which has led to applications in conservation [375]. While researchers are beginning to apply similar techniques to analyze and classify animal sounds[376], there are currently no simple tools for zookeepers and researchers to gain insight into the sounds of the animals they care for or to use such insights to improve their care. Caution and humility are also needed in the development of new sonic interventions for animals. Our understanding of animals' sonic worlds and communication is limited and assumptions can lead to detrimental consequences, as this anecdote from a professional working at an avian incubation center attests: Caregivers in charge of hand-rearing chicks used to play static recordings of the bird species during feeding times to familiarize them with the sounds, in preparation for housing them with conspecific birds once they matured. However, after years of using the same recordings, they realized they were playing alarm calls. One can only start imagining the cognitive consequences for the birds.

[366] David J Shepherdson. Environmental enrichment: past, present and future. *International Zoo Yearbook*, 2003

[367] Robert Yanofsky et al. Changes in general behavior of two mandrills (papio sphinx) concomitant with behavioral testing in the zoo. *The Psychological Record*, 1978

[368] Deborah L Wells. The effects of animals on human health and well-being. *Journal of Social Issues*, 2009

[369] AS Chamove. Cage design reduces emotionality in mice. *Laboratory Animals*, 1989

[370] Elaine N Videan et al. Effects of two types and two genre of music on social behavior in captive chimpanzees (pan troglodytes). *Journal of the American Association for Laboratory Animal Science*, 2007; G Gvaryahu et al. Filial imprinting, environmental enrichment, and music application effects on behavior and performance of meat strain chicks. *Poultry Science*, 1989; and Deborah L Wells et al. Auditory stimulation as enrichment for zoo-housed asian elephants (elephas maximus). *Animal Welfare*, 2008

[371] Nikki S Rickard et al. The effect of music on cognitive performance: Insight from neurobiological and animal studies. *Behavioral and Cognitive Neuroscience Reviews*, 2005; and K Uetake et al. Effect of music on voluntary approach of dairy cows to an automatic milking system. *Applied animal behaviour science*, 1997

[372] L Brent and O Weaver. The physiological and behavioral effects of radio music on singly housed baboons. *Journal of medical primatology*, 1996; and RB Jones. Environmentai enrichment: the need for practicai strategies to improve pouitry weifare. *Welfare of the laying hen*, 2004

[373] Nicky NE Kim-McCormack et al. Is interactive technology a relevant and effective enrichment for captive great apes? *Applied animal behaviour science*, 2016

[374] John Joseph Valletta et al. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 2017

[375] Roberto Salguero-Gómez et al. Comadre: a global data base of animal demography. *Journal of Animal Ecology*, 2016

[376] Peter J Dugan et al. Phase 1: Dcl system research using advanced approaches for land-based or ship-based real-time recognition and localization of marine mammals-hpc system implementation. 2016; and Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 2017

In order to explore these ideas, we entered into a collaboration between the MIT Media Lab and the San Diego Zoo, to design and build interactive sonic enrichment systems for animals in managed care. Our approach is based on the potential of animal-animal and human-animal relationships as an environmental enrichment for the welfare of zoo-housed animals, specifically in terms of animal vocal communication. We developed a set of vocal and sonic enrichment interventions and studies designed to expand our knowledge in the field of intra-species, inter-species, and human-animal communication in the context of non-domesticated animals. By doing so, we explore to what extent animal and human voices can be understood and transformed to create new types of connection in sonic enrichment interventions.

6.1.2 - Four principles

In this chapter, we introduce a provocative approach to envision changes in sound treatment and design for zoo-housed animals. We propose considerations and early explorations toward the design of sonic and vocal interventions for animals in managed care. The goal of this work is to lay the groundwork for the design of a zoo of the future, with a focus on sounds, for the benefit of the animals. In this chapter, we identify four key guiding principles for rethinking sounds in zoos:

- 1) Listening to animals collectively—sounds heard by animals
- 2) Respecting animal-animal communication—accepting our human limitations
- 3) Listening to animals individually—sounds produced by animals
- 4) Giving agency to animals—letting go of our control

For each principle, we present an initial analysis of the challenge presented, early explorations in the form of novel systems, intervention or design thinking, as well as insights to guide future development in the field.

The organization of this chapter follows the four principles of our approach. In the section **"6.2 - Listening to animals collectively with the Sonic Diversity endeavor"**, we present a brief summary of preliminary thinking related to global sound designs and sonic diversity in zoos. The ongoing Sonic Diversity effort is a series of questions and ideas designed to grasp the importance of soundscape study and design for zoo-housed animals.

In section **"6.3 - Respecting animal-animal communication & the TamagoPhone project"**, we propose rationale and strategies to alleviate our

human limitations in designing systems for animals by using interactive real-time audio systems for animal connection and enrichment. We then present a possible instance of such intervention in the design of the TamagoPhone project, an enriched incubator to maintain vocal interaction between bird parents and egg during artificial incubation.

In section **"6.4 - Listening to animals individually & the Panda Project"**, we expand on the potential for deepening interspecies understanding and connection through the voice. We explore how current technology can help us understand individual animals better. We propose an example of this type of technology in the Panda Project, a real-time, deep learning-based tool for acoustic monitoring, to provide caregivers with meaningful information from panda vocalizations.

Finally, in section **"6.5 - Giving agency to animals & the JoyBranch project"**, we explore the importance of giving animals more control over their sonic environment. We present insights into unique ergonomic, ethical, and agency-related challenges in designing interactive sonic enrichment systems for animals. We present the JoyBranch project, an interactive intervention deployed at the San Diego Safari Park to allow Sampson, a music-savvy hyacinth macaw, to control his sonic environment.

6.1.3 - Acknowledgement

Though this initiative was primarily led by the author, the conception and development of the various interventions and explorations were conducted in collaboration with multiple stakeholders.

The design of the **TamagoPhone project** was conceived in collaboration with Janelle Sands.

For the **Sonic Diversity Project**, the recordings were deployed in collaboration with Gabriel Miller and analyzed with the help of George Stefanakis.

For the **Panda Project**, the realization was done in collaboration with Clement Duhart (ESILV) and with guidance from Gabriel Miller and Megan Owen (SDZG). Tracking collars with acoustic recording units (ARU) were deployed and recordings of behavior (bamboo feeding, cub suckling, rest) and cub vocalizations were manually decoded by research staff at the CCRCGP-Heatuaping Reintroduction Base. Research staff included Xiao Yan, Mengmeng Sun, Wu Daifu, Liu Xiaogiang, Zhou Shiqiang, Mou Shifjie, and He Shengshan. Shotgun microphone recordings of adult panda vocalizations were made by Ben Charlton, SDZG Research Fellow. The collar-mounted acoustic recording project was conceived by Megan Owen (SDZG), Zhang Hemin (CCRCGP), Wu Daifu (CCRCGP) and Li Disheng (CCRCGP), and managed by Wu Daifu (CCRCGP) and Huang

Yan (CCRCGP).

The **JoyBranch** device and interventions were done in collaboration with David Su, Akito van Troyer, Janet Baker and Gabriel Miller, and with the help of Anne Harrington and Lydia Yu. We also wish to thank Jenna Duarte and Michelle Handrus for contributing their extensive support. The data analysis was done in collaboration with Anne Harrington and Lydia Yu.

All the research aspects of the projects mentioned in this chapter are the personal contribution of the author. The projects were sparked and developed following several visits at the San Diego Zoo between 2018 and 2019. Through observing animals and meeting with animal researchers and caregivers, the need for sonic and vocal enrichment interventions arose. Throughout our design phases, we consulted with specialists, including Irene Pepperberg, Arno Klein, Wenfei Tong, Danika Oriol-Morway, and Alwyn Wils. The various projects in this chapter emerged from these conversations and present possible ways to approach the challenges outlined in this chapter. This initiative was made possible by the animals from the San Diego Zoo and we wish to thank them for inspiration and collaboration. All procedures described were approved by the Zoological Society of San Diego IACUC under proposal 19-002.

6.2 - Listening to animals collectively & the Sonic Diversity endeavor

We believe that zoos need to adopt a comprehensive approach to rethinking their global soundscapes. Those are the sounds heard by the animals, the sonic environment in which they evolve, and which they often cannot affect or turn off. Sound is so paramount for many species that it should become an integral part of exhibit design. Such an approach could help tackle important questions related to the experience of zoo animals. What is the everyday sonic experience of the animal, and how does it differ from their counterparts in the wild? What are the effects on animals' physical and mental wellbeing? How does it affect their cognitive development, ability to interact with other individuals from the same species, and more generally their perception of the world?

In most zoos and sanctuaries, animals are exposed to sounds from four different sources: 1) Geophonic: natural weather-based sounds that can include rain, wind, or thunder; 2) Anthropophonic: sounds resulting from human activities that can include voices, screams, construction sounds, sounds played on speakers, or airplanes passing by; 3) Heterospecific: sounds generated from animals from a different species; 4) Conspecific and self-generated: sounds coming from animals from the same species or from the individuals themselves.

Ongoing and future work in this domain include analysis of ambient sounds throughout various exhibits in zoos and developing approaches to balance anthropophonic, heterospecific and conspecific sounds depending on the specific needs of the species on exhibits.

6.3 - Respecting animal-animal communication & the TamagoPhone project

In addition to looking at the complex symphony of animal voices in zoos, it is also crucial to acknowledge our limitations in understanding the messages carried vocally between animals and to envision ways to alleviate the risks of creating damaging human interference. In this regard, we can find some perspectives in the mysteries still persisting in our understanding of bird vocalizations.The complete study of bird vocalizations is outside the scope of this thesis work. In the context of this chapter we simply present information that supports the idea that, despite our best efforts, humans are currently not capable of obtaining a true and complete understanding of the vocal experiences of animals.

Bird vocalizations are extremely diverse, and humans have tried to decode them for a very long time, often in an anthropocentric and symbolic manner. Ornithomancy (from Greek ornis—"bird"—and manteia—"divination") is the practice of reading omens from the actions and cries of birds. It was practiced in ancient Greek [377], Roman [378] and Jewish [379] traditions. At the crossroad between meaning and music, bird songs are full of symbolism and often considered for their pleasantness to the human ear. The common distinction between bird "calls" versus bird "songs" is an example of this anthropocentric view. One very ancient instance of the relationship between bird and human language can be found in Tuvan throat singing traditions that incorporate piercing overtones representing bird noises. In Tuvan culture, bird vocalizations are thought to be the original inspiration for human music[380].

In recent centuries, ornithologists have made tremendous breakthroughs in decrypting specific contexts of bird interactions. Birds vocalizations are thought to have diverse functions, ranging from mate attraction, to territory defense, learning, alarm raising, and more[381]. One beautiful and seemingly simple way some have tried to use human words to render/transcribe bird calls is as, “Hi, I am here! Are you here?” There are many ways to express such a message. Indeed, even humans, when saying a simple sentence vocally, cannot help but introduce context and unicity through our individual voices, tones, vocal postures, etc. Each time a person verbalizes such a sentence, it is as unique as when a bird sings their morning calls: as a way to express their presence, that they have survived the night, and are inquiring about their peers’ presence and state.

Another more playful phenomenon in connecting bird vocalization and human language can be seen in the mnemonics and written forms used by birders to classify and recognize bird vocalizations. Some bird calls are

[377] Homer. *The odyssey*

[378] Cyril Bailey. *Phases in the religion of ancient Rome*. Univ of California Press, 1932

[379] Gerrit Bos. Jewish traditions on divination with birds (ornithomancy). *Gen*, 2015

[380] David McNamee. Hey, what’s that sound: Throat singing, 2010. URL theguardian.com/music/2010/jun/02/throat-singing

[381] Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003

so distinctive as to be integrated into the species' common names, such as the cuckoo, the chiffchaff, or the chickadee. Writers have demonstrated admirable ingenuity in writing down bird sounds as demonstrated in the very poetic book from John Bevis: *Aaaaw to Zzzzzd: The Words of Birds*[382]. The book presents both a lexicon and examples of mnemonics. The lexicon proposes onomatopoeic words that attempt to replicate the sounds birds make. Covering more than 1,500 species of birds from North America and Europe, it proposes plausible notations for the calls of each species. Mnemonics catch the rhythms and emphases of bird calls into phrases from the English language, as in the call of the brown thrasher, which can be heard as: "Drop it, drop it, pick it up, pick it up." Or the song of the white-throated sparrow: "O sweet Canada, Canada, Canada." Or the Carolina wren: "Teakettle, teakettle, teakettle." Similarly to the way in which one often sees objects in the shape of clouds, humans can not help but hear words in the songs of birds. This poetic phenomenon underlines once again our propensity to interpret other animals' voices in a human way.

[382] John Bevis. Aaaaw to zzzzzd: The words of birds, 2010

The scientific study of bird vocalizations can be found in the fields of neurology[383], medicine[384] and evolutionary biology[385]. However, no one could plausibly argue that they can translate bird sounds into human meaning. Some, like Noam Chomsky, explain this limitation by hypothesizing a lack of "real" language in birds and tenaciously set human language fundamentally apart from all other animal behavior[386]. Others argue for the development of more thorough biolinguistics to investigate these questions[387]. We believe that the problem here is twofold. First, the definition of language itself seems to have been created and to have evolved to exclude specific populations considered uncivilized, including non-human animals [388]. Second, our human languages simply do not possess the correct vocabulary for what needs to be expressed as we are set in a different umwelt.

[383] Erich D Jarvis. Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences*, 2004

[384] Michael S Brainard and Allison J Doupe. Translating birdsong: songbirds as a model for basic and applied medical research. *Annual review of neuroscience*, 2013

[385] Masakazu Konishi et al. Contributions of bird studies to biology. *Science*, 1989

[386] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014

[387] Tiffany C Bloomfield et al. What birds have to say about language. *Nature neuroscience*, 2011

[388] Steven R Fischer. *History of language*. Reaktion Books, 2001

Von Uexküll (born in 1864, died in 1944) introduced the concept of umwelt, or phenomenal world, to address human biases in the study of animal perception[389]. By emphasizing the extreme phylogenetic contrasts that exist in the sensory worlds of different animal species, he exposes that animal experiences cannot be understood without considering the animals' environment and their perception of the environment. Uexküll introduced the fundamental idea that one needs to investigate life processes from the point of view of the specimen. To quote him: "Although the descriptions of the animals' worlds give the reader a feeling for their experiences, this empathy is illusory and sometimes misleading." At the species level, each animal also communicates in the context of its own personal umwelt, surrounded by the particulars of its own life, which include individual conspecifics, heterospecifics, and the micro- and macro-habitats in which it lives"[390].

[389] Jakob Von Uexküll. A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, 1992

[390] Sarah Partan and Peter Marler. The umwelt and its relevance to animal communication: introduction to special issue. *Journal of Comparative Psychology*, 2002

Following Uexküll, Nagel goes a step further by insisting on the possible insolubility of the mind-body problem[391]. In his famous article "What Is It Like to Be a Bat?", he affirms the inaccessible nature of animal experience. Taking the experience of a bat and its unique apparatus for sound perception, he proposes the conundrum that beyond animals' experience of sounds, the entire concept of otherness in perception seems to be inaccessible to humans.

[391] Thomas Nagel. What is it like to be a bat? *The philosophical review*, 1974

If qualias cannot be shared, they ought to be respected as such. Bird songs have inspired artists, musicians, poets, and philosophers, but they mainly need to inspire respect. It is with this in mind that we envisioned the TamagoPone project. The TamagoPhone is a conceptual and design exercise to bring awareness of subtle, though fundamental, vocal exchanges in the animal kingdom and think about ways humans can limit perturbations when caring for wildlife and livestock. This project specifically targets the still-mysterious and understudied phenomenon of vocal communication between bird parents and their offspring who are still within the egg. This project was conceived in collaboration with Janelle Sands.

6.3.1 - The TamagoPhone - Proposed Future Project

Figure 53: White hen with chickens by Anton Ignaz Hamilton (Austrian, 1696–1770)

It is common practice for zoos and bird conservationists to incubate bird eggs in artificial incubators to maximize hatching rates. In the wild or even in zoos within enclosures, eggs can be vulnerable to predators[392], diseases[393] or temperature problems[394], and the use of artificial incubators during part or the entirety of the incubation time can greatly improve the chances of survival of the chicks and support preservation efforts for endangered species.

However, some species exhibit important prenatal vocal interaction while within the egg. Indeed, parents birds often produce vocalisations directed to their eggs and in some species, chicks already produce calls from within the egg a few days before hatching. Common artificial incubation techniques deprive embryonic chicks of integral parent-offspring vocal

[392] Elke Schüttler et al. Vulnerability of ground-nesting waterbirds to predation by invasive american mink in the cape horn biosphere reserve, chile. *Biological Conservation*, 2009

[393] MC Lábaque et al. Microbial contamination of artificially incubated greater rhea (rhea americana) eggs. *British Poultry Science*, 2003

[394] AM King'Ori et al. Review of the factors that influence egg fertility and hatchability in poultry. *International Journal of Poultry Science*, 2011

communications during early development. Recent research has shed light on specific behavioral contexts associated with vocal pre-hatching events for specific species[395]. Led by such research, one could think of using curated static recordings during artificial incubation to alleviate the lack of vocal interactions. However, our understanding of those vocal interactions is still in its infancy, and might never be understood well enough to synthesize meaningful replacement or select relevant recordings. Instead, we propose the idea of supplementing existing egg incubation techniques with a two-way, real-time audio system to allow mother bird and unhatched eggs to communicate with each other remotely and in real time during incubation. With this approach, when the real egg is removed from the nest it would be immediately replaced by an augmented "dummy" egg, containing a microphone and speaker, which would be cared for by the parent birds. The augmented artificial incubator would also contain a system of microphone and speaker, in addition to the traditional temperature, humidity, and motion control systems. Both sides, the parent and the egg, would be connected by a two-way audio streaming platform, with all of audio components integrated as inconspicuously as possible. We believe that such a system could increase connectivity between the mother and her young while acknowledging our human limitations in understanding the possible meanings and functions of the vocal signals exchanged. This system could also be used by scientists to enable new research on pre-hatching parent-chick interactions and their consequences post-hatching.

As of the writing of this thesis, the project hasn't yet been reduced to practice. This is due to both a lack of time and administrative difficulties in testing such systems with live animals. However, as we still consider this work relevant for this dissertation, we present the motivation and rationale for the project, as well as potential applications and preliminary methodology design for evaluation.

6.3.1.1 - Background on avian pre-hatching vocal communication

Previous research on responsiveness and vocalization in bird embryos supports our investigations. Nice et al. described the early stages of development of a wide number of avian species[396]. It is now well established that avian prenatal sensory experience affects development and has long-term consequences on postnatal behavior[397], which varies between altricial and precocial species. Species of birds who are able to feed themselves, are covered with down, have their eyes open, and leave the nest days following hatching, are classed as precocial. Birds with closed eyes, very little down, and are unable to leave the nest for some time are classified as altricial. Previous work has highlighted the development of auditory sensitivity of bird embryos prior to hatching. For example, auditory sensitivity appears at day 11 in the domestic fowl[398] and day 14 in the duck[399].

[395] Diane Colombelli-Négrel et al. Embryonic learning of vocal passwords in superb fairy-wrens reveals intruder cuckoo nestlings. *Current Biology*, 2012; and Mylene M Mariette. Prenatal acoustic communication programs offspring for high posthatching temperatures in a songbird. *Science*, 2016

[396] Margaret Morse Nice et al. Studies in the life history of the song sparrow. 1964; and Margaret Morse Nice. *Development of behavior in precocial birds*, volume 8. New York:[Linnaean Society], 1962

[397] Barry Metcalfe Freeman et al. *Development of the avian embryo: a behavioural and physiological study*. Springer, 1974

[398] JC Saunders. The development of auditory evoked responses in the chick embryo. *Minerva Otolaryngol*, 24:221–229, 1974

[399] Masakazu Konishi. Development of auditory neuronal responses in avian embryos. *Proceedings of the National Academy of Sciences*, 70(6):1795–1798, 1973

Some known functions of parental vocal signals: There is growing evidence of a behavioral continuity between the embryo and the chick. As Hinge puts it "hatching is not a zero-point for the development of behavior"[400]. Indeed, previous field studies have shown that communication between embryo and bird parents can affect the future relationship between parents and chicks. In the guillemot, chicks learn pre-hatching to discriminate between their own parents' vocalizations and those of other birds[401]. Another pre-hatching exposure of the chick to parental vocalization is linked with long-term feeding behavior appropriateness as seen in the laughing gull. When repetitively exposed to the parental "crooning" call before hatching, the young exhibit more pecking behavior on their parent's beak to request food. This hypothesis has been verified by playing recordings of parental "crooning" calls during artificial incubation[402]. Other birds, such as fairy-wrens, sing code-words to their unhatched young to later repeat after hatching, so the mother can identify and feed her own young instead of imposter cuckoo young dropped into her nest[403]. In the zebra finch, an altricial species and one of the most-studied species of songbird, the parents will have different incubation calls depending on the temperature, which affects the growth and development of the young. This suggests that the calls may act as a sort of vocal meteorological warning preparing the chicks for warm conditions after they hatch[404]. In regards to social connections, exposure to vocalizations pre-hatching has been shown to mediate post-hatching attachments and imprinting in precocial species[405].

Functions of embryonic vocal signals: Prior works have shown evidence of pre-hatching embryo vocalizations in various precocial species, including domestic fowl[406], ducks[407], gulls[408] and quails[409]. Some researchers have hypothesized that altricial species cannot vocalize prior to hatching[410]; however, others have suggested that for very small eggs, vocalizations might be produced but too quietly to be audible[411]. Studies have shown that embryo vocalizations can be triggered by sounds, changes in temperature, or movement of the egg. An increase in vocalization in bird embryos in response to the maternal call has been shown in ducks[412] and gulls[413]. At late stages of incubation, chick vocalizations also seem to respond to siblings' vocalizations[414], and embryo distress calls can be stopped by the sound of clucking [415]. In several species, parent birds respond to the chick vocalization. Sibling vocalizations are also linked to increased survival rates and hatchability. In the yellow-legged gull, embryos who hear adults 'warning calls communicate them to their siblings through egg vibrations. The entire clutch can then perceive the vibratory cues of predation risk from their more advanced clutch mates[416].

[400] Robert A Hinde. Animal behaviour: A synthesis of ethology and comparative psychology. 1970

[401] Beat Tschanz. *Trottellummen: die Entstehung der persönlichen Beziehungen zwischen Jungvogel und Eltern*. Parey, 1968

[402] Monica Impekoven. Prenatal experience of parental calls and pecking in the laughing gull. *Animal Behaviour*, 1971

[403] Diane Colombelli-Négrel et al. Embryonic learning of vocal passwords in superb fairy-wrens reveals intruder cuckoo nestlings. *Current Biology*, 2012

[404] Mylene M Mariette. Prenatal acoustic communication programs offspring for high posthatching temperatures in a songbird. *Science*, 2016

[405] L James Shapiro. Pre-hatching influences that can potentially mediate post-hatching attachments in birds. *Bird Behavior*, 1981; and Roger M Evans. The development of learned auditory discriminations in the context of post-natal filial imprinting in young precocial birds. *Bird Behavior*, 1982

[406] Frederick Stephen Breed. *The development of certain instincts and habits in chicks*. Number 1. Pub. at Cambridge, Boston, Mass., 1912

[407] Eckhard H Hess. " imprinting" in a natural laboratory. *Scientific American*, 1972

[408] Monica Impekoven. Prenatal experience of parental calls and pecking in the laughing gull. *Animal Behaviour*, 1971

[409] MA Vince. Embryonic communication, respiration and the synchronization of hatching. *Avian Incubation, Behavior, Environment, and Evolution*, pages 88–99, 1969

[410] Ronald W Oppenheim. Prehatching and hatching behaviour in birds: a comparative study of altricial and precocial species. *Animal Behaviour*, 1972

[411] Barry Metcalfe Freeman et al. *Development of the avian embryo: a behavioural and physiological study*. Springer, 1974

[412] Gilbert Gottlieb. Prenatal auditory sensitivity in chickens and ducks. *Science*, 1965

[413] Monica Impekoven. The response of incubating laughing gulls to calls of hatching chicks. *Behaviour*, 1973

[414] Friedrich Goethe. Beobachtungen bei der aufzucht junger silbermöwen. *Zeitschrift für Tierpsychologie*, 1955

[415] NE Collias. The development of social behavior in birds. *The Auk*, 1952

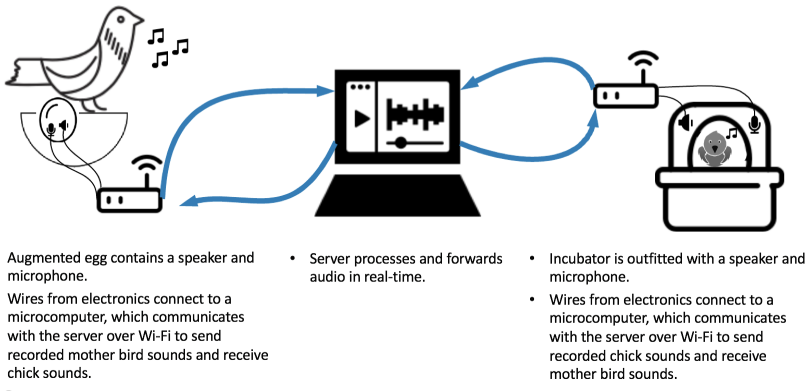
[416] Jose C Noguera et al. Bird embryos perceive vibratory cues of predation risk from clutch mates. *Nature ecology & evolution*, 2019

6.3.1.2 - The TamagoPhone intervention

In the context of zoos, endangered species are most often artificially incubated to support preservation efforts. Although modern incubators maintain temperature, humidity, and egg motion at optimal levels and increase survivability, they do not preserve the potential pre-hatching vocal connection between parents and embryos. Our proposed system aims to re-establish this auditory link between mother and her incubating young with a two-way, streaming audio system.

The system diagram of the TamagoPhone project is presented below. It is composed of an augmented dummy egg with a wireless two-way audio connection, a central system for filtering and control, and an augmented incubator with a two-way audio connection.

Figure 54: Overview of the TamagoPhone system



6.3.1.3 - Possible factors for Evaluation

Such a system would need to be evaluated both in terms of engineering and in terms of impact on the bird, specifically environmental improvement for the wellbeing and ontogenic development of the birds.

Factors needed to evaluate the engineering of the system include audio streaming quality and latency, artificial egg durability, system invasiveness, and power demands. It is important to have a low streaming audio latency to offer the most natural communication for the birds. High artificial egg durability is desirable, as we need the artificial eggs to remain intact and functioning despite natural bird behaviors such as egg turning[417], and interactions with the elements, such as water and heat. The system invasiveness would need to be personalized for each species, as birds have different thresholds for the ability to recognize their own eggs. Lastly, power demands are important in the case of the dummy egg; in smaller species where a battery wouldn't fit inside the egg, an induction coil might

[417] DAT New. A critical period for the turning of hens' eggs. *Development*, 5(3):293–299, 1957

need to be inserted into the nest/nest box to allow continuous charging during the time of incubation.

To evaluate the impact of the system on the wellbeing of the mother and her young, as well as their relationship, one would need to compare our system to the effect of traditional incubators or natural incubation on the behavior of the offspring and their social connections with the parents and other conspecific adults. Testing the system initially on common precocial species such as chickens could allow quick turnover and relatively low stakes compared to using endangered species. To evaluate potential differences in behaviors, one could look to measure wellbeing and connection in ways inspired by previous work in the field in evaluating bird health and connectivity. In previous work, mother-chick connectivity has been measured by blowing a gentle puff of air in the young birds' faces and assessing the duration and extent of maternal distress by measuring the mother's heart rate, eye temperature, and duration of time spent making distressed sounds and preening, being careful to follow animal safety protocols[418]. Observing the quantity of and types of communications between the mother and young can also be illustrative of the strength of the mother-chick relationship. One could also evaluate chick behaviors such as fearfulness and activity level. A young bird's flight response and amount of time walking around versus perching or frozen immobile after a human stands up nearby as a startling effect can be used as markers of fearfulness[419]. To evaluate activity level, we could compare the amount of time pecking at the floor and bathing in the dust.[420] To evaluate the health of young chicks, it is common to track the bird's development, particularly weight over time, as relative heaviness is considered an indicator of better health in several bird species[421]. To assess potential effects on the hen, one could measure the frequency of their egg-laying, frequency of natural egg turning behaviors, food consumption, and extent of distinct distress behaviors such as feather-pecking. Comparing these metrics between the groups would be useful in evaluating the effect of for our system, which we believe could become a tool in incubation techniques for captive bird-rearing. More extensive behavioral testing could look at the parental behaviors that the chicks later develop with their own eggs.

[418] Joanne Edgar, Suzanne Held, Charlotte Jones, and Camille Troisi. Influences of maternal care on chicken welfare. *Animals*, 2016

[419] R Bryan Jones. Fear and adaptability in poultry: insights, implications and imperatives. *World's Poultry Science Journal*, 52 (2):131–174, 1996

[420] Joanne Edgar, Suzanne Held, Charlotte Jones, and Camille Troisi. Influences of maternal care on chicken welfare. *Animals*, 2016

[421] Christine Careghi, Kokou Tona, Okanlawon Onagbesan, Johan Buyse, Eddy Decuypere, and Veerle Bruggeman. The effects of the spread of hatch and interaction with delayed feed access after hatch on broiler performance until seven days of age. *Poultry science*, 2005; and Sharon L Deem, Andrew J Noss, Rosa Leny Cuéllar, and William B Karesh. Health evaluation of free-ranging and captive blue-fronted amazon parrots (amazona aestiva) in the gran chaco, bolivia. *Journal of Zoo and Wildlife Medicine*, 2005

6.3.1.4 - Potential applications

If well-engineered, the TamagoPhone system could open the door to new applications in the context of preservation, research, and livestock management. In regards to preservation programs for rare avian species, our system could be a useful tool for zoos and conservation centers. In addition to increasing the species population, zoos could also preserve behavioral characteristics of the species and raise birds more adept at social inter-

actions with conspecifics. Our system could alleviate the risk posed by artificial incubation of depriving embryos of parental contact, which may be required to establish normal species identity and behaviors.

In terms of potential for research, the TamagoPhone could be a useful tool for scientists and researchers who study birds. Until now, it has been acknowledged that the normal effects of embryonic calling can only be observed in the wild and in conditions where eggs are incubated by the parents[422]. However, we believe that a sonically augmented incubator could allow researchers to test novel hypotheses regarding pre-hatching avian vocalization.

[422] Barry Metcalfe Freeman et al. *Development of the avian embryo: a behavioural and physiological study*. Springer, 1974

Finally, we believe that this project could offer a different perspective on current intensive farming practices. The chicken industry especially uses the most sophisticated and optimized technologies, which currently focus on enhancing the production of meat and eggs which deprives birds natural environment. The TamagoPhone aims to use technology to restore important vocal interactions in the ways we raise poultry and potentially improve animal welfare..

6.4 - Listening to animals individually & the Panda project

Legends about humans speaking animal languages or animals speaking human languages have been part of folklore around the globe for millennia. Biblical narratives include instances of beastly talks, such as the dialogues between Eve and the snake (Genesis 3 NIV) or the interaction between Balaam and his donkey (Numbers 22:21-38 KJV). Some have conjectured that all of the animals in the story of the Garden of Eden were able to talk[423]. In folktales around the globe, snakes and dragons are often associated with speech. Tasting the blood of a dragon in the Norse Nibelungen tale[424], or eating the heart of a snake in Indian legends[425] brings the power to speak with animals.

Although we may never reach a complete mutual understanding, there is potential for deepening interspecies connections through the voice. And this is true on both sides. We have many examples of adaptative vocal behavior from animals when interacting with humans. Adult cats do not meow in the wild, nor to other domesticated cats[426]. An orangutan named Tilda would use two vocal sounds, the "raspberry" and the extended grunt, exclusively when addressing her human[427]. Alex, Irene Pepperberg's star grey parrot, would only mimic human speech in a learned manner when interacting with humans[428]. During her contact with humans, a white whale named Noc began to spontaneously make sounds, which were recorded for comparison to human speech [429]. And elephants are also known to imitate human speech[430]. More and more experts agree that many animal species are capable of metacognition (thinking about one's own thoughts) [431], and we believe that this can be expressed through vocal interactions.

On the human side, some people are particularly sensitive to animals' vocalizations and can reach a profound understanding with non-human animals, especially when they have spent years interacting. However, most pet owners, as well as a non-negligible part of the scientific community, have often relied on the animal's intelligence and pushed their abilities to understand our human languages rather than the other way around. The training of Koko, the female gorilla who was taught to recognize and use over 1,000 words in sign language, has raised various controversies[432]. Some experts said her language skills were inaccurate[433]. "Sure, Koko could pair an impressive number of words to objects and phenomena, but when she signed 'happy' or 'love,' did she really feel those things the way we do?" We believe that this is not truly the concern. How many humans have a true understanding and honesty in their use of the words "happy" or "love"? Who are we to assess Koko's understanding of such complex social constructs? We believe that the real debate with such experimentation is that non-human primates already have their own languages, and their

[423] George Savran. Beastly speech: intertextuality, balaam's ass and the garden of eden. *Journal for the Study of the Old Testament*, 19(64):33–55, 1994

[424] Karl Blind. Wagner's" nibelung" and the siegfried tale. *The Cornhill magazine*, 1882

[425] JG Frazer. The language of animals.(continued). *The Archaeological Review*, 1888

[426] Penny L Bernstein and Erika Friedmann. Social behaviour of domestic cats in the human home. In *The Domestic Cat: The Biology of its Behaviour*. Cambridge University Press, Cambridge, 2014

[427] William D Hopkins, Jared P Taglialatela, and David A Leavens. Chimpanzees differentially produce novel vocalizations to capture the attention of a human. *Animal behaviour*, 2007

[428] Irene M Pepperberg. Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science*, 11(3): 83–87, 2002

[429] Sam Ridgway, Donald Carder, Michelle Jeffries, and Mark Todd. Spontaneous human speech mimicry by a cetacean. *Current Biology*, 22(20):R860–R861, 2012

[430] Angela S Stoeger et al. An asian elephant imitates human speech. *Current Biology*, 2012b

[431] Nate Kornell. Metacognition in humans and animals. *Current Directions in Psychological Science*, 18(1):11–15, 2009

[432] Francine GP Patterson and Ronald H Cohn. Language acquisition by a lowland gorilla: Koko's first ten years of vocabulary development. *Word*, 41(2):97–143, 1990

[433] Molly Roberts. We wanted to believe in koko, and so we did, 2018. URL https://www.chicagotribune.com

own types of intelligence. The anthropocentric approach is interesting but shouldn't prevail over respect for the animals and their dignity. The language most often taught to animals was created by humans, for humans.

Some modern companies claim to have achieved the holy grail of animal communication, selling so-called animal "translators" such as the ones offered by Zoolingua[434], or claiming to be able to decode dog barks, like the Family Dog Project[435]. Such companies have been the focus of heated controversy in the field of animal communication, as their approach reduces animal cognition and language to a mere decodable projection of their human counterparts.

[434] Zoolingua. Zoolingua, 2018. URL http://zoolingua.com/

[435] Etologia Tanszek. The family dog project, 2019. URL https://familydogproject.elte.hu

However, some species do have very well-documented acoustic ecologies and, in some cases, behavioral contexts are associated with sonic events, each offering a unique acoustic signature that allows for classification and contains additional fine details about the individual emitting those sounds. We believe that a worthy goal for learning to "listen to animals individually" is to disseminate knowledge about those known vocalizations and to provide researchers and caregivers with tools not to translate animal sounds but to quickly and easily associate specific known vocalizations with scientifically established associated behavioral contexts. This could at least enable the monitoring of animal vocalizations in a more comprehensive way and at most spark the creation of databases for further explorations of animal vocalizations in controlled settings.

Driven by these considerations, we developed the Panda Project, a new bio-acoustic, deep learning-based system for panda monitoring in collaboration with the San Diego Zoo. In the following sections, we present the rationale, context, methods, results, and direct applications of the project.

6.4.1 - The Panda Project - Panda Monitoring

Figure 55: Panda cub Xiao Liwu "Little Gift" and his mother Bai Yun at the San Diego Zoo

Understanding panda behavior is crucial to continued conservation management, yet our capacity to monitor behavior in the wild, using GPS and accelerometer data, is limited. Giant pandas are an ideal species for vocal, incidental, and ambient acoustic monitoring. The acoustic ecology of the species is well documented, and different behavioral contexts are associated with sonic events, each offering a unique acoustic signature allowing classification and also containing additional fine details about the individual emitting those sounds (sex, size, age, identity, intent, hormone level). Acoustic monitoring offers a promising approach to the detection and classification of behaviors and life history events, including reproduction, predator encounters, and feeding, sleeping, and nursing behaviors. Application of acoustic monitoring to conservation management requires the development of automated techniques for data extraction from recordings. Here, using an open source deep-learning approach, we demonstrate the successful real-time identification and classification of seven panda vocalizations, in addition to chewing and nursing sounds.

6.4.1.1 - Pandas

The giant panda (*Ailuropoda melanoleuca*) is a solitary bear species that forages exclusively on bamboo[436]. The species is vulnerable to extinction[437], and there are approximately 2,000 remaining individuals in the wild[438]. Pandas are rare, ranging over six mountain ranges in the south-central provinces of Sichuan, Gansu, and Shaanxi in China. Giant pandas have been the focus of intensive conservation efforts over the past two decades, including a conservation breeding and translocation program that has relied heavily on scientific research focused on the species' behavior. Translocation efforts began in 2013, and thus far the survival rate appears to be high[439]. However, to date, there is no evidence of successful recruitment into the wild population. As with other species, survival and successful reproduction require behavioral and social competence, which can be lacking in captive-bred individuals. Thus understanding the behavior of released, mentored, and free-ranging individuals is critical.

[436] George B. Schaller et al. The giant pandas of wolong. *The Quarterly Review of Biology*, 1990

[437] IUCN. The international union for conservation of nature's red list of threatened species. version 2016, 2016

[438] RR Swaisgood et al. Developmental stability of foraging behavior: evaluating suitability of captive giant pandas for translocation. *Animal conservation*, 2018

[439] Zhisong Yang et al. Reintroduction of the giant panda into the wild: A good start suggests a bright future. *Biological Conservation*, 2018

6.4.1.2 - Traditional methods for tracking pandas

Panda abundance has been assessed via extensive field surveys, and individuals have been tracked using radio and GPS tracking. From these data, habitat selection and seasonal movements have been inferred. However, an understanding of the behavioral mechanisms that underpin population parameters is lacking. Limited use of collar-mounted accelerometers have provided some very coarse behavioral data (i.e., active/inactive), and early studies of wild panda behavior were accomplished via direct observations of a very few individuals[440]. While these methods have provided valuable insights regarding the movements, range, and habitat selection of wild giant pandas [441], finer-scale behaviors that are critical to survival and social interactions, such as foraging, nursing or affiliative vocalizations, cannot be reliably observed. Thus, to obtain a better understanding of panda behavior, activity budget, and interactions, researchers have turned toward acoustic monitoring of the species[442].

[440] George B. Schaller et al. The giant pandas of wolong. *The Quarterly Review of Biology*, 1990

[441] Vanessa Hull et al. Space use by endangered giant pandas. *Journal of Mammalogy*, 2015; and Dunwu Qi et al. Different habitat preferences of male and female giant pandas. *Journal of Zoology*, 2011

[442] Xiao Yan et al. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support translocations. *Scientific reports*, 2019

6.4.1.3 - Panda general activities and sonic signatures

Based on limited studies of wild[443], and extensive studies of captive[444], giant pandas, the behavioral repertoire of the species is well-described, with non-social behavior largely consisting of feeding, resting, and directed locomotion. While these activities are non-vocal, they still present an identifiable sonic signature[445]. The low caloric and nutritional value of bamboo mean that pandas must spend upwards of 14 hours a day feeding. Bamboo feeding is sonically distinctive, as can be the "slurping" sound of drinking water. During the breeding season (between March and May [446] and during the period of maternal care, the nature of daily activity

[443] Jindong Zhang et al. Activity patterns of the giant panda (ailuropoda melanoleuca). *Journal of Mammalogy*, 2015

[444] DG Kleiman and G Peters. Auditory communication in the giant panda: motivation and function. In *Proceedings of the Second International Symposium on the Giant Panda. Tokyo Zoological Park Society*, 1990; and B D Charlton et al. Coevolution of vocal signal characteristics and hearing sensitivity in forest mammals. *Nature communications*, 2019

[445] Xiao Yan et al. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support translocations. *Scientific reports*, 2019

[446] George B. Schaller et al. The giant pandas of wolong. *The Quarterly Review of Biology*, 1990

changes[447], and context-distinctive acoustic communication becomes more prominent[448].

Courtship and breeding: As a solitary, wide-ranging, seasonally monoestrous and spontaneously ovulating species, coordination of courtship and breeding is dependent on temporally and contextually distinctive patterns of scent and acoustic communication. Scent communication is most prominent during mate-search, and pandas will advertise their reproductive status via scent marking, thus facilitating conspecific assessment and localization. Once pandas are in proximity, acoustic communication gains prominence. Studies with captive pandas have determined the functional significance of several panda vocalizations [449], including bleats—the most common vocalization emitted by both males and females during mate-search and signaling non-aggressive intent; chirps—high-pitched tonal vocalization emitted by females during the estrous period; barks, growls, and roars—aggressive calls produced during agonistic encounters; squeals—emitted by submissive individuals, often after a lost fight; moans—denoting ambivalent intent or slightly aggressive behavior; and finally, honks—mostly emitted when an individual is alone, the function of this sound is still unknown but has been associated with stress from strange or unfamiliar surrounding noise [450].

Maternal care and cub development: Maternal care behavior is also highly dependent on acoustic communication between mothers and cubs. Additionally, differences in cub vocalizations have been identified as associated with plaintive or content behaviors, as well as indicating arousal level in infants [451]. For maintenance behavior (e.g., feeding and resting), we see a pronounced seasonal variation in levels. Finally, cubs nurse 6–14 times a day for up to 30 minutes each time and are only fully weaned at 8–9 months [452]. Suckling activity has a very specific sonic signature and can be monitored from collar microphones on the mother.

By recording and analyzing sounds, we can non-invasively map behaviors of individuals, either in managed care or by using collar-based recordings of free-ranging individuals, whether fully wild or translocated. This approach could offer new insight into the lives, individual histories, and energy budgets of the animals. In addition to better understanding panda behavior, sound analysis can also detect information about species nearby, including birds, insects, and humans, as well as human activities, any of which might impact panda behavior in unknown ways.

[447] Megan A Owen et al. Dynamics of male–female multimodal signaling behavior across the estrous cycle in giant pandas (ailuropoda melanoleuca). *Ethology*, 2013

[448] M A Owen et al. Signalling behaviour is influenced by transient social context in a spontaneously ovulating mammal. *Animal Behaviour*, 2016

[449] Benjamin D Charlton et al. Vocal behaviour predicts mating success in giant pandas. *Royal Society open science*, 2018

[450] David M Powell et al. Effects of construction noise on behavior and cortisol levels in a pair of captive giant pandas (ailuropoda melanoleuca). *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2006

[451] Angela S Stoeger et al. Acoustic features indicate arousal in infant giant panda vocalisations. *Ethology*, 2012a

[452] Xiaojian Zhu et al. The reproductive strategy of giant pandas (ailuropoda melanoleuca): infant growth and development and mother–infant relationships. *Journal of Zoology*, 2001

6.4.1.4 - Automatic sound classifications & Identification

Besides manual sound identification by human listeners, some automatic methods have been used in animal monitoring for the past few decades. These techniques are often based on detecting specific acoustic features such as amplitude tracking, spectrogram scanning, or call duration. The application of acoustic monitoring to wildlife management has been used in diverse contexts. For example, sonic monitoring has been used with farm pigs and chickens with the ultimate goal of decreasing animal stress and increasing productivity. Marx et al. (2003) detected and characterized call types of piglet vocalization and determined that parameters of energy emission, main frequency, and call duration were particularly appropriate to characterize call types [453]. Moura et al. (2008) used those parameters and confidence intervals through visual analysis of the spectrogram to identify calls in near real-time [454]. This elementary method worked well in a farm setting but would not be well-suited to more complex calls and more diverse contexts.

[453] G Marx et al. Analysis of pain-related vocalization in young pigs. *Journal of sound and vibration*, 2003

[454] DJ Moura et al. Real time computer stress monitoring of piglets using vocalization analysis. *Computers and Electronics in Agriculture*, 2008b

6.4.1.5 - Machine learning in bioacoustics

In order to tackle more complex sonic signatures and less curated environments, researchers have been using machine learning models. The recent development of deep learning has opened doors to new opportunities in bioacoustics. For acoustic identification of wildlife, deep learning-based models have been developed and tested for a variety of species, from amphibians[455] to bats[456], insects[457], or birds [458]. Such lab-based results are essential for accelerating the development of approaches to field deployments. Previous work from collaborator Clement Duhart has tackled similar challenges[459].

[455] Julia Strout et al. Anuran call classification with deep learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017

[456] Oisin Mac Aodha et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, 2018

[457] Ivan Kiskin et al. Mosquito detection with neural networks: the buzz of deep learning. 2017

[458] Ilyas Potamitis. Deep learning for detection of bird vocalisations. 2016

[459] Clement Duhart et al. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018

6.4.1.6 - ML In Panda context

To date, most work has focused on species presenting acute and frequent vocal communications or sonic behaviors. In the context of pandas, one challenge comes from the rarity of their sonic behavior, justifying the need for a robust automated system. In addition, simple spectrographic analyses fail in distinguishing some of the sonic activities or the subtle details in the sounds. Indeed, subtle variation in sounds can potentially reveal important information, such as what part of the bamboo is being eaten (e.g., shoots, culm, or leaves), and thus its nutritional content [460], or critical hormonal information [461]. Through refinement cycles, deep learning offers high potential for generating highly detailed results (if given a large enough database), and eventually presenting fine-grained classifications of sonic events.

[460] Xiao Yan et al. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support translocations. *Scientific reports*, 2019

[461] Benjamin D Charlton et al. Vocal behaviour predicts mating success in giant pandas. *Royal Society open science*, 2018

6.4.1.7 - Attached vs non-attached audio sensors

In addition to detecting subtleties in behaviors, the deep learning approach also offers the potential to understand subtleties in recording contexts. Indeed, in our case, we are training our system with recordings coming from both stationary microphones and collar-mounted audio sensors. Consequently, our system can potentially learn to recognize both types of recordings„ which makes it more modular and robust for use with different types of audio streams coming from various sources.

6.4.1.8 - ML in outdoor environments

Outdoor environments are dynamic and require continuous learning to increase classifiers' robustness to both episodic and permanent acoustic changes throughout the year. In addition, managed care environments can present a high level of noise pollution, containing anthropophonic and geophonic sounds, such as weather conditions (e.g., rain, wind), human activities, and other bio-acoustic sound produced by nearby animals, including insects. In order to detect and identify sounds coming from a target animal, the classifier must be able to isolate its specific sounds from the ambient acoustic environment. Filtering approaches are widely explored in lab-based experiments; however, their robustness is limited in real-field deployment. Initial system parameters change over time because of equipment attrition and environmental changes. Instead of classic filtering of the acoustic signal, our system tackles this challenge by learning the different environment acoustics, including the animal target and equipment attrition. This is done through a semi-automatic database augmentation mechanism which automatically extracts sounds that the system is not able to identify, based on a confidence function detailed in [462]. A flow controller limits the recording volume and parameterizes the extraction balance between unidentified and uncertain predictions. A custom-made online platform allows human experts to annotate and discuss these recordings while building a local acoustic database used to refine the classifiers iteratively.

[462] Clement Duhart et al. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018

6.4.2 - Audio Contexts

In this project, we tested our system to recognize 17 sonic events, including four vocal and sonic signatures behaviors of giant pandas (feeding, nursing, squawking, honking) as well as anthropophonic (e.g. airplanes, human voices), geophonic (e.g., rain, wind, quiet), and other bio-acoustic scenes (four species of birds and two species of insects). Our non-exclusive multiclass prediction model outputs real-time classification amongst 17 classes (among which four are subcategories of panda sounds and four others are subcategories of bird sounds).

Our choice of four panda-specific sounds is an effort to represent the three

main contexts for which automatic recognition of panda sounds is feasible (either vocal, incidental or ambient) and can yield important applications in understanding and conserving the species. Those three contexts are:

- **Daily activities**, for which sounds of feeding and resting can help infer activity budget (estimate % time engaged in feeding).
- **Caring for cubs**, for which sounds of nursing, as well as cub vocalizations (both plaintive sounds and those denoting contentment), can provide information regarding development and the quality of maternal care.
- **Mating context**, in which affiliative and aggressive vocalizations, as well as their specific acoustic characteristics, can help with breeding management in conservation contexts.

In the discussion section, we come back to these three contexts and discuss different ways our system of bioacoustic monitoring can help increase understanding of giant pandas and support the conservation effort.

6.4.3 - Methods

This system offers three contributions to the field of bioacoustics. First, the use of an initial general purpose bio-acoustic classifier (called T0) optimizes the initial data liberalization steps by reducing human labor in the generation of a clean annotated database. Second, once the system is in use, it presents an intuitive online web platform for continuous use on diverse and multiple audio streams. Third, in only four iteration steps, our system reached robust results in classifying four panda vocalizations with an accuracy of 94 percent

6.4.3.1 - Overview

In order to create a robust bio-acoustic model of panda vocalizations, we initially trained the model through an iterative process of database augmentation (illustrated in Figure XX). Classically, this step is very labor-intensive and requires human listeners to precisely sample and accurately label a large amount of audio data. Indeed, one major challenge in machine learning techniques is the constitution of high quality, diverse, and large labeled datasets. However, with recordings captured from live animals, both in captivity or in the wild, it is often impossible to obtain clean, organized sounds, and it is very time-consuming to clean and manually label the database.In our case, we reduce this labor-intensive step by using an initial (naive) general-purpose bio-acoustic classifier called T0. T0 has no prior knowledge about panda sounds at the first iteration but is already trained on a selected series of ambient sounds (such as birds, planes, etc.). This classifier is used for a pre-liberalization step to reduce human actions, turning what was an extensive annotation step into a faster confirmation step. Indeed, instead of labeling every sound of interest from the audio,

human experts only need to confirm or infer the presence and nature of panda sounds from specific samples extracted by the system.

Starting from T0, we introduced our training dataset in four successive cycles. In the first cycle, T0 processes and classifies our first set (25 percent of the database) and separates all the unknown sounds (panda sounds) from the ones it knows (car, ambient sounds, birds, human voices, etc.). All panda sounds are automatically edited and placed in the "I do not know" category. Expert listeners then use an online web platform to label those sounds, which are used to train our T1 classifier, thus finishing the first cycle. This operation is repeated four times, introducing another 25 percent of our original database each time. At each cycle, additional samples are correctly pre-labeled, reducing human labor.

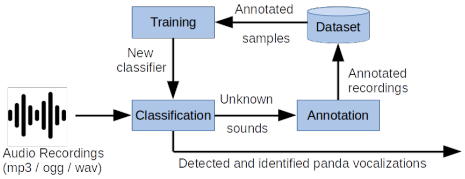


Figure 56: Processing Cycle

To measure classifier improvement, we cross-validate the new classifier (Tn+1) version based on the recordings previously annotated by the expert listeners. At this stage, improvement is measured by accounting for previously mis-classified samples being accurately identified.

To summarize, there are three steps repeated in each iteration cycle. The first step (**classification**) extracts samples of interest, the second step (**annotation**), presents the samples of interest to a human expert through an online platform for manual labeling, and finally, the third step (**training**) refines the classifiers based on the new database augmentation to create a more accurate model at each iteration.

6.4.3.2 - Database Organization

The system uses two different sources of sounds for its database generation: panda sounds and ambient sounds. For panda sounds, we used two different sources for panda vocalization combined with a general-purpose bio-acoustic dataset from the Tidmarsh project[463]. Dataset 1 contains collar-mounted recordings of relatively low quality but taken very close to the source, and Dataset 2 contains recordings from high-quality shotgun microphones.

[463] Clement Duhart et al. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018

The Tidmarsh database is used in two ways: 1) it is used to train the classifier to be able to analyze the general acoustics, such as geophony from weather, anthropophony from human activities, and other bio-acoustics from birds in order to detect panda vocalization; 2) it is used to augment the panda vocalization dataset by combining it with background noise, such as a panda sound on top of a rainy or windy background. In addition to providing additional samples for the training, it increases significantly the robustness of the classifier, as observed in the deep learning community. Some additional database augmentation techniques have been used,

such as adding noise, time translation, spectrogram blurring, and minimal frequency shifting. The final dataset is composed of 400,000 samples from the general bio-acoustic background and 15,000 samples from panda vocalizations, which are combined during the training stage.

The two panda datasets were initially composed of long recordings (1s to 1000s) containing the targeted sound signature along with silence, rest, and other ambient sounds either by themselves or layered with the panda sound. Those recordings were not usable as-is and had to be processed into a cleaner and more systematic organization. Our system is designed to account for this. The approach taken is to start with an initial model T0. T0 has formerly been trained using 400,000 extracted 500 ms recordings (called samples) distributed over 66 classes, including geophonic scenes (e.g., rain, wind, quiet), anthrophonic sounds (e.g., cars, airplanes, human voices) and bio-acoustic events (e.g., crickets, cicadas, amphibians, spring peepers, birds). Figure 57 presents the online annotation platform used by human experts to label the small proportion of samples prelabeled by Tn-1 manually.

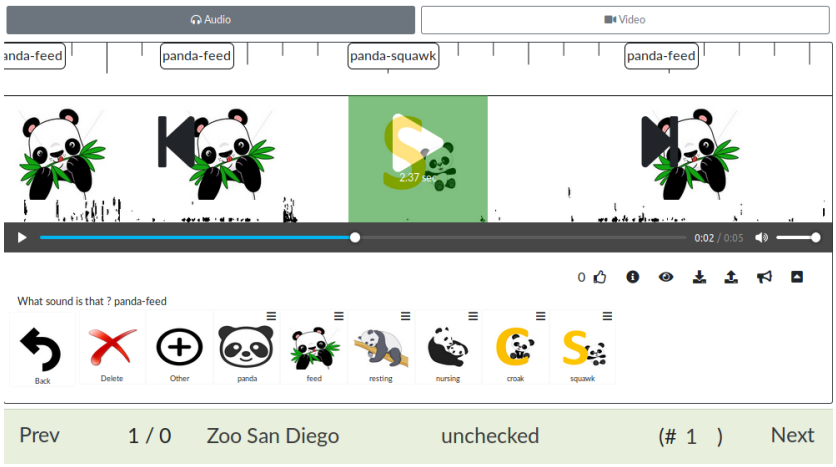


Figure 57: The online interface allows the user to play the audio recording as well as visualize the audio spectrum. When a known panda sound is heard, the user places a visual marker

6.4.3.3 - Panda Vocalization Classification

Our system offers high flexibility in the source of incoming audio and can work on pre-recorded audio as well as live streams coming from direct microphones. When received by the system, input audio streams are split into 500ms samples with an overlapping Hamming window. For each sample, a Mel-Filter Banks (MFB) spectrogram is computed to obtain a visual 2D representation of the sound, followed by background noise reduction using a medial filter on the spectrum. Finally, samples are normalized and filtered with frequency bandpass on the panda vocalization range, between 50 Hz and 12 kHz. Our choice of a simple preprocessing step reflects a trade-off between the real-time constraints and available computing resources.

The classifier used is a convolutional neural network (CNN) with an expert architecture, as illustrated in Figure 58. The classifier layer is separated into three interdependent sub-classifiers: an Acoustic Scene Classifier (ASC), an Acoustic Panda Classifier (APC) and an Acoustic Bird Classifier (ABC). The ASC computes a general classification that weighs in the inhibition of the APC according to its probabilistic estimation of panda presence. The classifiers are trained by turns and share the same stack of convolutional layers responsible for acoustic feature learning. The cost function is a non-exclusive, multi-class cross-entropy with parameter regularization. The ASC (T0) was trained on seven different acoustic scene samples (quiet, wind, rain, airplanes, crickets, cicadas, and human voice) as well as brief generic sonic events for expert classifier inhibition control for the bird and panda subcategories. The ABC was trained on four common bird species (crow, sparrow, robin, and blackbird). Finally, the APC was trained on four panda sounds of unique acoustic signatures: feeding, nursing, squawking, and honking.

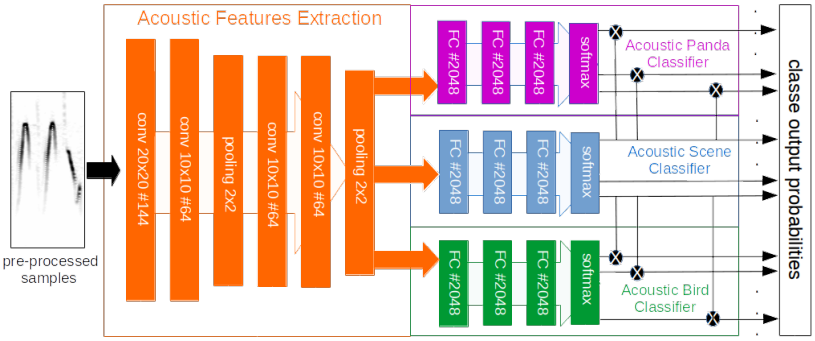


Figure 58: Architecture

6.4.4 - Results of the Panda Project

In order to evaluate the accuracy of our classifier, the original datasets have been balanced (each class was given an equal number of sample) and split into training (80 percent of samples) and testing (20 percent of samples) datasets. Final training and testing datasets are a combination of both. Hence the evaluation weights panda sounds equally to the other bio-acoustic classifiers.

Our final training reached an accuracy of 94.2 percent, with a balanced error rate between the different classes as observed in Figure 60 showing the confusion matrix on the testing dataset. Figure 59 presents the different confusion matrices obtained after each of the four training database augmentation cycles. Even though it has no knowledge of panda sounds, we can see that the classifier T0, as initially used for extracting panda

backgrounds, helps us by bootstrapping the system for the next training iteration, thus refining the classifier improvement until reaching the final training cycle, number 5

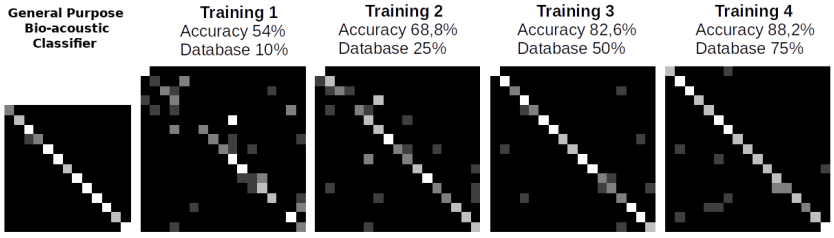


Figure 59: Accuracy in classification after 1, 2, 3, and 4 cycles of training

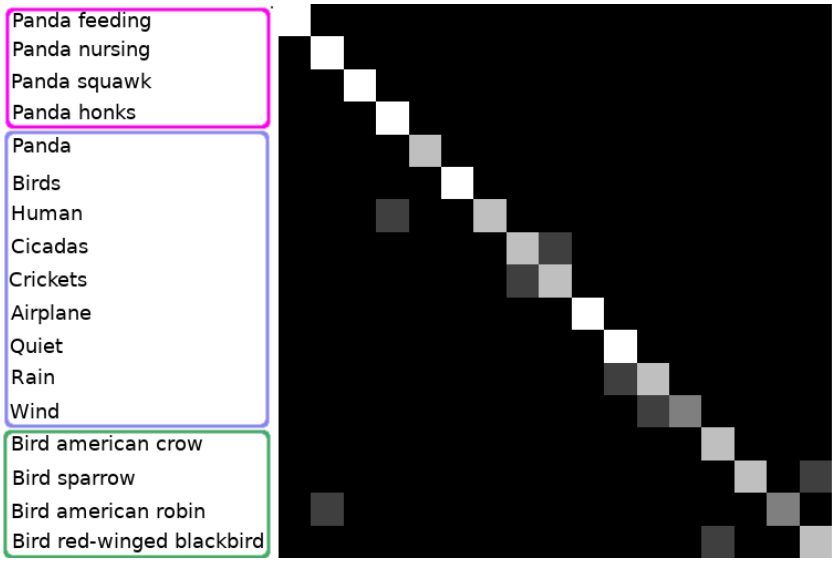


Figure 60: Accuracy in classification after the 5th cycle of training

6.4.5 - Application and Discussions of the Panda Project

6.4.5.1 - Potential applications

The method described above and the initial positive results for accuracy and robustness in classification opens a new door toward understanding, studying, and managing giant pandas. Given prior knowledge in panda acoustic ecology, we describe three behavioral contexts in which the application of our system of automated bio-acoustic pattern recognition can help increase understanding of giant pandas and support conservation efforts:

Feeding and Resting: The classification of resting and feeding behaviors is important to infer activity budget (estimate percentage of time engaged in feeding). By providing an accurate estimation of feeding vs eating

behaviors, our system could help detect changes in appetite or outbreaks of disease in zoo contexts, or inadequate food consumption in the wild.

Visual monitoring is often not sufficient to infer the quantity of food consumed by each individual, especially if several animals are sharing the same enclosure. In addition, in the wild, location sensors alone offer limited potential to classify feeding and drinking behaviors. Wijers has demonstrated the potential of audio to help classify drinking behaviors that are often misclassified with location sensors alone[464]. In this context, acoustic monitoring offers unique potential, as feeding and resting have very different sonic signatures in giant pandas.

[464] Matthew Wijers et al. Listening to lions: animal-borne acoustic sensors improve bio-logger calibration and behaviour classification performance. *Frontiers in Ecology and Evolution*, 2018

Nursing and caring for cubs: Acoustic monitoring can also provide a unique understanding of mother/cub activities. In the captive setting, suboptimal maternal care has been noted in cases where very young cubs have difficulties suckling [465] and human intervention was needed to increase the likelihood of cub survival. In such cases, careful recording of how many times the cub nursed daily and when it suckled is needed. When a mother panda wears a collar-based microphone, acoustic monitoring can help detect suckling from its acoustic signature, and our system can provide such information in real time from an audio stream.

[465] Xiangming Huang et al. Human assistance in a giant panda mother for rearing her baby. *Sichuan journal of zoology*, 2005

The recognition of suckling behavior can also help in cases of twin births. In giant pandas, twins occur in 46.4 percent of births and often result in the mother choosing to care for only one baby and reject the other [466] Successful results have been obtained by alternating the maternal and human care for each cub [467]. In such cases, acoustic monitoring of sucking time for each cub could help to ensure adequate feeding for each cub.

[466] Rebecca J Snyder et al. Giant panda maternal care: A test of the experience constraint hypothesis. *Scientific reports*, 2016

[467] Y Huang et al. Use of artificial insemination to enhance propagation of giant pandas at the wolong breeding center. 2002

Playing recordings or appropriate cub vocalizations have also shown to help facilitate maternal care in females who reject their cubs right after birth, suggesting the importance of monitoring and generating a better understanding of those vocalizations [468]

[468] GQ Zhang et al. A method for encouraging maternal care in the giant panda. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2000

Cub vocalization can also provide important data to caregivers. Researchers have identified two distinct types of cub vocalization more often associated with plaintive vs content behaviors and have manually labeled such audio recordings in their databases. In addition, Charlton et al. [469] have provided evidence of experienced multiparous mothers spending more time engaging in key maternal behaviors (nursing, grooming, and holding cubs) resulting in less vocal cubs than those from new, inexperienced mothers. This further supports the idea of monitoring cub vocalizations to assess the mother's quality of care.

[469] Rebecca J Snyder et al. Giant panda maternal care: A test of the experience constraint hypothesis. *Scientific reports*, 2016

Furthermore, as pandas are the smallest placental mammal at birth (weighing only about 100 grams) compared with adult size, vocal communication to convey distress or arousal states to the mother might be essential for survival in the wild [470] a better understanding and monitoring of those behaviors might help with the preservation effort.

[470] Angela S Stoeger et al. Acoustic features indicate arousal in infant giant panda vocalisations. *Ethology*, 2012a

Mating: Giant pandas are solitary for the greatest part of the year, except during the mating season when they interact more with conspecifics. During this season, their vocalizations are more varied and can reveal meaningful information about their intentions, fertility, and aggressiveness. At least one translocated male panda is presumed to have died during the mating season from injuries received during intrasexual competition [471]. However, in most cases, pandas' distinctive vocal behaviors reflect appropriate courtship and breeding behavior and can help with the preservation effort.

For instance, both males and females emit bleating sounds during the mating season that contain encoded information about the caller's sex and age [472], as well as size and identity [473]. These sounds might be interesting to monitor in assessing appropriate response behavior or understanding the contextual acoustic surroundings, both in terms of calls emitted by an individual wearing a collar, for instance, or calls heard by that individual.

In addition, adult vocalizations can also contain information on their hormone levels: female giant panda chirps have the potential to signal the caller's precise estrous stage [474]; for males, bleat duration is a marker of androgen levels [475]

The three contexts described above and their associated vocal, incidental, and ambient sonic signatures can play an essential role in assessing and improving management strategies.

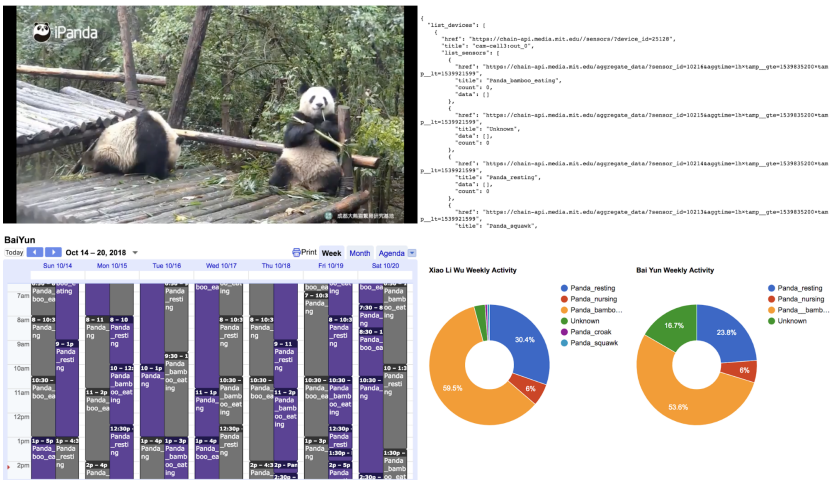


Figure 61: A secure online platform allows caregivers and researchers to access the real-time feed and classification of panda vocalizations, as well as the back-logged data organized as a personal calendar of each specimen

6.4.5.2 - A case for ambient sounds

In previous works, other ambient sounds (self-vocalizations, vocalizations emitted by other species, anthropogenic sources, or environmental sources) are often seen as interferences from the targeted behaviors and are seen as a challenge for classification[476]. In our work, such soundscape elements

[471] Xiao Yan et al. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support translocations. *Scientific reports*, 2019

[472] B D Charlton et al. The information content of giant panda, ailuropoda melanoleuca, bleats: acoustic cues to sex, age and size. 2009b

[473]

[474] B D Charlton et al. Female giant panda (ailuropoda melanoleuca) chirps advertise the caller's fertile phase. *Proceedings of the Royal Society B: Biological Sciences*, 2009a

[475] B D Charlton et al. Vocal cues to male androgen levels in giant pandas. *Biology Letters*, 2010

[476] Matthew Wijers et al. Listening to lions: animal-borne acoustic sensors improve bio-logger calibration and behaviour classification performance. *Frontiers in Ecology and Evolution*, 2018

help improves the robustness and the potential of our system in four ways.

First, we trained our system with both types of sounds for increased accuracy and to make it more modular and robust to different types of audio streams coming from various sources[477].

[477] Clement Duhart et al. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018

Second, in providing contextual soundscape, our system can potentially help researchers better understand the drivers of panda behaviors. Previous work has shown the possible use of acoustic monitoring and automatic classification of vocalization, as well as contextual audio classification for the study of freely behaving Eurasian jackdaws in both captivity and the field [478]. Our system further automates the recognition of contextual scenes by allowing it to learn novel sounds on the fly.

[478] D Stowell et al. On-bird sound recordings: automatic acoustic recognition of activities and contexts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017

Third, some types of ambient sound might potentially increase hormonal stress in giant pandas in a long-lasting way [479], especially for females, and might affect their reproductive conditions. In consequence, automatic monitoring of ambient sounds and their quality might help with the reproduction effort in captivity.

[479] M A Owen et al. Monitoring stress in captive giant pandas (ailuropoda melanoleuca): behavioral and hormonal responses to ambient noise. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2004

Fourth, the system can help researchers recognize the animal's exposure to human activities and identify anthropogenic events. Once understood through live learning and added to the classification collection, such sounds can also help managers in detecting possibly threatening activities (gunshots, construction, logging, livestock encroachment) and acting in a timely way to prevent harm to the animals. Previous examples of using acoustic bio-loggers for demonstrating the risks of human activities on wild harbor porpoises, right whales, etc., include [480].

[480] D M Wisniewska et al. High rates of vessel noise disrupt foraging in wild harbour porpoises (phocoena phocoena). *Proceedings of the Royal Society B: Biological Sciences*, 2018; and R M Rolland et al. Evidence that ship noise increases stress in right whales. *Proceedings of the Royal Society B: Biological Sciences*, 2012

6.4.6 - Conclusion of the Panda Project

Our project offers a tool for panda caregivers and researchers to help with the understanding of panda behavior for continued conservation management. Acoustic monitoring offers a promising approach to the detection and classification of behaviors and life history events, especially as giant pandas are an ideal species for vocal, incidental and ambient acoustic monitoring.

Application of acoustic monitoring to conservation management requires the development of automated techniques for data extraction from recordings. Our open source deep-learning approach offers a successful tool for real-time identification and classification of seven panda vocalizations, in addition to chewing and nursing sounds.

6.5 - Giving agency to animals & the JoyBranch project

Once we have learned to appreciate and question the symphony of voices in the zoo; once we have acknowledged our human limitations in decoding animal-animal vocal interactions; and once we have developed new ways to listen to individual animals' voices, we have to think of ways to improve animals' living conditions and wellbeing in regard to their sonic and vocal environments.

Enrichment is a methodology for caregivers to offer zoo animals improved psychological and physiological well-being. Although many species rely on auditory senses, sonic enrichment is rarely implemented. Zoo soundscapes are dominated by human-generated noises and do not respond meaningfully to animals' behavior. Designing interactive sonic enrichment systems for animals presents unique ergonomic, ethical, and agency-related challenges. The following project presents a case study of such design. We deployed two novel interventions at the San Diego Zoo to allow Sampson, a music-savvy hyacinth macaw, to gain control over his sonic environment. Our results suggest that (1) the bird uses, understands, and benefits from the system, and (2) visitors play a major role in Sampson's engagement with this technology. With his new agency, the bird seemingly gains more control over his interactions with the public, creating an interspecies experience mediated by technology. The resulting animal-human-computer interaction may inform mediated interspecies experiences in the future.

6.5.1 - The JoyBranch Project - Context

Figure 62: the Hyacinth Macaw Sampson interacting with the JoyBranch. The bird controls the music in his exhibit by holding the stick with his foot and beak

Sampson is a solitary, 18-year-old hyacinth macaw (Anodorhynchus hyacinthinus) living at the San Diego Safari Park. According to his expert caregivers, the bird likes music and has some favorite songs and specific musical tastes depending on his mood. The caregivers provide music for him occasionally during care sessions but would like to do so more frequently, and in ways that the bird can control himself. With the help of animal experts and zoo professionals, we designed and deployed two systems to allow Sampson to control and play music for and by himself from a series of five curated songs. Our design objectives were for our system to

be ergonomic, ethical, engaging, understandable and agency-enhancing. The first intervention, JoyBranch, is a physical joystick embedded within a tree branch that the bird manipulates with his beak or feet to play a song. The second intervention, BobTrigger, relies on the bird's head bobbing – a natural behavior associated with positive engagement – as a visual cue to trigger music. We deployed and tested the two systems over five days. The presence of the experimenter and zoo visitors during the sessions revealed important factors regarding the bird's engagement with the system. We also collected comments and structured interviews from Samson's expert caregivers, Jenna Duarte and Michelle Handrus. This project was approved by the Zoo IACUC committee, which oversees the ethics and animal wellbeing during research.

Sampson's specific situation, both in terms of location within the zoo and personality, was central to this implementation and to the interpretation of the results. Sampson's enclosure is about 4x8m and is surrounded by a waist-high fence (see figure 63), preventing guests from approaching too closely. The enclosure contains a large perch that allows him to move freely toward the four walls of his enclosure without going down to the ground.

Figure 63: Sampson's enclosure is at the entrance of the zoo and visitors often stop by and focus their attention on the bird. Sampson is particularly sensitive to the attention of children.

Every day, Sampson is brought in to his enclosure at around 5:30am and brought to his night chambers at around 3pm. Sampson's exhibit is located at the entrance of the zoo, making Sampson the first animal visible to visitors. The San Diego Safari Park was visited by 1.5 million guests in 2018, which means that an average of 4,000 guests passed by Sampson daily. The entrance avenue is large, and only a fraction of visitors actually stop to pay attention to the bird[481]. As most parrots do[482], he likes to "show off" anyway. He will display his feathers, vocalize, and parade in front of his keepers and visitors to attract their attention. We are told that

[481] Rébecca Kleinberger, Janet Baker, and Gabriel Miller. Initial observation of human-bird vocal interactions in a zoological setting. *PeerJ Preprints*, 2019a

[482] Andrew U Luescher. *Manual of parrot behavior*. 2006

he is very sensitive to audience attention, and if visitors spend too much time looking at other nearby animals, he will express his frustration by vocalizing loudly at them. He expresses specific signs of excitement (head bobbing, head nodding, positive vocalizations) when seeing volunteers or patrons he knows well. The bird also particularly likes children, and he exhibits more signs of arousal when young children and students gather around his enclosure.

6.5.2 - Background for the JoyBranch project

6.5.2.1 - Sonic Enrichment in Zoos

From the historical, human-focused menageries of the 18th century to the more animal-focused preservation sanctuaries of today, zoos have evolved to respect and protect species, as well as to educate the public about them[483]. Today, caregivers are specially trained and highly sensitive to the needs of the animals under their care. Formally introduced in zoo husbandry in the 1940s, behavioral enrichment offers a framework for zookeepers and caregivers to enhance the quality of life of captive animals, enable them to express their natural behaviors, and reduce boredom and stereotypic behaviors. Stereotypic behaviors are abnormal behaviors frequently seen in captive animals, especially in hand-reared animals[484]. They are indicative of poor psychological wellbeing and may include pacing, rocking, swimming in circles, excessive sleeping, self-mutilation, and feather-picking. The introduction of enrichment interventions aims at improving the quality of captive animal care by identifying and providing "the environmental stimuli necessary for psychological and physiological well-being"[485]. One of the first evaluations of enrichment apparatuses dates back to 1978[486].

Most enrichment techniques used today involve enclosure design, toys, food delivery, adapted puzzles, co-housing of different species, and introduction of sensory stimuli. Concerning sensory-based enrichment, Wells[487] considers that the greatest benefits for animal welfare are obtained through enrichments that target the dominant sense of animals. Hearing is a dominant sense for many species, and a multitude of studies have evaluated the welfare benefits of music—including natural sounds[488], classical music[489], country music[490], and radio broadcasts[491]—in enhancing the welfare of captive animals. However, most of the existing studies only evaluate the effects of static recordings played and controlled by experimenters and thus do not provide insights into how sonic interactiveness (i.e., customizing sounds based on responses from animals) might improve the efficacy of such sonic stimuli. In this project, we tackle this gap in research on animal welfare by demonstrating instances of interactive sonic stimuli for a captive animal at the San Diego Safari Park. Kim-McCormack[492] shows the grow-

[483] Vernon N Kisling. *Zoo and aquarium history: Ancient animal collections to zoological gardens*. CRC press, 2000

[484] Isabelle Williams et al. The effect of auditory enrichment, rearing method and social environment on the behavior of zoo-housed psittacines (aves: Psittaciformes); implications for welfare. *Applied Animal Behaviour Science*, 2017

[485] David J Shepherdson. Environmental enrichment: past, present and future. *International Zoo Yearbook*, 2003

[486] Robert Yanofsky et al. Changes in general behavior of two mandrills (papio sphinx) concomitant with behavioral testing in the zoo. *The Psychological Record*, 1978

[487] Deborah L Wells. The effects of animals on human health and well-being. *Journal of Social Issues*, 2009

[488] AS Chamove. Cage design reduces emotionality in mice. *Laboratory Animals*, 1989

[489] Elaine N Videan et al. Effects of two types and two genre of music on social behavior in captive chimpanzees (pan troglodytes). *Journal of the American Association for Laboratory Animal Science*, 2007; G Gvaryahu et al. Filial imprinting, environmental enrichment, and music application effects on behavior and performance of meat strain chicks. *Poultry Science*, 1989; and Deborah L Wells et al. Auditory stimulation as enrichment for zoo-housed asian elephants (elephas maximus). *Animal Welfare*, 2008

[490] Nikki S Rickard et al. The effect of music on cognitive performance: Insight from neurobiological and animal studies. *Behavioral and Cognitive Neuroscience Reviews*, 2005; and K Uetake et al. Effect of music on voluntary approach of dairy cows to an automatic milking system. *Applied animal behaviour science*, 1997

[491] L Brent and O Weaver. The physiological and behavioral effects of radio music on singly housed baboons. *Journal of medical primatology*, 1996; and RB Jones. Environmentai enrichment: the need for practicai strategies to improve pouitry weifare. *Welfare of the laying hen*, 2004

[492] Nicky NE Kim-McCormack et al. Is interactive technology a relevant and effective enrichment for captive great apes? *Applied animal behaviour science*, 2016

ing relevance of interactive digital applications for captive primates and insists on the importance of giving control to the animal. We believe that giving animals a similar kind of agency in shaping their sonic surroundings, especially in turning systems on or off, may yield substantial benefits.

Very few past instances of enrichment interventions have targeted the animal's sonic environment, although auditory input is a major way in which most species perceive the world. In addition, current zoo sonic environments are often limited, are highly disrupted by human-generated noise, and do not respond meaningfully to animal behavior. Seventy-four percent of zoos surveyed by Hoy and colleagues never provide auditory enrichment to their captive animals, even though more than half of caregivers report that it is "important" or "very important"[493]. The few existing programs in sonic enrichment lack interactivity and do not respond meaningfully to animal behavior: for example, simple looped recordings of natural environments coming out of loud-speaker systems do not meaningfully respond to animals' attempts to interact. The ability to interact seems indeed paramount in assuring mental safely and maximum benefit for the animal, and a captive animal's relationship with humans is a major factor that influences the way it interacts with its environment[494]. Agency can be defined as the propensity to engage actively with the environment with the main purpose of gathering knowledge and enhancing its skills[495]. The ability to interact seems indeed paramount in assuring realistic interaction and maximum benefit for the animal.

[493] Julia M Hoy et al. Thirty years later: Enrichment practices for captive mammals. *Zoo Biology*, 2010

[494] Anna M Claxton. The potential of the human–animal relationship as an environmental enrichment for the welfare of zoo-housed animals. *Applied Animal Behaviour Science*, 2011

[495] Marek Špinka, Françoise Wemelsfelder, et al. Environmental challenge and animal agency. *Animal welfare*, pages 27–43, 2011

6.5.2.2 - Animal Music

Previous work has also targeted animals' ability to produce music. Gupfinger presents a review of non-human musical expression[496] and proposes a phenomenology of animal music classifying instances into: Animal Movement as Control Source; Unconscious Performers; Trained Musicians; and Voluntary Musicians. We were inspired by their methodology for creating animal-centered musical interaction design, taking into account musical capacities as well as physical and cognitive abilities. Our objective differs as we are not trying to make the bird more musical, but to provide an enrichment system based on music and agency. Pons explores choice in sonic enrichment for orangutans in captivity by manipulating objects[497]. They use sounds (instead of human-made music) to allow them to create their own "music." This is an important step in the design of interactive enrichment, as it tackles the question of sonic agency for animals. Our designs and interventions further the work by testing and analyzing how an interactive auditory enrichment system performs in practice. French's work[498] on interactive auditory enrichment for elephants gives precedent for interventions with a physical trigger in non-primate captive animals.

[496] Reinhard Gupfinger et al. Animals make music: A look at non-human musical expression. *Multimodal Technologies and Interaction*, 2018

[497] Patricia Pons et al. Sound to your objects: a novel design approach to evaluate orangutans' interest in sound-based stimuli. In *ACI'16*. ACM, 2016

[498] Fiona French et al. High tech cognitive and acoustic enrichment for captive elephants. *Journal of neuroscience methods*, 2018a

The "Sound Jam" workshop[499] demonstrates momentum within the ACI field to creating interactive auditory enrichment systems. Our work greatly benefits from the current climate of innovation in the field — a climate that is sensitive to issues of agency and ethics in animal technologies. Further motivating the need for interactiveness, Rivto found that there may be differences in how orangutans and humans experience music[500]. When given the choice, orangutans often chose silence over listening to sound. The animals exercise their right to choose not to have a sound played. Thus, interactive systems that give animals a choice in what they hear, as our system aims to do, is essential to creating an intervention that is truly enriching.

6.5.2.3 - Human-animal relationship (HAR) as enrichment

Hosey lays the groundwork for understanding human-animal relationships in zoos and presents a model for understanding how past experiences with humans inform relationships in the present[501]. His work strengthens our argument that it is necessary to consider the role of humans in the life of zoo animals especially during interventions. Indeed, human presence can often be detrimental to the wellbeing of animals in managed care. For instance, the presence of human visitors increased distress levels of wolves[502], pandas[503], orangutans[504], and koalas[505]. Anthropogenic noise pollution, such as construction noise, has also been shown to increase stress and reduce healthiness in big cats[506] as well as laboratory, domestic, and free-living animals[507]. However, certain human-animal interactions offer possible benefits, especially for zoo animals. The uniqueness of each human-animal diad can help explain the complexity of the connections between animals in managed care and their primary caregivers[508]. In her work, Claxton explores the effects of daily contact with both familiar caregivers and unfamiliar visitors and concludes that those interactions can lead to positive outcomes for the animals if the interaction with the humans is intentionally designed to address environmental enrichment aims[509]. She also highlights the importance of tailoring the human contact on a species-by-species basis. Understanding the interactions between animals in managed care and zoo visitors can allow visitor characteristics and behaviors that are most appealing to animals to be determined and lead to higher levels of animal-human interaction[510], playfulness[511], and energy expenditure[512].

2018 Ig Nobel anthropology prize winners Persson et. al showed that zoo-housed chimpanzees imitate human visitors as often and as well as visitors imitate the chimpanzees[513]. The ability and interest that the primates have in engaging in so-called imitative games are thought to help maintain social engagement. Not only can they imitate, but they can also recognize

[499] Fiona French et al. Soundjam: acoustic design for auditory enrichment. 2018b

[500] Sarah Elizabeth Ritvo. Music preference and discrimination in three sumatran orangutans. 2013

[501] Geoff Hosey. A preliminary model of human–animal relationships in the zoo. *Applied Animal Behaviour Science*, 2008; and Geoff Hosey et al. Are we ignoring neutral and negative human–animal relationships in zoos? *Zoo biology*, 2015

[502] María Pifarré et al. The effect of zoo visitors on the behaviour and faecal cortisol of the mexican wolf (canis lupus baileyi). *Applied Animal Behaviour Science*, 2012

[503] Sandra Quadros et al. Zoo visitor effect on mammal behaviour: Does noise matter? *Applied Animal Behaviour Science*, 2014

[504] L Birke. Effects of browse, human visitors and noise on the behaviour of captive orang utans. *Animal Welfare*, 2002

[505] Megan J Larsen et al. Number of nearby visitors and noise level affect vigilance in captive koalas. *Applied Animal Behaviour Science*, 2014

[506] Julia Chosy et al. Behavioral and physiological responses in felids to exhibit construction. *Zoo biology*, 2014

[507] Caitlin R Kight et al. How and why environmental noise impacts animals: an integrative, mechanistic review. *Ecology letters*, 2011

[508] Samantha J Ward et al. Keeper-animal interactions: Differences between the behaviour of zoo animals affect stockmanship. *PloS one*, 2015

[509] Anna M Claxton. The potential of the human–animal relationship as an environmental enrichment for the welfare of zoo-housed animals. *Applied Animal Behaviour Science*, 2011

[510] Shelley Cook. Interaction sequences between chimpanzees and human visitors at the zoo. *Zoo Biology*, 1995

[511] C Owen. Do visitors affect the asian short-clawed otter in a captive environment. In *Zoo Research Symposium*, 2004

[512] AJ Nimon et al. Cross-species interaction and communication: a study method applied to captive siamang and long-billed corella contacts with humans. *Applied Animal Behaviour Science*, 1992

[513] Tomas Persson et al. Spontaneous cross-species imitation in interactions between chimpanzees and zoo visitors. *Primates*, 2018

when they are being imitated. In this work, joint attention is key to a rich, mutual, cognitively beneficial multispecies shared experience. Other previous work on primates also highlights the acute awareness of animals for the experimenters. Various species of primates and birds have been shown to be seemingly aware of the impact of their actions on others' behavior[514] and are able to grasp the experimenter's intention[515] or, at least, to engage socially as well as gazing directly into the experimenter's face[516]. Far from only serving as entertainment for the animals, such human-animal relationship (HAR) interactions appear to be related to general social and communicative needs.

Parrots especially, as a social and communicative species, are known for their need for interaction and attention, often interpreted as "showing off" or "doing a performance"[517]; Duarte reported that "Sampson loves to show off, especially to kids." To attract visitors, he will look at them, call, nod, bob, move back and forth in his enclosure, and flap his wings. "He wants compliments, he knows when people are praising him." Such performative behaviors reflect the importance of attention and companionship.

6.5.2.4 -The rival/model procedure

The rich literature on parrot learning and intelligence, specifically from the work of pioneer Irene Pepperberg, further highlights the incredible social intelligence of birds, their awareness of the interspecies interactions in their surroundings, and their acute need for attention. Before her work, behaviorists mainly used Skinner-operant conditioning and had deemed it impossible to teach parrots how to talk, but by using the model/rival method developed by Todt in 1975[518], Pepperberg was able to teach grey parrot Alex to identify and name more than 50 different objects and understand quantities up to six, among other abstract concepts[519]. The model/rival procedure used for training involves two people; one is the trainer and gives instructions, and the other is the model who gives correct and incorrect responses and acts as the student's rival for the trainer's attention. In some cases, the role of the rival can be played by another previously trained bird. The parrot, in the role of student, tries to reproduce the correct behavior motivated by gaining the attention of the human trainer[520]. This situation creates a triangular interaction between the bird, the trainer and the rival. The success of the model/rival procedures demonstrates the importance of the social context for parrots. In our case, we are less interested in training than in genuinely sparking the bird's interest and providing him with a playful tool and an experience that is intrinsically rewarding. It might be impossible to truly separate intrinsic (pleasure) and extrinsic (human attention) rewards in this context. This historical approach in the context of teaching helps us to not only understand the complex role of the

[514] Richard W Byrne and Lucy A Bates. Primate social cognition: uniquely primate, uniquely social, or just unique? *Neuron*, 2010

[515] Thomas Suddendorf and Andrew Whiten. Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological bulletin*, 2001

[516] Mark Nielsen. The imitative behaviour of children and chimpanzees: A window on the transmission of cultural traditions. *Revue de primatologie*, 2009; and Malinda Carpenter and Josep Call. Comparing the imitative skills of children and nonhuman apes. *Revue de primatologie*, (1), 2009

[517] Stephanie Spinner. *Alex the parrot: no ordinary bird*. Knopf Books for Young Readers, 2012; and Irene M Pepperberg and Mary A McLaughlin. Effect of avian–human joint attention in allospecific vocal learning by grey parrots (psittacus erithacus). *Journal of Comparative Psychology*, 1996

[518] Dietmar Todt. Social learning of vocal patterns and modes of their application in grey parrots (psittacus erithacus) 1, 2, 3. *Zeitschrift für Tierpsychologie*, 1975

[519] Irene M Pepperberg and Irene M Pepperberg. *The Alex studies: cognitive and communicative abilities of grey parrots*. Harvard University Press, 2009

[520] Irene M Pepperberg. Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science*, 11(3): 83–87, 2002

human relationship for the parrot, but also points to the importance of the sociological and psychological aspects underlying the human experience in the animal-computer-human triad.

[521] Julia M Hoy et al. Thirty years later: Enrichment practices for captive mammals. *Zoo Biology*, 2010

6.5.2.5 - ACI/HCI

Advances in technology can play an important role in creating enrichment systems for zoo animals, especially in the acoustic domain. This need arises from the limited time that caregivers and volunteers can spend with each animal[521]. However, we also believe that it is less desirable to provide a digital system for the animal to use in an isolated way than to design experiences that function in a social context while giving more control to the animals.

[522] Austin Stewart. second livestock. URL http://www.theaustinstewart.com/secondlivestock.html

[523] Lisa J Wallis et al. Utilising dog-computer interactions to provide mental stimulation in dogs especially during ageing. In *Animal-Computer Interaction*. ACM, 2017

[524] Andrea W Clay et al. The use of technology to enhance zoological parks. *Zoo biology*, 2011; and Becky Scheel. Designing digital enrichment for orangutans. In *Animal-Computer Interaction*. ACM, 2018

In terms of technological innovation and exploration, the field of animal-computer interactions has produced inspiring previous work in solo enrichment systems. Technology for animals is a growing industry. Pets and livestock have inspired designers, artists and entrepreneurs to create new technologies and concepts, from fictional VR headsets for chickens[522] to off-the-shelf automatic feeders[523]. In the context of zoo-housed animals, computers and tablets have been used to provide primates with a variety of enriching applications, from face-matching games to "Tinder for orangutans" in breeding research[524] In those examples, the main role of the human volunteers was to hold the devices. Keepers have expressed the need for enrichment systems that are "hands-off" because tablets often have to be held by caregivers, which is time-consuming and can cause safety issues. With the orangutans, frustration was observed because they were not able to hold the devices themselves[525].

[525] Marcus Carter et al. Naturalism and aci: augmenting zoo enclosures with digital technology. In *Advances in Computer Entertainment Technology*. ACM, 2015

[526] Ilyena Hirskyj-Douglas et al. Where hci meets aci. In *Nordic Conference on Human-Computer Interaction*. ACM, 2016; and Sarah Webber et al. Hci goes to the zoo:[workshop proposal]. In *CHI*. ACM, 2016

However, some of the most interesting works involving animals and digital technologies are the ones that acknowledge the key role played by humans in animal environments. This statement is not to argue that animals do not have rich inner and outer lives independent from the human race, but when we introduce human technologies into their umwelt, the human context should not be overlooked. Accordingly, there is a richness in exploring the junction where HCI and ACI meet[526]. The CHI community has proposed interesting instances of digital technologies used in the context of interspecies interactions between humans and non-human animals (or rather, between animal and human-animal). Previous inspiring work from the CHI community involving animals has acknowledged and curated the human's role and behavior to appeal to the specific animal or species, from a challenging play partner[527], to an empathetic audience[528], a conversational partner[529], or a provider of remote petting[530]. Such an approach has the potential to lead to higher levels of animal-human interaction, playfulness,

[527] Frank Noz and Jinsoo An. Cat cat revolution: an interspecies gaming experience. In *CHI*. ACM, 2011; Michelle Westerlaken and Stefano Gualeni. Felino: The philosophical practice of making an interspecies videogame. In *The Philosophy of Computer Games*, 2014; and Robert Yanofsky et al. Changes in general behavior of two mandrills (papio sphinx) concomitant with behavioral testing in the zoo. *The Psychological Record*, 1978

[528] Sarah Webber et al. Kinecting with orangutans: Zoo visitors' empathetic responses to animals? use of interactive technology. In *CHI Conference on Human Factors in Computing Systems*. ACM, 2017

[529] Donghyeon Ko et al. Bubbletalk: Enriching experience with fish by supporting human behavior. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 2018

[530] Ping Lee et al. A mobile pet wearable computer and mixed reality system for human–poultry interaction through the internet. *Ubicomp*, 2006

and energy expenditure[531] especially when the animal is provided with increased agency through the digitally mediated interspecies interaction. In[532], mandrills housed in the Washington Park Zoo could initiate a speed race game with a visitor by pushing a lit circle.

Human-animal interactions have been of interest to researchers both from a sociological and an ethnographic standpoint. Interactions with zoo animals have been seen as potential ways to foster caring attitudes toward individual animals and species[533]. The incorporation of digital technologies in these interactions can potentially increase empathy and engagements in a time of multimedia and vicarious representations of the living world[534]. Repeated meaningful interactions, especially with non-domesticated species, might help humans perceive a glimpse of the animal's otherness and their unique Ümwelts[535]. In addition, our work follows the theoretical foundation proposed by Westerlaken and Gualeni in[536] to create "Digitally Complemented Zoomorphism." Our work acknowledges and engages with the risks of focusing on the perception of animal needs, based on subjective human judgments and the human end of the animal-human relationship.

6.5.3 - Methods

In this section, we first present our design choices, ideation process, and the design of our two systems (JoyBranch and BobTrigger) before going over elements of mapping and musical choices. Finally, we present the deployment methodology.

6.5.3.1 - Approach and Design Choices

Our design objectives were for our system to be ergonomic, ethical, engaging, understandable and to increase the agency of the bird. This project is shaped by existing enrichment practices, interviews, constraints, related work, and feedback from zoo professionals. Before designing and deploying the systems, we had extensive discussions and interviews with over ten Zoo professionals during a preliminary 1-week preparation trip to the Zoo. During that trip, we gathered information about the bird's needs and specific character from two bird experts working in the avian reproduction center, as well as from his caregivers, Handrus and Duarte. Handrus has been caring for Sampson everyday for over a year. Duarte has extensive expertise on parrots and has worked with Sampson for 5 years. We also met and presented our design for review to other animal experts who were involved in ideation and helped decide between design alternatives.

We had the chance to discuss our research with internationally renowned experts in this field such as Irene Pepperberg who gave us advice on how to use the rival/model method as an early priming mechanism to grasp

[531] Anna M Claxton. The potential of the human–animal relationship as an environmental enrichment for the welfare of zoo-housed animals. *Applied Animal Behaviour Science*, 2011

[532] Robert Yanofsky et al. Changes in general behavior of two mandrills (papio sphinx) concomitant with behavioral testing in the zoo. *The Psychological Record*, 1978

[533] Michael D Kreger and Joy A Mench. Visitor-animal interactions at the zoo. *Anthrozos*, 1995

[534] Patricia G Patrick and Sue Dale Tunnicliffe. The zoo voice: Zoo education and learning. In *Zoo Talk*. Springer, 2013

[535] Thomas Nagel. What is it like to be a bat? *The philosophical review*, 1974

[536] Michelle Westerlaken and Stefano Gualeni. Felino: The philosophical practice of making an interspecies videogame. In *The Philosophy of Computer Games*, 2014

the animal interest toward a device rather than using it for training. Four professionals working at the New England Exotic Wildlife Sanctuary, a parrot sanctuary also gave us feedback on how to understand bird attention, stress and engagement. Moderate physical activity, locomotion toward objects and physical interaction with toys and other branches can be used as metrics for engagement. Stress can be observed through pacing, feather plucking, and excessive grooming. Boredom manifests through long periods of inactivity and immobility. Focused staring may suggest attention and assessment of possible threats, while repeated gazing towards people suggests interest and intrigue. Such metrics were used in our methodology to assess the ethical and engaging aspects of our systems. Wenfei Tong, a researcher on bird vocalizations, recommended us to pay close attention to the bird's relationship ith researchers and the audience, which prompted us to record bird-visitor interactions and use gaze, vocalization, locomotion and body orientation of the bird and of the visitors to assess their engagement and level of rapport. David Rothenberg, musician and expert in interaction and in playing music with animals inspired us to think about the extent to which the visitors may play the role of an audience in motivating the bird's behavior.

All those discussions highly informed our design (interaction, twig-aspect, trigger, easy "off button," bob triggers) and methods used (shared audio space, human-bird-experimenter dynamics) as well as deciding which data to gather (continual video, logs) and how to conduct our protocol and analysis. This helped us tailor our design objectives toward a natural-looking object on which the bird has immediate agency based on physical contact/trigger and gesture that presents an easy "off button" by not interacting with it for 10 seconds. Those discussions also drove us into using interactive musical elements that create a shared auditory space in which to observe human-bird-experimenter social dynamics.

Duarte recommended using the bird's head bobbing to trigger music. When she plays Sampson music, the bird often bobs, and when she stopped the music the bird stops, looks at her and bobs a couple more times. She interprets this as the bird wanting her to play more: "he wants the music back". This might create a positive bias that needs to be acknowledged and assessed in future studies.

It should be noted that changes in the bird's environment are routine. Indeed, in accordance with the zoo rules, Sampson's environment is modified every few weeks and enriched with new branches, ropes and passive enrichment objects such as bells and wooden objects. Those were a source of inspiration for our design as we used the same texture, material and external aspect as those regular enrichment elements for our JoyBranch design.

6.5.3.2 - System Design

Guided by those interactions with experts, we developed two interactive systems to potentially give Sampson control over his sonic environment. The first system is a physical device called the JoyBranch, which consists of a joystick mechanism embedded within a wooden log (fig 64) placed inside his enclosure. Each time the bird manipulates the joystick, music begins to play. When the stick is released, music continues for 10 seconds and then fades out linearly over 2s (fig 67). For the second intervention, called BobTrigger, music is activated manually by an experimenter each time the bird bobs his head. If bobbing stops for more than 10s, the music fades out over 2s; if the bobbing is interrupted for less than 10s and restarted, the music stays on (fig 67). This is controlled by a custom app.

The goal of the JoyBranch was to provide an interface to naturally entice the bird to interact and create a clear connection between his actions and the music activation. The interface design was inspired by existing low-tech enrichment techniques familiar to the bird (bells, ropes, wooden branches). The JoyBranch is designed to look as natural as possible: a section of a wooden log with a standard wooden perch attached. No electronics, sensors, cabling, or display screens are accessible or visible to the animal, and we used animal-grade wood designed for birds. Since parrots are very destructive, we designed our system to be breakable without endangering the bird. Inside the JoyBranch, the perch is attached to a joystick connected to an embedded Linux computer powered by a portable battery (fig 64). The bird only needs to push the stick five degrees to trigger the music, requiring only 1 Newton of force. The audio output triggered by the movement of the branch is sent by a Bluetooth transmitter to a receiver outside the enclosure. The sound is then played by a portable speaker. When the stick is released, it comes back to its neutral position. If the JoyBranch is not re-triggered within 10s, the music fades out. Within the enclosure,

Figure 64: JoyBranch closed (left) and opened (right) with hardware components visible

the JoyBranch is securely attached with zip ties to a permanent metal tray bolted to a perch. Sampson can approach the device from several angles while standing on nearby branches or the perch attached to the tray. The dimensions of the JoyBranch are designed to be ergonomic by matching Sampson's size and allowing him to hold the branch with his beak at different heights.

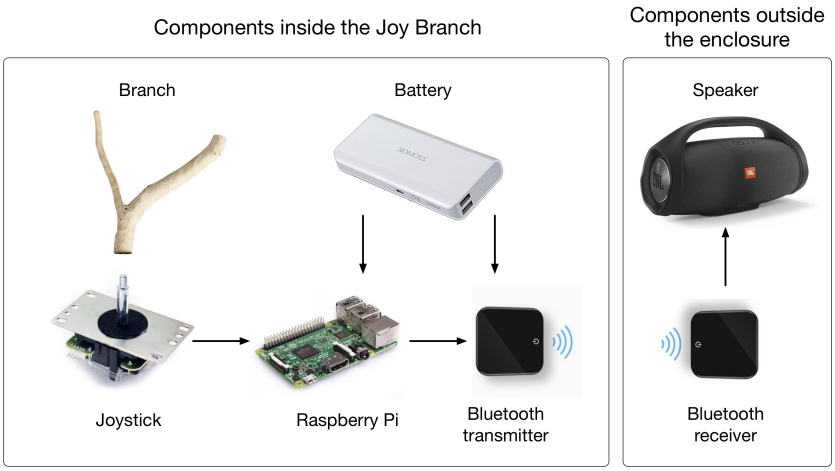


Figure 65: JoyBranch system diagram

The BobTrigger intervention doesn't require any additional elements in the enclosure and is engaging to the extent that it reinforces the bird bobbing behavior, commonly associated with positive engagement. Ethics are considered by only playing and maintaining the music when the bird actively bobs and expresses engagement. Although the bird has real agency over the music, understandability may be less obvious for the bird as the human in the loop is still technically in control and the relationship between his behavior and the music might be less clear for the bird.

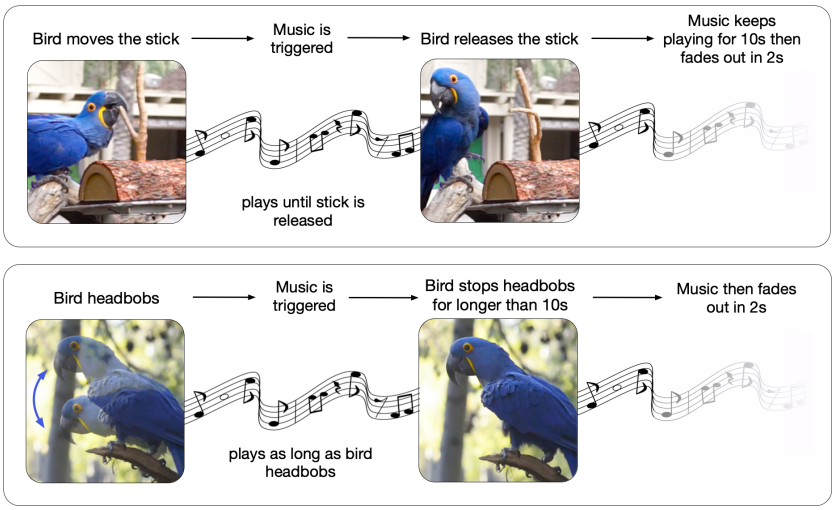


Figure 66: Interaction Design for the JoyBranch (top) and the BobTrigger (bottom) interventions

6.5.3.3 - Mapping

For each session, we selected one of five songs. The choice of having the music fade out in the absence of interaction was to provide the bird with a metaphorical "off button," so the music would not play for long periods of time. This method was approved by the animal experts consulted. This choice has clear limitations, as the bird has to retrigger continuously if he wishes the music to continue, which might cause frustration if the music keeps fading out, but we believe that this trade-off is needed to assure an ethical experience.

The system was originally designed to allow the bird to choose between different musical tracks depending on the orientation of the joystick. However, for this first iteration, we started with the simplest mapping of only one track at the time regardless of the joystick orientation. Future work will explore more complex mappings.

The musical tracks used were five beat-heavy, up-tempo, popular dance songs: "Billie Jean" by Michael Jackson, "Karma Chameleon" by Culture Club, "I Like to Move It" by Erick Morillo, "Get Lucky" by Daft Punk, and "Jump In The Line" by Harry Belafonte. We rotated through the different tracks for every session, so the experience would not be too repetitive, but we used similar genres to limit the songs' influence on the bird. Sampson is already familiar with the genre and some of the songs and is known for engaging positively with such music when played by his keepers.

Parrots are often mentioned in research on animals and music, as they exhibit two key indicators for musicality in animal species: vocal learning and entrainment[537]. Entrainment involves the ability to synchronize movements to a beat[538]. It has been thought to be unique to species that can produce vocal mimicry (elephants, lovebirds, parrots, and in particular a famous sulfur-crested cockatoo named Snowball) but has also been observed in non-mimics such as sea lions[539]. When music is playing, Sampson sometimes exhibits entrainment through head-bobbing and head-whipping behaviors. Those behaviors are interpreted positively by his keepers. Previous work has explored the capability of parrots to generate musical content through interactive instruments[540] including a swing that creates a sound modulated by swinging and a joystick that produces single notes when triggered. Such work is important in understanding the complex musicality of the animals. In our case, we were more interested in reactions to complete upbeat songs.

The choice of using music instead of natural soundscapes containing bird sounds was motivated by ethical concerns. Indeed, there are ethical issues[541] surrounding the use of so-called "audio playback"—the technique of playing back bird calls to engage birds —as it creates unfulfilled expectations of the presence of other birds, and may contain poorly understood bird calls that could have unexpected consequences.

[537] Marisa Hoeschele, Hugo Merchant, Yukiko Kikuchi, Yuko Hattori, and Carel ten Cate. Searching for the origins of musicality across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015

[538] Aniruddh D Patel, John R Iversen, Micah R Bregman, and Irena Schulz. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current biology*, 2009

[539] Peter Cook, Andrew Rouse, Margaret Wilson, and Colleen Reichmuth. A california sea lion (zalophus californianus) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *Journal of Comparative Psychology*, 2013

[540] Reinhard Gupfinger and Martin Kaltenbrunner. Sonic experiments with grey parrots: A report on testing the auditory skills and musical preferences of grey parrots in captivity. *Animal Computer Interactions*, 2017

[541] David Sibley. The proper use of playback in birding, 2011. URL https://www.sibleyguides.com/2011/04/the-proper-use-of-playback-in-birding/; and American Birding Association. Code of ethics, 2012. URL http://www.aba.com

System	JoyBranch	BobTrigger
Ergonomy	- Size & Material - Simple degrees of freedom - Low force required	- Natural bobbing behavior used for trigger
Ethics	- Only plays if actively triggered - Music stops if not re-triggered - Short intervention time	- Only plays if bobs - Music stops if not re-triggered - Short intervention time
Engagement	- Inspired by branch nibbling - Need to be physically active - No reward other than music	- Based on a behavior of positive engagement
Understandability	- Immediate relationship between gesture and music	- Emerges from previously observed behavior of bobbing when Samson wants music

Figure 67: Table of Design Objectives for JoyBranch and BobTrigger.

6.5.3.4 - Deployment

The evaluation aims to assess the enrichment potential for Sampson (frequency, understanding, ergonomy, agency) and the influence of visitors/experimenter on his use of the systems. Empirical methods used in the evaluation include interviews and observation. The testing lasted five consecutive days. On the morning of day 0 we ran three initial baseline sessions without any intervention. Then, from day 1 to day 4, we ran three sessions (one JoyBranch session and two BobTrigger sessions) each day between 6:30 am and 12 pm. We installed the JoyBranch on a tray with zip ties before 7 am, when Sampson is moved from his night enclosure to his exhibit. At 7 am, we ran a JoyBranch intervention session. At 9 am, we ran a first BobTrigger session (S2), and finally at 11 am we ran a second BobTrigger session (S3). The schedule was chosen to allow for at least an hour break between sessions. The sessions ran from 45 minutes to an hour each depending on the caregivers' schedule, who had to be present during the setup and breakdown. The resulting 16 sessions were observed by experimenters and facilitators and videotaped from two cameras at different angles for analysis and further inter-observer review.

On day 1, at the start of the initial JoyBranch intervention,we introduce the JoyBranch to Sampson through a 5-minute priming session. During this priming session, the experimenter showed Duarte how to interact with the system in front of Sampson. In doing so, we grasp the animal's interest toward the device and reassure the bird that the object is safe. For each BobTrigger intervention, the session was preceded by five seconds of music to announce the beginning of the session to the bird. In the analysis, we did not include times when Sampson's familiar caregivers were present because the bird is more naturally engaged and his attention is on the familiar faces in these situations. Moreover, we designed the two interventions specifically to enrich him during times when his keepers are not present.

6.5.4 - Analysis

6.5.4.1 - Tools and labels

In order to gather feedback on the animal's reaction to the interventions, we observed and labeled a series of different events, actions, and behaviors during the baseline recordings, the JoyBranch interventions, and the Bob-Trigger interventions, using a semi-assisted manual labeling interface[542]. For data collection, we categorized observations of events and behaviors. From the videos and field notes, we recorded the start and end times of a total of nine different observational behaviors from the bird: preening, feather puffing, stretching, standing on one leg, vocalizing, head bobbing, head nodding, eating, and moving within the enclosure. We also recorded instances of music being triggered and for how long. We collected a total of 1,645 events and behaviors.

[542] Olivier Friard and Marco Gamba. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 2016

6.5.4.2 - Attention labelling

We also collected information on the attention between the three agents, the visitors, the bird, and the experimenter/system through four metrics: bird attention toward the visitors (BAV), bird attention toward the experimenter (BAE), bird attention toward the JoyBranch (BAJB), and visitor attention toward the bird (VAB). To assess when the bird was focusing his attention toward the experimenter (BAE), we used field notes as well as the video recordings of all of the sessions, and considered moments when the bird's head was oriented perpendicularly to the experimenter's location, his eye fixed and focused toward the experimenter with relatively infrequent blinking, and he displayed little to no body motion or locomotion within the enclosure. We used the same metrics to assess when the bird was orienting his attention toward visitors (BAV). Contrary to the experimenter, who mainly remained still, visitors walk around the enclosure, and Sampson following them with his gaze is an additional indicator of his attention toward them. Because Sampson's enclosure is located at the only entrance to the Safari Park, an average of 10 visitors per minute pass him, but not every visitor actually pays attention to him. To assess visitor attention toward the bird (VAB), we used field notes and video recordings to see when people approach the enclosure, interrupt their walk, look at the bird, and stay for at least a few seconds. We also used clues such as when the visitors take photos, talk with each other about the bird ("Look! A macaw!", "Look mama, the big bird!", etc.) or vocally address the bird directly ("Hi, macaw!", "Hello, bird," "You are a beautiful bird!", etc.) We collected a total of 851 moments of attention.

6.5.4.3 - Caregiver interviews

The bird's caregivers Handrus and Duarte provided insight and help at all stages of the project. To not influence the bird during the interventions, they were not present during the sessions. We ran independent 1-hour long interview sessions with each of them. At the end of the last deployment day, 12 short 2-minute video clips of specific interesting behaviors from the sessions were shown to the caregivers. During the interviews, each clip was shown to the caregivers and they were asked 3 questions. For each clip, they were asked (1) to describe what they see, (2) whether they had seen this behavior before, and (3) how they would interpret this behavior. The recorded interviews were then transcribed into text and relevant information is presented in the Results section.

6.5.5 - Results

6.5.5.1 - Valence/arousal map

In order to reach a higher understanding of S, we analyzed and classified the diversity of observable behaviors. Valence/arousal models for humans are used extensively in affective computing and HCI. Using previous research in the behavioral ecology of parrots, keeper feedback and interviews, as well as our own observations, we established a behavioral valence/arousal map for S. Here we come back to each observable behavior and explain the motivations behind their classification.

Preening and **stretching** are grooming behaviors that occur when the bird is calm and not deeply engaged with its environment (low arousal). A parrot would not self-groom if it were anxious about possible danger (medium/high valence). **Feather puffing** can have a range of meanings, from being cold to being frustrated or anxious. It is generally associated with slightly negative experiences (low/medium valence, variable arousal). **Standing on one leg** is a very low-energy activity—a restful pose that might indicate that the bird is tired (low arousal, medium/low valence). **Eating** sessions are positive and often medium-energetic. Parrots are very dexterous at peeling nuts with their beaks and tongues (medium arousal, high valence). **Nibbling** is a playful, vigorous activity consisting of chewing and rubbing the beak with a stick or branch. We observed that during the intervention, Sampson sometimes nibbled on regular branches and sometimes on the JoyBranch stick. **Head nodding** is a series of small, up-and-down head motions (high arousal) whereas **bobbing** is a more vigorous up-and-down motion that at times includes the whole body (very high arousal). We defined one bob as a full up and down motion. Bobbing also encompassed moments when the bird included a leg lift while motioning and when he bobbed at an angle/a little sideways. **Head whips** are very

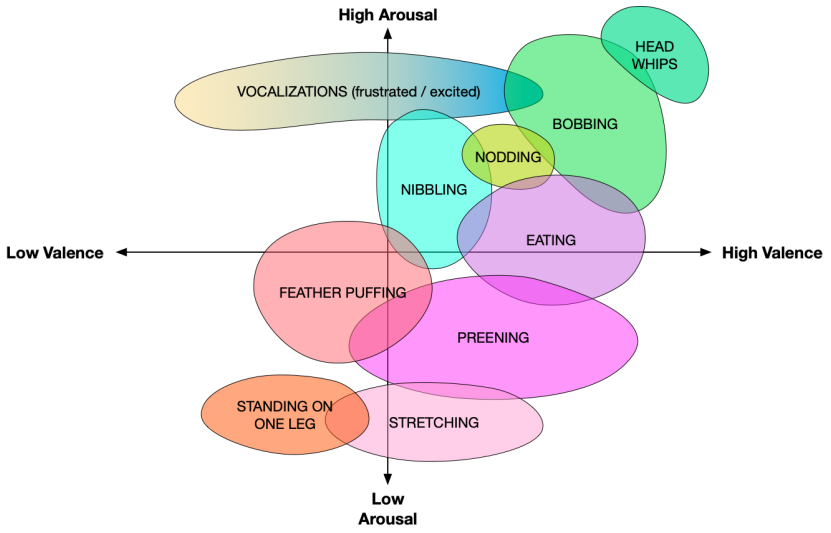


Figure 68: Arousal/valence map of Sampson's observational behaviors

vigorous throws of the bird's head back and forward (highest arousal). Sampson often repeats this motion rhythmically, but instances of one whip were counted. All three head motions are seen as positive or very positive and suggest the bird's engagement with the music (high valence). Finally, **vocalizing** describes any sound the bird made, be it soft or loud. Both soft and strong vocalizations can express contentment and engagement but in other cases, harsh vocalization can express frustration. Figure 68 represents Sampson's arousal/valence map of observational behaviors.

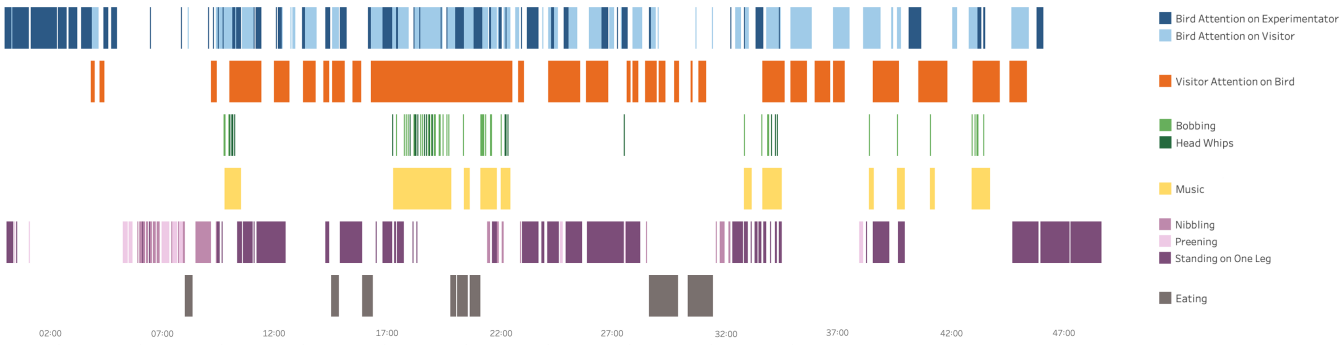


Figure 69: Representation of a typical example of behavioral correspondences during a session. (Day 2 Session 2, duration approx 50m.)

6.5.5.2 - Session Overview

In this section, we present a typical example of behaviors and suggest insights into some of the bird's critical behaviors and correspondences

that we observed (Fig 69). We notice that bird attention is relatively brief and tends to alternate between the experimenter and the visitors. Visitor attention is more maintained. The correspondence between bobbing and music is causal as, in this case, the music was triggered by the head bobs. We can observe apparent non-co-occurrence between eating and resting/grooming behaviors (nibbling, preening, and standing on one leg). We also observe that bobbing behaviors tend to occur in groups and seem to be reinforced by the music. Longer attention of visitors coincides with music playing. In the following sections, we go into more detail in the analysis of correspondences and potential causalities.

6.5.5.3 - JoyBranch evaluation

We were interested in assessing if and how the bird used the systems. For the JoyBranch, we noted each time the bird triggered the music, how long he triggered the music for, and in which way. Figure 70 shows all the times the bird triggered the branch (either by poking at it, nibbling on it, holding it with his beak, or holding it with one foot) and the resulting intervals when the music plays. The third row shows every occurrence of head bobbing or head whips.

The bird triggered the music a total of 33 times over four days, for a total of 31m33 s, representing 20 percent of the total session time. The bird used a fifth of his time interacting with the new device, and the duration of each trigger got longer over time, from an average of 8s for the first day to an average of 2m26s for the last day, which suggests increased interest (Fig 70).

Without any instructions, the bird found four ways to interact with the branch. Initially, he only triggered the music by nibbling on the stick. Nibbling on branches is a natural parrot behavior, as attested by Michelle Handrus: "He turned it on, is also playing with it in his beak as he would

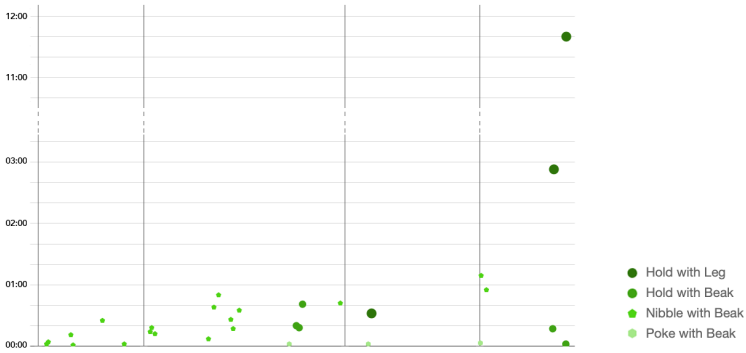


Figure 70: Day by day, the bird finds new ways to activate the JoyBranch, and the bob-triggers, hold-triggers, and feet-triggers increase in duration, suggesting learning and exploration

normally with sticks." However, the time he spent nibbling the branch (5 percent) was higher than the time spent nibbling on branches during the baseline (3.2 percent) or during the BobTrigger sessions (1.15 percent), suggesting a particular interest from the bird for the device. During the first day, Sampson nibbled on the branch and bobbed when the music would play, and once the music stopped, he would stop dancing and resume the nibbling and so on. It seemed that he hadn't yet made the connection between the music and his actions on the stick, according to Jenna Duarte: "I think he's just enjoying the music." However, during the second day, we observed several instances when the bird simply poked the stick with his beak and then turned his head toward the speaker and seemed to listen to the resulting music. Handrus interpreted this behavior as, "He's like, 'wait!' Did I do that?" According to Duarte, it can be seen as, "Like that speaker over there makes noise when I pull on the stick. Let's try it again and see if it does again." Then, he exhibited a novel behavior in holding the branch with his beak for extended amounts of time. Holding branches is uncommon for parrots, and both Duarte and Handrus told us that they had never seen this behavior before: "He like just went up and pulled it back. He didn't like...go up to it, like...trying to break it off. He didn't go up to it, and like, rub his beak on it. He just...literally pulled it back. Like he knew that if he pulled it back, I mean I'm just... yeah... He just literally, pulled it back, which is just kind of weird." According to Duarte: "I think that he has figured it out. They don't generally do that. They do that, you know...when they're playing they'll pull on things." Indeed, macaws are very strong and enjoy breaking branches and nuts, however, both keepers noticed that the bird was being very gentle with the device: "He could just break it, normally if he wanted, like, he would break it very easily, but he doesn't want to, he is being very careful." On the third day, the bird continued holding the branch with his beak and then started holding it with one of his feet. Foot-holding behavior enabled him to also look around while the music was playing.

The total time spent interacting with the device suggests a sustained interest in the JoyBranch, and newfound ways to interact with the branch day-by-day suggests learning, enjoyment, and interest. According to both our subjective observations and the caregiver's judgment, the bird understood the connection on day 2.

Bobbing always occurred during the music or shortly after, suggesting a strong connection between the two events as well as a positive experience for the bird. However, bobbing events became shorter and less frequent as the bird used more efficient ways to keep the music playing (longer nibbling, beak and foot holding). As parrots are thought to enjoy bobbing as a way to entrain to the music, we interpreted this trend as justifying the

need for a no-contact, more ergonomic way to trigger the music. Additional considerations regarding novelty effect and possible adaptation of the bird to environment changes are explored in the Discussion.

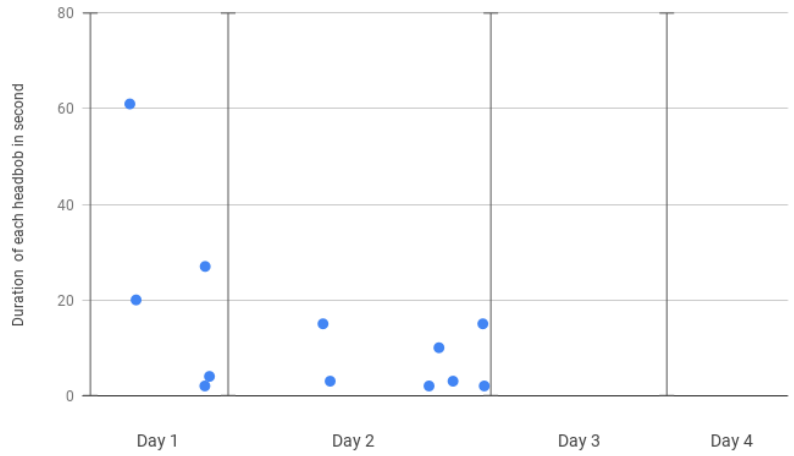


Figure 71: During the JoyBranch intervention, although occurrence and duration of triggers increase over time, bobbing become less frequent (0.18 occurrences of bobs per minute for day 1, 0.13 for day 2, and 0 on day 3 & 4) and their duration becomes shorter with time. This might be due to the ergonomics of the branch as the bird finds more efficient ways to keep the music playing.

6.5.5.4 - Attention analysis

The data collected during the BobTrigger sessions give insight into (1) the use of the system, and (2) the bird's perception of the intervention, as well as (3) the way he uses and directs his new agency.

Contrary to the JoyBranch intervention, the BobTrigger intervention doesn't involve a tangible object. Instead, each bob from the bird is manually observed and recorded by the experimenter and used to trigger music in real-time. Even though the experimenter was not located directly next to the bird, the bird noted their presence quickly and spent 19 percent of his time focused on the experimenter through short glimpses, at an average of 1.1 glimpses per minute.

It also appears that the bird quickly made the connection between the presence of the experimenter, bobbing, and the music. We distinguish headbob-triggers that occur while no music is playing, causing the music to start, versus entrained head bobs that occur while the music is already on. Within the eight intervention sessions, we recorded 94 instances of headbob-triggers, and in 73 percent of the cases (68 out of 94), the bird had his attention focused on the experimenter when triggering. We can also observe what can be interpreted as a learning curve for this phenomenon, as seen in Figure 72, which shows the evolution over the eight sessions.

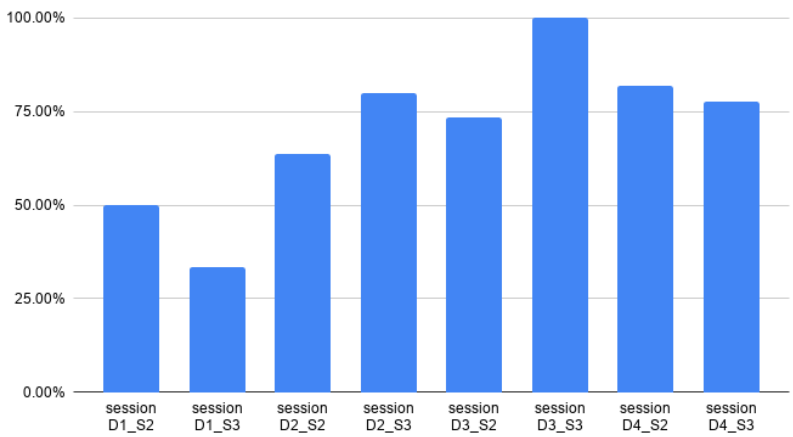


Figure 72: Over time, the bird more frequently looks at the experimenter when head-bobbing, suggesting a form of learning

This trend suggests that the bird made a connection between the human and the music. This might be interpreted as the bird assuming that the experimenter is in control and asking them for music. On the other hand, it could also indicate that, in order to gain the attention of the experimenter, the bird used the music to amplify his presence.

In addition, we observed 32 instances where the music was playing, then stopped, then the bird looked at the experimenter and bobbed within 10 seconds of the music stopping, prompting the music to resume. This suggests that the bird also understands the connection between bobs and music. In this context, instead of playing an invisible system, our results suggest the bird is actively "asking" the experimenter to play music. We could say that Sampson is "playing the experimenter" as an instrument to obtain the music. Indeed, once the music played, the bird generally stopped paying attention to the experimenter until the music stopped.

Zoo visitors also represent an important component of Sampson's life while he's on exhibit. During the baseline recordings, we noted that 47 percent of his time is spent observing and paying attention to the visitors. Although the attention of the flow of visitors is also directed toward the bird about 51 percent of the time, only 22 percent of the total time is composed of shared attention between visitors and the bird. This can be explained by several factors. The bird seems to have preferences for which visitors or groups of visitors he dedicates attention to. For instance, he is more active when children and frequent visitors come by. However, the bottleneck also comes from visitors, as they have the agency to walk away or look at the map or other exhibits while the bird is focused on them. Indeed, during baseline, 80 percent of the instances of shared attention (12 out of 15) ended with the visitors walking away while the bird was still looking at them.

To test whether the attention dynamic was modified by the technological intervention, we first compared the average duration of visitor attention toward the bird during baseline, then during the intervention while music was not playing, then when the music was playing. A Mann-Whitney U-test indicated that the duration of visitor attention was greater when the bird triggered music (mean=160s, std=107s) than when no music was playing (mean=32s, std=23s), p=3.3494e-17.

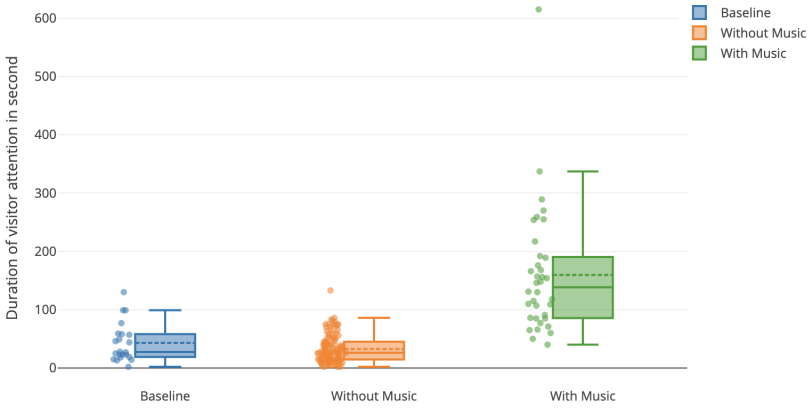


Figure 73: The average duration of visitor attention is comparable between baseline and when the bird doesn't trigger music during BobTrigger intervention. However, when the music is playing, the average duration of visitor attention is increased by a factor of 4. This suggests that the bird might be using his new agency to control the visitor's attention.

In addition, the trigger of the music also influenced the distribution of shared attention and timing dynamics. Indeed, in the baseline, as stated above, 80 percent of shared-attention occurrences were ended by visitors. During the intervention, when no music was playing, the percentage of shared attention occurrences ended by visitors while the bird was still paying attention to them was 63 percent(71 out of 112). This is comparable to the 80 percent during baseline. However, when the bird was playing music, this number dropped to 25 percent. This suggests that the bird had gained more control over the interaction and was now the one who decided when to end the interaction. Playing music is a successful way to keep the public interested.

6.5.6 - Discussion of the JoyBranch project

6.5.6.1 - Future Work

This preliminary analysis of the critical behaviors and correspondences observed through the intervention have brought insights into the perceived understanding and enjoyment of the animal and the influence of the experimenter and visitors. In the future, ongoing work will provide a statistical understanding of the inter-dependencies of the various behaviors. Future iterations of the project would also need to take into account additional parameters that may influence the bird's and visitors' behavior (i.e. weather,

temperature, precise time of day, what the bird ate, etc.) We know that Sampson tends to seek children's attention, but there may be other attributes of the visitors influencing his behaviors. In[543], experimenters noticed that primates were more likely to respond to men carrying objects. In other cases, visitor density and general noise were a key factor in the animal's responses[544].

[543] Shelley Cook. Interaction sequences between chimpanzees and human visitors at the zoo. *Zoo Biology*, 1995

[544] Alexandra Farrand. The effect of zoo visitors on the behaviour and welfare of zoo mammals. 2007

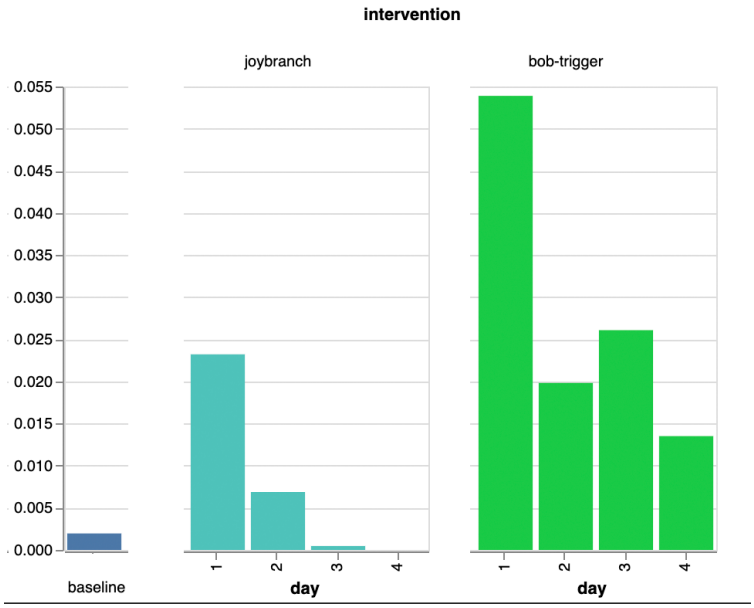


Figure 74: Fraction of session time spent "dancing" (exhibiting head-bobbing, head-nodding and head-whipping behaviors) for each intervention. Sampson triggered playback of music between 15 and 25 percent of the time by head-bobbing, and up to 45 percent of the time by use of the JoyBranch. Curiously, during day 3, JoyBranch usage was low—however, this was the session during which we remained out of sight for a significant time.

This work also has clear limitations in the deployment to only one bird. Future work on a larger set of species and individuals is needed to generalize the results. However, the focused and personalized character of this work is what allowed us to take advantage of the individual experience of Sampson and take the time to engage in a dialogue with the animal. Opening up such a dialogue might be the most valuable outcome of the project for the researchers as well as, potentially, an inspiration for the bird to establish deeper contact.

The BobTrigger was based on the idea of using an automated system for bobbing recognition. Using the footage recorded during the deployment, we later implemented such an automated system that we will include in future deployments. However, this raises additional ethical questions as the experimenter was also monitoring the stress level of the bird. In addition, we believe, and in view of our results, that such a system should be deployed incorporating human collaboration. One possible future version involves visitors co-triggering music. To create an ultimate musical enrichment device for Sampson, we are working on a hybrid interface that incorporates the benefits of a physical system (tangibility, immediate connection

action/sound) with the benefits of the deep-learning-based visually aware system to track the bird's behavior (no need to learn a new behavior, easy to deploy, fewer safety concerns). Such a system will also keep a log of the triggers and behaviors to provide more complete insights.

6.5.6.2 - Novelty effect, attachment & ethics

We believe that the negative effects of introducing and removing interventions are limited. Indeed, from regular zoo policies, the bird is used to having novel objects regularly added and removed from the enclosure for enrichment and Sampson is also familiar to having music played for him by his caregivers. This hopefully limits the risks of Sampson to becoming overly attached to the device. In terms of ethics and companionship, Sampson is housed at night with other parrots with whom he routinely communicates from across the zoo. The zoo is very attentive to the question of animal companionship, which will be addressed in future deployments. During the intervention, we calibrated the music loudness such as not to add to current ambient dB (above existing human voices, ambient noise, announcements, etc). The possible implications, benefits, and risks of long-term deployment can be thoroughly explored through longitudinal testing.

6.5.6.3 - Implication for HCI

Our results suggest a relationship between attention dynamic and lead—follower influences in the triangle interactions between the bird, the experimenter, and the visitor. By gaining more control over his sonic environment, Sampson might have effectively gained more control over his interactions with humans resulting in an interspecies experience mediated by technology.

This work offers insights for the field of HCI. By highlighting the experimenter's influence, this work supports the need for a human in the loop in ACI. This could also have potential implications if using technology to help visitors understand and read animal intent to lead to better interspecies understanding. Our design journey may also provide insights to future interactive technologies to engage humans and birds in zoos. The parallel use of physical vs gesture-driven systems for birds could expand the tangible vs virtual system in humans. In addition, our use of fully versus semi-automated systems broadens the discussion of such systems for humans. Our project also tackles the question of designing to understand and affect the attention ecology in zoos. Additionally, our approach also touches on the importance of animal involvement in interactive design. Finally, this work also may provide insight for more diverse HCI agents. Zoo visitors often wish to see interesting behaviors. Here, they can interact cooperatively with another species to play music. Our system enables novel interaction. It is an example of communication between different animals

where the system itself becomes an agent to enrich the experience.

Anecdotal events during deployment also shed light on Sampson's personality. On day 3, we wanted to see if the bird would use the JoyBranch if left alone. The experimenter went to hide and the zoo was not yet open, so there were no visitors. Not only did the bird show no interest in the device anymore, but he also started a long series of loud calls and repeated head rotations as if he were looking for the experimenter. This interpretation was confirmed by Duarte and Handrus: "I think he's surprised ... hey that girl's not here" and "Well, he's looking for you." This behavior lasted about five minutes until the bird walked to the JoyBranch and activated it while still calling. After release, he resumed looking around. Once the music stopped, he then used his feet for the first time to hold the branch, allowing him to keep the music playing while looking around. The experimenter then came back from hiding, and when Sampson spotted her, he released the branch and stopped calling. More than a demonstration of a need to perform for an audience, this episode spoke to the bird's interests in companionship and establishment of rapport.

The bird's use of the device suggests that he may combine the simple enjoyment of listening to music with a more elaborate schema to attract and maintain public attention. The device seems to be used as a means to an end, and the bird's agency appears to be sometimes directed toward careful control of the visitors' and experimenter's attention.

In a more distant future, as more species facing extinction can only be preserved in managed care, we wish for zoos to integrate animal agency in all aspects of their design. We hope for a zoo of the future where a better balance is established between animals and human-animals. Indeed, until animals can be the lead of co-creative projects with humans, we will remain egocentric in our view of ACI. If, however, we can reach the point where animals create their own understanding of all the parts of a human system, we can then truly start envisioning a meaningful interspecies internet.

6.5.7 - Conclusion of the JoyBranch project

We created naturalistic interactive systems for a solitary music-savvy macaw to gain control over his sonic environment. Very few past instances of enrichment interventions have targeted the animal's sonic environment, even though for most species, auditory input is a major way in perceiving the world. The need for technological devices arises from the limited time caregivers can spend with each animal[545]. The deployment of the two interventions brought insight into the acute social awareness of the animal and the triangular interaction between the experimenter, the bird, and the

[545] Julia M Hoy et al. Thirty years later: Enrichment practices for captive mammals. *Zoo Biology*, 2010

visitors. Our results suggest not only that the bird understands, enjoys, and makes use of the systems, but also that the visitors play a major role in the animal's motivations and engagement with the technology. By gaining more control over his sonic environment, Sampson effectively gained more control over his interaction with the public. The interaction created became an interspecies experience mediated by technology. The resulting triangle interaction between the animal, the visitors, and the computer may bring insights into the potential of technology for future interspecies enrichment and communication.

6.6 - Conclusion on Sonic and Vocal Enrichment at the Zoo

This chapter presented a new approach to tackling to question of Sonic and Vocal Enrichment for Animals in Managed Care. Our research aimed to improve the care of zoo animals through a four-pronged approach: 1) to propose directions to understand the effect of the general sonic environment on animals and ideas on how to include it in exhibit design; 2) to develop tools that recognize our current human limitations in decoding animal language and respect the intricacy of animal communication; 3) to build tools that will not only allow us to further understand animal vocal behaviors and enrich their lives in managed care, but also empower keepers and researchers to monitor, analyze, and provide care in a more data-driven, evidence-based manner; 4) to give control back to the animals by providing agency in the control of their environments and developing experiences that respond to the animals' changing behaviors and needs.

Following those four directions, we presented a series of projects at various stages of development. The Sonic Diversity endeavor inquires about the effects of ambient soundscapes in zoos. The TamagoPhone proposal is an augmented egg incubator that preserves vocal connections between parent birds and embryo while within the egg. The Panda Project is a bio-acoustic, deep learning-based system for panda monitoring to assist keepers in better understanding behaviors of giant pandas. Finally, the JoyBranch project describes the design and deployment of interventions to allow a hyacinth macaw to gain control over his sonic environment.

These projects yielded new insights in the understanding of vocal connections and the roles humans may play to respect, enrich and better understand animal and human vocal and sonic experiences. Through these projects, we have:

- applied soundscape ecology thinking to the zoo environment and revealed new questions that could affect zoo soundscape design.
- revealed the potential of modeling animal voice to support the preservation efforts for vulnerable species.
- imagined ways and technologies to alleviate some of the possible negative influence of human management of animals and to influence animals' wellbeing and the preservation of species identity and behaviors.
- uncovered the acute social awareness some animals have during interactions with human visitors and presented them with technologically mediated ways to alleviate this awareness to create meaningful interspecies interactions driven by the animals.

Each of those projects sheds light on the additional potential of Vocal Connection by extending our approach beyond a unique species. The Sonic Diversity project touches the experiential and connected paradigm as it interrogates the effects of shared sonic and vocal space between species who did not coevolve together. It ranges from personal to interpersonal to interspecies contexts because the experience itself may affect each specimen differently, and may also affect their ability to communicate with their conspecifics and also to heterospecific.

The TamagoPhone touches on the voice as a whole by targeting pre-hatching vocalizations that we—as humans—may not yet understand. It also touches on the connection paradigm, and specifically the connections between parent and an offspring that is literally locked in a closed box forbidding tactile, visual and even scent-based interactions. This project covers both the interpersonal and interspecies contexts as it is intended to allow communication between conspecific (parent and offspring), but we humans are also intervening as witnesses.

The Panda Project arises from a holistic approach to the voice: we can only map what we already know of the acoustic ecology of the species but those are meaningful acoustic components on another dimension than words or semantics. This project spans over the three contexts because it is the personal expression of individual animals, used for interpersonal signaling and us humans are listening.

Finally, while the JoyBranch system is about connection at several levels, because of Sampson's specific situation, in being alone in his exhibit, the context is purely cross-species. The introduction of such technologies assists in creating togetherness with visitors. The interventions resulted in intentional meaningful interactions that are made more special because the intent comes from the bird more than the humans.

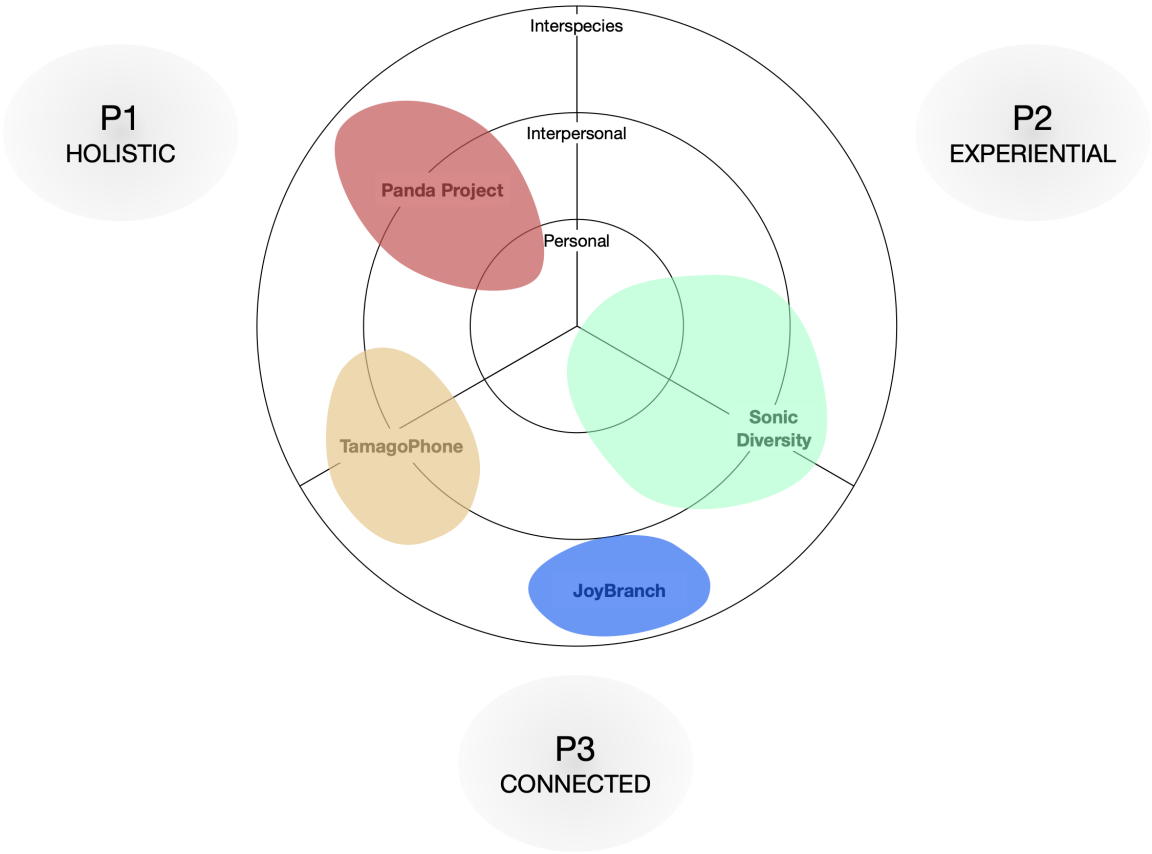


Figure 75: Mapping of the various Zoo projects onto the Vocal Connection space

7 - Conclusion & Future Directions

This chapter starts with a summary of the research completed in this dissertation. I then present an overview of the contribution of this work and conclude with suggestions for future directions.

7.1 - Summary of research

This dissertation introduced a new approach for looking at the voice holistically, in its experiential nature, based on its propensity to connect.

Looking at the voice holistically allowed us to remove semantics from the equation and focus on the vocal signal, be it acoustic, behavioral, or neurological. Indeed, when interacting, we are sharing thoughts, but we are also sharing parts of who we are as of a moment in time. Besides words and emotions, the voice also contains clues about where we come from, who we are, and where we might be heading. In this dissertation, I focused on the potential of the voice to access more profound levels of connection and understanding. I tackled this question by presenting an overview of the evolution of the voice in time at different scales. From this overview, I presented our first archetype of the voice as a holistic object. This was our first paradigm.

Considering the voice in its experiential nature brought me to explore the journey of the voice throughout the mind and to construct a phenomenology of inner voices based on their potential for connection. This exploration was anchored in understanding the experience of complex verbalized thoughts as not only linked with language but also as shaped by our more visceral experience of the voice and our motor ability to produce vocal sounds. The second archetype proposes that our experience of voices, both interior and exterior, informs our experience of the world. This framed the voice as a permeable membrane between the personal inner life and the social space we share with others. The way inner and outer voices interface affects our perception and processing of inner experiences (thoughts, mental models, subconscious, mental states) and outer information (text, data, discussions). This journey led to the development of our second archetype of the voice

as an experience. This was our second paradigm.

Exploring how the voice is anchored in social behavior led me to delve into stories of the voice across people. Through this approach, I learned to frame the voice in its fundamental social nature, to build and regulate companionships originating in social grooming. Those behaviors appear to be the direct origin of our vocal messages and also to remain one of the most important parts of the transmitted message. This third archetype frames the voice in its tentacular propensity to create bridges and connections, to transform otherness into togetherness. This was our third paradigm.

Through this novel framing of the voice, I established and defined our problem space as an ecosystem to navigate when considering Vocal Connection. I then presented a series of initial personal projects, namely the ORB, the MiniDuck book, the SIDR interface, and the Nebula platform described in Chapter 3. Finally, I presented a body of work to support and illustrate this approach by establishing connections at three levels: individual, interpersonal, and extending beyond human languages.

The Memory Music Box establishes a sense of connection across space and time and is designed to enhance a sense of connectedness for and to the elderly. We created this new form factor and user experience for people at any level of cognitive ability or memory loss. This approach uses techniques from reminiscence therapy and emotive smart objects to alleviate the sense of isolation often associated with memory loss. By providing a very easy-to-use form factor—a jewelry music box—we lead the user to engage with a multisensory portal to reduce feelings of isolation and establish connections with the physical object, the past, and a remote correspondent.

With the Mumble Melody project, we extract musicality from everyday speech to help people who stutter gain increased fluency. Stuttering affects about 1 percent of the adult population worldwide. Adults who stutter are often fluent when singing and lateralization tendencies [546] support the theory of some largely independent neural processes between music and speech, which provides us with an opportunity to combine modulated auditory feedback with music. Therefore, we created a system that shifts attention during speech from the words themselves to the musicality of the spoken voice to affect the neural basis of speech production. Stuttering is associated with a combination of factors. Our approach works across domains to reconfigure the neural basis of speech and reconcile the different mind-agents at play in those problem areas.

[546] Steven Brown, Michael J Martinez, and Lawrence M Parsons. Music and language side by side in the brain: a pet study of the generation of melodies and sentences. *European journal of neuroscience*, 23(10): 2791–2803, 2006

Finally, with the Sonic Enrichment at the Zoo project, we developed ways to improve connections within and between species—and between humans and animals—by exploring sonic and vocal enrichment interventions at the San Diego Zoo. Without interactive sonic and vocal stimulation, many species demonstrate an inability to mate, raise young, or exhibit other natural engagement behaviors while in human care. Current sonic environments in zoos are often limited, lack references from nature, are disrupted by human-generated noise, and do not respond meaningfully to the animal's behavior. Our approach is guided by data about the meanings of environmental sounds or vocal behaviors with the goal of designing and building interactive sonic and vocal enrichment systems for animals in managed care.

This approach and each of our projects yielded meaningful contributions and opened the door to future directions.

7.2 - Contributions

7.2.1 - Research and Project Contributions

This work introduced a new approach for thinking about the voice and tackling experience and interface design that uses the power of the voice to create connections between individuals, within different parts of the self, and across species. The contributions from this dissertation include:

- A cross-referenced reading of diverse fields related to voices as well as methods and examples of how to use knowledge from those different fields to create experiences of connection.
- A history of prior arts, projects, and interventions that embrace the characteristics of the Vocal Connection. This includes scientific exploration, technical application, architectural techniques, and societal phenomena, as well as artistic, musical, and performative pieces.
- A collection of Design Studies that explore different aspects of the meaning of voice and connection in various contexts.
- Three novel systems demonstrating the potential of Vocal Connection in different contexts: interpersonal, personal, and interspecies, in addition to design, implementation, deployment, and evaluation of each system.
  - With the **Memory Music Box** project,
    * we introduce the new concept of Cognitively Sustainable Design.
    * we present the design of a device to help older adults remain connected with loved ones through several prototypes.

    - offers an evaluation of the potential use of the device by target grandparents and grandchildren users.
- Through the **Mumble Melody** initiative,
    - we introduce the idea of combining music and vocal auditory feedback to affect the neural basis of speech production.
    - we created a series of modes of such Musically Modulated Auditory Feedback.
    - we tested the potential of such feedback in increasing fluency for adults with persistent developmental stuttering.
    - we demonstrated the fluency-evoking effect of some of our modes while people are using the system and their superiority to state-of-the-art techniques (FAF and DAF).
    - we obtained a slight—although significant—increase in fluency shortly after using the feedback, suggesting possible longer-term effects of such methods, which need to be confirmed through longitudinal testing.
- With our explorations for **Sonic and Vocal Enrichment at the Zoo**
    - we explore the potential of using sounds, voices, and soundscape design as tools to create enrichment for animals in managed care.
    - we identify four key guiding principles for rethinking sounds in zoos by (1) listening to animals collectively, (2) respecting animal-animal communication, (3) listening to animals individually, and (4) giving sonic agency to animals.
    - we created a series of projects and interventions to explore the potential of each approach

7.2.2 - Insights

The journey of this dissertation also yielded a series of more general insights that I collected within the following three themes.

Voice, Technology, and Estrangement: Technology is not essential in creating experiences of Vocal Connection. We create some version of such experiences each time we have a simple conversation or discuss the weather with a neighbor. However, there is power in creating surprising experiences, especially around an object as familiar as the voice. In my designs, I use the concept of estrangement to help people change perspective about their everyday voice. In this work, I reversed the common connotations associated with the word "estrangement" and define it as the freeing act of embracing distance and looking at the world with "the eyes of a horse or a child."[547] This is the act of making "foreign" something that has become so familiar that we don't even see it anymore. If technology is not essential to pursue this goal, it can become a very useful tool to help us enrich our

[547] Carlo Ginzburg. À distance. Neuf essais sur le point de vue en histoire. 1998

perspectives. Using digital systems and powerful computations can help in creating such experiences, that ultimately can help change perspective and eventually make the technology obsolete. Maybe the zookeepers using our Panda Project biomonitoring systems can use it to learn to detect specific vocalizations to develop a more powerful relationship and dialogue with the animals within their care. Maybe, after using the SIDR interface to obtain real-time feedback on the use of the shared vocal space during meetings, coworkers can reach better group emotional intelligence and won't need the system anymore. Maybe after getting used to hearing their voices as music, the adults using our Mumble Melody system will gain control over the neural basis of vocal perception, which could make the system obsolete. This is how I hope for those devices and technologies to be used.

Learning to listen, learning to observe, toward an epistemology of the voice: The various projects presented in this thesis have each enriched the way I listen. The Orb project taught me the potential of listening beyond sounds. The Music Box project entirely revolved around the importance of listening even when we think that there is nothing to listen to anymore. The Zoo project taught me to listen both globally and locally at the same time, to understand both the context and the individual creature. It also opened my ears to the voices I cannot understand. But the most important lesson was from the Mumble Melody project. This project gave me the opportunity to meet many people who stutter (PWS). Conversing with people who stutter taught me a lot. People who do not stutter often instinctively try to finish the sentences of PWS. We feel uneasy with the saccaded tempo of the conversation, and want to fill the blanks, accelerate the conversation. We think we are helping. However, this is wrong on many levels. Finishing someone else's sentence is a clear sign of impatience and may even deny people who stutter the space to express themselves at their own rhythm. In the first few months of working on this project, when I would converse with PWS, my brain would try to fill the blanks by mentally predicting their next words. But, to my surprise, I learned that, most of the time, my predictions were wrong. This was a powerful lesson in humility and brought me to change my way of listening—to everyone. This is, unfortunately, a widespread phenomenon in conversations. Instead of truly listening, we think about what we will say next or how we could steer the discussion toward another topic. We play around with our inner voice. But listening to someone is about being present with them. For many of us, this is an acquired skill. This dichotomy between experiencing and observing also relates to the epistemological problem of the relation between the observing subject and the external world. The subject/object problem is a long-standing philosophical issue in the study and analysis of human experience. For Descartes, the subject is a thinking thing that is not extended, and the object is an extended thing

which does not think. Heidegger rejects this distinction between subject and object by arguing that there is no subject distinct from the external world of things because Dasein is essentially being-in-the-world[548]. This may call for the development of an epistemology of the voice.

[548] A Kadir Çüçen. Heidegger's reading of descartes' dualism: The relation of subject and object. In *The Paideia Archive: Twentieth World Congress of Philosophy*, 1998

Voice, context, and society: A voice is never isolated. It evolves surrounded and influenced by other voices. This is illustrated by the dark history of speech and voice deprivation experiments when scientists and thought leaders would experiment on infants to see what happens when children grow up deprived of any vocal contact from other human beings. The results are highly unscientific and unclear but most children seem to perish at a very young age. Looking at the voice in isolation would be similar to looking at Sampson's musical skills without considering the presence and reaction of his audience of visitors. It could be like studying stuttering without the environmental and social factors. Context is key. Malinovski's differentiation between "primal" and "intellectual" languages can be seen again as a context difference. The voice also depends and influences the social architecture of the population studied. The societal context and the way individuals or species in a cohabitating group co-evolve dictate the journey of their voices. It takes space and time to create an ecosystem of voices. Could there be some similarities in the difference of animal voice between the wild and zoo environment and the difference of human voices between "primal" vs "intellectualized" societies? What happens to the great animal orchestra when species that did not co-evolve together are relocated to a zoo environment where they have to share the same sonic environment? Maybe their population will be preserved but what becomes of their voices? How do they adapt to sonic cohabitation?

7.3 - Future Directions

Vocal interactions are ubiquitous which has led to the development of research in a very broad spectrum of disciplines with well-defined normative boundaries. There is however a lot of room for cross-pollination between these fields which could benefit various researchers as well as professionals. In this work, I have applied multidisciplinary technology-based research methodologies to further our understanding and create novel instances of vocal experiences. I believe that people in specific fields and professions could extend their impact based on this approach.

Health professionals would benefit from a clearly defined, and clinically tested, use of the voice as a diagnostic tool. Diseases such as Lou Gehrig's disease (ALS), Parkinson's disease, Alzheimer's disease, or multiple system atrophy (MSA) affect muscle activity and have important effects on the voice. Could the voice be used as early detection biomarker?

The voice could also be used as a marker for other conditions such as depression or dementia. However, most existing voice diagnostic tasks are still experimental and speech-based. Inspired by the approach taken in this dissertation, we could imagine the development of diagnostic tests based on the voice beyond words. Vocal tasks based on music or other sound imitation have the advantage of dissociating speech pathways from pure voice pathways. One could even imagine the potential of such an approach for nonverbal people or people on the autistic spectrum.

Endocrinologists could also consider the voice as an ally. The larynx is a target organ for a lot of hormones including thyroid hormones, estrogens, progesterone, androgynes, and testosterone. Hormones are known to affect mood but their action on the voice may create even more complex cascading experiences. We have explored throughout this work the potential of the voice to deeply affect our sense of self, sometimes purely subconsciously. This may help us look at hormone-related mental health in a new light and offer possible directions for both diagnostic and treatments. For instance, postpartum depression could correlate with a drastic change in voice due to hormonal change. The voice changes but we are not aware of it, leading to possible dissociation with oneself and a sense of not recognizing oneself without understanding why. Voice-based therapies or better education of such phenomena may help alleviate the negative impact of such changes. Hormone-based vocal changes also affect people who menstruate and may have critical consequences for professional singers.

Acousticians and audio device manufacturers are transforming the field of communication through novel experiences of vocal communication. Wearing headphones or earphones used to signify a wish to not engage in conversations with people around. The establishment of an entirely new type of audio device that can isolate you acoustically or become transparent on the click of a button is transforming the everyday experience of sounds and the way it intertwines with the acoustic and social environment. However, most such devices still have limitations in the way they render the user's own voice. On the one hand, they could benefit from a more holistic look at the voice beyond the simple sound signal and look at the voice as truly embodied, including physical vibrations and containing a complex auditory and somatosensory feedback. On the other hand, the use today of such modulated feedback associated with those devices may already be creating a new experience of the voice, unique to those situations. Indeed, users might be creating a new voice, a specific vocal experience associated with specific brain pathways due to this phenomenon. Either way, better study of voice perception and neurology of the voice could help in understanding and curating such vocal experiences.

AI voice assistants are becoming more common and designers of voice services are curating new types of vocal experience between humans and machines. Voices are becoming very realistic and companies are pushing toward the development of agents resembling human conversational partners as closely as possible. However, when considering the experience of machine-directed speech in the framing of Vocal Connection, this approach might be lacking crucial factors and present important blind spots in terms of user experience. The study of infant-directed speech and animal-directed speech may help better understand the specific phenomena at play in conversation with machines and could be used as inspiration for the design of better voice assistants and conversational AI. One first step for a better design of voice assistants would be to get a sense of crucial non-verbal clues contained in the user's voice. Such meta-acoustic information could then be used for the synthesis of the assistant's voice in response. This could greatly influence the establishment of rapport between users with machines.

Voice Command today enables computers and machines to be controlled without any physical contact. Computers, cars, lighting systems, smart homes can be turned one and off, and controlled solely by the user's voice. Similarly to voice assistants, these systems often only take into account the semantic meaning present in the vocal message. However, it could be possible to create systems in cars that also estimate the driver's alcohol or drug intake from their voice and only allow manual driving if they fall below a certain degree of estimated intoxication. This could be accomplished using a simple microphone. Smart home systems could not only detect stress and emotions from the voice when the user is talking to them but could also create a complex mapping of social dynamics within a household or at work. It could go a long way to know that the older sibling responded aggressively to their little brother, or that the father-in-law is regularly being passive-aggressive when talking to their spouse, or who instigates fights and what is the source of negative communication patterns. Although this would raise critical ethical concerns, smart homes could also provide a certain level of couple or family therapy based on people's vocal meta information.

Musicians and professional singers often develop a very unique relationship with their instrument. Learning to play an instrument can be seen as embodying and integrating the feedback loop between the musician's motor control intent, and the auditory—and sometimes tactile/vibratory—feedback signals. A similar loop is at the origin of the voice. This loop may also be the source of the inner voice. This may bring us to ask questions regarding the experience of the inner instrument for musicians. If the inner voice allows us to have complex thoughts and to process external

information such as texts, numbers, etc., can the inner language of musicians also allow them to enrich their inner lives? In discussions on our second paradigm, we established the experience of inner music as also deriving from our ability to produce sounds from our bodies. Musicians may use such approaches to consider their practice as including an element of inner musical dialogue or inner musical spirituality.

Anyone who experiences an inner voice may benefit from a better understanding of this mesmerizing phenomenon. The realization that our ability to develop a rich sonic inner life might derive from our mechanical ability to produce sounds may lead to fascinating new research directions. Indeed, we can understand our experience of complex thoughts as not only linked with language but also as shaped by our more visceral experience of the voice. In addition to providing a connection between thoughts and actions, voices also underlie our experience of text, data, and information, as every piece of textual information read silently passes through the inner voice. One can wonder what this means in terms of the omnipresence of texts in our environment, or go a step further and query whether the inner voice may exist independently from language, or at least from human language. Do other vocal mammals experience an inner voice? Do birds rehearse their songs silently? Far from denying that non-verbal animals have complex inner lives, this theory might suggest that any species capable of producing vocal sounds would potentially experience an inner life or at least an inner sonic life.

7.4 - Final Words

This work started with seeing the beauty in what is familiar and with the desire to share it. The voice gives us infinite wonder at the tip of our tongue. It is so simple and yet so rich, an endless source of fascination and discovery. It bridges and unites the creatures of the Earth, allowing us all to communicate while also inviting us to think. Our voice is with us, indeed at our command every second of every day. It possesses the key to the secret garden of our inner lives and sheds light on the profundity of the existence of others. In this work, I have sought to celebrate the voice, and to share tools that leverage its potential. My ultimate goal has been to reveal the ultimate, exquisite power of the voice to connect, and in so doing, to uncover some of the many ways in which it enriches our lives. It is a quest that I am likely to continue for many years to come.

Bibliography

Jean Abidbol. Hormones and the voice. *The Singer's Guide to Complete Health*, 2013.

Francisco Aboitiz. A brain for speech. evolutionary continuity in primate and human auditory-vocal processing. *Frontiers in neuroscience*, 2018.

Thomas A Ala et al. Using the telephone to call for help and caregiver awareness in alzheimer disease. *Alzheimer Disease & Associated Disorders*, 2005.

Ben Alderson-Day et al. Inner speech: development, cognitive functions, phenomenology, and neurobiology. *Psychological bulletin*, 2015.

Jessica D Alexander and Lynne C Nygaard. Reading voices and hearing text: Talker-specific auditory imagery in reading. *Journal of Experimental Psychology: Human Perception and Performance*, 34(2):446, 2008.

Edgar F Allin. Evolution of the mammalian middle ear. *Journal of Morphology*, 1975.

Per A Alm. Stuttering and the basal ganglia circuits: a critical review of possible relations. *Journal of communication disorders*, 2004.

Ofer Amir and Tal Biron-Shental. The impact of hormonal fluctuations on female vocal folds. *Current opinion in otolaryngology & head and neck surgery*, 2004.

Gavin Andrews et al. Stuttering: Speech pattern characteristics under fluency-inducing conditions. *Journal of Speech, Language, and Hearing Research*, 1982.

Joy Armson and Michael Kiefte. The effect of speecheasy on stuttering frequency, speech rate, and speech naturalness. *Journal of Fluency Disorders*, 2008.

Luc H Arnal, Adeen Flinker, Andreas Kleinschmidt, Anne-Lise Giraud, and David Poeppel. Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, 2015.

Barry Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 1992.

Dimitrios Assimakopoulos et al. Highlights in the evolution of diagnosis and treatment of laryngeal cancer. *The Laryngoscope*, 2003.

American Birding Association. Code of ethics, 2012. URL `http://www.aba.com`.

Joseph S Attanasio. The dodo was lewis carroll, you see: Reflections and speculations. *Journal of fluency disorders*, 1987.

Jean-Julien Aucouturier et al. Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 2016.

Saint Augustine. *The confessions*. Clark, 1876.

MM Babiker. Development of dependence on aerial respiration in polypterus senegalus (cuvier). *Hydrobiologia*, 1984.

Stefano Baccianella et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, 2010.

Alan Baddeley. Working memory. *Science*, 1992.

Cyril Bailey. *Phases in the religion of ancient Rome*. Univ of California Press, 1932.

Lucie Bailly. *Interaction entre cordes vocales et bandes ventriculaires en phonation: exploration in-vivo, modélisation physique, validation in-vitro.* PhD thesis, 2009.

James Baker. The dragon system–an overview. *IEEE Transactions on Acoustics, speech, and signal Processing*, 1975.

Juliana V Baldo et al. The role of inferior parietal and inferior frontal cortex in working memory. *Neuropsychology*, 2006.

Domna Banakou and Mel Slater. Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking. *Proceedings of the National Academy of Sciences*, 2014.

Deanna M Barch. The cognitive neuroscience of schizophrenia. *Annu. Rev. Clin. Psychol.*, 2005.

M Elizabeth Barnes. Ernst haeckel's biogenetic law (1866). *Embryo Project Encyclopedia*, 2014.

MF Bassett and CJ Warne. On the lapse of verbal meaning with repetition. *The American Journal of Psychology*, 1919.

Shari S Bassuk et al. Social disengagement and incident cognitive decline in community-dwelling elderly persons. *Annals of internal medicine*, 1999.

C Philip Beaman and Tim I Williams. Earworms (stuck song syndrome): Towards a natural history of intrusive thoughts. *British Journal of Psychology*, 2010.

Georg V Békésy. The structure of the middle ear and the hearing of one's own voice by bone conduction. *The Journal of the Acoustical Society of America*, 21(3):217–232, 1949.

Michel Belyk, Shelly Jo Kraft, and Steven Brown. Stuttering as a trait or state–an ale meta-analysis of neuroimaging studies. *European Journal of Neuroscience*, 2015.

Daryl J. Bem. Self Perception Theory, 1972.

Stan Bennett. The process of musical creation: Interviews with eight composers. *Journal of research in music education*, 1976.

D Frank Benson, William A Sheremata, Remi Bouchard, Joseph M Segarra, Donald Price, and Norman Geschwind. Conduction aphasia: a clinico-pathological study. *Archives of Neurology*, 1973.

Jonathan Berger and Song Hui Chon. Simulating the sound of one ' s own singing voice. 2003.

Penny L Bernstein and Erika Friedmann. Social behaviour of domestic cats in the human home. In *The Domestic Cat: The Biology of its Behaviour*. Cambridge University Press, Cambridge, 2014.

John Bevis. Aaaaw to zzzzzd: The words of birds, 2010.

L Birke. Effects of browse, human visitors and noise on the behaviour of captive orang utans. *Animal Welfare*, 2002.

Sarah-Jayne Blakemore. Why can't you tickle yourself? In *The Anatomy of Laughter*. Routledge, 2017.

Steven K. Blau. Musicality of speech changes with mood. *Physics Today*, 2010.

Karl Blind. Wagner's" nibelung" and the siegfried tale. *The Cornhill magazine*, 1882.

Gordon W Blood and Ingrid M Blood. Bullying in adolescents who stutter: Communicative competence and self-esteem. *Contemporary Issues in Communication Science and Disorders*, 2004.

O Bloodstein and N Bernstein Ratner. A handbook on stuttering new york. *NY: Thomson-Delmar*, 2008.

Tiffany C Bloomfield et al. What birds have to say about language. *Nature neuroscience*, 2011.

Susan Bluck and Linda J Levine. Reminiscence as autobiographical memory: A catalyst for reminiscence theory development. *Ageing & Society*, 1998.

Paul Boersma et al. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

James R Booth, Lydia Wood, Dong Lu, James C Houk, and Tali Bitan. The role of the basal ganglia and cerebellum in language processing. *Brain research*, 2007.

Gerrit Bos. Jewish traditions on divination with birds (ornithomancy). *Gen*, 2015.

Anne K Bothe et al. Stuttering treatment research 1970–2005: I. systematic review incorporating trial quality assessment of behavioral, cognitive, and related approaches. *American Journal of Speech-Language Pathology*, 2006.

Robert M Bradley and Charlotte M Mistretta. Fetal sensory receptors. *Physiological Reviews*, 1975.

Michael S Brainard and Allison J Doupe. Translating birdsong: songbirds as a model for basic and applied medical research. *Annual review of neuroscience*, 2013.

AR Braun et al. Altered patterns of cerebral activity during speech and language production in developmental stuttering. an h2 (15) o positron emission tomography study. *Brain: a journal of neurology*, 120, 1997.

T Braun Janzen and MH Thaut. Cerebral organization of music processing, 2018.

Frederick Stephen Breed. *The development of certain instincts and habits in chicks*. Number 1. Pub. at Cambridge, Boston, Mass., 1912.

L Brent and O Weaver. The physiological and behavioral effects of radio music on singly housed baboons. *Journal of medical primatology*, 1996.

Margot Brereton et al. The Messaging Kettle : Prototyping Connection over a Distance between Adult Children and Older Parents. *Proc. CHI*, 2015.

Geraldine Bricker-Katz, Michelle Lincoln, and Steven Cumming. Stuttering and work life: An interpretative phenomenological analysis. *Journal of fluency disorders*, 2013.

K Brodmann et al. Neue ergebnisse über die vergleichende histologische lokalisation der grosshirnrinde mit besonderer berücksichtigung des stirnhirns. *Anatomischer Anzeiger*, 1912.

Steven Brown, Michael J Martinez, and Lawrence M Parsons. Music and language side by side in the brain: a pet study of the generation of melodies and sentences. *European journal of neuroscience*, 23(10): 2791–2803, 2006.

Bradley R Buchsbaum, Juliana Baldo, Kayoko Okada, Karen F Berman, Nina Dronkers, Mark D'Esposito, and Gregory Hickok. Conduction aphasia, sensory-motor integration, and phonological short-term memory–an aggregate analysis of lesion and fmri data. *Brain and language*, 2011.

Theresa A Burnett et al. Voice f0 responses to pitch-shifted auditory feedback: a preliminary study. *Journal of Voice*, 1997.

Theresa A Burnett et al. Voice f0 responses to manipulations in pitch feedback. *The Journal of the Acoustical Society of America*, 1998.

Richard W Byrne and Lucy A Bates. Primate social cognition: uniquely primate, uniquely social, or just unique? *Neuron*, 2010.

Shanqing Cai, Jason A Tourville, Deryk S Beal, Joseph S Perkell, Frank H Guenther, and Satrajit S Ghosh. Diffusion imaging of cerebral white matter in persons who stutter: evidence for network-level anomalies. *Frontiers in human neuroscience*, 2014.

Francisco EC Cardoso et al. Cocaine-related movement disorders. *Journal of the Movement Disorder Society*, 1993.

Christine Careghi, Kokou Tona, Okanlawon Onagbesan, Johan Buyse, Eddy Decuypere, and Veerle Bruggeman. The effects of the spread of hatch and interaction with delayed feed access after hatch on broiler performance until seven days of age. *Poultry science*, 2005.

Malinda Carpenter and Josep Call. Comparing the imitative skills of children and nonhuman apes. *Revue de primatologie*, (1), 2009.

Marcus Carter et al. Naturalism and aci: augmenting zoo enclosures with digital technology. In *Advances in Computer Entertainment Technology*. ACM, 2015.

Herbert Newton Casson. *The history of the telephone*. AC McClurg & Company, 1910.

Clive K Catchpole and Peter JB Slater. *Bird song: biological themes and variations*. Cambridge university press, 2003.

William Cayley. Diagnosing the cause of chest pain. *Am Fam Physician*, 2005.

AS Chamove. Cage design reduces emotionality in mice. *Laboratory Animals*, 1989.

Soo-Eun Chang, Emily O Garnett, Andrew Etchell, and Ho Ming Chow. Functional and neuroanatomical bases of developmental stuttering: current insights. *The Neuroscientist*, 2018.

B D Charlton et al. Female giant panda (ailuropoda melanoleuca) chirps advertise the caller's fertile phase. *Proceedings of the Royal Society B: Biological Sciences*, 2009a.

B D Charlton et al. The information content of giant panda, ailuropoda melanoleuca, bleats: acoustic cues to sex, age and size. 2009b.

B D Charlton et al. Vocal cues to male androgen levels in giant pandas. *Biology Letters*, 2010.

B D Charlton et al. Coevolution of vocal signal characteristics and hearing sensitivity in forest mammals. *Nature communications*, 2019.

Benjamin D Charlton et al. Vocal behaviour predicts mating success in giant pandas. *Royal Society open science*, 2018.

Jennifer Chesters, Ladan Baghai-Ravary, and Riikka Möttönen. The effects of delayed auditory and visual feedback on speech production. *The Journal of the Acoustical Society of America*, 2015.

Kai-Jo Chiang et al. The effects of reminiscence therapy on psychological well-being, depression, and loneliness among the institutionalized aged. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences*, 2010.

Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

Julia Chosy et al. Behavioral and physiological responses in felids to exhibit construction. *Zoo biology*, 2014.

Cecilia P Chow et al. Vocal turn-taking in a non-human primate is learned during ontogeny. *Proceedings of the Royal Society B: Biological Sciences*, 2015.

Jakob Christensen-Dalsgaard. Amphibian bioacoustics. In *Handbook of signal processing in acoustics*. Springer, 2008.

Anna M Claxton. The potential of the human–animal relationship as an environmental enrichment for the welfare of zoo-housed animals. *Applied Animal Behaviour Science*, 2011.

Andrea W Clay et al. The use of technology to enhance zoological parks. *Zoo biology*, 2011.

Amon Cohen and AD Berstein. Acoustic transmission of the respiratory system using speech stimulation. *Biomedical Engineering, IEEE*, 1991.

NE Collias. The development of social behavior in birds. *The Auk*, 1952.

Theresa M Collins et al. *Thomas Edison and Modern America*. Boston: Palgrave Macmillan, 2002.

Diane Colombelli-Négrel et al. Embryonic learning of vocal passwords in superb fairy-wrens reveals intruder cuckoo nestlings. *Current Biology*, 2012.

Peter Cook, Andrew Rouse, Margaret Wilson, and Colleen Reichmuth. A california sea lion (zalophus californianus) can keep the beat: motor entrainment to rhythmic auditory stimuli in a non vocal mimic. *Journal of Comparative Psychology*, 2013.

Shelley Cook. Interaction sequences between chimpanzees and human visitors at the zoo. *Zoo Biology*, 1995.

Martin Cooke et al. Robust automatic speech recognition with missing and unreliable acoustic data. *Speech communication*, 2001.

Raymundo Cornejo et al. Enriching in-person encounters through social media: A study on family connectedness for the elderly. *International Journal of Human-Computer Studies*, 2013.

Dirk Corstens, Eleanor Longden, Simon McCarthy-Jones, Rachel Waddingham, and Neil Thomas. Emerging perspectives from the hearing voices movement: implications for research and practice. *Schizophrenia bulletin*, 2014.

Ângelo Costa et al. Multi-agent personal memory assistant. *Trends in practical applications of agents and multiagent systems*, 2010.

Jean Costa et al. Regulating feelings during interpersonal conflicts by changing voice self-perception. In *CHI*. ACM, 2018.

Robert Costanza et al. Sustainability or collapse: what can we learn from integrating the history of humans and the rest of nature? *AMBIO: A Journal of the Human Environment*, 2007.

Beverly J Cowart. Development of taste perception in humans: sensitivity and preference throughout the life span. *Psychological bulletin*, 1981.

Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal processing magazine*, 2001.

Robert G Crowder. Perception of the major/minor distinction: I. historical and theoretical foundations. *Psychomusicology: A Journal of Research in Music Cognition*, 1984.

A Kadir Çüçen. Heidegger's reading of descartes' dualism: The relation of subject and object. In *The Paideia Archive: Twentieth World Congress of Philosophy*, 1998.

Meagan E Curtis and Jamshed J Bharucha. The minor third communicates sadness in speech, mirroring its use in music. *Emotion*, 2010.

Cycling74. Max MSP, 1919. URL https://cycling74.com/.

Ayelet Dassa et al. The role of singing familiar songs in encouraging conversation among people with middle to late stage alzheimer's disease. *Journal of music therapy*, 2014.

Nico F Declercq and Cindy SA Dekeyser. Acoustic diffraction effects at the hellenistic amphitheater of epidaurus: Seat rows responsible for the marvelous acoustics. *The Journal of the Acoustical Society of America*, 2007a.

Nico F Declercq and Cindy SA Dekeyser. The acoustics of the hellenistic theatre of epidaurus: the important role of the seat rows. *Canadian Acoustics*, 2007b.

Volker B Deecke, John KB Ford, and Paul Spong. Dialect change in resident killer whales: implications for vocal learning and cultural transmission. *Animal behaviour*, 2000.

Sharon L Deem, Andrew J Noss, Rosa Leny Cuéllar, and William B Karesh. Health evaluation of free-ranging and captive blue-fronted amazon parrots (amazona aestiva) in the gran chaco, bolivia. *Journal of Zoo and Wildlife Medicine*, 2005.

ME Delany. Sound propagation in the atmosphere: a historical review. *Acta Acustica united with Acustica*, 1977.

Diana Deutsch, Trevor Henthorn, and Rachael Lapidis. Illusory transformation from speech to song. *The Journal of the Acoustical Society of America*, 2011.

Anthony Steven Dick et al. The frontal aslant tract (fat) and its role in speech, language and executive function. *cortex*, 2019.

Dragon. Naturallyspeaking, 2019. URL https://www.nuance.com/dragon.html.

Nina F Dronkers. A new brain region for coordinating speech articulation. *Nature*, 1996.

Homer Dudley and Thomas H Tarnoczy. The speaking machine of wolfgang von kempelen. *The Journal of the Acoustical Society of America*, 1950.

Robert Dudley and A Stanley Rand. Sound production and vocal sac inflation in the túngara frog, physalaemus pustulosus (leptodactylidae). *Copeia*, 1991.

Peter J Dugan et al. Phase 1: Dcl system research using advanced approaches for land-based or ship-based real-time recognition and localization of marine mammals-hpc system implementation. 2016.

Clement Duhart et al. Deep learning locally trained wildlife sensing in real acoustic wetland environment. In *International Symposium on Signal Processing and Intelligent Recognition Systems*. Springer, 2018.

Scott Durling and Jo Lumsden. Speech recognition use in healthcare applications. In *International conference on advances in mobile computing and multimedia*, 2008.

Joanne Edgar, Suzanne Held, Charlotte Jones, and Camille Troisi. Influences of maternal care on chicken welfare. *Animals*, 2016.

Victor Egger. *La parole intérieure: essai de psychologie descriptive*. Alcan, 1904.

Hillary Anger Elfenbein and Nalini Ambady. On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological bulletin*, 128(2):203, 2002.

Lise Eliot. *What's going on in there?: how the brain and mind develop in the first five years of life*. Bantam, 2000.

Wolfgang Enard, Molly Przeworski, Simon E Fisher, Cecilia SL Lai, Victor Wiebe, Takashi Kitano, Anthony P Monaco, and Svante Pääbo. Molecular evolution of foxp2, a gene involved in speech and language. *Nature*, 2002.

Roger M Evans. The development of learned auditory discriminations in the context of post-natal filial imprinting in young precocial birds. *Bird Behavior*, 1982.

Vasileios Exadaktylos, Mitchell Silva, Daniel Berckmans, and H Glotin. Automatic identification and interpretation of animal sounds, application to livestock production optimisation. *Soundscape Semiotics-Localization and Categorization*, pages 65–81, 2014.

Grant Fairbanks. Selective vocal effects of delayed auditory feedback. *Journal of Speech & Hearing Disorders*, 1955.

Grant Fairbanks and Newman Guttman. Effects of delayed auditory feedback upon articulation. *Journal of Speech & Hearing Research*, 1958.

Alexandra Farrand. The effect of zoo visitors on the behaviour and welfare of zoo mammals. 2007.

Bronwyn S Fees et al. A model of loneliness in older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 1999.

David R Feinberg et al. Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice. *Hormones and behavior*, 2006.

Charles Fernyhough and H Pashler. Inner speech. *The encyclopedia of the mind*, 2013.

William P Fifer and Christine M Moon. The role of mother's voice in the organization of brain function in the newborn. *Acta Paediatrica*, 83(s397): 86–93, 1994.

Steven R Fischer. *History of language*. Reaktion Books, 2001.

Anne L Foundas et al. Anomalous anatomy of speech–language areas in adults with persistent developmental stuttering. *Neurology*, 2001.

Anne L Foundas et al. The speecheasy device in stuttering and nonstuttering adults: Fluency effects while speaking and reading. *Brain and language*, 2013.

JG Frazer. The language of animals.(continued). *The Archaeological Review*, 1888.

Barry Metcalfe Freeman et al. *Development of the avian embryo: a behavioural and physiological study*. Springer, 1974.

Fiona French et al. High tech cognitive and acoustic enrichment for captive elephants. *Journal of neuroscience methods*, 2018a.

Fiona French et al. Soundjam: acoustic design for auditory enrichment. 2018b.

Olivier Friard and Marco Gamba. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution*, 2016.

HSPaJT Fruhstorfer and year=1970 publisher=Elsevier others, journal=Electroencephalography and clinical Neurophysiology. Short-term habituation of the auditory evoked response in man.

Ellen C Garland, Jason Gedamke, Melinda L Rekdahl, Michael J Noad, Claire Garrigue, and Nick Gales. Humpback whale song on the southern ocean feeding grounds: implications for cultural transmission. *PloS one*, 2013.

Asif A Ghazanfar and Drew Rendall. Evolution of human vocal production. *Current Biology*, 2008.

Kathleen Gibson. Tools, language and intelligence: Evolutionary implications. *Man*, 1991.

Carlo Ginzburg. À distance. Neuf essais sur le point de vue en histoire. 1998.

Lewis Glinert. Golem! the making of a modern myth. In *Symposium: A Quarterly Journal in Modern Literatures*. Taylor & Francis, 2001.

Friedrich Goethe. Beobachtungen bei der aufzucht junger silbermöwen. *Zeitschrift für Tierpsychologie*, 1955.

Julio Gonzalez et al. Early effects of smoking on the voice: a multidimensional study. *Medical Science Monitor*, 2004.

Luc Goossens et al. The genetics of loneliness: linking evolutionary theory to genome-wide genetics, epigenetics, and social science. *Perspectives on Psychological Science*, 2015.

Maria Luisa Gorno-Tempini, Simona Marina Brambati, Valeria Ginex, Jennifer Ogar, Nina F Dronkers, Alessandra Marcone, Daniela Perani, Valentina Garibotto, Stefano F Cappa, and Bruce L Miller. The logopenic/phonological variant of primary progressive aphasia. *Neurology*, 2008.

Charles Mayo Goss. Galen on anatomical procedures (de anatomicis administrationibus). *Translation of the surviving books with introduction and notes by Charles Singer. Oxford University Press, New York, The Anatomical Record*, 1958.

Gilbert Gottlieb. Prenatal auditory sensitivity in chickens and ducks. *Science*, 1965.

Jonathan Graff-Radford et al. The neuroanatomy of pure apraxia of speech in stroke. *Brain and language*, 2014.

Harold P Greeley et al. Detecting fatigue from voice using speech recognition. In *Signal Processing and Information Technology*, 2006.

Frank Guenther. *Neural Control of Speech*. 2016a. ISBN 978-0-262-03471-5.

Frank H Guenther. *Neural control of speech*. Mit Press, 2016b.

Barry Guitar. *Stuttering: An integrated approach to its nature and treatment*. Lippincott Williams & Wilkins, 2013.

Reinhard Gupfinger and Martin Kaltenbrunner. Sonic experiments with grey parrots: A report on testing the auditory skills and musical preferences of grey parrots in captivity. *Animal Computer Interactions*, 2017.

Reinhard Gupfinger et al. Animals make music: A look at non-human musical expression. *Multimodal Technologies and Interaction*, 2018.

G Gvaryahu et al. Filial imprinting, environmental enrichment, and music application effects on behavior and performance of meat strain chicks. *Poultry Science*, 1989.

Stephen Halliwell et al. *Aristotle's poetics*. University of Chicago Press, 1998.

Hadi Harb and Liming Chen. Voice-based gender identification in multimedia applications. *Journal of intelligent information systems*, 2005.

William A Hargreaves et al. Voice quality in depression. *Journal of Abnormal Psychology*, 1965.

Meredydd LL Harries, Judith M Walker, David M Williams, S Hawkins, and IA Hughes. Changes in the male voice at puberty. *Archives of disease in childhood*, 77(5):445–447, 1997.

Robert J Hartsuiker and Herman HJ Kolk. Error monitoring in speech production: A computational test of the perceptual loop theory. *Cognitive psychology*, 2001.

Elaine Hatfield and Christopher Hsee. The impact of vocal feedback on emotional experience and expression. 1995.

Graeme Hawthorne. Measuring social isolation in older adults: Development and initial validation of the friendship scale. *Social Indicators Research*, 2006a.

Graeme Hawthorne. Measuring social isolation in older adults: development and initial validation of the friendship scale. *Social Indicators Research*, 2006b.

E Charles Healey et al. Factors contributing to the reduction of stuttering during singing. *Journal of Speech, Language, and Hearing Research*, 1976.

Christopher L Heavey et al. The phenomena of inner experience. *Consciousness and cognition*, 2008.

Catherine Helmer et al. Marital status and risk of alzheimer's disease a french population-based cohort study. *Neurology*, 1999.

Laurence Henry et al. Social coordination in animal vocal interactions. is there any evidence of turn-taking? the starling as an animal model. *H., Casillas, M., Levinson, SC, eds.(2016). Turn-Taking in Human Communicative Interaction. Lausanne: Frontiers Media. doi: 10.3389*, 2016.

Peter G Hepper and B Sara Shahidullah. Development of fetal hearing. *Archives of Disease in Childhood-Fetal and Neonatal Edition*, 71(2): F81–F87, 1994.

Peter G Hepper et al. Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, 1993.

Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 2012.

Ingo Hertrich et al. The role of the supplementary motor area for speech and language processing. *Neuroscience & Biobehavioral Reviews*, 2016.

Eckhard H Hess. " imprinting" in a natural laboratory. *Scientific American*, 1972.

Kate Hevner. The affective character of the major and minor modes in music. *The American Journal of Psychology*, 1935.

Gregory Hickok. The functional neuroanatomy of language. *Physics of life reviews*, 2009.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 2007.

Gregory Hickok, John Houde, and Feng Rong. Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, 2011.

Theodore Hiebert. The tower of babel and the origin of the world's cultures. *Journal of Biblical Literature*, 2007.

Robert A Hinde. Animal behaviour: A synthesis of ethology and comparative psychology. 1970.

Ilyena Hirskyj-Douglas et al. Where hci meets aci. In *Nordic Conference on Human-Computer Interaction*. ACM, 2016.

Marisa Hoeschele, Hugo Merchant, Yukiko Kikuchi, Yuko Hattori, and Carel ten Cate. Searching for the origins of musicality across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015.

Charles Holbrow, Elena Naomi Jessop, and Rébecca Kleinberger. Vocal vibrations: A multisensory experience of the voice. In *NIME*, 2014.

Julianne Holt-Lunstad et al. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on psychological science*, 2015.

Tjalling Jan Holwerda et al. Feelings of loneliness, but not social isolation, predict dementia onset. *J Neurol Neurosurg Psychiatry*, 2012.

Brigitte Holzinger. Conversation between stephen laberge and paul tholey, july, 1989. *Lucidity Letter*, 9(1), 1990.

Homer. *The odyssey*.

William D Hopkins, Jared P Taglialatela, and David A Leavens. Chimpanzees differentially produce novel vocalizations to capture the attention of a human. *Animal behaviour*, 2007.

Geoff Hosey. A preliminary model of human–animal relationships in the zoo. *Applied Animal Behaviour Science*, 2008.

Geoff Hosey et al. Are we ignoring neutral and negative human–animal relationships in zoos? *Zoo biology*, 2015.

John F Houde and Michael I Jordan. Sensorimotor adaptation of speech i: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, 45(2), 2002.

John F Houde et al. Modulation of the auditory cortex during speech: an meg study. *Journal of cognitive neuroscience*, 2002.

Julia M Hoy et al. Thirty years later: Enrichment practices for captive mammals. *Zoo Biology*, 2010.

Xiangming Huang et al. Human assistance in a giant panda mother for rearing her baby. *Sichuan journal of zoology*, 2005.

Y Huang et al. Use of artificial insemination to enhance propagation of giant pandas at the wolong breeding center. 2002.

A James Hudspeth. How the ear's works work. *Nature*, 1989.

Vanessa Hull et al. Space use by endangered giant pandas. *Journal of Mammalogy*, 2015.

Russell T Hurlburt et al. Toward a phenomenology of inner speaking. *Consciousness and Cognition*, 2013.

Russell T Hurlburt et al. Exploring the ecological validity of thinking on demand: neural correlates of elicited vs. spontaneously occurring inner speech. *PLoS One*, 2016.

Monica Impekoven. Prenatal experience of parental calls and pecking in the laughing gull. *Animal Behaviour*, 1971.

Monica Impekoven. The response of incubating laughing gulls to calls of hatching chicks. *Behaviour*, 1973.

Kori Inkpen et al. Experiences2go: sharing kids' activities outside the home with remote family members. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 2013.

Alice M Isen and Joyce M Gorgoglione. Some specific effects of four affect-induction procedures. *Personality and Social Psychology Bulletin*, 1983.

IUCN. The international union for conservation of nature's red list of threatened species. version 2016, 2016.

Marjan Jahanshahi et al. A fronto–striato–subthalamic–pallidal network for goal-directed and habitual inhibition. *Nature Reviews Neuroscience*, 2015.

Anthony Jahn and Andrew Blitzer. A short history of laryngoscopy. *Logopedics Phoniatrics Vocology*, 1996.

Petr Janata. The neural architecture of music-evoked autobiographical memories. *Cerebral Cortex*, 2009.

Lisa M Jaremka et al. Cognitive problems among breast cancer survivors: loneliness enhances risk. *Psycho-Oncology*, 2014.

Erich D Jarvis. Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences*, 2004.

James S Jenkins. The voice of the castrato. *The Lancet*, 1998.

Jack Jiang, Emily Lin, and David G Hanson. Vocal fold physiology. *Otolaryngologic Clinics of North America*, 2000.

FL Jones. Poor breath sounds with good voice sounds. a sign of bronchial stenosis. *CHEST Journal*, 93(2):312–313, 1988.

R Bryan Jones. Fear and adaptability in poultry: insights, implications and imperatives. *World's Poultry Science Journal*, 52(2):131–174, 1996.

RB Jones. Environmentai enrichment: the need for practicai strategies to improve pouitry weifare. *Welfare of the laying hen*, 2004.

Simon R Jones and Charles Fernyhough. Thought as action: Inner speech, self-monitoring, and auditory verbal hallucinations. *Consciousness and cognition*, 2007.

Camil Jreige, Rupal Patel, and H Timothy Bunnell. Vocalid: personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 259–260. ACM, 2009.

Tejinder K Judge et al. The family window: the design and evaluation of a domestic media space. In *CHI*. ACM, 2010.

Judy Kuster. Some apps for Stuttering. URL https://www.mnsu.edu/comdis/kuster/appsforstuttering.html.

Lejla Junuzović-Žunić et al. Voice characteristics in patients with thyroid disorders. *The Eurasian journal of medicine*, 2019.

David Kahn and Allan Hobson. Theory of mind in dreaming: Awareness of feelings and thoughts of others in dreams. *Dreaming*, 2005.

Joseph Kalinowski et al. Stuttering amelioration at various auditory feedback delays and speech rates. *International Journal of Language & Communication Disorders*, 1996.

Tobias Kalisch and other. Cognitive and tactile factors affecting human haptic performance in later life. 2012.

Anne Marie Kanstrup et al. Designing connections for hearing rehabilitation. In *DIS*, 2017.

Fares Kayali et al. Elements of play for cognitive, physical and social health in older adults. In *Human Factors in Computing and Informatics*. Springer, 2013.

Helen Keller and Annie Sullivan. *The story of my life*. Doubleday, 1904.

Ray D Kent. The uniqueness of speech among motor systems. *Clinical linguistics & phonetics*, 2004.

Keruve. Keruve alzheimer's gps tracking device, 2008. URL http://www.keruve.com/.

Caitlin R Kight et al. How and why environmental noise impacts animals: an integrative, mechanistic review. *Ecology letters*, 2011.

James M Kilner, Karl J Friston, and Chris D Frith. Predictive coding: an account of the mirror neuron system. *Cognitive processing*, 2007.

Nicky NE Kim-McCormack et al. Is interactive technology a relevant and effective enrichment for captive great apes? *Applied animal behaviour science*, 2016.

William H Kimbel and Lucas K Delezene. "lucy" redux: A review of research on australopithecus afarensis. *American journal of physical anthropology*, 2009.

AM King'Ori et al. Review of the factors that influence egg fertility and hatchability in poultry. *International Journal of Poultry Science*, 2011.

Ivan Kiskin et al. Mosquito detection with neural networks: the buzz of deep learning. 2017.

Vernon N Kisling. *Zoo and aquarium history: Ancient animal collections to zoological gardens*. CRC press, 2000.

DG Kleiman and G Peters. Auditory communication in the giant panda: motivation and function. In *Proceedings of the Second International Symposium on the Giant Panda. Tokyo Zoological Park Society*, 1990.

Rébecca Kleinberger. Vocal musical expression with a tactile resonating device and its psychophysiological effects.

Rébecca Kleinberger. Singing about singing: using the voice as a tool for self-reflection, 2014.

Rébecca Kleinberger, Janet Baker, and Gabriel Miller. Initial observation of human-bird vocal interactions in a zoological setting. *PeerJ Preprints*, 2019a.

Rébecca Kleinberger, Stefanakis George, and Sebastian Franjou. Speech companions: Evaluating the effects of musically modulated auditory feedback on the voice. 2019b.

Rebecca Kleinberger, Alexandra Rieger, Janelle Sands, and Janet Baker. Supporting elder connectedness through cognitively sustainable design interactions with the memory music box. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. ACM, 2019c.

Rebecca Kleinberger, George Stefanakis, Satrajit Ghosh, Tod Machover, and Michael Erkkinen. Fluency effects of novel acoustic vocal transformations in people who stutter: An exploratory behavioral study, 2019d.

Bernard MW Knox. Silent reading in antiquity. *Greek, Roman, and Byzantine Studies*, 9(4):421–435, 1968.

Donghyeon Ko et al. Bubbletalk: Enriching experience with fish by supporting human behavior. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 2018.

Masakazu Konishi. Development of auditory neuronal responses in avian embryos. *Proceedings of the National Academy of Sciences*, 70(6):1795–1798, 1973.

Masakazu Konishi et al. Contributions of bird studies to biology. *Science*, 1989.

Nate Kornell. Metacognition in humans and animals. *Current Directions in Psychological Science*, 18(1):11–15, 2009.

Michael D Kreger and Joy A Mench. Visitor–animal interactions at the zoo. *Anthrozos*, 1995.

Thomas S Kuhn. The structure of scientific revolutions. *Chicago and London*, 1962.

Filipa MB Lã, William L Ledger, Jane W Davidson, David M Howard, and Georgina L Jones. The effects of a third generation combined oral contraceptive pill on the classical singing voice. *Journal of Voice*, 21(6): 754–761, 2007.

MC Lábaque et al. Microbial contamination of artificially incubated greater rhea (rhea americana) eggs. *British Poultry Science*, 2003.

René Théophile Hyacinthe Laennec. *De l'auscultation médiate: ou, Traité du diagnostic des maladies des poumons et du coeur; fondé principalement sur ce nouveau moyen d'exploration*, volume 2. Culture et civilisation, 1819.

René Théophile Hyacinthe Laennec. *Traité de l'auscultation médiate, et des maladies des poumons et du coeur*. Société Typographique Belge, 1837.

Matti Laine et al. Left hemisphere activation during processing of morphologically complex word forms in adults. *Neuroscience Letters*, 1999.

José L Lanciego, Natasha Luquin, and José A Obeso. Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine*, 2012.

Marge E Landsberg. *The genesis of language: a different judgement of evidence*, volume 3. Walter de Gruyter, 2011.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (iaps): Technical manual and affective ratings. *NIMH Center for the Study of Emotion and Attention*, 1997.

Richard A Lanham. *The economics of attention: Style and substance in the age of information*. University of Chicago Press, 2006.

Megan J Larsen et al. Number of nearby visitors and noise level affect vigilance in captive koalas. *Applied Animal Behaviour Science*, 2014.

Marianne Latinus and Pascal Belin. Human voice perception. *Current biology : CB*, 21, 2011.

William Lauder et al. A comparison of health behaviours in lonely and non-lonely populations. *Psychology, Health & Medicine*, 2006.

John Laver. *Principles of phonetics*. Cambridge University Press, 1994.

Amanda Lazar et al. Designing for the third hand: Empowering older adults with cognitive impairment through creating and sharing. In *DIS*, 2016.

Jean-Pierre Lecanuet and Benoist Schaal. Sensory performances in the human foetus: A brief summary of research. *Intellectica*, 2002.

Chaiwoo Lee et al. Perspective: Older adults' adoption of technology: an integrated approach to identifying determinants and barriers. *Journal of Product Innovation Management*, 2015.

Ping Lee et al. A mobile pet wearable computer and mixed reality system for human–poultry interaction through the internet. *Ubicomp*, 2006.

Robert E Lemon. How birds develop song dialects. *The Condor*, 1975.

Morris Michael Lewis. *Infant speech: A study of the beginnings of language*. Routledge, 2013.

Philip Lieberman. Can chimpanzees swallow or talk? a reply to falk. *American Anthropologist*, 1982.

Michelle Lincoln, Ann Packman, and Mark Onslow. Altered auditory feedback and the treatment of stuttering: A review. *Journal of fluency disorders*, 2006.

Siân E Lindley, , et al. Desiring to be in touch in a changing communications landscape: attitudes of older adults. In *CHI*. ACM, 2009.

M. a. Little et al. Testing the assumptions of linear prediction analysis in normal vowels. 2006.

Hung-Huan Liu et al. Mobile guiding and tracking services in public transit system for people with mental illness. In *TENCON IEEE*, 2009.

Gill Livingston et al. Dementia prevention, intervention, and care. *The Lancet*, 2017.

Hélène Loevenbruck. What the neurocognitive study of inner language reveals about our inner space. *Langage Intérieur/Espaces Intérieur, Inner Speech/Inner Space*, 2018.

Edward Loper and Steven Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

F Luc, Robert M Kroll, Shitij Kapur, and Sylvain Houle. A positron emission tomography study of silent and oral single word reading in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research*, 2000.

Alvin Lucier. I am sitting in a room. 2000.

Andrew U Luescher. *Manual of parrot behavior*. 2006.

Tanya M Luhrmann, Ramachandran Padmavati, Hema Tharoor, and Akwasi Osei. Differences in voice-hearing experiences of people with psychosis in the usa, india and ghana: interview-based study. *The British Journal of Psychiatry*, 2015.

Aleksandr Romanovich Luria. *Higher cortical functions in man*. Springer Science & Business Media, 2012.

Oisin Mac Aodha et al. Bat detective—deep learning tools for bat acoustic signal detection. *PLoS computational biology*, 2018.

Bronislaw Malinowski. The problem of meaning in primitive languages. *Language and literacy in social practice: A reader*, pages 1–10, 1994.

M Manjutha, J Gracy, P Subashini, and M Krishnaveni. Automated speech recognition system–a literature review. *Computational Methods, communication techniques and informatics*, 2017.

Philippe Manoury. *La musique du temps réel*. Editions MF, 2012.

Peter Mariën et al. Consensus paper: language and the cerebellum: an ongoing enigma. *The Cerebellum*, 2014.

Mylene M Mariette. Prenatal acoustic communication programs offspring for high posthatching temperatures in a songbird. *Science*, 2016.

Mika H Martikainen et al. Suppressed responses to self-triggered sounds in the human auditory cortex. 2005.

Maryanne Martin. Speech recoding in silent reading. *Memory & Cognition*, 1978.

G Marx et al. Analysis of pain-related vocalization in young pigs. *Journal of sound and vibration*, 2003.

Ludo Max, Frank H Guenther, Vincent L Gracco, Satrajit S Ghosh, and Marie E Wallace. Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary issues in communication science and disorders*, 2004.

Brian Mayton et al. The networked sensory landscape: Capturing and experiencing ecological change across scales. *PRESENCE: Teleoperators and Virtual Environments*, 2017.

Simon McCarthy-Jones. *Hearing voices: The histories, causes and meanings of auditory verbal hallucinations*. Cambridge University Press, 2012.

James S McCasland and Masakazu Konishi. Interaction between auditory and motor activities in an avian song control nucleus. *Proceedings of the National Academy of Sciences*, 78(12):7815–7819, 1981.

Karen McComb et al. Elephants can determine ethnicity, gender, and age from acoustic cues in human voices. *Proceedings of the National Academy of Sciences*, 2014.

Peter K. McGregor. Revealing the acoustic mysteries of Byzantine churches, 2019. URL https://faithandform.com/.

David McNamee. Hey, what's that sound: Throat singing, 2010. URL theguardian.com/music/2010/jun/02/throat-singing.

FJ Meiland et al. Cogknow: development of an ict device to support people with mild dementia. *Journal on Information Technology in Healthcare*, 2007.

Jeffrey Meyers. *The genius and the goddess: Arthur Miller and Marilyn Monroe*. University of Illinois Press, 2012.

Rachel M Miller, Kauyumari Sanchez, Lawrence D Rosenblum, James W Dias, and Neal Dykmans. Talker-specific accent: Can speech alignment reveal idiolectic influences during the perception of accented speech? *The Journal of the Acoustical Society of America*, 127(3):1958–1958, 2010.

Marvin Minsky. *Society of mind*. Simon and Schuster, 1988.

Shalini Misra, Lulu Cheng, Jamie Genevie, and Miao Yuan. The iphone effect: the quality of in-person social interactions in the presence of mobile devices. *Environment and Behavior*, 48(2):275–298, 2016.

Karyn Moffatt, , et al. Connecting grandparents and grandchildren. In *Connecting Families*. 2013.

Jay P Mohr et al. Broca aphasia: pathologic and clinical. *Neurology*, 1978.

A Morin. Inner speech. encyclopedia of human behavior, w. hirstein, 2012.

Kellie Morrissey et al. The value of experience-centred design approaches in dementia research contexts. In *CHI 2017*.

Daniella Jorge de Moura, Irenilza de Alencar Nääs, Elaine Cangussu de Souza Alves, Thayla Morandi Ridolfi de Carvalho, Marcos Martinez do Vale, and Karla Andrea Oliveira de Lima. Noise analysis to evaluate chick thermal comfort. *Scientia Agricola*, 2008a.

DJ Moura et al. Real time computer stress monitoring of piglets using vocalization analysis. *Computers and Electronics in Agriculture*, 2008b.

Ruth Mugge, Jan PL Schoormans, and Hendrik NJ Schifferstein. Emotional bonding with personalised products. *Journal of Engineering Design*, 20 (5):467–476, 2009.

Nadia Müller et al. Listen to Yourself: The Medial Prefrontal Cortex Modulates Auditory Alpha Power During Speech Preparation. *Cerebral cortex*, 2014.

James C Mundt et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 2007.

Kevin G Munhall et al. Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, 2009.

Raymond LH Murphy Jr et al. Visual lung-sound characterization by time-expanded wave-form analysis. *New England Journal of Medicine*, 1977.

Carlos G Musso. Imhotep: the dean among the ancient Egyptian physicians. an example of a complete physician. *Humane Medicine Health Care*, 2005.

Elizabeth D Mynatt et al. Digital family portraits: supporting peace of mind for extended family members. In *CHI*. ACM, 2001.

Edward D Mysak. Pitch and duration characteristics of older males. *Journal of Speech and Hearing Research*, 1959.

Thomas Nagel. What is it like to be a bat? *The philosophical review*, 1974.

Ulrich Natke and Karl Theodor Kalveram. Effects of frequency-shifted auditory feedback on fundamental frequency of long stressed and unstressed syllables. *Journal of Speech, Language, and Hearing Research*, 2001.

Ulrich Natke, Juliane Grosser, and Karl Theodor Kalveram. Fluency, fundamental frequency, and speech rate under frequency-shifted auditory feedback in stuttering and nonstuttering persons. *Journal of Fluency Disorders*, 2001.

Michael Nees. Hearing ghost voices relies on pseudoscience and fallibility of human perception. 2015.

VE Negus. The mechanism of swallowing, 1942.

Deborah G Kemler Nelson et al. How the prosodic cues in motherese might assist language learning. *Journal of child Language*, 1989.

DAT New. A critical period for the turning of hens' eggs. *Development*, 5 (3):293–299, 1957.

Margaret Morse Nice. *Development of behavior in precocial birds*, volume 8. New York:[Linnaean Society], 1962.

Margaret Morse Nice et al. Studies in the life history of the song sparrow. 1964.

NIH NIDCD. Stuttering. URL https://www.nidcd.nih.gov/health/stuttering.

Paula M Niedenthal. Embodying emotion. *science*, 2007.

Mark Nielsen. The imitative behaviour of children and chimpanzees: A window on the transmission of cultural traditions. *Revue de primatologie*, 2009.

AJ Nimon et al. Cross-species interaction and communication: a study method applied to captive siamang and long-billed corella contacts with humans. *Applied Animal Behaviour Science*, 1992.

Takeshi Nishimura, Akichika Mikami, Juri Suzuki, and Tetsuro Matsuzawa. Descent of the larynx in chimpanzee infants. *Proceedings of the National Academy of Sciences*, 2003.

Jose C Noguera et al. Bird embryos perceive vibratory cues of predation risk from clutch mates. *Nature ecology & evolution*, 2019.

Sam Norman-Haignere, Nancy G Kanwisher, and Josh H McDermott. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron*, 2015.

Stephen Nowicki and Peter Marler. How do birds sing? *Music Perception: An Interdisciplinary Journal*, 1988.

Arthur M. Noxon. *Understanding Church Acoustics*. Acoustic Sciences Corporation, 2001.

Frank Noz and Jinsoo An. Cat cat revolution: an interspecies gaming experience. In *CHI*. ACM, 2011.

Ruth Ogden and Catharine Montgomery. High time. *Psychologist*, 25(8), 2012.

Ronald W Oppenheim. Prehatching and hatching behaviour in birds: a comparative study of altricial and precocial species. *Animal Behaviour*, 1972.

World Health Organization. *World report on ageing and health*. World Health Organization, 2015.

C Owen. Do visitors affect the asian short-clawed otter in a captive environment. In *Zoo Research Symposium*, 2004.

M A Owen et al. Monitoring stress in captive giant pandas (ailuropoda melanoleuca): behavioral and hormonal responses to ambient noise. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2004.

M A Owen et al. Signalling behaviour is influenced by transient social context in a spontaneously ovulating mammal. *Animal Behaviour*, 2016.

Megan A Owen et al. Dynamics of male–female multimodal signaling behavior across the estrous cycle in giant pandas (ailuropoda melanoleuca). *Ethology*, 2013.

Donald H Owings, Eugene S Morton, et al. *Animal vocal communication: a new approach*. Cambridge University Press, 1998.

Sarah Partan and Peter Marler. The umwelt and its relevance to animal communication: introduction to special issue. *Journal of Comparative Psychology*, 2002.

Aniruddh D Patel. *Music, language, and the brain*. Oxford university press, 2010.

Aniruddh D Patel, John R Iversen, Micah R Bregman, and Irena Schulz. Experimental evidence for synchronization to a musical beat in a nonhuman animal. *Current biology*, 2009.

Patricia G Patrick and Sue Dale Tunnicliffe. The zoo voice: Zoo education and learning. In *Zoo Talk*. Springer, 2013.

Francine GP Patterson and Ronald H Cohn. Language acquisition by a lowland gorilla: Koko's first ten years of vocabulary development. *Word*, 41(2):97–143, 1990.

Mark W Pellowski and Edward G Conture. Characteristics of speech disfluency and stuttering behaviors in 3-and 4-year-old children. *Journal of Speech, Language, and Hearing Research*, 2002.

Wilder Penfield and Lamar Roberts. *Speech and brain mechanisms*, volume 62. Princeton University Press, 2014.

Irene M Pepperberg. Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science*, 11(3):83–87, 2002.

Irene M Pepperberg. Animal language studies: What happened? *Psychonomic bulletin & review*, 2017.

Irene M Pepperberg and Mary A McLaughlin. Effect of avian–human joint attention in allospecific vocal learning by grey parrots (psittacus erithacus). *Journal of Comparative Psychology*, 1996.

Irene M Pepperberg and Irene M Pepperberg. *The Alex studies: cognitive and communicative abilities of grey parrots*. Harvard University Press, 2009.

Isabelle Peretz, Dominique Vuvan, Marie-Élaine Lagrois, and Jorge L Armony. Neural overlap in processing music and speech. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2015.

Joseph S Perkell. Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of neurolinguistics*, 2012.

Marcela Perrone-Bertolotti, Jan Kujala, Juan R Vidal, Carlos M Hamame, Tomas Ossandon, Olivier Bertrand, Lorella Minotti, Philippe Kahane, Karim Jerbi, and Jean-Philippe Lachaux. How silent is silent reading? intracerebral evidence for top-down activation of temporal voice areas during reading. *The Journal of Neuroscience*, 32(49):17554–17562, 2012.

Marcela Perrone-Bertolotti et al. What is that little voice inside my head? inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural brain research*, 2014.

Tomas Persson et al. Spontaneous cross-species imitation in interactions between chimpanzees and zoo visitors. *Primates*, 2018.

Martin Pickford. Orientation of the foramen magnum in late miocene to extant african apes and hominids. *Anthropologie*, 2005.

María Pifarré et al. The effect of zoo visitors on the behaviour and faecal cortisol of the mexican wolf (canis lupus baileyi). *Applied Animal Behaviour Science*, 2012.

Anne Marie Piper et al. Exploring the accessibility and appeal of surface computing for older adult health care support. In *CHI*. ACM, 2010.

Ryan Pollard et al. Effects of the speecheasy on objective and perceived aspects of stuttering: A 6-month, phase i clinical trial in naturalistic environments. *Journal of Speech, Language, and Hearing Research*, 2009.

Patricia Pons et al. Sound to your objects: a novel design approach to evaluate orangutans' interest in sound-based stimuli. In *ACI'16*. ACM, 2016.

Ilyas Potamitis. Deep learning for detection of bird vocalisations. 2016.

James FA Poulet and Berthold Hedwig. A corollary discharge maintains auditory sensitivity during sound production. *Nature*, 2002.

David M Powell et al. Effects of construction noise on behavior and cortisol levels in a pair of captive giant pandas (ailuropoda melanoleuca). *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2006.

Robert Pracy. The infant larynx. *The Journal of Laryngology & Otology*, 97(10):933–947, 1983.

Jane E Prasse et al. Stuttering: an overview. *American family physician*, 2008.

Christine Preibisch et al. Evidence for compensation for stuttering by the right frontal operculum. 2003.

Daniel Pressnitzer, Jackson Graves, Claire Chambers, Vincent De Gardelle, and Paul Egré. Auditory perception: Laurel and yanny together at last. *Current Biology*, 2018.

Cathy J Price et al. A generative model of speech production in Broca's and Wernicke's areas. *Frontiers in psychology*, 2:237, 2011.

Wolfgang Prinz. Perception and action planning. *European journal of cognitive psychology*, 1997.

Melda Production. Mharmonizermb, 2019. URL https://www.meldaproduction.com/MHarmonizerMB.

Dunwu Qi et al. Different habitat preferences of male and female giant pandas. *Journal of Zoology*, 2011.

Sandra Quadros et al. Zoo visitor effect on mammal behaviour: Does noise matter? *Applied Animal Behaviour Science*, 2014.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *IEEE*, 1989.

Hayes Raffle et al. Hello, is grandma there? let's read! storyvisit: family video chat and connected e-books. In *CHI*. ACM, 2011.

Reaper. Digital audio workstation, 2019. URL reaper.fm.

Luiz Antonio de Lima Resende, Silke Anna Theresa Weber, Marcelo Fernando Zeugner Bertotti, and Svetlana Agapejev. Stroke in ancient times: a reinterpretation of psalms 137: 5, 6. *Arquivos de neuro-psiquiatria*, 2008.

Nikki S Rickard et al. The effect of music on cognitive performance: Insight from neurobiological and animal studies. *Behavioral and Cognitive Neuroscience Reviews*, 2005.

Sam Ridgway, Donald Carder, Michelle Jeffries, and Mark Todd. Spontaneous human speech mimicry by a cetacean. *Current Biology*, 22(20): R860–R861, 2012.

G Riley. Ssi-4 stuttering severity instrument fourth edition, 2009.

Sarah Elizabeth Ritvo. Music preference and discrimination in three sumatran orangutans. 2013.

Molly Roberts. We wanted to believe in koko, and so we did, 2018. URL https://www.chicagotribune.com.

Marcela D Rodríguez et al. Home-based communication system for older adults and their remote family. *Computers in Human Behavior*, 2009.

R M Rolland et al. Evidence that ship noise increases stress in right whales. *Proceedings of the Royal Society B: Biological Sciences*, 2012.

Paul Saenger. *Space between words: The origins of silent reading*. Stanford University Press, 1997.

Pierre Salamé and Alan Baddeley. Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology Section A*.

Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 2017.

Roberto Salguero-Gómez et al. Comadre: a global data base of animal demography. *Journal of Animal Ecology*, 2016.

Paul Sanden. Hearing glenn gould's body: Corporeal liveness in recorded music. In *Liveness in Modern Music*. Routledge, 2013.

JC Saunders. The development of auditory evoked responses in the chick embryo. *Minerva Otolaryngol*, 24:221–229, 1974.

Disa A Sauter, Frank Eisner, Paul Ekman, and Sophie K Scott. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, 107(6): 2408–2412, 2010.

George Savran. Beastly speech: intertextuality, balaam's ass and the garden of eden. *Journal for the Study of the Old Testament*, 19(64):33–55, 1994.

George B. Schaller et al. The giant pandas of wolong. *The Quarterly Review of Biology*, 1990.

Becky Scheel. Designing digital enrichment for orangutans. In *Animal-Computer Interaction*. ACM, 2018.

Klaus R Scherer. Expression of emotion in voice and music. *Journal of voice*, 1995.

Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

Florian Schiel et al. Rhythm and formant features for automatic alcohol detection. In *International Speech Communication Association*, 2010.

MB Schiff. Sex hormones and the female voice. *Journal of Voice-Official Journal of the Voice Foundation*, 13(3):424–424, 1999.

Arnold Schoenberg. *Moses und Aron: opera*, volume 8004. E. Eulenburg, 1984.

Gerd Schuller. Vocalization influences auditory processing in collicular neurons of the cf-fm-bat, rhinolophus ferrumequinum. *Journal of comparative physiology*, 1979.

Geralyn M Schulz and Megan K Grant. Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in parkinson's disease: a review of the literature. *Journal of communication disorders*, 2000.

Elke Schüttler et al. Vulnerability of ground-nesting waterbirds to predation by invasive american mink in the cape horn biosphere reserve, chile. *Biological Conservation*, 2009.

Tony Schwartz and Richard Kostelanetz. Interview with tony schwartz, american hörspielmacher. *Perspectives of New Music*, 1996.

David B Schwarz. *X: An Analytical Approach to John Chowning's Phone*. PhD thesis, Citeseer, 2010.

Sophie Scott. Theres a lot more to conversation than words. What really happens when we talk, Aeon Videos, 2015.

Marc Seal et al. Compelling imagery, unanticipated speech and deceptive memory: Neurocognitive models of auditory verbal hallucinations in schizophrenia. *Cognitive Neuropsychiatry*, 2004.

Mohamed L Seghier. The angular gyrus: multiple functions and multiple subdivisions. *The Neuroscientist*, 2013.

Paul C Sereno and Fernando E Novas. The complete skull and skeleton of an early dinosaur. *Science*, 1992.

Rahul K Shah et al. Relationship between voice quality and vocal nodule size. *Otolaryngology–Head and Neck Surgery*, 2008.

Claude E Shannon. Prediction and entropy of printed english. *Bell system technical journal*, 1951.

L James Shapiro. Pre-hatching influences that can potentially mediate post-hatching attachments in birds. *Bird Behavior*, 1981.

Marc Shell. Animals that talk. *Differences: A Journal of Feminist Cultural Studies*, 2004.

Marc Shell. Moses' tongue. *Common Knowledge*, 2006.

David J Shepherdson. Environmental enrichment: past, present and future. *International Zoo Yearbook*, 2003.

Sehar Shoukat. Rowan atkinson to mr. bean: A story of weakness to success-case study. *SJSS)*, 2019.

Waldo Shumway. The recapitulation theory. *The Quarterly Review of Biology*, 7(1):93–99, 1932.

David Sibley. The proper use of playback in birding, 2011. URL https://www.sibleyguides.com/2011/04/the-proper-use-of-playback-in-birding/.

James A Simmons, Ernest Glen Wever, and Joseph M Pylka. Periodical cicada: sound production and hearing. *Science*, 1971.

John A Sloboda. Music structure and emotional response: Some empirical findings. *Psychology of music*, 1991.

Maria L Slowiaczek and Charles Clifton Jr. Subvocalization and reading for meaning. *Journal of verbal learning and verbal behavior*, 1980.

Jonathan Smallwood and Jonathan W Schooler. The science of mind wandering: empirically navigating the stream of consciousness. *Annual review of psychology*, 66:487–518, 2015.

SmartSole. Smartsole:hidden, wearable monitoring and recovery solution for wandering. URL http://gpssmartsole.com/gpssmartsole/.

Rebecca J Snyder et al. Giant panda maternal care: A test of the experience constraint hypothesis. *Scientific reports*, 2016.

John Sparks. *Allogrooming in primates*. Aldine Chicago, 1967.

Marek Špinka, Françoise Wemelsfelder, et al. Environmental challenge and animal agency. *Animal welfare*, pages 27–43, 2011.

Stephanie Spinner. *Alex the parrot: no ordinary bird*. Knopf Books for Young Readers, 2012.

Vivek Kumar Rangarajan Sridhar et al. Detecting prominence in conversational speech: pitch accent, givenness and focus. In *Proceedings of Speech Prosody*. International Speech Communication Association Campinos, Brazil, 2008.

C Woodruff Starkweather. *Fluency and stuttering.* Prentice-Hall, Inc, 1987.

Elizabeth L Stegemööller et al. Music training and vocal production of speech and song. *Music Perception: An Interdisciplinary Journal*, 2008.

Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

Austin Stewart. second livestock. URL http://www.theaustinstewart.com/secondlivestock.html.

Angela S Stoeger et al. Acoustic features indicate arousal in infant giant panda vocalisations. *Ethology*, 2012a.

Angela S Stoeger et al. An asian elephant imitates human speech. *Current Biology*, 2012b.

D Stowell et al. On-bird sound recordings: automatic acoustic recognition of activities and contexts. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2017.

Ariel Stravynski and Richard Boyer. Loneliness in relation to suicide ideation and parasuicide: A population-wide study. *Suicide and Life-Threatening Behavior*, 2001.

Julia Strout et al. Anuran call classification with deep learning. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Andrew Stuart et al. Effect of monaural and binaural altered auditory feedback on stuttering frequency. *The Journal of the Acoustical Society of America*, 1997.

Andrew Stuart et al. Investigations of the impact of altered auditory feedback in-the-ear devices on the speech of people who stutter: initial fitting and 4-month follow-up. *International Journal of Language & Communication Disorders*, 2004.

Stuttering Foundation. A nonprofit organization helping those who stutter, 2018. URL https://www.stutteringhelp.org/.

Thomas Suddendorf and Andrew Whiten. Mental evolution and development: Evidence for secondary representation in children, great apes, and other animals. *Psychological bulletin*, 2001.

Hyewon Suh et al. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *CHI*. ACM, 2016.

RR Swaisgood et al. Developmental stability of foraging behavior: evaluating suitability of captive giant pandas for translocation. *Animal conservation*, 2018.

Diane Sweeney and Brad Williamson. *Biology: Exploring Life: Laboratory Manual*. Pearson Education, Incorporated, 2006.

Etologia Tanszek. The family dog project, 2019. URL `https://familydogproject.elte.hu`.

W Tecumseh Fitch and David Reby. The descended larynx is not uniquely human. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 2001.

Catherine Theys et al. A crucial role for the cortico-striato-cortical loop in the pathogenesis of stroke-related neurogenic stuttering. *Human Brain Mapping*, 2013.

Paul Tholey. Consciousness and abilities of dream characters observed during lucid dreaming. *Perceptual and Motor Skills*, 1989.

Xing Tian and David Poeppel. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in Human Neuroscience*, 2012.

Xing Tian et al. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, 2016.

Kaisa Tiippana. What is the mcgurk effect? *Frontiers in Psychology*, 2014.

Sarvada Chandra Tiwari. Loneliness: A disease? *Indian journal of psychiatry*, 2013.

Dietmar Todt. Social learning of vocal patterns and modes of their application in grey parrots (psittacus erithacus) 1, 2, 3. *Zeitschrift für Tierpsychologie*, 1975.

Jason A Tourville and Frank H Guenther. The diva model: A neural theory of speech acquisition and production. *Language and cognitive processes*, 2011.

Jason A Tourville, Kevin J Reilly, and Frank H Guenther. Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, 2008.

Beat Tschanz. *Trottellummen: die Entstehung der persönlichen Beziehungen zwischen Jungvogel und Eltern*. Parey, 1968.

Sherry Turkle. *Reclaiming conversation: The power of talk in a digital age*. Penguin, 2016.

Sherry Turkle. *Alone together: Why we expect more from technology and less from each other*. Hachette UK, 2017.

Lorraine K Tyler et al. Left inferior frontal cortex and syntax: function, structure and behaviour in patients with left hemisphere damage. *Brain*, 2011.

K Uetake et al. Effect of music on voluntary approach of dairy cows to an automatic milking system. *Applied animal behaviour science*, 1997.

Bob Uttl et al. Sampling inner speech using text messaging. *Proceedings of the Canadian Society for Brain, Behavior, and Cognitive Science*, 2012.

Vesa Välimäki et al. Digital audio antiquing-signal processing methods for imitating the sound quality of historical recordings. *Journal of the Audio Engineering Society*, 2008.

John Joseph Valletta et al. Applications of machine learning in animal behaviour studies. *Animal Behaviour*, 2017.

Daniel T van Bel et al. Social connectedness: concept and measurement. *Intelligent Environments*, 2009.

Akito van Troyer. Constellation: A tool for creative dialog between audience and composer.

Kleinberger Rebecca Van Troyer Akito. Miniduck webpage. URL http://web.media.mit.edu/~akito/Projects/MiniDuck/.

Emmett Velten Jr. A laboratory task for induction of mood states. *Behaviour research and therapy*, 1968.

Elaine N Videan et al. Effects of two types and two genre of music on social behavior in captive chimpanzees (pan troglodytes). *Journal of the American Association for Laboratory Animal Science*, 2007.

Ruvanee P Vilhauer. Inner reading voices: An overlooked form of inner speech. *Psychosis*, 2016.

MA Vince. Embryonic communication, respiration and the synchronization of hatching. *Avian Incubation, Behavior, Environment, and Evolution*, pages 88–99, 1969.

Nicole Vogelzangs et al. Urinary cortisol and six-year risk of all-cause and cardiovascular mortality. *The Journal of Clinical Endocrinology & Metabolism*, 2010.

Vokaturi. Vokaturi. URL https://developers.vokaturi.com/getting-started/overview.

Jakob Von Uexküll. A stroll through the worlds of animals and men: A picture book of invisible worlds. *Semiotica*, 1992.

Valentine Vox. *I can see your lips moving: The history and art of ventriloquism*. Kaye & Ward, 1981.

Lisa J Wallis et al. Utilising dog-computer interactions to provide mental stimulation in dogs especially during ageing. In *Animal-Computer Interaction*. ACM, 2017.

Samantha J Ward et al. Keeper-animal interactions: Differences between the behaviour of zoo animals affect stockmanship. *PloS one*, 2015.

Bronnie Ware. *The top five regrets of the dying: A life transformed by the dearly departing*. Hay House, Inc, 2012.

Richard M Warren. Perceptual restoration of missing speech sounds. *Science*, 1970.

Richard M Warren and Richard L Gregory. An auditory analogue of the visual reversible figure. *The American journal of psychology*, 1958.

David Watson et al. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 1988.

Sarah Webber et al. Hci goes to the zoo:[workshop proposal]. In *CHI*. ACM, 2016.

Sarah Webber et al. Kinecting with orangutans: Zoo visitors' empathetic responses to animals? use of interactive technology. In *CHI Conference on Human Factors in Computing Systems*. ACM, 2017.

David J. Weeks. A review of loneliness concepts, with particular reference to old age. *International Journal of Geriatric Psychiatry*, 1994.

Deborah L Wells. The effects of animals on human health and well-being. *Journal of Social Issues*, 2009.

Deborah L Wells et al. Auditory stimulation as enrichment for zoo-housed asian elephants (elephas maximus). *Animal Welfare*, 2008.

G Clare Wenger et al. Social isolation and loneliness in old age: review and model refinement. *Ageing & Society*, 1996.

Michelle Westerlaken and Stefano Gualeni. Felino: The philosophical practice of making an interspecies videogame. In *The Philosophy of Computer Games*, 2014.

Rainer Westermann, Kordelia Spies, Günter Stahl, and Friedrich W Hesse. Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of social psychology*, 1996.

Joseph Wherton et al. Designing technologies for social connection with older people. *Aging and the Digital Life Course*, 2015.

Matthew Wijers et al. Listening to lions: animal-borne acoustic sensors improve bio-logger calibration and behaviour classification performance. *Frontiers in Ecology and Evolution*, 2018.

Isabelle Williams et al. The effect of auditory enrichment, rearing method and social environment on the behavior of zoo-housed psittacines (aves: Psittaciformes); implications for welfare. *Applied Animal Behaviour Science*, 2017.

Robert S Wilson et al. Loneliness and risk of alzheimer disease. *Archives of general psychiatry*, 2007.

E Otha Wingo. *Latin punctuation in the classical age*, volume 133. Walter de Gruyter, 2011.

D M Wisniewska et al. High rates of vessel noise disrupt foraging in wild harbour porpoises (phocoena phocoena). *Proceedings of the Royal Society B: Biological Sciences*, 2018.

Frederic G Worden. Auditory habituation. In *Physiological Substrates*. Elsevier, 1973.

Peter Wright et al. Aesthetics and experience-centered design. *ACM Transactions on Computer-Human Interaction*, 2008.

Xiao Yan et al. Acoustic recordings provide detailed information regarding the behavior of cryptic wildlife to support translocations. *Scientific reports*, 2019.

Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2012.

Ying Yang et al. Detecting depression severity from vocal prosody. *Transactions on Affective Computing*, 2013.

Zhisong Yang et al. Reintroduction of the giant panda into the wild: A good start suggests a bright future. *Biological Conservation*, 2018.

Robert Yanofsky et al. Changes in general behavior of two mandrills (papio sphinx) concomitant with behavioral testing in the zoo. *The Psychological Record*, 1978.

Bo Yao et al. Silent reading of direct versus indirect speech activates voice-selective areas in the auditory cortex. *Journal of Cognitive Neuroscience*, 2011.

Aubrey J Yates. Delayed auditory feedback. *Psychological bulletin*, 1963.

Robert J Zatorre, Pascal Belin, and Virginia B Penhune. Structure and function of auditory cortex: music and speech. *Trends in cognitive sciences*, 6(1):37–46, 2002.

GQ Zhang et al. A method for encouraging maternal care in the giant panda. *Zoo Biology: Published in affiliation with the American Zoo and Aquarium Association*, 2000.

Jindong Zhang et al. Activity patterns of the giant panda (ailuropoda melanoleuca). *Journal of Mammalogy*, 2015.

Zhaoyan Zhang. Mechanics of human voice production and control. *The journal of the acoustical society of america*, 2016.

Zane Z Zheng, Ewen N MacDonald, Kevin G Munhall, and Ingrid S Johnsrude. Perceiving a stranger's voice as being one's own: A 'rubber voice'illusion? *PloS one*, 2011.

Xiaojian Zhu et al. The reproductive strategy of giant pandas (ailuropoda melanoleuca): infant growth and development and mother–infant relationships. *Journal of Zoology*, 2001.

Zoolingua. Zoolingua, 2018. URL http://zoolingua.com/.