ACOUSTIC COMMUNICATION:
AN INTERDISCIPLINARY APPROACH

Organized by
Emma Otta & Patrícia Ferreira Monticelli
Universidade de São Paulo
Pró-Reitoria de Pesquisa da USP

DOI: 10.11606/9786587596198

Organizing Committee of the ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH

Profa. Dra. Emma Otta (IPUSP)

Profa. Dra. Patrícia Ferreira Monticelli (FFCLRP)

Prof. Dr. Claudio Possani (IME-USP)

Dra. Tania Kiehl Lucci (IPUSP)

Dr. Ricardo Prist (IPUSP)

Cover Photo: Regina Macedo

Book formatting: Dra. Lilian Cristina Luchesi and Dra Aline Domingues Carneiro Gasco

English Editing Services: Michael Germain and Lisa Burger, MC TRADUÇÕES S/S LTDA

Catalogação na publicação
Serviço de Biblioteca e Documentação
Instituto de Psicologia da Universidade de São Paulo

Acoustic communication: an interdisciplinary approach / Organized por Emma Otta e Patrícia Ferreira Monticelli. -- São Paulo, Instituto de Psicologia da Universidade de São Paulo, 2021.

210 p.

E-book.

ISBN: 978-65-87596-19-8

DOI: 10.11606/9786587596198

1. Animal vocalization 2. Animal communication 3. Ethology I. Title

QL765

Ficha elaborada por: Elaine Cristina Domingues CRB5984/08

Funding Acknowledgements

Contents

Dedication

This book is dedicated to the memory of Professor César Ades who passed away on March 14, 2012. César was Professor of the Postgraduate Program of Experimental Psychology at the Institute of Psychology, University of São Paulo. He started the study of sound communication as part of animal behavior, from the perspective of ethology at the Department of Experimental Psychology. An analysis based on the Fonoteca Cesar Ades (FOCA) is presented in chapter 2. The author, Patricia Monticelli, did her master's and doctorate under his supervision studying the vocal repertoire of *Cavia aperea* and *Cavia porcellus*. The book is also dedicated to the memory of Edila Aparecida de Souza who worked for 23 years in the Ethology Lab. She was a motivated professional and gave her best to our University. Edila was one of thousands of people who have died from coronavirus. She passed away on June 3rd, 2020, at age 62. We remember her positive outlook on life, spontaneity, and willingness to work for a common goal and will continue working with this same attitude, in the face of enormous challenges. The Covid-19 pandemic has turned the world upside down. Vaccines are in development thanks to the efforts of scientists around the world. Science gives us hope for the future in our turbulent world.

Emma Otta and Patrícia Ferreira Monticelli

Acknowledgments

The organizers of ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH are grateful to Professor Sylvio Canuto, Dean of Research at the University of São Paulo for supporting the online event, the starting point for the book. We would also like to express our sincerest gratitude to the Organizing Committee, Professor Claudio Possani, Ricardo Prist, Ph.D., Tania Kiehl Lucci, Ph.D., and doctoral students Vinicius Frayze David and Bruna Campos Paula, in addition to all the contributors. Our colleague Regina Macedo, from the University of Brasília, painted *All the Voice*s [*Todas as Vozes*], which we also call *Brazilian Polyphony* [Polifonia Brasiliana] and became the cover of our book. At the closing session of the online session, Professor Regis Rossi Faria, colleague from USP, delighted the participants with SAPOS (2014, 6'30"), a guided auditory tour to an amphibious environment in concert, performed live. We are thankful for this ending experience. We thank the IPUSP Publishing Center [Núcleo de Publicações do Instituto de Psicologia da USP] for their help in editing this book.

About the Contributors

Aline D. Carneiro Gasco

Ph.D. in Science, honorary research fellow of the Ethology and Bioacoustics Laboratory (EBAC), Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Arnaldo Candido-Jr

Professor in the Department of Computer Science and Computational Mathematics, São Carlos Institute of Mathematical and Computer Sciences (ICMC), São Carlos, SP, Brazil.

Bruna Campos Paula

Ph.D. student at the Laboratory of Acoustics and Environment, Department of Mechanical Engineering, Polytechnic School of the University of São Paulo, São Paulo. Collaborating researcher of the Ethology and Bioacoustics Laboratory (EBAC), FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Bruna Lima Ferreira

Master Dissertation student of the Ethology and Bioacoustic Lab. (EBAC). Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Claudio Possani

Senior Professor at the Department of Mathematics, Mathematics and Statistics Institute (IME) of the University of São Paulo, São Paulo, SP, Brazil.

Daniel Bowling

Ph.D. Instructor at the Social Neurosciences Research Program, Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, CA, USA.

Domingos Hiroshi Tsuji

Professor at the Department of Otolaryngology of the Medical School Clinical Hospital (FMUSP/HC) of the University of São Paulo, São Paulo, SP, Brazil.

Edresson Casanova

Ph.D. student at the Department of Computer Science and Computational Mathematics, São Carlos Institute of Mathematical and Computer Sciences (ICMC), São Carlos, SP, Brazil.

Elodie Floriane Mandel-Briefer

Professor at the Behavioural Ecology group, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

Emma Otta

Professor at the Department of Experimental Psychology, Institute of Psychology (IPUSP), of the University of São Paulo, São Paulo, SP, Brazil.

Gabriel Francescoli

Professor at the Ethology Section, Faculty of Sciences, University of the Republic, Montevideo, Uruguay.

Lilian Cristina Luchesi

Ph.D. in Science, honorary research fellow of the Ethology and Bioacoustics Laboratory (EBAC), Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil.

Marcelo Finger

Professor at the Department of Computer Science, Mathematics and Statistics Institute (IME) of the University of São Paulo, São Paulo, SP, Brazil.

Maria Luisa da Silva

Professor of the Institute of Biological Sciences (ICB), Laboratório de Ornitologia e Bioacústica. Federal University of Pará (UFPA), Belém, PA, Brazil.

Patrícia Ferreira Monticelli

Professor at the Department of Psychology (FFCLRP), head of the Ethology and Bioacoustics Laboratory (EBAC), University of São Paulo, Ribeirão Preto, SP, Brazil.

Paulo do Canto Hubert Junior

Instructor at the School of Business Administration, Fundação Getúlio Vargas (FGV), São Paulo, SP, Brazil

Plínio Almeida Barbosa

Professor at the Department of Linguistics, Institute for Language Studies, University of Campinas, Campinas, SP, Brazil.

Regis Rossi Faria

Professor at the School of Arts, Sciences and Humanities (EACH), of the University of São Paulo, São Paulo, SP, Brazil.

Rogério Grassetto Teixeira da Cunha

Professor at the Institute of Natural Sciences (ICN), Federal University of Alfenas (UNIFAL), Alfenas, MG, Brazil.

Silvia Orlandi

Post-Doctoral Fellow at the Bloorview Research Institute, Holland Bloorview Kids Rehabilitation Hospital, Toronto, Canada.

Sylvia Corte

Professor at the Ethology Section, Faculty of Sciences, University of the Republic, Montevideo, Uruguay.

Vinicius Frayze David

Researcher at the Department of Experimental Psychology, Institute of Psychology (IPUSP), of the University of São Paulo, São Paulo, SP, Brazil.

About the book

The chapters in this eBook provide an overview of the scientific topics discussed at the online scientific meeting that inspired the book. The meeting *Acoustic Communication: An Interdisciplinary Approach* took place online on November 19-20, 2020, with the support of the Dean's Office for Research of the University of São Paulo, Brazil. We were experiencing a Covid-19 pandemic declared by the World Health Organization on March 11, 2020. Quarantine was declared in the State of São Paulo on March 13, 2020. In this context, we stress the importance of creating and preserving opportunities for the communication and exchange of ideas and research experiences among researchers, in addition to the University's initiative in fostering new ways of holding scientific meetings.

The meeting and the book were jointly organized by us, Professor Emma Otta, from the Department of Experimental Psychology of the University of São Paulo's Institute of Psychology (IPUSP), and Professor Patrícia Ferreira Monticelli, from the University of São Paulo's Department of Psychology at the Ribeirão Preto School of Philosophy, Science and Languages (FFCLRP-USP), as an extension of our research collaboration. Professor Monticelli coordinates the Laboratory of Ethology and Bioacoustics where research is carried out on reproductive, parental and communication behavioral aspects in terrestrial mammals. Professor Otta coordinates the Laboratory of Psychoethology, where research projects on Human Ethology are conducted. While studying nonverbal communication, she became interested in paralanguage, the non-verbal dimension of speech that contributes to its emotional quality.

During our collaborative research we noticed the need to consult specialists, given the interdisciplinary nature of the research topics under investigation. We systematize this experience here and share it with the readers through chapters that present the innovative research discussed in the talks and subsequent discussions, in the form of Peer Comments or Q&A transcription, prepared by the moderators of the presentations. We have divided the book into four parts: Animal Bioacoustics, Human Bioacoustics, Methods used in Bioacoustical Research and Analysis used in Bioacoustical Research.

Emma Otta & Patricia Ferreira Monticelli

Foreword

This book covers a fascinating topic: bioacoustics. After the course of events that guided me from an electronic engineer to a speech scientist, I thought my early dream of becoming a zoologist was over. This dream came true when I first met Patricia Monticelli and her lovely EBAC students back in 2016. Patricia's collaborations with psychologist Emma Otta and other colleagues were very fruitful, culminating recently in the Acoustic Communication: An Interdisciplinary Approach workshop. The book I have the honor of introducing is an outcome of this memorable event.

In fourteen chapters, the 22 contributors offer not only an inherently interdisciplinary approach to bioacoustics but give examples of several decades of scientific research developed for (human) speech. This is presented in the four parts: Animal Bioacoustics, Human Bioacoustics, Methods used in Bioacoustic Research, and Analysis used in Bioacoustic Research.

Part A opens this volume with a presentation of Bioacoustics as a subfield of Animal Communication, the latter a subfield of Ethology. Communication through sound is shown to be pervasive in both humans and non-humans, characterizing a social behavior that is crucial for each species. This is demonstrated when the authors investigate different intra- and inter-species behavior in primates, birds, guinea pigs, domestic and wild pigs, and domestic and wild horses, related or not to human interaction.

Part B describes what is known about speech production since Gunnar Fant’s work on Source-Filter theory, as well as what the study of the prosody of emotions in both verbal and non-verbal behaviors can offer to Bioacoustics, including the possibility of emotion recognition.

Part C presents software and algorithms developed for acoustic analysis in both human and non-human species, including infant cries. Praat, R, and BioVoice were used by the contributors of Part C, including presentations and examples of Machine Learning techniques for recognizing differences across bird species, different infant needs from their cries, and event detection.

Part D closes the book by presenting acoustic analysis applications to highlight the commonalities and differences in twins' speech, identify breathing issues in the case of voice disorders and build devices for speech synthesis and speaker identification, which is relevant for Forensic Phonetics in Speaker Comparison.

The researchers from Brazil, Canada, Denmark, and Uruguay that contributed to this wonderful book are prominent figures in the area of human and non-human sound communication. One major advantage of the 14 chapters is that they are written in a language that can be understood by both experts and people new to the area.

This book will be an important tool not only for students of Biology but also those in areas such as Computer Science, Electronic Engineering (including Telecommunications), Linguistics (including Phonetics), and Psychology. Experts from the same disciplines will also find a valuable resource for deepening their understanding of communication in all its shades and meanings by opening a window to a world where social networks must include non-human social networks aimed at a time where harmonic co-existence between species and nature will be a reality.

Plínio A. Barbosa
Department of Linguistics, University of Campinas
January 27th 2021

Chapter 1

Presenting Bioacoustics in Ethology

Gabriel Francescoli[1]

Abstract

Bioacoustic contributions to science are many and can be interpreted within the framework of several disciplines and subdisciplines from Biology to Social Sciences. The influence of Bioacoustics in Ethology, Biosemiotics, Physiology, Neurosciences, etc. and vice versa has modeled (and continues to model) in many ways the modern study of Animal Communication. Here, I will give a brief overview of these relationships and some of the research topics classically and currently addressed by researchers in these fields, underscoring their importance to Animal Behavior studies.

Keywords: Behavior, Bioacoustics, Ethology, research trends.

Bioacoustics is a cross-disciplinary science that combines biology and acoustics, usually referring to the investigation of sound production, transmission, and reception in animals (including humans). From this definition, we can deduce that Bioacoustics deals with the means of sound production and reception by animals, in a combination of sorts of sensory ecology and physics. However, from my point of view, Bioacoustics is much more than that because bioacoustic studies and tools influence many disciplines and subdisciplines, in addition to Biology and the Social Sciences (Figure 1.1).

In this paper, I will highlight many of these connections and mutual influences (obviously not all of them because there will always be new perspectives and influences I may not be aware of) to illustrate why I think Bioacoustics *per se* is not investigated by many researchers nowadays. As far as I know, this is because most Bioacoustic research is ultimately aimed at understanding other types of problems that are "in the orbit" of other disciplines and subdisciplines of Biology, Physics, Social Sciences, and even technology.

[1] Sección Etología, Facultad de Ciencias, Universidad de la República Montevideo, Uruguay. gabo@fcien.edu.uy

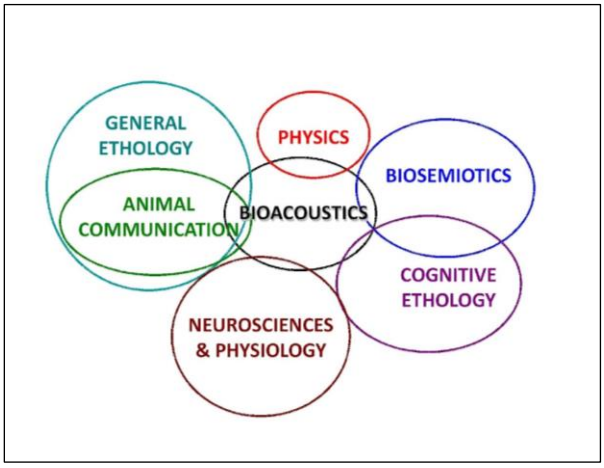


Figure 1.1. Diagram representing the relationships between Bioacoustics and other scientific fields, as discussed in this paper.

Mutual influences between Bioacoustics and Ethology

As mentioned before, Bioacoustics is the study of biological sounds from a general perspective, implying *any* biological sounds, irrespective of their use. This discipline involves studies dealing with animal communication problems, and other uses of biological sound production, such as sonar systems employed by many animals in traveling, hunting, etc. Most studies on the biological sounds emitted by animals are related to Ethology or Animal Behavior.

In Ethology, the main body of knowledge to which bioacoustic research results are linked is the subfield of Animal Communication. Communication studies in Ethology have a long tradition, sound communication being a very important subfield, and probably one of the oldest because of the interest in biological sounds among naturalists and researchers (Thorpe, 1979).

Sound signals are interesting and important to study because they are an extended means of sending messages and exchanging information between living beings. Sound signals are sometimes the best way to accomplish this, always depending on environmental characteristics, mainly due to their propagation capabilities and because sound signals (depending on their production mechanisms) are capable of quick and variable modulation, allowing the generation of adaptable and variable signals.

Initial efforts to understand and record sound signals (mostly bird songs) used musical notation and onomatopoeia, because of the lack of other recording media. Nevertheless, when the first recording apparatus appeared, most studies investigated stored sounds from physical and behavioral perspectives, using the best technologies available at

the time. The second step towards biological sound analysis linked to behavior was the adaptation of a device used to analyze human voice and language for general use in animal bioacoustics: sonography. These two steps resulted in what bioacoustics is today.

If we analyze the role played by acoustic signals and biological sounds in ethological studies, we can see that many subjects of general ethology can be explained and interpreted using communication studies and biological sounds. This is why I consider that animal communication, as a "sub-discipline" of Ethology, plays a fundamental role in the development of the discipline, and is perhaps the most relevant part of the relationship between Ethology and Bioacoustics.

Animal communication studies (and bioacoustics, for that matter) are important in Ethology because a significant part of behavioral studies dealing with social interactions as a broad concept needs a communication system that allows individuals to coordinate activities. This is a major point in ethological studies and can be confirmed by the fact that you can devise a course in general Ethology and explain a majority of ethological concepts using mainly communication examples and points of view.

Bioacoustics has contributed to Animal Communication studies in many ways, from the earlier descriptive research that showed the basic characteristics of biological sounds (Marler, 1977), to other more in-depth studies about signal design and their particular characteristics for certain uses (Bradbury & Vehrencamp, 2011). Later developments led researchers to study and interpret sound signals as part of repertoires or "languages," and the use of limited repertoires of sounds to generate signals conveying different messages, such as different signals for different receivers (Marler et al., 1986); differential repertoires for different sexes, ages, dominance status sub-groups (Bradbury & Vehrencamp, 2011); rhythm, tempo and duration of acoustic signals to communicate information (Francescoli, 2011); call combinations and compositional mechanisms generating different meanings from a restricted signal repertoire (Engesser et al., 2016); and different signals and combinations for predator identification and warning involving syntactic and semantic contents in animal signals (Seyfarth et al., 1980; Suzuki, 2013; Zuberbühler, 2018). These studies also connect general Ethology and Animal Communication with other disciplines and subdisciplines with which they share interests, research, and results.

Other related and interacting disciplines

One of the disciplines linked to these studies is Physics, because of our need to understand rules of sound production and reception that may act on animal sounds, limiting

their use and characteristics, such as performance constraints on the physical characteristics of emissions (Podos, 1997); "honesty" rules in vertebrate sound production (Fitch & Hauser, 2003); filters and sources of sound as a determinant of vocal signal characteristics (Taylor & Reby, 2010); and low-frequency enhanced audition related to bullar morphology in subterranean and desert rodents (Francescoli et al., 2012). These are very basic studies in terms of sound characteristics and hearing, but they contribute to understanding basic rules, identifying other more complex rules related to meaning and the use of sound signals. The aforementioned goals are common to other disciplines related to Ethology, Animal Communication and Bioacoustics, such as Cognitive Ethology, Biosemiotics, and Neurosciences.

Communication systems in any channel, and specifically bioacoustic signals, are potential gateways to understanding the cognitive abilities and capacities of many animals. This could be achieved not only through a complete understanding of some species' "language," which would probably allow direct communication between members of that species and us (there are some examples of these kinds of interactions, albeit very limited, such as the communicative and cognitive studies on parrots conducted by Pepperberg, 2006). Nonetheless, a number of studies have used sound signals in sophisticated experiments that provide insight into the mental capabilities of species such as different types of dog barks in different situations and towards different individuals (Fischer et al., 2001), and intentional vocalizations to attract the attention of others in dangerous situations (Crockford et al., 2015).

The link between stimulus, action, and neural "support" for cognitive and behavioral tasks relates Bioacoustics and Ethology to Neurosciences and Physiology because these disciplines allow in-depth analysis of some of the morpho-physiological bases of these behaviors. Some studies link human neural and hormonal systems to the regulation and performance of their animal counterparts while communicating, and determine some of the ways in which the relations between stimuli (external and internal) and responses can modulate communicative acts and adapt them to ongoing physical conditions. Another point regarding this relationship is memory studies, since they help understand some of the capabilities of communication systems and the implied social relationships.

The Biosemiotic approach, on the other hand, aims to understand the characteristics of animal (biotic) signals that allow subjects to generate meaningful utterances that can influence and even mislead receptors, because of the mental processes involved in generating signal meaning content and the repertoires used, such as deception and signal

misuse in fireflies (El-Hani et al., 2010); triadic relationships in behavior and signals (Francescoli, 2017); and agonism management using vocal signals in subterranean rodents (Francescoli & Schleich, 2018). Thus, what Biosemiotics attempts to understand is the generation of mental content or behavioral acts that allow the encoding and decoding of information into communication signals (sounds, in our case), and the processing and interpretation of external and internal information allowing a biological agent to interpret the world around it, interact with it and make the proper decisions. This process produces different levels of behavior (signals to other individuals) that are also the subject of ethological studies (von Uexküll, 1926; ethological sequence analysis as narrative analysis, Francescoli, 2019).

In my opinion, all of these disciplines and research fields combine different viewpoints and study methods that help us better understand Bioacoustics, animal communication in particular, and Animal Behavior (Ethology) in general.

References

Bradbury, J. W., & Vehrencamp, S. L. (2011). *Principles of animal communication*. (2nd ed.). London: Sinauer.

Crockford, C., Wittig, R. M., & Zuberbühler, K. (2015). An intentional vocalization draws other's attention: a playback experiment with wild chimpanzees. *Animal Cognition, 18*, 581-591.

El-Hani, C. N., Queiroz, J., & Stjernfelt, F. (2010) Firefly femmes fatales: A case study in the semiotics of deception. *Biosemiotics, 3*, 33-55.

Engesser, S., Ridley, A. R., & Townsend, S. W. (2016). Meaningful call combinations and compositional processing in the southern pied babbler. *PNAS, 113*, 5976-5981.

Fischer, J., Metz, M. Cheney, D. L., & Seyfarth, R. M. (2001). Baboon responses to graded barks. *Animal Behaviour*, 61, 925-931.

Fitch, W. T., & Hauser, M. D. (2003). Unpacking "honesty": vertebrate vocal production and the evolution of acoustic signals. *Acoustic communication, 16*, 65-137.

Francescoli, G. (2011). Tuco-tucos' vocalization output varies seasonally (*Ctenomys pearsoni*; Rodentia, Ctenomyidae): implications for reproductive signaling. *Acta ethologica, 14*, 1-6.

Francescoli, G. (2017). A semiotic interpretation of the Innate Releasing Mechanism concept and other ethological triadic relations. *Biosemiotics*, 10,461-468. https://doi.org/10.1007/s12304-017-9306-7.

Francescoli, G. (2019). "Evolutionary stories": narratives as evolutionary tools to describe and analyze animal behaviour and animal signals" In: Silvera-Roig, M. & López-Varela, A. (Eds) *"Cognitive and Intermedial Semiotics" (*chapter doi 10.5772/intechopen.89209). IntechOpen, London.

Francescoli, G., & Schleich, C. (2018). Agonism management through agonistic vocal signaling in subterranean rodents: A neglected factor facilitating sociality? *Biological Theory* doi 10.1007/s13752-018-0304-z

Francescoli, G., Quirici, V., & Sobrero, R. (2012). Patterns of variation in the tympanic bullae of tuco-tucos (Rodentia, Ctenomyidae, Ctenomys). *Acta Theriologica*, *57*, 153–163.

Marler, P. (1977). *The structure of animal communication sounds. Recognition of complex acoustic signals: report of Dahlem workshop*. Berlin: AbakonVerlagsgesellschaft.

Marler, P., Dufty, A., & Pickert, R. (1986). Vocal communication in the domestic chicken: II. Is a sender sensitive to the presence and nature of a receiver? *Animal Behaviour*, *34*, 194-198.

Pepperberg, I. M. (2006). Cognitive and communicative abilities of Grey parrots. *Applied Animal Behaviour Science, 100*, 77-86.

Podos, J. (1997). A performance constraint in the evolution of trilled vocalizations in a songbird family (Passeriformes: Emberizidae). *Evolution*, *51*, 537-551.

Seyfarth R.M., & Cheney, D.L. (2016). The origin of meaning in animal signals. *Animal Behaviour*, doi 10.1016/j.anbehav.2016.05.020

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour*, *28*, 1070-1094.

Suzuki, T. N. (2013). Communication about predator type by a bird using discrete, graded and combinatorial variation in alarm calls. *Animal Behaviour, 87*, 59-65.

Taylor, A. M., & Reby, D. (2010). The contribution of source-filter theory to mammal vocal communication research. *Journal of Zoology*, *280*, 221-236.

Thorpe, W. H. (1979). *Breve historia de la Etología.* Madrid: Alianza.

von Uexküll J. (1926). *Theoretical biology*. New York: Harcourt, Brace.

Zuberbühler, K. (2018). Combinatorial capacities in primates. *Current Opinion in Behavioral Sciences*, *21*, 138-144.

Peer Commentary

Bioacoustics as an interdisciplinary approach and the duality seismic-vocal signaling

By Lilian Cristina Luchesi

In this first chapter, Gabriel Francescoli discussed the insertion of Bioacoustics in Ethology. He shared some perspectives on the interdisciplinary approach and raised some issues to ponder in the Bioacoustics research area. In this section, I will comment on some of the aspects discussed by Francescoli, ending with the use of seismic signaling by rodents.

Not all zoologists and ethologists studying acoustical species practice Bioacoustics, and others focus only on the acoustic part of the signals themselves. In Francescoli's viewpoint, researchers normally go beyond acoustic signal analysis to other disciplines more connected with the behavioral and biosemiotics aspects of scientific studies. Some are only interested in the signals per se, mainly in the physics portion of the signal. Others are interested in acoustic signals but not interested in or able to cope with more advanced physics in analyzing acoustic signals per se. The challenge may be developing physics skills to deal with some of the problems with the signal itself. Professor Francescoli faces some of these problems in his approach once he is more interested in the communication part of animal behavior than the physical properties of signal transmission and reception. From what I see, it helps to establish partnerships with other parts of communication science, namely the physical properties of signals and their dispersion, to improve our overall knowledge of Bioacoustics.

Many books are available for those who might want to understand bioacoustics' interdisciplinarity and its approaches better. One of these is the general acoustic handbook edited by Thomas Rossing (Springer Handbook of Acoustics, 2014) and the book by Bradbury and Vehrencamp (2011) (Principles of Animal Communication), which focuses more on animal communication science. Some manuals describe vocal anatomy, sound production, and acoustic analysis for human speech, such as Ball and Rahilly (2000) and Harrington and Cassidy (1999). Brazilian professors Plínio Barbosa and Sandra Madureira (2015) published a Phonetic Acoustic Manual describing theory and experimental procedures used by Brazilian and European speakers. There are also many more manuals in different languages to better study human voice production, speaking, and singing.

The Bioacoustics referentiality debate of animal communication and the opposing views of functional referentiality and automatic, emotional response of vocal signals

In Francescoli's point of view, referentiality is a significant problem, and practical referentiality is the concept we are mainly striving for in our studies because it is the most widely accepted in the scientific community. However, referentiality as a whole acts through mental mechanisms whose existence has yet to be proven, and that should be our main goal. Many animal species exhibit referentiality in which mental content plays an important role. Still, how can we access this mental content to demonstrate that it does exist scientifically? We have to deal with this type of duality, in that many researchers firmly believe that there is mental referentiality in several higher vertebrate species. Nevertheless, for now, we must remain with the concept of practical referentiality because we cannot enter these animals' minds to show what occurs scientifically. That is the ultimate problem with these types of interactions.

I bring some elements to illustrate this debate and the ultimate problem highlighted by Francescoli. First, we could return to the starting point when the first known experimental observation of animal signaling was made by Darwin (1871) of the *Cercopithecus* vocal response to a snake (Townsend & Manser, 2013). After that, the pursuit of the signaling–meaning relation grew. Researchers seek to answer the evolutionary question: are animal communication signals full of meaning, or are they just automatic responses to constraints? Only 150 years later, Struhsaker (1967) described different alarm calls among vervet monkeys (*Cercopithecus aethiops*) to various predation risks. Nearly thirteen years later, Seyfarth et al. (1980a, 1980b) described three distinct signals for different classes of predators, including leopards, eagles, and snakes, which evoked different responses such as running up trees, looking up or running into dense bush, and looking at the ground around them, irrespective of the context in which the signals occurred. Furthermore, this association between predator class and type of alarm call appeared to improve from the infant to adult stages of a primate's life (Seyfarth et al., 1980b). These findings and their repercussions on animal communications studies were documented in a special issue devoted to the forty years of Seyfarth, Cheney's, and Marler's work: Communication in Nonhumans: The Fortieth Anniversary of Seyfarth, Cheney, and Marler (ABC, 7(2), 2020) presenting some of their research and the continued debate (Vonk, 2020).

Having different alarm calls for distinct classes of predators is not an exclusive trait of primates. Among rodents, diverse alarm calls for different classes of predators or differential urgency flight responses also evoked different responses in Gunnison's prairie dogs, *Cynomys gunnisoni* (Kiriazis & Slobodchikoff, 2006), ground squirrels (Owings & Leger, 1980), and Belding's ground squirrels, *Spermophilus beldingi*, (Robinson, 1980). Only six bird species exhibit functionally referential alarm calls (see Gill & Bierema, 2013 for a review). Several studies support the evidence of referential signals for different constraints evoking different responses. However, this theory does not support some signals or encounter obstacles in establishing a relation between signal and response and could be better classified as an emotional response (Rendall et al., 2009). For this reason, they could be redefined and referred to as signals that influence other actions rather than signals transmitting encoded information (Owren et al., 2010).

The dichotomy between the information or the emotional signals contained in animal alarm calls is false. Nowadays, it is widely accepted that both components (cognitive and emotional) are impossible to separate (Snowdon, 2020). As Francescolli said, the problem is how to access mental content using the scientific method. The search for a better understanding of the information in signals and how it can affect the conspecifics' behavior should be encouraged in future research (McRae, 2020).

Discussing the duality in the use of vocal or seismic signals; the transmission of the seismic signal, and its relation to different substrates

To Francescoli, when considering the seismic signaling transmission, the problem is that any solid substrate can be used to produce seismic sounds or signals. The problem is that the range over which signals can travel to other animals or the velocity of sound transmission depends on the composition of the substrate. Sand, highly or loosely compacted soil, or even rocks, are not the same because these seismic signals are very similar to those produced by some seismic events in nature, and their propagation characteristics are similar.

Concerning the seismic signal and vocal duality, the former are produced by tapping on the substrate with body parts such as limbs, teeth, or in many cases, the skull. In almost any case, they also produce sounds, but this sound probably does not propagate at long ranges as seismic signals do. There is also a problem with reception. Seismic signals may not be analyzed in a specialized brain but are still analyzed like any other sound. As such, are they seismic signals or sounds, depending on what part of the brain is analyzing the signal? Other researchers in the field of subterranean rodents, Francescoli's specialty,

proposed a neuro-specialized part of the animal's brain that analyzes these tapping patterns, thereby characterizing "real" seismic signals. Thus, to Francescoli, the problem has not yet been completely solved.

Adding information to what Francescoli said, many questions remain on how the substrate interferes with seismic signals. Seismic communication may have evolved to communicate to predators (Shelley and Blumstein, 2005). This signal modality is present in more than 320 species, distributed among nine major orders that use substrate-borne vibrations as an information source (Hill, 2008, 2009). Cocroft and Rodríguez (2005) estimated that over 195,000 taxa use vibrational signals alone or in combination with other mechanical signaling methods among insect species. There is some information about environmental interference in signaling transmission (Francescoli, 2017; Gordon & Uetz, 2012; Hill, 2008; see O'Connell-Rodwell, 2007 review), but the physical properties of seismic signaling are still poorly understood. As such, we return to the first point discussed here: Bioacoustics as an interdisciplinary research area.

The communication system of the tuco-tuco transmitting and identifying information

Francescoli says tuco-tucos may transmit information (age, sex) through vocal signals, and the recipients can identify it. Studying the *Ctenomys pearsoni* species, he found that female's long-range vocalizations (not all female tuco-tucos use long-range vocalizations) varied throughout the year. They are very short-range during the non-reproductive season, at the end of the reproductive season when they may be pregnant, and longer at the beginning of the reproductive season when females are "advertising" their condition. Males, on the other hand, vocalize long-distance calls all year long. Since males usually approach females by excavating a communication tunnel between their burrows (and do it apparently without significant errors in locating the female tunnels), avoiding high energy expenditure, Francescoli and colleagues postulate that long-range vocalizations make it possible to locate emitters in space inside the population. The length of the emission (and perhaps the repetition rate) may reveal the sex and reproductive condition of the emitter. These hypotheses are based on direct observation and radio-tracking of individuals in a population of more than 30 tuco-tucos during a two-year study. However, no playback experiments were conducted in the field to test them. Some of his yet unpublished results from laboratory tests show that females of different ages (estimated by their weight) reacted differently to the broadcast of male vocalizations during the reproductive period. The younger ones approach the sound source, and older ones display

ambiguous responses, depending on their reproductive condition. Still, all these captured females were probably older, more experienced individuals than their younger counterparts in their first reproductive period.

If we look at their auditory structures, their malleus in the middle ear is not enlarged and dense as far as we know. This anatomical specialization may be adapted to perceive substrate vibrations like in the golden mole. Enhanced low-frequency reception seems to work through different tympanic bulla inflation levels and internal partitioning. For more information on the topic, see Francescoli et al. (2012), where the subject is studied and discussed for many species of tuco-tucos belonging to almost all the genus distribution in South America.

In conclusion, there are many issues still waiting to be investigated on animal communication systems. I want to highlight that, specifically on rodent acoustic communication, Professor Francescoli wrote a review of the last three decades of subterranean acoustic communication, discussing some aspects of the tuco-tuco communication system (Schleich & Francescoli, 2018). This review is part of Rodent Bioacoustics (2018), covering several aspects of rodent communication systems. For example, different acoustic signals were observed between tuco-tuco males and females; females exhibited their signals (type C signals) in sexual encounters with *Ctenomys pearsoni* (Francescoli, 1999) and *C. talarum* (Schleich & Busch, 2002), which could be considered phenotypic plasticity (Francescoli, 2017). In addition, the naked mole-rat (*Heterocephalus glaber*) was recently included among mammals using acoustic communication to transmit social information about group membership through different colony dialects (Barker et al., 2021). This discovery shows how rich acoustic communication in nonhuman animals can be and the different paths taken within this field.

References

Ball, M. J., & Rahilly, J. (2000). *Phonetics: the science of speech*. Oxford, UK: Oxford University Press.

Barbosa, P. A., & Madureira, S. (2015). *Manual de fonética acústica experimental: aplicações a dados do português*. São Paulo: Cortez.

Barker, A. J., Veviurko, G., Bennett, N. C., Hart, D. W., Mograby, L., & Lewin, G. R. (2021). Cultural transmission of vocal dialect in the naked mole-rat. *Science, 371*(6528), 503–507. https://doi.org/10.1126/science.abc6588

Bradbury, J. W., & Vehrencamp, S. L. (2011). *Principles of animal communication* (2nd ed.). Sunderland: Sinauer Associates.

Cocroft, R. B., & Rodríguez, R. L. (2005). The behavioral ecology of insect vibrational communication. *BioScience*, *55*(4), 323–334. https://doi.org/10.1641/0006-3568(2005)055[0323:TBEOIV]2.0.CO;2

Darwin, C. (1871). *The descent of man, and selection in relation to sex* (1981 reprint). Princeton University Press.

Francescoli, G. (1999). A preliminary report on the acoustic communication in Uruguayan *Ctenomys* (Rodentia, Octodontidae): basic sound types. *Bioacoustics*, *10*(2–3), 203–218. https://doi.org/10.1080/09524622.1999.9753431

Francescoli, G. (2017). Environmental factors could constrain the use of long-range vocal signals in solitary tuco-tucos (*Ctenomys*; Rodentia, Ctenomyidae) reproduction. *Journal of Ecoacoustics*, *1*, R7YFP0. https://doi.org/10.22261/JEA.R7YFP0

Francescoli, G., Quirici, V., & Sobrero, R. (2012). Patterns of variation in the tympanic bulla of tuco-tucos (Rodentia, Ctenomyidae, *Ctenomys*). *Acta Theriologica*, *57*(2), 153–163. https://doi.org/10.1007/s13364-011-0064-7

Gill, S. A., & Bierema, A. M.-K. (2013). On the meaning of alarm calls: A review of functional reference in avian alarm calling. *Ethology*, *119*(6), 449–461. https://doi.org/10.1111/eth.12097

Gordon, S. D., & Uetz, G. W. (2012). Environmental interference: impact of acoustic noise on seismic communication and mating success. *Behavioral Ecology*, *23*(4), 707–714. https://doi.org/10.1093/beheco/ars016

Harrington, J., & Cassidy, S. (1999). *Techniques in Speech Acoustics* (Vol. 8). Netherlands: Springer. https://doi.org/10.1007/978-94-011-4657-9

Hill, P. S. M. (2008). *Vibrational communication in animals* (1st ed.). Cambridge, MA: Harvard University Press.

Hill, P. S. M. (2009). How do animals use substrate-borne vibrations as an information source? *Naturwissenschaften*, *96*(12), 1355–1371. https://doi.org/10.1007/s00114-009-0588-8

Kiriazis, J., & Slobodchikoff, C. N. (2006). Perceptual specificity in the alarm calls of Gunnison's prairie dogs. *Behavioural Processes*, *73*(1), 29–35. https://doi.org/10.1016/j.beproc.2006.01.015

McRae, T. R. (2020). A review of squirrel alarm-calling behavior: What we know and what we do not know about how predator attributes affect alarm calls. *Animal Behavior and Cognition*, *7*(2), 168–191. https://doi.org/10.26451/abc.07.02.11.2020

O'Connell-Rodwell, C. E. (2007). Keeping an "ear" to the ground: Seismic communication in elephants. *Physiology*, *22*(4), 287–294. https://doi.org/10.1152/physiol.00008.2007

Owings, D. H., & Leger, D. W. (1980). Chatter Vocalizations of California Ground Squirrels: Predator- and Social-role Specificity. *Zeitschrift Für Tierpsychologie*, *54*(2), 163–184. https://doi.org/10.1111/j.1439-0310.1980.tb01070.x

Owren, M. J., Rendall, D., & Ryan, M. J. (2010). Redefining animal signaling: influence versus information in communication. *Biology & Philosophy*, *25*(5), 755–780. https://doi.org/10.1007/s10539-010-9224-4

Rendall, D., Owren, M. J., & Ryan, M. J. (2009). What do animal signals mean? *Animal Behaviour*, *78*(2), 233–240. https://doi.org/10.1016/j.anbehav.2009.06.007

Robinson, S. R. (1980). Antipredator behaviour and predator recognition in Belding's

ground squirrels. *Animal Behaviour*, *28*(3), 840–852. https://doi.org/10.1016/S0003-3472(80)80144-8

Rossing, T. D. (2014). In T. D. (Ed.), *Springer Handbook of Acoustics*. New York: Springer. https://doi.org/10.1007/978-1-4939-0755-7

Shelley, E. L., & Blumstein, D. T. (2005). The evolution of vocal alarm communication in rodents. *Behavioral Ecology*, *16*(1), 169-177. https://doi.org/10.1093/beheco/arh148

Schleich, C., & Busch, C. (2002). Acoustic signals of a solitary subterranean rodent *Ctenomys talarum* (Rodentia: Ctenomyidae): Physical characteristics and behavioural correlates. *Journal of Ethology*, *20*(2), 123–131. https://doi.org/10.1007/s10164-002-0064-9

Schleich, C., & Francescoli, G. (2018). Three Decades of Subterranean Acoustic Communication Studies. In *Rodent Bioacoustics* (pp. 43–69). Cham: Springer. https://doi.org/10.1007/978-3-319-92495-3_3

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980a). Vervet monkey alarm calls: Semantic communication in a free-ranging primate. *Animal Behaviour*, *28*(4), 1070–1094. https://doi.org/10.1016/S0003-3472(80)80097-2

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980b). Monkey responses to three different alarm calls: evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803. https://doi.org/10.1126/science.7433999

Snowdon, C. T. (2020). Vervet monkey alarm calls: Setting the historical context. *Animal Behavior and Cognition*, *7*(2), 87–94. https://doi.org/10.26451/abc.07.02.02.2020

Struhsaker, T. T. (1967). Auditory Communication among Vervet Monkeys. In S. A. Altmann (Ed.), *Social Communication among Primates* (pp. 281–324). Chicago: University of Chicago Press.

Townsend, S. W., & Manser, M. B. (2013). Functionally Referential Communication in Mammals: The Past, Present and the Future. *Ethology*, *119*(1), 1–11. https://doi.org/10.1111/eth.12015

Vonk, J. (2020). Forty years on from the question of referential signals in nonhuman communication. *Animal Behavior and Cognition*, *7*(2), 82–86. https://doi.org/10.26451/abc.07.02.01.2020

Chapter 2

Exploring terrestrial mammals' acoustic communication as a web process

Patrícia Ferreira Monticelli[2]

Abstract

Social life involves a dynamic network of individuals of the same species interacting directly and indirectly. An audience that does not affect a dyadic interaction may use the exchange (not the signal alone) as an information source. Here, I take some examples of loud calls from the Emmons et al. audiobook of neotropical mammals as well as my own and explore their effects on a communication network. Beyond the emitter-recipient dyad, there may be others (third-parties) that can benefit from watching (or hearing) the dyad performance. Communicating through sound is energetically expensive and may pose a survival risk. Loud signaling may also benefit the emitter. In Caviinae rodents, for instance, female whining may attract other males' attention and challenge their performance, an opportunity for mate choice when the male offers no direct parental investment. In caviid societies, female whining is enough to promote mate choice, and broader home range females, such as the jaguar, must use loud calls to reach very distant mates.

Keywords: Animal behavior, Bioacoustics, crying, loud-call, web communication, whine.

In 2005, four years after the XXVII International Ethological Conference (Tübingen, Germany), a book edited by Peter McGregor containing articles by prominent authors in bioacoustics was published. Animal Communication Networks (ACN) resulted from the communication network symposium held at the annual event. Like previous books on animal communication, it described signaling as "one of the most conspicuous behaviors" in communication interactions, with consequences for reproduction and survival (McGregor, 2005, Introduction, p. 1). Unlike the others, however, the ACN book breaks with the traditional (or didactical) dyadic characterization of communication involving information transmission between an emitter and a recipient (Hauser, 1998). McGregor (2005) describes communication as "inherently social behavior," and should therefore be considered under a species' social life issues.

[2] Ethology and Bioacoustic Laboratory, Department of Psychology FFCLRP of the University of São Paulo, Brazil monticel@usp.br

Social life involves a dynamic network of individuals of the same species interacting directly or indirectly. An audience that does not affect a dyadic interaction may use the interaction (not the signal alone) as a source of cues or information. Using information for self-convenience (the operational level of selective forces in natural selection) constitutes eavesdropping behavior (McGregor, 1993). Eavesdropping here does not include the illegitimate receiver (Otte, 1974) or the signal interceptor (e.g., a bat locating an anuran by its advertisement call; Ryan, 1988).

In social eavesdropping (Peake, 2005; or merely eavesdropping, as proposed by Searcy & Nowicki, 2005), a third-party benefits from monitoring other members of the species, memorizing the participants' identities and performance, and making decisions accordingly (McGregor, 2005; Searcy & Nowicki, 2005). Passerines exhibit a wide range of eavesdropping behaviors, observed in nature or experimental paradigms (McGregor, 2005). For instance, a third-party male that has observed two other males interacting in a cage will be more likely to challenge the "loser" than the "winner" in later opportunities. In nature, a third-party receiver would benefit from invading the territory of the less aggressive singing male instead of one of its more aggressive neighbors (McGregor, 2005).

Loud calling in mammals from the audience perspective

Mammalian species communicate through acoustic signals that can travel long distances, expanding the concept of an audience. Whether in more local, territorial species or those in which individuals are dispersed over broad areas, in the so-called "fluid social system" (McComb & Reby, 2005), whenever an individual emits a loud call, any conspecific in the broadcast area is a potential receiver. Even softer calls can reach traveling individuals (third parties). For instance, if a maned-wolf reproductive couple calls each other after hunting in different territories (Kleiman, 1972; Dietz, 1984), what effect would it have on a third party that was just passing by, whether male or female? It could be advantageous for all of them, supposing that the third party is not looking for hunting territory or a sexual partner (Dietz, 1984). A young animal that might be injured if caught in a defended territory would also benefit by leaving silently. The couple's investment in producing the loud call would not negatively affect them, as observed by McComb and Reby when studying lions, elephants, and cervids (McComb & Reby, 2005). What if the traveling individual is a healthy male seeking hunting territory? The acoustic structure of the loud "extended-bark" (Dietz, 1984) of the resident male could inform the silent approaching male about its ability to defend its territory, benefiting the third-party male, and perhaps the resident male if it was advertising itself (using the communication network

to its advantage). This would characterize a "social eavesdropping" episode, as defined by McGregor (1993), with consequences for the individuals' reproductive success. If we ignore the fact that social eavesdropping behavior is a potential selective force over signal design and use, we may not formulate a robust hypothesis about signal evolution (Cheney & Seyfarth, 1999). Thus, I will start by presenting the data collected on my research interest, namely terrestrial neotropical mammals. Which of their calls should be analyzed from the audience's perspective?

Louise Emmons and colleagues compiled a sound library to serve as a field guide to rainforest mammals in association with Conservation International and published by the Cornell Laboratory of Ornithology (Emmons et al., 1997). They divided the collection of mammal calls into six broad categories, according to the recording context and the observed reactions, as follows: (1) disturbance or alarm calls that are usually sequences of snorts, barks, and chucks (or the seismic signal of foot-stomping), ranging from less intense to very loud calls; (2) loud or long calls, harsh species-specific roars, whistles or screams, are found in almost all 54 neotropical primates in their audio guide, and many of the non-primate species (foxes, dogs, raccoons, cats, tapirs, and spiny rats). The authors suggest that some of these were equivalent to bird songs, used in territorial defense (e.g., in spacing groups or individuals in natural areas; such as in howler monkeys, Kitchen et al., 2015) and reproduction (e.g., promoting contact between sexual partners such as in the *Puma concolor*, Potter, 2002, and the maned-wolf, *Chrysocyon brachyurus*, Brady, 1981). This includes species-specific medium-to-short-distance courtship calls, which are rarely observed in mammals (Magrini & Monticelli, 2012). (3) Defensive/aggressive or threat calls include non-vocal sounds, such as tooth grinding or rattling and growling or hissing-like signals. These may function to repel or intimidate a threatener and, unlike species-specific loud calls, are shared by very distant taxa, such as small birds and canids (Emmons et al., 1997). (4) Distress calls (agony cries) are soft or loud cries or screams uttered in extremely dangerous situations, such as when being caught by a predator or handled by a human, or after being injured in a dispute with a conspecific. (5) The social calls category is diverse and may be associated with mood (Eisenberg, 1974). It consists of variable squeals, whines, or grunts uttered during minor conflict interactions before injuries occur, during food or other resource disputes or close social contact (e.g., naso-anal contact between caviomorph rodents, Eisenberg, 1974; Barros et al., 2011; Verzola-Olivio & Monticelli, 2017; Alencar-Jr & Monticelli, *no prelo*; in Carnivora such as giant otters, tapirs, raccoons, among many others; Emmons et al., 1997; Gasco, Ferro & Monticelli, 2019; and the canids studied by Brady, 1981). Finally, (6) the isolation call is a medium-

intensity peep-like sequence of notes uttered by infants separated from their mother or in danger (e.g., in Caviidae: domestic and wild cavies, Monticelli et al., 2004; capybaras, Barros et al., 2011; and pacas, Lima et al., 2018; and in many primates and non-primate species, Emmons et al., 1997).

Under this proposed classification, mammals communicate at longer distances when disturbed or warned, separated from others (an infant from its mother, a mother from its infant, reproductive pairs from each other, or members from the rest of the group) or in danger. Some of these calls are presented in Figures 2.1 and 2.2, extracted from the Emmons et al. field guide and EBAC sound library collection (FOCA)[3].

Note the similarity in the acoustic structure of loud calls used in corresponding situations (Figure 2.1), comparing the calls produced by isolated infants of different species (first line in the figure) and the disturbance/alarm calls on the second line. There may be a difference between the communication function of infants calling their mother and conspecifics warning others about a detected risk, which may be reflected in the type of signal. In both cases, predators or competing individuals may constitute the third parties that also exert selective pressure on signaling behavior. A social eavesdropper would benefit by perceiving its species' alarm call. With respect to infant cries, the third-party may ignore the attack (in the case of an infanticide male with access to females when killing her young), or act in its indirect reproductive success (supposing genetic partnership and the possibility of helping the infant).

Terrestrial mammals uttered the loud infant calls presented in Figure 2.1, all inhabiting the Atlantic Forest biome. These calls cover a large frequency band and are composed of high-pitched units that are not exactly the same (notes that are not stereotyped may reflect internal state variations, Eisenberg, 1974). On the other hand, long-distance alarm calls occupy a narrower frequency band and are formed by rapidly repeated lower-pitched notes of the same structure. The difference in structure can be explained in terms of sound transmission and the utility of the calls: a lost infant needs to be located, while an individual that shouts danger would have to be hidden. In terms of sound transmission, at ground level (the calling location of a terrestrial mammal) in tropical forest habitats, sound attenuation is lower between 500 and 2,000 Hz (Marten et al., 1977). This means that to travel over longer distances with minimum attenuation, the sound should concentrate energy in this small frequency band, as seen in the alarm calls represented on the second line of Figure 2.1.

[3] Fonoteca César Ades (FOCA).

Some infant calls may also follow this rule, but they may get lost closer to their mother. Thus, for binaural location, wider-band spectra may favor sound location, since they contain a larger number of frequencies for comparison purposes than narrower-band sounds (Marler, 1955; Vencl, 1977). In summary, there are two distinct form-and-function models for loud communication without visual contact, one favoring locality, and the other based on nonlocality (Marler, 1967). A lost infant calling for its mother needs to be found quickly, and an "easy-to-locate signal" would help the mother. But how to avoid infanticide by males and predators finding the infant first? This may be regulated by calling duration and infant behavior. In wild cavies ("wild guinea-pigs"), isolated infants utter a short sequence of whistles looking for the mother while remaining undercover in captivity. Seconds later, it would find its mother or repeat the series one more time before staying silent (Monticelli & Ades, 2013). If the infant is isolated in a cage, following an experimental paradigm to obtain long sequences of whistles from guinea-pig infants, the wild infant will stay silent. I interpret this as a hierarchical motivational process (Monticelli et al., 2004; Corat et al., 2012). To freeze would be more urgent than looking for the mother in this open arena.

Another noteworthy detail in Figure 2.1 is related to the mocó, a caviomorph rodent, like capybaras and cavies, that inhabits a diverse environment. It is endemic to the Caatinga, the open, hot and dry Brazilian biome. Even in such different habitats, one can find a suggestive convergence in signal structure in the long-distance alarm calls. The evolution of signals has long been addressed by Peter Marler and Eugene Morton and recently reviewed in Magrath et al. (2020). The similarity in alarm call structure creates opportunities for heterospecific recognition and self-centered response (e.g., marmosets and birds: Vencl, 1977; among birds, reciprocally: Magrath et al., 2007).

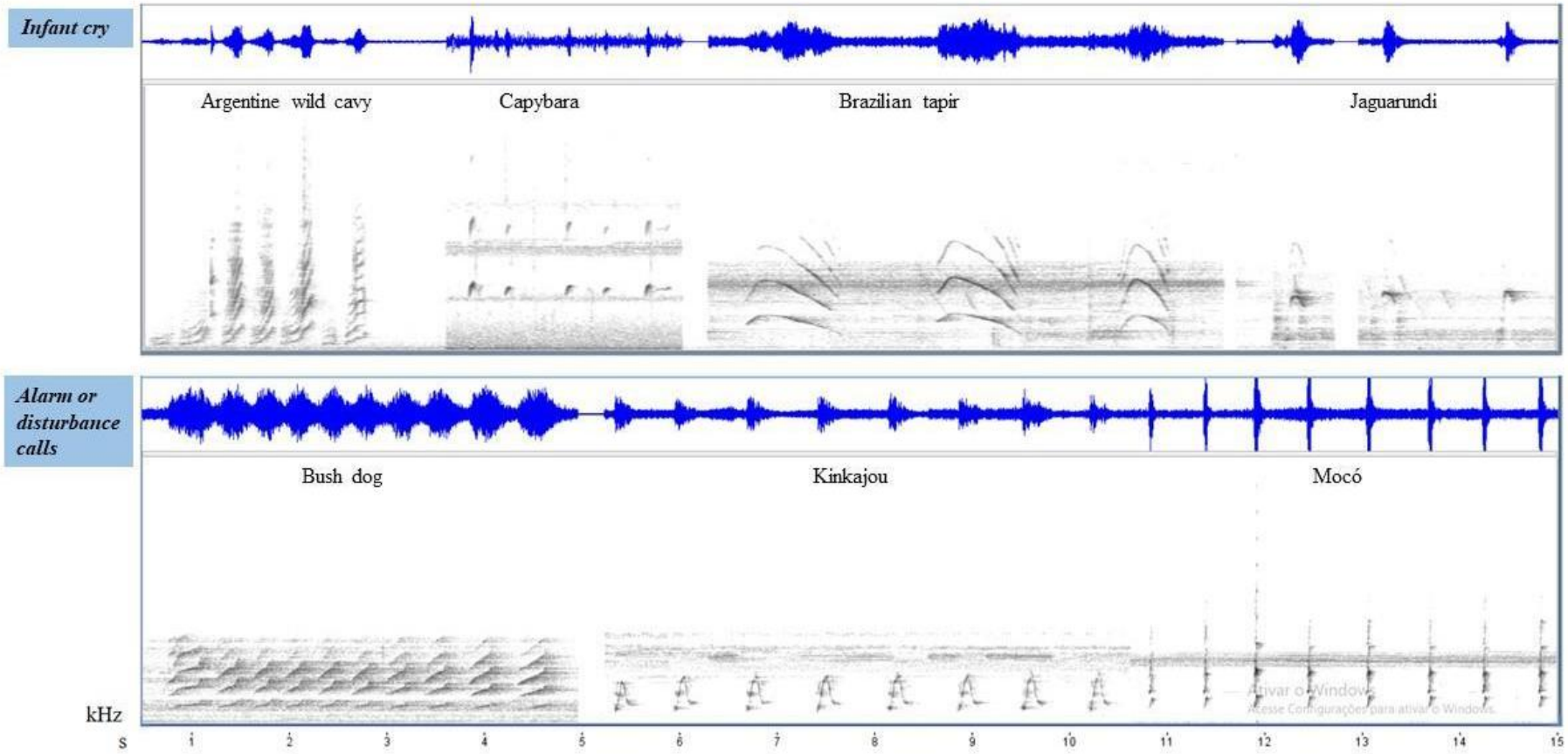


Figure 2.1. Loud calls of disturbed or warned terrestrial mammals: infants separated from the mother (above) and alarmed or disturbed individuals. Note the similarity in their structure: a wide frequency band of high-pitched units in infant isolation calls, and narrower in frequency band and pitch, and composed of rapidly repeated stereotyped short units. Records used in the spectrograms were obtained from Emmons et al. audiobook (1997): *Tapirus terrestris*, *Herpailurus yaguarondi*, *Speothos venaticus* juvenile and *Potos flavus*; and FOCA: *Cavia aperea*, *Hydrochaeris hydrochoerus*, and *Kerodon rupestris*. All images were prepared in Raven 1.5 Hann window using 1024 FFT and overlap 90.

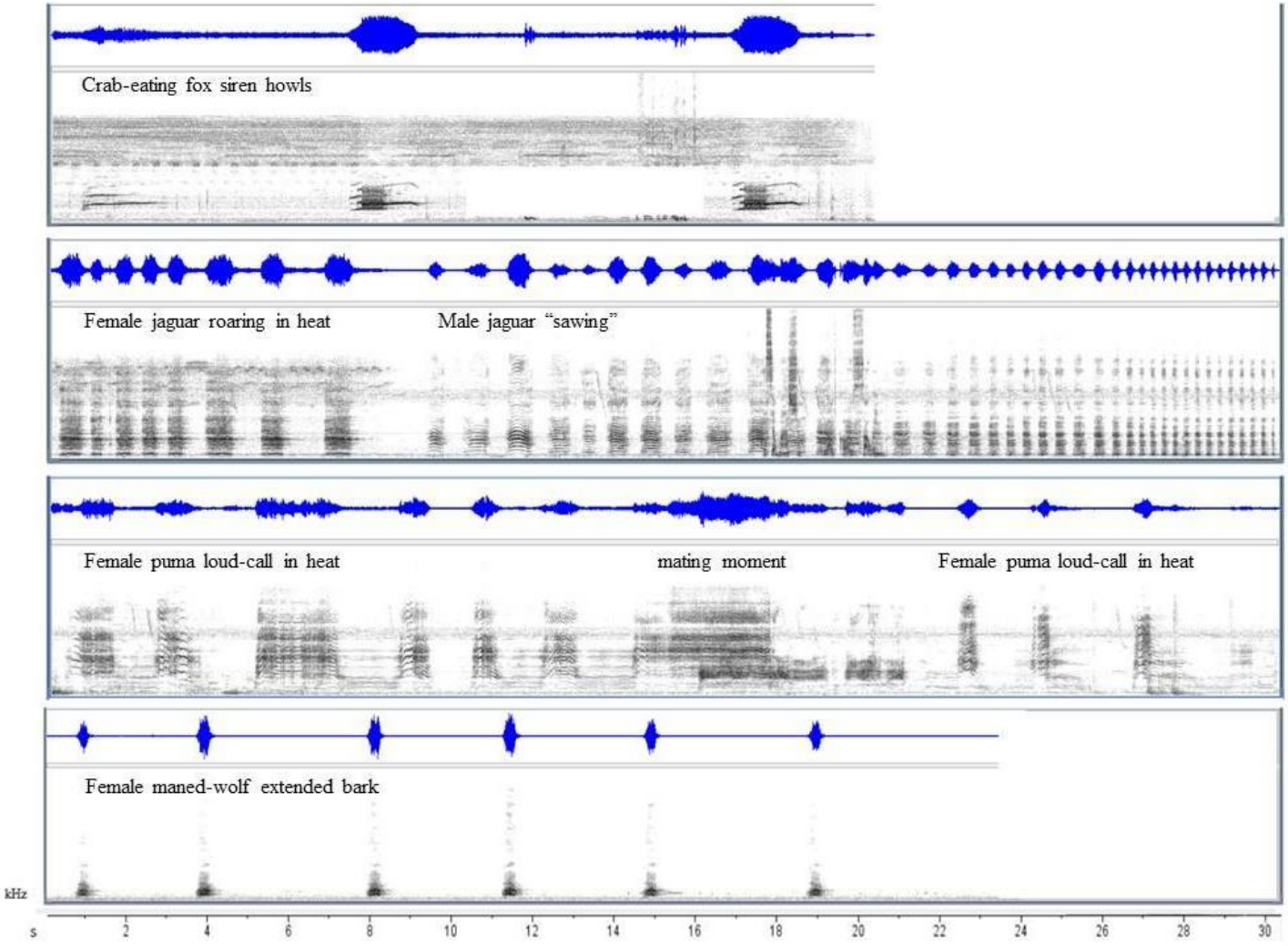


Figure 2.2. Long-distance contact promoting calls: female estrus calls in felids and the extended-bark of the maned-wolf. Records from the Emmons et al. audiobook (1997): *Cerdocyon thous* (probably the "siren howls" of Brady, 1981), *Panthera onca* (female in heat roaring followed by male "sawing", as denominated by Emmons et al.), *Puma concolor* (female in heat call before and after mating); and FOCA: *Chrysocyon brachyurus* recorded by Flora Balieiro. All images were prepared in Raven 1.5 Hann window using 1024 FFT and overlap 90.

Another category of loud calls uttered by neotropical terrestrial mammals are those used in reproductive contexts: signaling estrous and reuniting sexual partners (Figure 2.2). These may be the loudest or longest-distance calls in non-primate terrestrial mammals. Notes are longer, more chaotic, and concentrate energy over a thin band in the lower frequencies. They are used by top predators that may pay lower trade-offs for being heard by predators other than human beings. Playback jaguar calls can be made by an aluminum instrument prepared by hunters to attract males.

Nevertheless, our interest here is the benefit of a third party in other conspecific signal exchange interactions. Almost any male jaguar or puma hearing a female of their species calling may achieve mating if they are the first to arrive, but what happens if they are not? Would they benefit from noting the presence of another male? They could use this information to remain quiet, avoid injury, and wait for a distraction of the first male, or, if they know they are larger, force the female to mate. What if the third party is a female jaguar or puma? Could they benefit from hearing other females' interactions with males? I have no answer to these questions, and am not sure if anyone has. But we may be able to determine these and other species' communicative and cognitive abilities by studying their reactions as an eavesdropping phenomenon in a communication web.

The crying guinea-pig

Female guinea pigs emit a low whistle and whining sound when approached by a male and when interacting with other females (Berryman, 1976; Figure 2.3). Taken together, it sounds like a cry. Males also cry during moderate arousal levels when interacting with a larger or dominant male (Berryman, 1976; Coulon, 1982), and infants emit high-pitched cries and whistles when alone (Arvola, 1974; Berryman, 1976). According to Berryman's observations, receptive females would also squeal (a series of higher-pitched notes, with variably modulated frequency) while whining during courtship.

A complete description of the guinea-pig's acoustic repertoire was provided by John Arvola (1974) and Julia Berryman (1976) before the digital recording and analysis of sounds. My students and I extended these descriptions to other caviid rodent species and adopted the existing technology, which is significantly more advanced than in the 1970s. A mix of subsounds and whines are uttered almost continuously by guinea-pigs in social conditions and they are silent only while resting (Berryman, 1976).

The guinea-pig *Cavia porcellus* is a domestic form of the wild Andean cavy (*Cavia tschudii*) (Spotorno et al., 2004). In Brazil, other wild cavy species are present, including

C. magna, *C. intermedia*, and *C. aperea*, all of which we have recorded. Despite being more economical in sound production, the communicative behavior of wild species is similar to that of *C. porcellus* in social conditions. When in groups, wild cavies will also approach each other after resting or arriving from another location, using subsounds and whines (Figure 2.3). Domestication exaggerated sound production (Monticelli & Ades, 2011, 2013), but subsounds and cries seem to be ancestral traits in *Cavia* taxa, and are shared among the four species we studied. They may be even more ancestral: the mocó *Kerodon rupestris* (subfamily Hydrochaerinae), sister of Caviinae, also cries in similar contexts (Figure 2.3). The role of female whining-squealing, or merely crying during courtship in Caviidae remains unknown.

The web communication approach provided me with insights. Consider a group of guinea-pigs kept in $30m^2$ (Jacobs, 1976) or $12m^2$ (Verzola-Olivio, 2016) outdoor enclosures. After a few days together, the mixed group of males and females will become structured in hierarchical dominance ranks (not precisely linear, Jacobs, 1976; Verzola-Olivio, 2016), divided into affinity clusters consisting of male-female or female-female dyads (Verzola-Olivio, 2016). Higher-ranking males monitor females within their cluster, safeguarding against the approach of other males, and establish stronger associations with one "preferred female" (Jacobs, 1976). The female-defense-polygyny may occur via male courtship behavior throughout the year, independently of estrous and the male's ability to repel other males. Jacob's naturalistic study showed nuances in group behavior: at the end of pregnancy, the associating males court their cluster females more than any other males, even when he was not the "group's normal alpha male"; on the day the male's cluster female gives birth, which coincides with postpartum estrus, in 10 out of 18 episodes, non-alphas rose in ranking order when compared to their modal daily rank during pregnancy. Male-female associations usually endure through successive pregnancies (an association pair lasted more than 13 months); nevertheless, male replacements occurred most commonly during female estrous, after severe fighting between the original association male and the challenger. Half of the eight supposed replacements involved multiparous females, two occurred during the female's first pregnancy, and the other two between pregnancies. None of these interactions took place in silence. Males purred while displaying a varied courtship repertoire of circling, rumbling, moving their feet up and down from the substrate, pursuing and pressing the female's back with their chin, and so on (Rood, 1972). The female being courted moves, urinates, turns her body around to face the male, and eventually stops, allowing him to mount and displaying lordosis, while

vocalizing intermittently, stopping when he moves away, to chase other males, for instance. When he chases her or tries to mount her, the cries become modulated in frequency and intensity. The female cries aloud, stops and permits mounting, and then faces and hits the male before moving on again and being followed by him. The same behavior is seen in the wild species.

This is an opportunity for mate choice. The cry can be heard by other males, even in free-living populations. It is clearly associated with the male courtship call (purr) in domestic and wild species (Monticelli & Ades, 2011); female cries usually overlap with the purr in our recordings. In the Laboratory, cry emission affects the other males in the same cage or room; they will all move frenetically, purring, growling, or squealing, and some will risk approaching the defended female in the outdoor enclosures (Rood, 1972; Jacobs, 1976; Verzola-Olivio, 2016). The concurrent males may teeth-chatter and display agonist activities against each other, such as growling.

According to Jacobs, the association with a male is not permanent. The male may be replaced by another, due to apparent female disinterest (immediate or later) and increased interest in the second male. The author observed differences in male courtship repertoire according to its affinity with the female. During their courtship display, association males exhibit circling, rumping, and swaying. This suggests that this enriched display plays a role in mate choice and male competition. Thus, I presume that females' relatively loud calls are a form of eavesdropping on cavy males. A set of experiments could be conducted to test for the presence of eavesdropping and the performance of non-association males (satellite or subordinate males, according to Asher et al., 2008), in line with McGregor's (2005) paradigms. How does female choice correlate with the distinct qualities of crying?

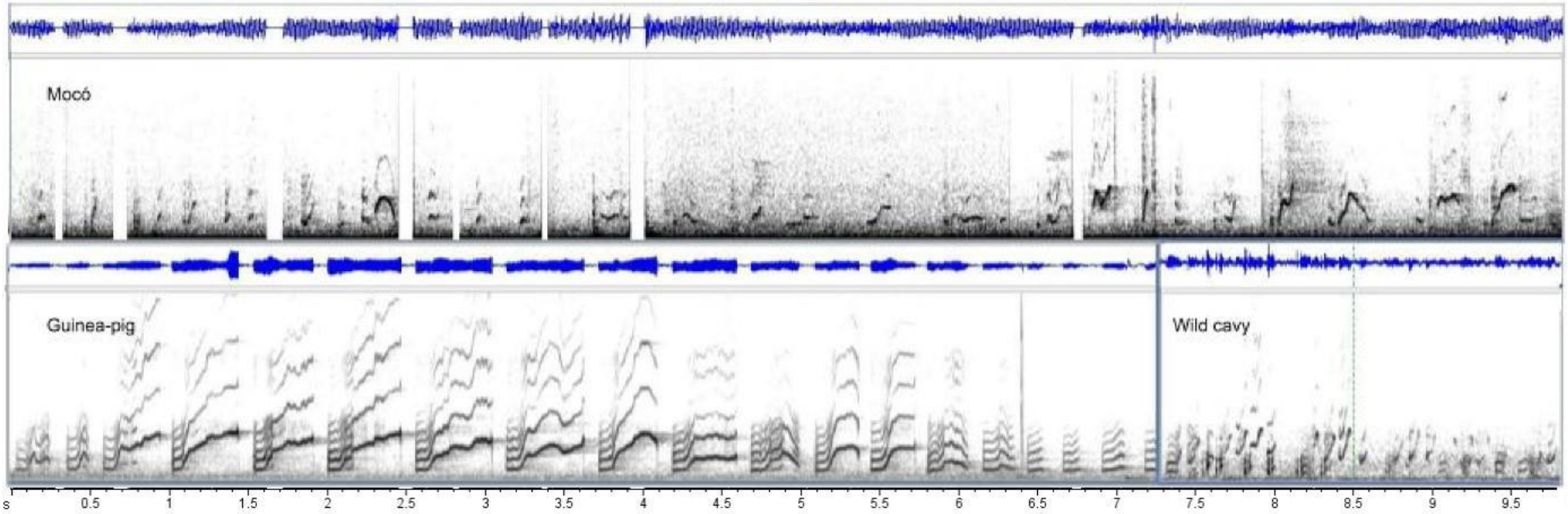


Figure 2.3. The caviid rodent cry consists of low whistles, whines and squeals that sound like a cry. At the top of the figure is an edited sequence of subsounds, low-whistles and whines ending with squeals (the last four notes) uttered by *Kerodon rupestris*, popularly known as mocó. At the bottom, whines naturally change to higher-pitched screams and squeals, changing again to a sequence of whines, produced by a guinea-pig (*C. porcellus*) when a human scratched its back; and two sequences of whines and squeals uttered by a free-living wild cavy (*Cavia intermedia*) female during social foraging.

While I aimed to explore the web communication theory applied to the neotropical species of terrestrial mammals, the search does not end here. The ACN book inspires us to investigate further. It considers that intermediate recipients may also transfer (retransmit) information they acquire to others; the third-party would acquire self-use and transmit data to third connections in the network. The next steps are to obtain evidence to support or challenge the web communication theory, in line with Peter McGregor's suggestion, extending to the contribution of McComb and Reby (2005) to Old World terrestrial mammal species.

References

Arvola, A. (1974). Vocalization in the guinea-pig, *Cavia porcellus* L. In *Annales Zoologici Fennici*, 11, 1-96. Societas Biologica Fennici Vanamo.

Asher, M., Lippmann, T., Epplen, J. T., Kraus, C., Trillmich, F., & Sachser, N. (2008). Large males dominate: ecology, social organization, and mating system of wild cavies, the ancestors of the guinea pig. *Behavioral Ecology and Sociobiology*, *62*(9), 1509-1521. https://doi.org/10.1007/s00265-008-0580-x

Barros, K. S., Tokumaru, R. S., Pedroza, J. P., & Nogueira, S. S. (2011). Vocal repertoire of captive capybara (*Hydrochoerus hydrochaeris*): structure, context and function. *Ethology, 117(1),* 83-94. https://doi.org/10.1111/j.1439-0310.2010.01853.x

Berryman, J. C. (1976). Guinea-pig vocalizations:Their structure, causation and function. *Zeitschrift für Tierpsychologie, 41(1),* 80-106. https://doi.org/10.1111/j.1439-0310.1976.tb00471.x

Brady, C. A. (1981). The vocal repertoires of the bush dog (*Speothos venaticus*), crab-eating fox (*Cerdocyon thous*), and maned wolf (*Chrysocyon brachyurus*). *Animal Behaviour, 29(3),* 649-669. https://doi.org/10.1016/S0003-3472(81)80001-2

Cheney, D. L., & Seyfarth, R. M. (1999). Mechanisms underlying vocalizations of nonhuman primates (pp. 629-644). In Marc D. Hauser and Mark Konishi (eds.) T*he design of animal communication.* Cambridge, MA: MIT Press. https://doi.org/10.3758/BF03199083

Corat, C., Tarallo, R. C. R. B., Savalli, C., Tokumaru, R. S., Monticelli, P. F., & Ades, C. (2012). The whistles of the Guinea pig: an evo-devo proposal. R*evista de Etologia, 11(1),* 46-55. Retrieved 20 May 2021 from: http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1517-28052012000100006&lng=pt&tlng=en

Coulon, J. 1982. La communication acoustique du cobaye domestique comparaison avec quelques rongeurs. *Journal de Psychologie, 79(1-2),* 55–78. Retrieved 20 May 2021 from: http://pascal-

francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=PASCAL83X0041193

Dietz, J. M. (1984). Ecology and social organization of the maned wolf (Chrysocyon brachyurus). *Smithsonian Contributions to Zoology*, 1-51.

Eisenberg, J. F. (1974). The function and motivational basis of hystricomorph vocalizations. In *Symposia of the Zoological Society of London*, 34, 211-224

Emmons, L. H., Ross, D. L., & Whitney, B. M. (1997). *Sounds of Neotropical Rainforest Mammals: An Audio Field Guide.* Conservation International. Editorial Ithaca, NY, Library of Natural Sounds, Cornell Laboratory of Ornithology, US.

Gasco, A., Ferro, H. F., & Monticelli, P. F. (2019). The communicative life of a social carnivore: acoustic repertoire of the ring-tailed coati (*Nasua nasua*). *Bioacoustics*, *28*(5), 459-487. https://doi.org/10.1080/09524622.2018.1477618

Gentry, K. E., Lewis, R. N., Glanz, H., Simões, P. I., Nyári, Á. S., & Reichert, M. S. (2020). Bioacoustics in cognitive research: Applications, considerations, and recommendations. *Wiley Interdisciplinary Reviews: Cognitive Science, 11(5)*, e1538. https://doi.org/10.1002/wcs.1538

Hauser, M. D. (1998). Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. *Animal Behaviour*, *55*(6), 1647-1658. https://doi.org/10.1006/anbe.1997.0712

Jacobs, W. W. (1976). Male-female associations in the domestic guinea pig. *Animal Learning & Behavior*, *4(1)*, 77-83. https://doi.org/10.3758/BF03211991

Kitchen, D. M., da Cunha, R. G. T., Holzmann, I., & de Oliveira, D. A. G. (2015). Function of loud calls in howler monkeys. In Martín M. Kowalewski, Paul A. Garber, Liliana Cortés-Ortiz, Bernardo Urbani & Dionisios Youlatos (eds.) *Howler Monkeys* (pp. 369-399). New York, NY: Springer. https://doi.org/10.1007/978-1-4939-1957-4_14

Kleiman, D. G. (1972). Social behavior of the maned wolf (*Chrysocyon brachyurus*) and bush dog (Speothos venaticus): a study in contrast. *Journal of Mammalogy*, *53*(4), 791-806. https://doi.org/10.2307/1379214

Lima, S. G., Sousa-Lima, R. S., Tokumaru, R. S., Nogueira-Filho, S. L., & Nogueira, S. S. (2018). Vocal complexity and sociality in spotted paca (*Cuniculus paca*). *PloS one, 13(1)*, e0190961.https://doi.org/10.1371/journal.pone.0190961

Magrath, R. D., Haff, T. M., & Igic, B. (2020). Interspecific communication: gaining information from heterospecific alarm calls. In *Coding strategies in vertebrate acoustic communication* (pp.287-314). Cham: Springer. https://doi.org/10.1007/978-3-030-39200-0_12

Magrath, R. D., Pitcher, B. J., & Gardner, J. L. (2007). A mutual understanding? Interspecific responses by birds to each other's aerial alarm calls. *Behavioral Ecology*, *18(5)*, 944-951. https://doi.org/10.1093/beheco/arm063

Magrini, L. & Monticelli, P. F. (2012). Evolução de caracteres associados ao comportamento de corte em Hystricognathi (Rodentia: Mammalia): o chamado de corte e a exibição visual do macho. [Evolution of characters associated with courtship behavior in Hystricognathi (Rodentia: Mammalia): the courtship call

and the visual display of the male.]In: III Simpósio Latino-Americano de Etologia, 2012, Ribeirão Preto, SP. III Simpósio Latino-Americano de Etologia, 2012. v. 11. p. 87-90.

Marler, P. (1955). Characteristics of some animal calls. *Nature*, *176*(4470), 6-8. Retrieved 20 May 2021 from: https://www.nature.com/articles/176006a0.pdf?origin=ppub

Marler, P. (1967). Animal Communication Signals: We are beginning to understand how the structure of animal signals relates to the function they serve. *Science, 157(3790)*, 769-774. https://doi.org/10.1126/science.157.3790.769

Marten, K., Quine, D., & Marler, P. (1977). Sound transmission and its significance for animal vocalization: II. Tropical forest habitats. *Behavioral Ecology and Sociobiology, 2(3),* 291-302. Retrieved 20 May 2021 from: https://www.jstor.org/stable/4599137

Mccomb, K., & Reby, D. A. V. I. D. (2005). Vocal communication networks in large terrestrial mammals. In McGregor (ed.) *Animal communication networks.* Cambridge, MA: *Cambridge University Press*, 372-389.

McGregor, P. K. (1993). Signalling in territorial systems: a context for individual identification, ranging and eavesdropping. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *340*(1292), 237-244. https://doi.org/10.1098/rstb.1993.0063

McGregor, P. K. (Ed.). (2005). *Animal communication networks.* New York: Cambridge University Press.

Monticelli, P. F., & Ades, C. (2011). Bioacoustics of domestication: alarm and courtship calls of wild and domestic cavies. *Bioacoustics, 20(2),* 169-191. https://doi.org/10.1080/09524622.2011.9753642

Monticelli, P. F., & Ades, C. (2013). The rich acoustic repertoire of a precocious rodent, the wild cavy *Cavia aperea*. *Bioacoustics, 22(1),* 49-66. https://doi.org/10.1080/09524622.2012.711516

Monticelli, P. F., Tokumaru, R. S., & Ades, C. (2004). Isolation induced changes in guinea pig *Cavia porcellus* pup distress whistles. *Anais da Academia Brasileira de Ciências, 76(2),* 368-372. http://dx.doi.org/10.1590/S0001-37652004000200027

Otte, D. (1974). Effects and functions in the evolution of signaling systems. *Annual Review of Ecology and Systematics*, *5*(1), 385-417. https://doi.org/10.1146/annurev.es.05.110174.002125

Peake, T. (2005). Eavesdropping in communication. In Peter McGregor (ed.) *Animal communication networks* (pp. 13-37). Cambridge: Cambridge University Press.

Potter, J. (2002). Acoustical and functional analysis of Mountain lion (*Puma concolor*) vocalizations. *The Journal of the Acoustical Society of America,* 111(5), 2393-2393. https://doi.org/10.1121/1.4778134

Rood, J. P. (1972). Ecological and behavioural comparisons of three genera of Argentine cavies. *Animal Behaviour Monographs, 5,* 1–83. Retrieved 20 May 2021 from: https://psycnet.apa.org/record/1974-29608-001

Ryan, M. J. (1988). Constraints and patterns in the evolution of anuran acoustic communication. In Bernd Fritzsch (ed.) The evolution of the amphibian auditory system, pp. 637-677. New York: Wiley.

Searcy, W., & Nowicki, S. (2005). *The evolution of animal communication: Reliability and Deception in Signaling Systems*. Princeton, NJ: Princeton University Press.

Spotorno, Á. E., Valladares, J. P., Marín, J. C., & Zeballos, H. (2004). Molecular diversity among domestic guinea-pigs (*Cavia porcellus*) and their close phylogenetic relationship with the Andean wild species *Cavia tschudi*i. *Revista Chilena de Historia Natural, 77(2)*, 243-250.

Vencl, F. (1977). A case of convergence in vocal signals between marmosets and birds. *The American Naturalist, 111(980)*, 777-782.

Verzola-Olivio, P., & Monticelli, P. F. (2017). The acoustic repertoire of Cavia intermedia as a contribution to the understanding of the Caviidae communication system. *Bioacoustics*, *26*(3), 285-304. https://doi.org/10.1080/09524622.2016.1278401

Verzola-Olivio, P. (2016). *Associações preferenciais e o papel da fêmea nas relações intersexuais em cobaias* (*Cavia porcellus*). [*Preferential associations and the role of females in intersexual relationships in guinea pigs (Cavia porcellus)*] Master 's Dissertation, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, University of São Paulo, Ribeirão Preto. doi:10.11606/D.59.2017.de-12012017-105950. Retrieved 27 october, 2020, from: www.teses.usp.br

Peer commentary

By Gabriel Francescoli

In her lecture, Patrícia Monticelli presented the web communication approach to bioacoustics using terrestrial Neotropical mammals as a model. She shared her research and perspectives based on McGregor's book about web Communication. In her opinion, the book promoted the use of the network approach. Animals use the information they obtain from a vocal interaction, not for the signal alone, but for their benefit, since it generates long-lasting fame for the participants (the traditional emitter and recipient of a communication interaction).

Considering the possibility of eavesdropping on animal communication, she starts with the advantages for an eavesdropper. If an animal hears others vocally competing for food, it receives two important pieces of information: food is available and the mood and size conditions of the competitors. Based on its own experiences in previous disputes, it can judge if it will steal the items. It is a little more interesting to think about the adaptive consequences for the emitter. If someone can hear me interacting with a recipient, and I am aware of this, I may manipulate the signal to intimidate both competitors. If they can identify me and memorize my performance, they may avoid competing with me in future situations. I save time and energy. Let us take the puma and the jaguar (and I suppose this also occurs in cavies and coatis). Females will probably still call when interacting with a male, as I saw in a video posted on the internet. If she already has a partner, why is she still calling? It may be a reproductive strategy to attract another, perhaps better, male (a third party) to the scene. I can see how this can work from the emitter's perspective; this female, for instance, gains an advantage when she still communicates while in estrus, even if she has a sexual partner.

The classical playback experiments may reveal the aspects of the sound that serve as cues to eavesdropping. We always wonder about the "messages" in the call. Individual recognition, sex, mood, and age are aspects that have been found in mammal calls, which are also informative to eavesdroppers. Based on this information, they can plan their next move by recalling the result of their performance in previous experiences with a known individual; they can decide between cooperating or taking

advantage of others, and so on. In summary, individuals in a communication web have opportunities to evaluate the consequences of their next actions and make better decisions for themselves.

Chapter 3

Vocal mimicry in parrots

Maria Luisa da Silva & Leiliany Negrão de Moura[4]

Abstract

Parrots are attractive, colorful, smart and interactive birds; since they can imitate the human voice, parrots are frequently trafficked and some species are critically endangered. How they mimic is still unknown. We offer a number of hypotheses to explain their ability to mimic human voice and present some of the results of 10 years of research on the orange-winged Amazon in nature to support our theories.

Keywords. Parrots, Orange-winged Amazon, Vocal learning, Vocal mimicry.

The order Psittaciformes, comprising the families Strigopidae, Nestoridae, Cacatuidae, Psittacidae, Psittrichasidae and Psittaculidae, which include birds such as the Amazon parrots, cockatoos, lorikeets, lories, and parakeets (Joseph et al. 2012), is a very large group of easily recognized birds, generally restricted to the tropics worldwide. There are 33 species of Amazon parrots in Brazil, called *papagaios* in Portuguese. Parrots' size varies from 27cm (*Amazona xanthops*) to 40 cm (*Amazona farinosa*). The Psittacidae family has the largest number of endangered species (Colar & Juniper, 1992) and recording groups of these species in one location can give a false impression of stability, masking populations of endangered species that are not reproducing (Sick, 1997; Moura et al. 2008; Moura et al. 2010). The monogamous reproductive system with altricial nestlings (born without plumes, eyes closed and depending on their parents), that take more than 4 months to emancipate, also contributes to the threat of these charismatic birds, a situation we observed in the orange-winged Amazon *Amazona amazonica* (Moura et al. 2011, 2014). Because they are attractive, colorful, smart, interactive, and imitate the human voice these birds make desired pets; they are frequent victims of wildlife trafficking. Their ability to mimic human voices is not fully understood, but we offer a number of hypotheses to explain this ability and present some of the results of 10 years of research on the

[4] Federal University of Pará, Belém, PA, Brazil. mlsilva@ufpa.br

orange-winged Amazon in nature to support our theories (Moura et al. 2010, 2011, 2014).

Vocal learning

The conceptual basis of animal acoustic communication is related to the biological species concept. Each species performs a vocalization that represents the species-specific song, a way to achieve successful reproduction and bear offspring with high quality genes (Silva & Vielliard, 2010). The vocal learning model has been extensively studied in Oscines (Kroodsma & Miller 1996, Marler 2004, Todt 2004). As in human beings, vocal learning depends on hearing, and the process of birdsong learning is considered to be mediated by the memorization and vocal coordination skill. This implies the precise movement of both sides of the internal tympaniformis membrane of the syrinx, the unique avian vocal organ, and the resulting response from the sensorimotor process. Song learners exhibit functional songs only in the presence of a model (Vielliard 2004; Silva & Vielliard, 2010). The syrinx provides the potential ability to produce sounds independent of resonance, without the physical limitations related to body size expected in other vertebrates that use vocal cords (Silva & Vielliard *op cit*.). Sound utterance in birds corresponds to low-cost metabolic energy that has been demonstrated in distress calls, the most costly (Jurisevic et al. 1999, Vielliard, 2000). We recorded a small 9-grams manakin (*Machaeropterus regulus*) uttering an 800 Hz song, a frequency too low for a small larynx and tiny body, considering other similar vertebrates (Silva et al. 2001).

Songbirds, parrots and hummingbirds have evolved vocal learning and associated brain structures independently. In mammals, cetaceans, bats and primates also achieved vocal learning (Jarvis et al., 2000; Silva & Vielliard, 2010). The study of vocal learning has advanced in recent years with research on the neurophysiological bases that underlie this communication strategy in birds. Investigations on vocal behavior in neotropical birds have shown that complex and unpredictable songs can be produced even by hummingbirds, scarcely investigated in terms of sound communication (Jarvis et al. 2000; Silva & Vielliard 2006).

A molecule identified as forkhead box P2 - FOXP2, which is related to the motor-control mechanism through the auditory response of vocal communication in humans, was studied in a vocal learner bird, the zebra finch (*Taeniopygia guttata*) (Teramitsu & White 2006). When adult males sing alone, FOXP2 mRNA is strongly

inhibited in area X of the brain nuclei related to the song process, but this does not occur when females are nearby. Authors have concluded that FOXP2 plays an important role in vocal control and depends on social context. Both behavioral and neurobiological studies reveal the importance of social contact in the ontogenesis of sound learning, at different levels of exposure to various contexts in periods determined by specific features.

Vocal mimicry

The ability to mimic artificial or human sounds or other species' songs has been recorded in a number of bird species. Interspecific vocal mimicry is common among birds, and earlier studies have suggested that 15-20% of passerine species mimic to some degree (Marshall, 1950; Hindmarsh, 1984); however, this is almost certainly an underestimate, since sonographic analysis may reveal unsuspected mimics. Examples are the mockingbird (Marshall, 1950), Drongo *Dicrurus paradiseus* (Goodale & Kotagama, 2006), Menuridae, Ptilonorhynchidae and Atrichornithidae (Lyrebirds, Satin bowerbird, and Scrub-birds), Starling *Sturnus vulgaris* (Hindmarsh, 1984) and rare recordings of the grey parrot in the wild (Cruickshank *et al.*,1993) or other Psittaciformes, such as the budgerigar *Melopsittacus undulatus*, whose contact call imitation in adults likely contributes to pair bond formation and maintenance (Hile *et al.* 2000). Neotropical regions harbor some species of the genera *Sporophila* (*S. violacea and S. laniirostris*), which imitate alarm calls from other species inhabiting surrounding areas and mimic sounds with reduced significance for the model (Morton, 1976).

The drongo's behavior demonstrates that alarm-associated calls may have learned components, and that birds can learn the appropriate usage of calls that encode different types of information. Like humans, they have also developed the rare trait of vocal learning, that is, the ability to acquire vocalizations through imitation rather than by instinct. Mimicry is normally suppressed by the need for specific identification, but may emerge as a displacement activity, where a species is exposed to predation or disturbance as a result of the need to make a loud and continuous noise that can be expressed in song if it is biologically advantageous (Goodale & Kotagama, 2006).

However, the species-specific code must be maintained. Songs mediated by vocal learning are identified by variations in populations (dialects) or individuals, to allow recognition at the same levels. The species-specific code can be placed in any

parameter of the song, and imitating strangers in a temporal structure may be a code for specific recognition. This is what must have occurred with the well-known polyglot thrush (*Turdus lawrencei*) from the southwestern Amazon: its song consists exclusively of imitations, but still provokes a specific territorial defense reaction to its playback. In addition to not being clear where the specific recognition code is located, because there is no single sound element of its own, its ability to memorize imitations is astounding. The 15-minute song performed by a songbird in Acre, Brazil, contained reliable imitations of the complete and complex songs of 52 different species, some not vocalizing at that time of the year (Vielliard, 2004).

Amazon Parrots

Our field work observations of orange-winged Amazon *Amazona amazonica* behavior show that this species is highly social; we even observed flocks of parrots having "conversations". We studied the population of Parrot Island (Ilha dos Papagaios), a roosting site for *A. amazonica*, a common species in the region. It is located in Guajara Bay, south of Belém, Brazil (01º 31' 37'' S, 48º 30' 22'' W), and covers an area of 7.4 ha. The number of parrots increased from April/2004 (3,899) to July/2004 (8,539), and began to decline in August/2004 (5,351). This decrease was presumably due to the onset of the breeding season, when paired individuals leave the roost in search of a nest, where they breed and rear young until the nestlings can fly (Moura *et al.* 2008; 2010).

We have studied the complex communication of *A. amazonica,* and identified a repertoire of nine different vocalizations, such as maintaining pair and group contact (two types of calls), predation risk (three types of alarm call) and agonistic situations (Moura *et al*., 2011). Certain calls are used in specific behavioral contexts and can elicit appropriate answers. The species exhibits a complex vocal repertoire during breeding, suggesting the importance of these signals for its survival. The social organization and prolonged ontogenetic development of this parrot may explain these sophisticated acoustic communication systems. We also described population dialects, recorded in 8 to 10 individuals per population in Camará (Ilha do Marajó), Magalhães Barata, Moju, Salinópolis, Santa Bárbara and Tailândia do Pará, all Pará state municipalities, as well as Palmas in Tocantins state, Brazil. The distances between the areas are not proportional to the connection distance obtained in cluster analysis. We concluded that the sound parameters vary independently of distance, and that species-

specific recognition parameters are maintained despite the significant variation within and between populations. Our findings on vocalizations of the present species confirm the existence of vocal learning and population cultural transmission. We also observed gestural communication, related to parental care, when the species was near the nest. Despite *A. amazonica* displays a sophisticated vocal repertoire, their behavioral gestures may represent a survival strategy, and clever defense of the nest, reducing the risk of attracting the attention of predators (Moura *et al.* 2014). We recorded and observed the breeding, vocal and gestural behavior of this species for over ten years without recording or hearing any imitation of strange sounds or those of other species.

How can we explain parrot speakers? Irene Pepperberg studied a special case, perhaps the smartest parrot in the world, the grey *Psittacus erithacus* by the name of Alex. This parrot species is known for its intelligence and vocal mimicry ability. Alex could perform various cognitive tasks and spoke English at a level comparable to a very young child (Pepperberg 1999, 2002). Although the Psittacinae brain is organized very differently from that of mammals, studies of avian cognition have produced surprising results. Some parrot species speak more than others, which may be related to the nature of their own species-specific song, usually contact calls. *Amazona aestiva* can repeat more understandable words and music than *Amazona amazonica*, for example (Vielliard, 1994 and personal communications). Based on our personal experiences and the studies and references presented here, we can surmise why parrots speak to their human companions in captivity but rarely mimic other bird species in the wild. We personally observed how important social contact is for Psittaciformes species. They are gregarious and live in crowded roosting sites, exchanging important information about foraging and how to avoid predators. When a person raises a parrot at home, it is typically a young bird; otherwise, there is significant likelihood that the specimen will die. The bird then starts to establish an interaction by mimicking words, which may raise its chances of survival. This hypothesis is speculative and further studies with other parrot species in the wild and in captivity may provide an answer.

Acknowledgments

We thank Jacques Vielliard *(in memoriam)* for his invaluable assistance and all our current and former students. We also thank the Federal University of Pará, CNPq, Boticário Foundation for the Protection of Nature and CAPES for funding this project.

References

Colar, N. J. & Juniper, A. T. (1992). Dimensions and causes of the Parrot conservation crisis. In: S. R. Beissinger & N. F. R. Snyder (eds.). *New world parrots in crisis* (pp. 1-24). Washington: Smithsonian Institution Press.

Cruickshank, A. J., Gautier, J. & Chappuis, C. (1993). Vocal mimicry in wild African Grey Parrots *Psittacus erithacus*. *Ibis, 135(3),* 293-299. https://doi.org/10.1111/j.1474-919X.1993.tb02846.x

Goodale E. & Kotagama S. W. (2006). Context-dependent vocal mimicry in a passerine bird. *Proceedings of the Royal Society* B: Biological Sciences, 27 3875–880. http://doi.org/10.1098/rspb.2005.3392

Hile, A. G., Plummer, T. K., & Striedter, G. F. (2000). Male vocal imitation produces call convergence during pair bonding in budgerigars, *Melopsittacus undulatus*. *Animal Behaviour*, 59(6), 1209–1218. https://doi.org/10.1006/anbe.1999.1438

Hindmarsh, A.M. (1984). Vocal Mimicry in Starlings. *Behaviour, 90(4),* 302–324. https://doi.org/10.1163/156853984X00182

Jarvis, E. D., Ribeiro, S., Silva, M. L., Ventura, D., Vielliard, J., & Mello, C. V. (2000). Behaviourally driven gene expression reveals song nuclei in hummingbird brain. *Nature,* 406, 628–32. https://doi.org/10.1038/35020570

Jarvis, E. D. & Mello, C. V. (2000). Molecular mapping of brain areas involved in parrot vocal communication. *Journal of Comparative Neurology, 419*(1), 1-31. 10.1002/(SICI)1096-9861(20000327)419:1<1::AID-CNE1>3.0.CO;2-M

Joseph, L, Toon, A., Schirtzinger,E. Wright, T. Schodde, R. (2012). A revised nomenclature and classification for family-group taxa of parrots (Psittaciformes). *Zootaxa, 3205*, 26–40. http://dx.doi.org/10.11646/zootaxa.3205.1.2

Jurisevic, M., Sanderson, K. & Baudinete, R. (1999). Metabolic Rates Associated with Distress and Begging Calls in Birds. P*hysiological and Biochemical Zoology, 72*(1), 38–43. https://doi.org/10.1086/316636

Kroodsma, D. E. & Miller, E. H. (1996). *Acoustic communication in birds*. New York: Academic Press.

Marler, P. (2004). Innateness and the instinct to learn. *Anais da Academia Brasileira de Ciências, 76*(2), 189-200. https://doi.org/10.1590/S0001-37652004000200002

Marshall, A.J. (1950). The function of Vocal Mimicry in Birds, *Emu–Austral. Ornithology, 50*(1), 5-16. https://doi.org/10.1071/MU950005

Morton, E. (1976). Vocal Mimicry in the Thick-Billed *Euphonia*. *The Wilson Bulletin, 88*(3), 485-487.

Moura, L. N.; Vielliard, J.; Silva, M. L. (2008). Flutuação populacional e comportamento reprodutivo do Papagaio-do-mangue *Amazona amazonica* (pp. 223-238). In Jaime Martinez & Nêmora P. Prestes (Ed.): *Biologia da Conservação: estudo de caso com o Papagaio-charão e outros papagaios brasileiros.* Passo Fundo: UPF Editora.

Moura, L. N.; Vielliard, J.; Silva, M. L. (2010). Seasonal fluctuation of the Orange-Winged Amazon at a roosting site in Amazonia. *The Wilson Journal of Ornithology*, 122(1), 88 - 94. https://doi.org/10.1676/09-013.1

Moura, L. N., Silva, M. L., Vielliard, J.M.E. (2011). Vocal repertoire of wild breeding Orange-winged Parrot Amazona amazonica in Amazonia. *Bioacoustics, 20,* 331–340. https://doi.org/10.1080/09524622.2011.9753655

Moura, L. N., Silva, M. L., Garotti, M. F., Rodrigues, A. L. F., Santos, A. C. Ribeiro, I.F. (2014). Gestural communication in a new world parrot. *Behavioural Processes 105* (2014) 46–48. https://doi.org/10.1016/j.beproc.2014.03.003

Pepperberg, I. M. (1999). *The Alex studies*. Cambridge, MA: Harvard University Press.

Pepperberg, I. M. (2002). Cognitive and communicative abilities of grey parrots. *Current Directions in Psychological Science, 11*(3), 83-87.

Sick, H. (1997). *Ornitologia brasileira* [Brazilian ornithology] (2nd. ed.). Rio de Janeiro: Nova Fronteira.

Silva, M. L., Baudet, G., Sigrist, T. & Vielliard, J. (2001). – Descrição do comportamento de corte do Dançarino-de-coroa-vermelha, *Machaeropterus regulus* (Aves, Pipridae). *Bol. Mus. Biol. Mello Leitão (N.Sér.)* 11/12: 171-188.

Silva, M. L.; Vielliard, J. A. (2006). Entropy calculations for measuring bird song diversity: the case of the white-vented violet-ear (*Colibri serrirostris*) (Aves, Trochilidae*). Razprave IV. razreda SAZU*, XLVII-3.

Silva, M. L.; Vielliard, J. A. (2010). A aprendizagem vocal em aves: evidências comportamentais e neurobiológicas (pp. 177-197). In Alda Loureiro Henriques, Grauben José Alves de Assis, Regina Célia Souza Brito & William Lee Berdel Martin (Eds.) Estudos do Comportamento II. Belém: Editora da UFPA..

Teramitsu, I. & White, S. A. (2006). FoxP2 Regulation during undirected singing in adult songbirds. *Journal of Neuroscience*, 26(28), 7390-7394. https://doi.org/10.1523/JNEUROSCI.1662-06.2006

Todt, D. (2004). From birdsong to speech: a plea for comparative approaches. *Anais da Academia Brasileira de Ciências*, *76*(2), 201-208. https://doi.org/10.1590/S0001-37652004000200003

Vielliard, J. (2004). A diversidade de sinais e sistemas de comunicação sonora na fauna brasileira. *I Seminário Música Ciência Tecnologia: Acústica musical*. USP, São Paulo.

Vielliard, J. (2000). Bird community as an indicator of biodiversity: results from quantitative surveys in Brazil. *Anais da Academia Brasileira de Ciências*, *72*(3). https://doi.org/10.1590/S0001-37652000000300006

Vielliard, J. (1994). Bioacoustics and phylogeny among Amazona Parrots. *The Ornithological Notebook of the International Ornithological Congress*. P634. Hofburg, Vienna. August 20-25.

Peer Commentary

By Patricia Ferreira Monticelli and Aline D. Carneiro Gasco

I had the satisfaction of chairing a session of my good friend Maria Luisa da Silva - "Malu" - on the phenomenon of voice mimicry by parrots and commenting about this with Aline's hand and heart. Maria Luisa da Silva (Silva; hereafter) is an Associate Professor at the Federal University of Pará (UFPA) in Belém, surrounded by the Brazilian Amazonia biome. She holds a Ph.D. in Neuroscience and Behavior and is internationally recognized in ornithology and bioacoustics. One of her most celebrated works is the secret life of the orange-winged Amazon and was documented in a movie in 2011 (access here). In addition, Silva has done outstanding work on the International Bioacoustics Council (IBAC), which has dubbed her the *Dame of Brazilian Bioacoustics*.

Vocal mimicry in birds is a quite well-known phenomenon related to the ability to learn sounds. Three avian taxa can learn sounds of their species: parrots, hummingbirds, and passerines. Among the Passerines, Lawrence's Thrush (*Turdus lawrencii*) has the amazing ability to mimic many sounds. *T. polyglottus* goes further: Jacques Vielliard (2004) counted up to 52 sequences of other species' songs in its repertoire.

Parrots are the most cherished bird and sought to have at home. Both wild and companion parrots form everlasting pair bonds with conspecifics that may facilitate the mimicking process by changing the acoustic repertoire or due to a song-learning adaptive function (Hile et al., 2000). The pair bond formation in wild parrots also involves cross-modal communication. During a period of up to 4 months, the parents will look together after the nestling and communicate mostly through the visual channel to each other. The male parent will be most of the time occupied with guarding the nest entrance and communicating with the female through head movements (Moura et al., 2014). The focus of the parrot pair will be to misdirect any chance of revealing the nest entrance to humans and other potential predators (Moura et al., 2011; watch the video here).

Concerning companion parrots, it could confuse whether the vocal mimicry would be a by-product of cognitive abilities or not. Silva presented two arguments to

solve the confusion. Firstly, mimicry is always linked to social bonds because parrots are highly social even without the companionship of conspecifics. It would prompt them to use their natural ability to repeat what they hear from humans. Parrots living with both conspecifics and human companions can either communicate with each other or mimic speech. Silva once observed four parrots living under the companionship of a human tutor. The parrot that was already uttering words taught the others how to do that. Supposedly, the function of vocal mimicry may be to develop social bonds and increase the reproductive repertoire of the parrots. Males with a larger repertoire may offer better genes and then influence females' reproductive choices.

Secondly, a standalone parrot that mimics humans may be misdirecting its need to bond with a conspecific by bonding with a human tutor. Alternatively, this behavior would be a response to a stressful environment. Silva disagrees with authors who consider speech mimicry a natural behavior that means good welfare when performed in captivity (Broom and Kirkden, 2004; Dawkins, 1990). In her opinion, the assumption of human companionship providing enough social interactions to a parrot is an anthropocentric view. Mimicry is inherent in parrots, and therefore, they can hear, repeat, retain and learn speech. The vocal feedback parrots receive from human companions is inadequate and, at the same time, is all they have to mimic.

The impressive potential of birds for learning alien sounds may lead one to wonder about the extensive comparative studies. Is birds' ability to communicate a step toward the evolution of language? There are astonishing examples of birds' social and vocal complexity. Firstly, the chickadees (*Poecile* sp), a small passerine, combine different syllable structures forming new call categories with different communicative functions. Secondly, the *Guira guira* repertoire is composed of 27 signals that can also be recombined in a wider variety of contexts (Mariño, 1989). *G. guira* is a gregarious species that form groups of up to 12 individuals with distinct roles in the social group. Some of them are "helpers" who emit three different alarm calls according to the approaching danger, whether on land (a snake) or the air (a bird of prey) (Mariño, 1989). Lastly, the Rufous-bellied thrush (*Turdus rufiventris*) has a vocal communication system individually distinguished (Silva, Piqueira & Vielliard, 2000). All of those three examples may compel readers to see that comparative studies should not equate bird communication to human language in terms of grammar.

Reaching out to the end, Silva and Leiliany Moura opened up about their mutual experience hearing an extremely sad song uttered by a parrot mother facing a

dead nestling. The death was attributed to a spider they found in the nest. She approached the nest and left immediately, performing a deep dive into the air while uttering a unique whine-like call. Silva and Moura speculated if the song derived from a changed mood caused by her dreadful experience. We can speculate about the emotional content on that call and urge for further investigation.

References

Broom, D. M., Kirkden, R. D., 2004. Welfare, stress, behaviour and pathophysiology (pp. 337–369). In Dunlop, R. H., Malbert, C. H. (Eds.), *Veterinary Pathophysiology*. Ames, IA: Blackwell.

Dawkins, M. S. (1990). From an animal's point of view: motivation, fitness and animal welfare. *Behavioural and Brain Sciences, 13*, 1-61.

Hile, A. G., Plummer, T. K., & Striedter, G. F. (2000). Male vocal imitation produces call convergence during pair bonding in budgerigars, *Melopsittacus undulatus*. *Animal Behaviour, 59*(6), 1209-1218. https://doi.org/10.1006/anbe.1999.1438

Mariño, H. F. (1989). *Comunicação sonora do anu-branco:* Avaliações eco-etológicas e evolutivas. Editora: Unicamp, 1 ed.

Moura, L. N., Silva, M. L., Garotti, M. M., Rodrigues, A. L., Santos, A. C., & Ribeiro, I. F. (2014). Gestural communication in a new world parrot. *Behavioural Processes, 105*(2014), 46-48. https://doi.org/10.1016/j.beproc.2014.03.003

Moura, L. N.; Gogala, M.; Doolittle, E. & Silva, M. L. (2011). *A vida secreta do Papagaio-do-mangue* [The secret life of Orange-winged Amazon, Documentary]. Available on: https://drive.google.com/file/d/1h9ILLbz4ffkTIQzK3S1eFllwU0kUbDqc/view?usp=sharing

Silva, M. L., Piqueira, J. R. C., & Vielliard, J. M. (2000). Using Shannon entropy on measuring the individual variability in the rufous-bellied thrush *Turdus rufiventris* vocal communication. *Journal of Theoretical Biology, 207*(1), 57-64. https://doi.org/10.1006/jtbi.2000.2155

Vielliard, J. (2004). A diversidade de sinais e sistemas de comunicação sonora na fauna brasileira. *I Seminário Música Ciência Tecnologia*: Acústica musical. USP, São Paulo.

Chapter 4

The evolution of vocal expression of emotions: evidence from a long-term project on ungulates

Elodie F. Mandel-Briefer[5] and Aline D. Carneiro Gasco[6]

Abstract

This chapter aims at presenting a summary of the main result of a long-term project, in which Elodie Briefer and collaborators have compared how several species of domestic and wild ungulates express and also perceive emotions, not only within but also between species. The species included in this long-term project were the following: goats (*Capra hircus*), cows (*Bos taurus*), domestic horses (*Equus caballus*), Przewalski's horses (*Equus przewalskii*), wild boars (*Sus scrofa*) and pigs (*Sus scrofa domesticus*). Overall, a clear picture that arises from both the expression and perception parts of the project is that expression of emotional arousal seems to have been well conserved throughout evolution, whereas the expression of emotional valence seems more species-specific.

Keywords: Affective state; arousal; domestication; motivation; source–filter theory; valence.

Studying animal emotion

The dimensional approach

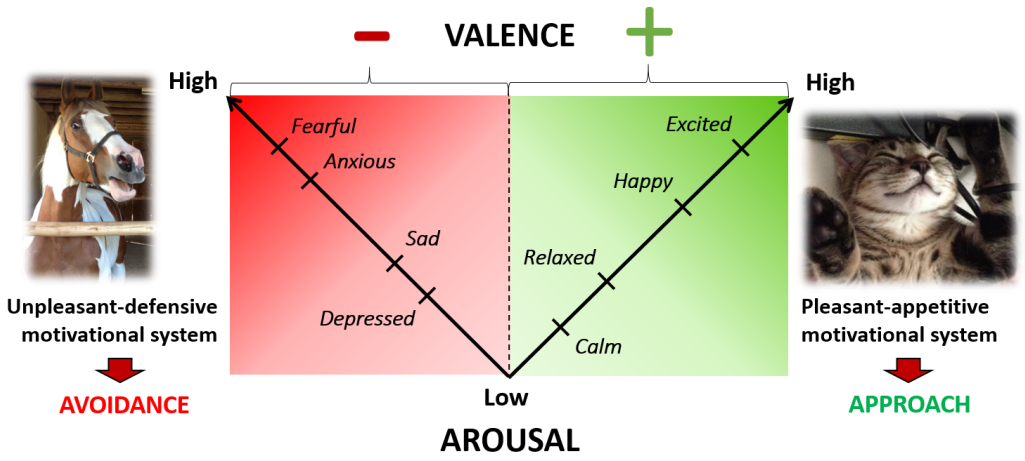
We will first discuss how to study animal emotions using the dimensional approach, which consists in categorizing emotions according to their two main dimensions: the first is valence, which can be positive (e.g., excited, happy, relaxed, calm) or negative (e.g., fearful, anxious, sad, depressed); and the second is arousal, which can be defined as bodily activation (Figure 4.1) (Russell 1980; Bradley et al. 2001; Mendl et al. 2010).

[5] Behavioural Ecology group, Department of Biology, University of Copenhagen, Copenhagen, Denmark. elodie.briefer@bio.ku.dk

[6] Ethology and Bioacoustics Laboratory (EBAC), Department of Psychology, FFCLRP, University of São Paulo, Ribeirão Preto, SP, Brazil. Dr. Aline Gasco transcripted the lecture to this text and edited it with Elodie. agasco.coati@gmail.com

Emotions of negative valence are part of the unpleasant-defensive motivational system (Figure 4.1), which triggers avoidance from the emotion-eliciting stimulus. This usually arises in contexts associated with a decrease in fitness. By contrast, emotions of positive valence are part of the pleasant-appetitive motivational system (Figure 4.1), which triggers an approach towards the emotion-eliciting stimulus, and is typical of contexts associated with an increase in fitness (Bradley et al. 2001).

Figure 4.1. Illustration of the dimensional approach (adapted from Briefer 2020). The



valence differentiates the negative emotions, shown in the red box, from the positive ones, shown in the green box. Arousal can be defined as the intensity of the valence. The negative emotions are part of the unpleasant-defensive motivational system that triggers avoidance. The positive emotions are part of the pleasant-appetitive motivational system that triggers the approach.

The arousal dimension, which can be considered as the intensity of the valence, ranges from low to high levels within both negative and positive valences (Figure 4.1) (Russell, 1980; Mendl et al., 2010). Therefore, both low-arousal negative emotions, such as depressed and sad, and high-arousal negative emotions, such as fearful and anxious, can occur (Figure 4.1). The same applies to positive emotions; positive emotions can be both of low arousal (e.g., calm and relaxed) and high arousal (e.g., happy and excited).

The observable components

Since verbal reports of emotions, which assess the subjective component of emotions ('feeling') cannot be obtained in animals, assessments of their emotions must be based on other observable components. These are the neuro-physiological, behavioural and cognitive components (Mendl et al., 2010). The neuro-physiological component can comprise, among others, changes in brain activity or in heart-rate, which are mostly linked to the arousal of the emotion. The cognitive component is reflected by changes in appraisal, attention, memory and judgments such as the cognitive biases (Mendl et al., 2010). Finally, the behavioural component describes how the animal reacts (e.g., changes in the ear, tail, body or head posture).

The behavioural component also comprises the expression of emotions, which can be mainly facial or vocal. Since these expressions are aimed at regulating social interactions (Briefer, 2018), they can be predicted to be rather conspicuous and allow a human observer to perceive and measure them. Expressions of emotions are thus very promising indicators of animal emotions for further investigations.

Evidence for vocal expression of emotions

According to the source-filter theory of vocal production, the human voice is produced through the following mechanism: air coming from the lungs triggers the vocal folds to vibrate, which determines the source of the sound and the lowest frequency (or fundamental frequency; 'F0'). This source sound is then filtered in the vocal tract that comprises the vocal and nasal cavities. According to the shape and properties of the vocal tract, some frequencies will be amplified and result in formant frequencies, while others will be dampened (Fant, 1960). Overall, the experience of an emotion will result in changes in the vocal apparatus during voice production, such as a more tensed tract, rapid respiration, and decreased salivation. Those modifications will, in turn, affect the acoustic structure of the voice (Juslin & Scherer, 2005).

In non-human animals, since vocalisations are produced in a fundamentally similar way as in humans, we can expect similar effects of emotions on the acoustic structure of their vocalisations (Briefer, 2020). In addition, since the majority of the mammalian species have relatively less control over vocalisations compared to humans (Jürgens, 2009), we can expect an even more direct link between emotions and the structure of vocalisations in animals than in humans (Briefer, 2012).

In line with the above-mentioned predictions, the literature shows evidence for vocal expression of emotions across species (reviewed in Briefer, 2012, 2020). The expression of emotional arousal has been studied in a wide range of species, particularly in contexts related to pain, hunger, and predation. By contrast, the evidence for vocal expression of valence is still limited to a few species. Such limited evidence suggests that, across species, calls tend to be shorter with lower and less variable fundamental frequencies within positive emotions in comparison to negative ones (Briefer, 2012, 2020) (Figure 4.2). This is also what Briefer and colleagues have found in most species studied in their long-term project that included goats, domestic horses, pigs, cows, Przewalski's horses and wild boars (Figure 4.2).

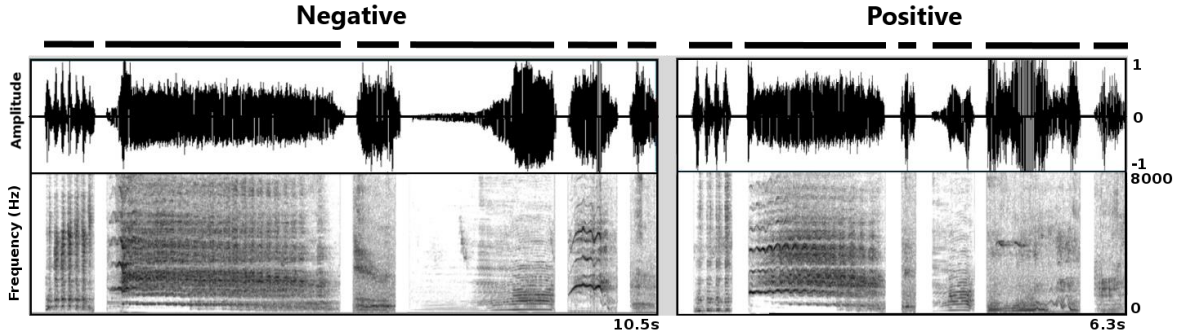


Figure 4.2. Oscillograms (above) and spectrograms (below) of emotionally negative (left) and positive (right) contact calls produced by all the species included in Briefer and collaborators' long-term project; (from left to right) goat, domestic horse, pig, cow, Przewalski's horse and wild boar.

Evidence for vocal contagion of emotions

The encoding of emotional valence and arousal in the producer's vocal signal can be perceived by the receiver of the signal. This receiver can decode the producer's information and be affected by it, leading to a process termed 'emotional contagion', or 'state-matching' (de Waal, 2008; Briefer, 2018). In addition, this process can occur not only between one producer and one receiver, but also between several individuals in a social group (Figure 4.3). As a result, the contagion of either low or high negative (Figure 4.3a) or positive (Figure 4.3b) emotions can take place (Briefer, 2018), and facilitate communication, coordination and cooperation among group members (Preston & de Waal, 2002; de Waal, 2008; Spinka, 2012).

Vocalisations are salient and discrete events that are hard for the surrounding listeners to avoid. Since vocalisations travel over very long distances, through obstacles and even in the dark; they can be predicted to play a crucial role in the contagion of emotions (Briefer, 2018). In non-human animals, there is strong evidence suggesting that vocal contagion of emotional arousal occurs. In comparison, there is little evidence for vocal contagion of emotional valence.

The contagion of emotional arousal results in matched emotional arousal between producer and receiver of the vocalisations. Evidence has been provided using playback experiments mainly in contexts of predation, aversion, aggression, stress and hunger (reviewed in Briefer, 2018). Those studies suggested that playbacks of higher arousal calls trigger behavioural responses suggesting higher arousal emotions in the receivers during the experiments. By contrast, there is only limited evidence showing that calls associated with positive emotions trigger more positive emotions in receivers when compared with calls linked to negative emotions; and vice-versa (Briefer, 2018). Indeed, the results of most studies published to date might lack proper control of the arousal dimension due to the use of call types of different functions. Therefore, further studies investigating vocal contagion of emotions are required and should account for these confounding factors, particularly regarding contagion of emotional valence (Briefer, 2018).

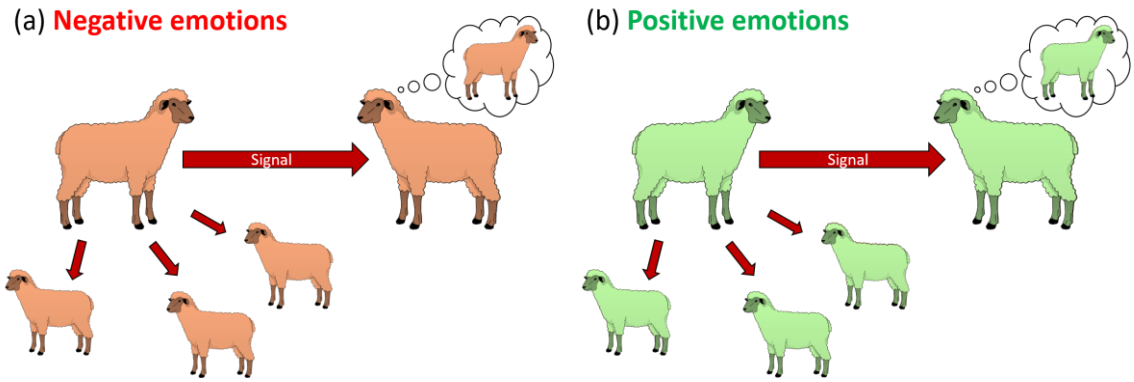


Figure 4.3. Vocal contagion of (a) negative emotions and (b) positive emotions from a producer to several receivers.

Goals of the Cross-species Study

In the long-term study of Briefer and collaborators focusing on the expression and perception of emotional valence within and between ungulate species, three main factors were considered to explain similarities and differences in emotion expression, as well as cross-species perception of emotions: (1) phylogeny; (2) domestication; and (3) familiarity with the species. In all experiments, the effect of emotional arousal was carefully controlled for.

Methods and Hypotheses

Similarities among ungulates in vocal expression of emotions

To investigate the effect of phylogeny and domestication on vocal expression, Briefer and colleagues have compared how domestic and wild Equidae (domestic and Przewalski's horses), and domestic and wild Suidae (domestic and wild pigs) express emotional valence. To this aim, they recorded these four species in various positive and negative situations of various arousal levels, assessed via heart-rate measurements (in the domestic species) and via locomotion (in the wild species). After controlling for changes related to arousal, variations in acoustic parameters related to valence were investigated and compared between domestic and wild species.

The following hypotheses and predictions were made; if phylogeny played a role in the expression of emotions, pigs and wild boars would be expected to express emotions in a rather similar way, and domestic and Przewalski's horses to do so as well. If some indicators of emotions were highly conserved throughout evolution, as suggested by Darwin (1872), it would be expected that Suidae and Equidae express emotions in a similar way. If domestication played a role, it would be expected that pigs and wild boars, as well as domestic and Przewalski's horses, differ substantially following the artificial selection that those species underwent during domestication.

Cross-species perception of emotions among ungulates and between ungulates and humans

To investigate the cross-species perception of emotional valence and the effect of familiarity, phylogeny and domestication on this phenomenon, the negative and positive animal vocalisations recorded during the first part of the project, as well as

human actor voices from a validated database (GEMEP corpus; Bänziger and Scherer, 2010), were played back to the subjects. Domestic and Przewalski's horses were thus tested with conspecific whinnies, whinnies of each other and human voice. Similarly, domestic and wild pigs were tested with conspecific grunts, each other´s grunts and human voice. In addition, a large online questionnaire was conducted to test people's ability to rate the valence of ungulate's contact calls correctly. To this aim, the sounds of the four ungulate species mentioned above, in addition to cow and goat calls, were included in that survey. We also collected data on participants' demography, familiarity with the various species and empathy.

The following hypotheses and predictions were made. If familiarity plays a role in the cross-species perception of emotional valence, it would be expected that the species familiar with humans could perceive the expression of emotions in the human voice. By contrast, the closely-related domestic and wild ungulates had never heard each other's vocalisations. Therefore, it was not expected that they could perceive each other's expression of emotions.

Similar predictions can be proposed about the human perception of animal vocalisations. Briefer and collaborators predicted that humans should perceive the vocal expression of emotions in familiar species more easily than in those unfamiliar to us. If phylogeny played a role in the cross-species perception of emotional valence, it would be expected that all ungulates, and particularly closely-related species, perceive each other's expression of emotions. However, it was expected that these species would not perceive the expression of emotions in our voice because humans are more distantly related to them. Alternatively, it would be assumed that some indicators of emotions could be very well conserved throughout evolution. Finally, if domestication played a role in the cross-species perception of emotional valence, it was expected that only the domestic ungulates, but not their wild counterpart, could perceive the expression of emotions in human voice, and vice-versa.

Results

In the following sections, we will describe only the results regarding the similarities in vocal expression of emotions among ungulates that are already published in Briefer et al. (2015), Maigrot et al. (2017), Maigrot et al. (2018), and Briefer et al. (2019). To have a glimpse of the results about the cross-species

perception of emotions among ungulates and between ungulates and humans that are to be published, please refer to the online presentation delivered during the conference.

Vocal expression of emotions

Equidae

Domestic horse's whinnies were shown to present a rather rare phenomenon among mammals; the presence of two fundamental frequencies, hereafter called 'F0' (the lowest) and 'G0' (the highest), as shown in Figure 4.4a. The acoustic parameters of the whinnies that changed according to the emotional valence were the energy quartiles, the amplitude modulation, the duration and the value of G0 (Briefer et al., 2015). The main indicators of valence, after controlling for arousal, were a shorter duration and a lower G0 (Briefer et al., 2015) (Figure 4.4a).

Whinnies of Przewalski's horses were found to have a structure similar to that of domestic horses, with two fundamental frequencies. However, apart from a similar decrease in energy quartiles and amplitude modulation rate between negative and positive valence, other indicators of valence differed between the two species. In particular, Przewalski's horses did not produce shorter whinnies with a lower G0 in the positive contexts (Maigrot et al., 2017) (Figure 4.4b).

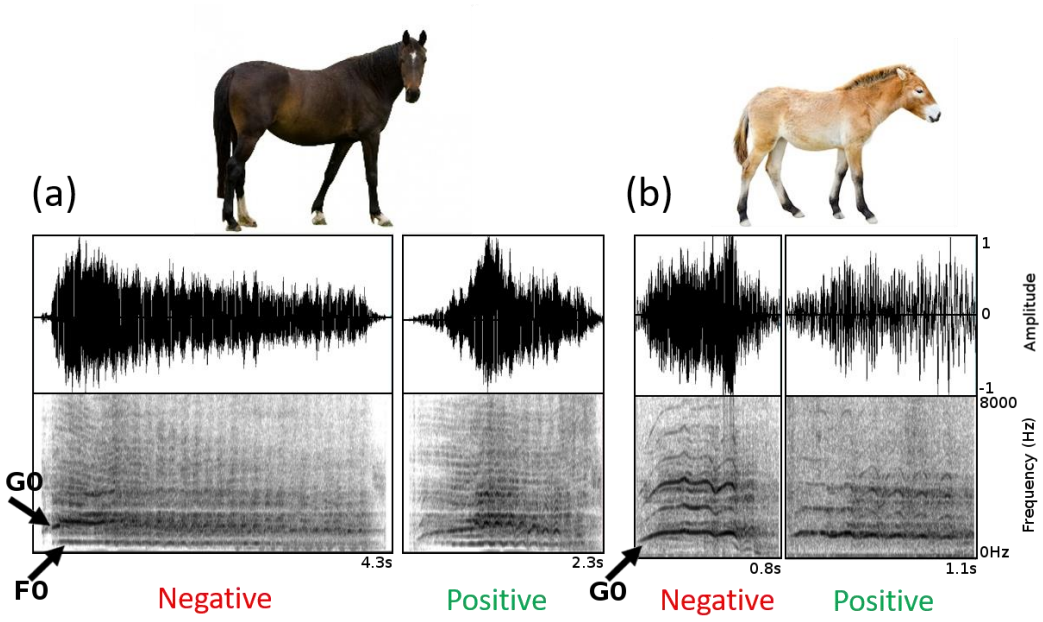


Figure 4.4. Oscillograms (above) and spectrograms (below) of negative and positive whinnies of (a) a domestic horse and (b) a Przewalski's horse.

Suidae

Domestic pigs produced grunts with notably higher formants, a narrower range of the third formant and shorter duration in positive compared to negative situations (Briefer et al., 2019) **(**Figure 4.5a). Similar changes were observed in the range of the third formant and duration between negative and positive situations in wild boars. However, the opposite pattern of formant change was observed, with formant frequencies decreasing in negative situations in comparison to flat formants in positive ones (Maigrot et al., 2018) (Figure 4.5b).

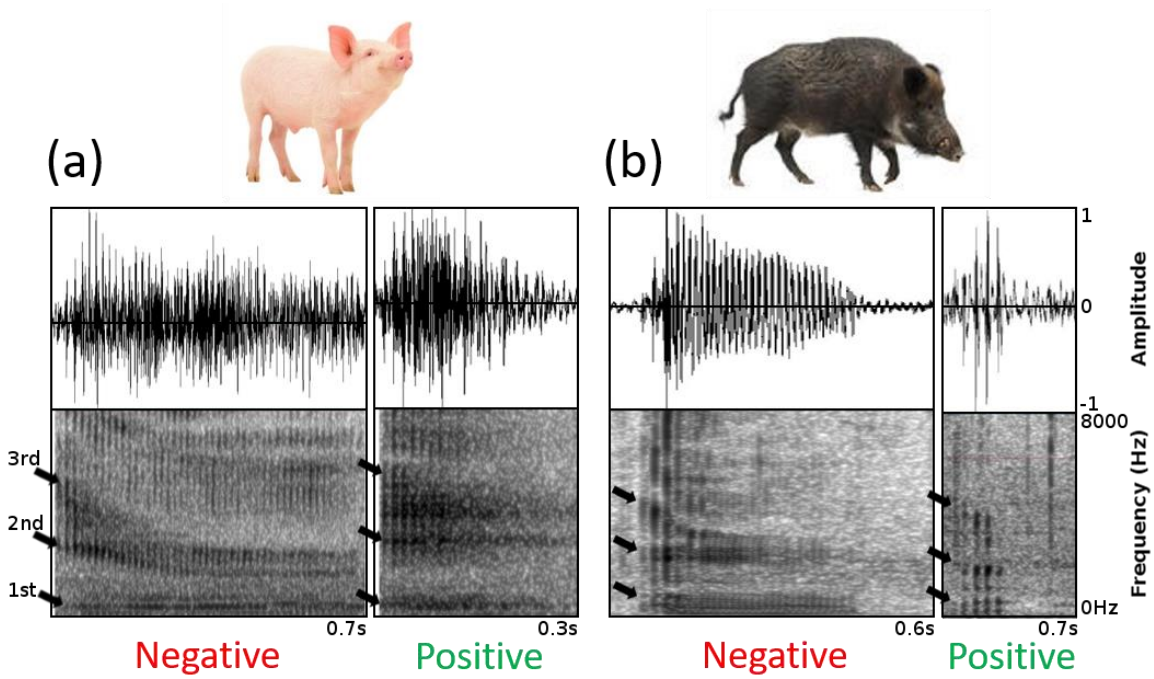


Figure 4.5. Oscillograms (above) and spectrograms (below) of negative and positive grunts of (a) a domestic pig and (b) a wild boar. The arrows indicate the 1st, 2nd and 3rd formants.

Cross-species perception of emotions

Domestic and wild ungulates

Detailed results of the playback experiment conducted with the four ungulate species will be disclosed in an upcoming publication (Maigrot et al., submitted). The results by Maigrot and collaborators suggest that all species except wild boars perceive

indicators of emotional valence across species (including human voice). By contrast, wild boars only reacted differently to negative and positive domestic pig grunts. Wild boars displayed more behavioural indicators suggesting negative emotions during playbacks of positive pig grunts compared to negative pig grunts. This might suggest that wild boars were prompted to perceive positive pig grunts as negative due to the difference in emotion expression between these two species (Maigrot et al., submitted).

Humans

The results of the online questionnaire, which will also be described in more detail in an upcoming publication (Sowerby Greenall et al., close to submission), suggest that participants were able to recognise the emotional arousal of vocalisations in a relatively similar way across species. However, valence recognition depended on the species, with some that were correctly classified above chance level and some even below chance level. Other effects on our ability to recognise the emotions of animal sounds were revealed, such as how often participants were in contact with the various species and their empathic tendencies. Lastly, the duration and spectral center of gravity were found to affect how correctly emotional arousal and valence were rated (Sowerby Greenall et al., close to submission).

Concluding remarks

To conclude, Briefer and collaborators' long-term project on ungulates contributes to our understanding of the evolution of emotion expression and the role that various factors (e.g., domestication) play in this process. In terms of expression, the project showed that both phylogeny and domestication might have played a role in the evolution of vocal expression of emotions. Indeed, similarities but also important differences can be found in the way domestic and wild ungulates express emotional valence. In terms of perception, the project revealed that phylogeny might have played a role in the cross-species perception of emotions in Equidae; the horse species of this study seemed to perceive indicators of emotional valence across species, suggesting high conservation of some indicators throughout evolution. By contrast, in Suidae, domestication might have affected the perception of emotions since domestic pigs, but not wild boars, seemed to perceive indicators of emotional valence across species, including humans. The results of the online questionnaire suggested that human

perception of the cross-species perception of emotions depended on familiarity with the species (contact frequency), empathy and several vocal parameters.

Overall, a clear picture that arises from both the expression and perception parts of the project is that, as suggested previously (e.g., Briefer et al., 2012; Filippi et al., 2017), expression of emotional arousal seems to have been well conserved throughout evolution, whereas the expression of emotional valence seems more species-specific.

Acknowledgements

This research was funded by a Swiss National Science Foundation grant awarded to E.F.B. (Grant No. PZ00P3 148200). We are thankful to all the people who participated both in the playback experiments and on the online questionnaire. We also thank the attendees of the online conference who participated in the Q&A session. Thanks to the Institute of Psychology of the University of São Paulo, the pandemic did not prevent us from gathering virtually.

References

Bänziger, T., & Scherer, K. R. (2010). Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus. In K. R. Scherer, T. Bänziger, & E. B. Roesch (Eds.), *Blueprint for affective computing: A sourcebook* (pp. 271–294). Oxford: Oxford University Press.

Bradley, M. M., Codispoti, M., Cuthbert, B. N., & Lang, P. J. (2001). Emotion and motivation I: Defensive and appetitive reactions in picture processing. *Emotion*, 1(3), 276-298. https://doi.org/10.1037/1528-3542.1.3.276

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1-20. https://doi.org/10.1111/j.1469-7998.2012.00920.x

Briefer, E. F., Padilla de la Torre, M., & McElligott, A. G. (2012). Mother goats do not forget their kids' calls. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1743), 3749-3755. https://doi.org/10.1098/rspb.2012.0986

Briefer, E. F., Maigrot, A. L., Mandel, R., Freymond, S. B., Bachmann, I., & Hillmann, E. (2015). Segregation of information about emotional arousal and valence in horse whinnies. *Scientific Reports*, *5(1),* 1-11. https://doi.org/10.1038/srep09989

Briefer, E. F. (2018). Vocal contagion of emotions in non-human animals. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1873), 20172783. https://doi.org/10.1098/rspb.2017.2783

Briefer, E. F., Vizier, E., Gygax, L., & Hillmann, E. (2019). Expression of emotional valence in pig closed-mouth grunts: Involvement of both source- and filter-related parameters. *Journal of the Acoustical Society of America*, *145*(5), 2895-2908. https://doi.org/10.1121/1.5100612

Briefer, E. F. (2020). Coding for dynamic information: Vocal expression of emotional arousal and valence in non-human animals. In T. Aubin, & N. Mathevon (Eds), *Coding strategies in vertebrate acoustic communication* (pp. 137-162). Cham: Springer.

Darwin, C. R. (1872). *The expression of the emotions in man and animals*. London: John Murray. 1st edition.

de Waal, F. B. M. (2008). Putting altruism back into altruism: The evolution of empathy. *Annual Review of Psychology*, *59*, 279-300. DOI: https://doi.org/10.1146/annurev.psych.59.103006.093625

Fant, G. (1960). *Acoustic theory of speech production*. New York: Mouton Publishers, The Hague.

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., ... & Güntürkün, O. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1859), 20170990. https://doi.org/10.1098/rspb.2017.0990

Jürgens, U. (2009). The neural control of vocalization in mammals: A review. *Journal of Voice*, *23*, 1-10. https://doi.org/10.1016/j.jvoice.2007.07.005

Juslin, P., & Scherer, K. R. (2005). Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds), *The new handbook of methods in nonverbal behavior research* (pp. 65-135). Oxford: Oxford University Press.

Maigrot, A. L., Hillmann, E., Anne, C., & Briefer, E. F. (2017). Vocal expression of emotional valence in Przewalski's horses (*Equus przewalskii*). *Scientific Reports*, *7(1)*, 1-11. https://doi.org/10.1038/s41598-017-09437-1

Maigrot, A. L., Hillmann, E., & Briefer, E. F. (2018). Encoding of emotional valence in wild boar (*Sus scrofa*) calls. *Animals*, *8*(6), 85. https://doi.org/10.3390/ani8060085

Mendl, M., Burman, O. H. P. & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences*, *277*, 2895-2904. https://doi.org/10.1098/rspb.2010.0303

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences*, *25*(1), 1-20. DOI: 10.1017/s0140525x02000018

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. https://doi.org/10.1037/h0077714

Špinka, M. (2012). Social dimension of emotions and its implication for animal welfare. *Applied Animal Behavior Science*, *138*(3-4), 170-181. https://doi.org/10.1016/j.applanim.2012.02.005

Peer Commentary

By Aline D. Carneiro Gasco

The nature versus nurture controversy is likely solved for Elodie Briefer, as she states in her chapter. Next, we tackle the emotions encoded in the calls as triggers of emotional reactions, perception, and expression. In Briefer's assumption, innate emotions in wild and domestic animals are associated with releasing the cortisol hormone in the blood (its circulating amount is usually an indicator of the emotional arousal experienced). Domestication lowered the frightened levels of animals with effects on the expression and experience of their emotions.

From Briefer's chapter, it is clear that emotion perception plays an important role in regulating social interactions. Moreover, nonhuman animals learn to perceive others' emotions in both arousal and valence dimensions. For example, animals associate these calls with negative emotions when the negative utterances are produced in an agonistic context of repeatedly biting and knocking. Therefore, perception of emotions allows the individuals to regulate approaching and separation behaviors according to the context.

Regarding the representative call types with unique emotional valence in the species, Briefer states that different calls are often associated with different valence. For instance, in negative situations, horses squeal, and in positive, they knickers; in the same fashion, human laughter usually indicates positive emotions and crying, negative ones. There are also call types uttered in positive and negative contexts, like the widespread contact calls (horses' whinnies, goats' bleats, and pigs' grunts) with different functions that sometimes are misled.

Briefer tracked behavioral reactions of human listeners to playback trials with emotional valence sounds. During the recordings of negative and positive calls, every individual's behavioral and physiological responses were observed to validate the arousal and valence of the emotion triggered. The selection of the contexts triggering the emotional vocalizations respected the assumption that positive situations were associated with increased fitness in the wild. In contrast, the opposite ones showed a decrease. Briefer used the separation or isolation paradigms to the negative emotions experienced and the reunion and food reward to positive emotions. The emotional

valence was assessed by the arousal intensity and heart rate (in the domestic species) or locomotion rate (in the wild species).

The computation of the behavioral responses was possible due to the procedure described above. Hence, Briefer collected reliable data to measure the differences in the behavioral responses among species as a function of the valence of the context. She illustrated the behavioral data collection showing the cases of horses and goats. In horses, the main indicator of valence was the height of the head being higher in negative than in positive situations. By contrast, goats displayed their tails more often in positive situations. However, Briefer pointed to the procedure's limitations for studying the vocal contagion of emotions in goats and horses. She shared her concerns about the difficulties of matching behavioral responses with the indicators of positive emotions during playback trials with listeners of the positive calls. The same difficulties occurred with the indicators of negative emotions during playback of negative calls.

As a final statement, Briefer described the main accomplishment in her bioacoustic work. She highlighted the vocal parameters found to change with valence in each ungulate species of her long-term study. Looking ahead, Briefer also plans to artificially modify the acoustic parameters to investigate which one is more determinant for valence perception. She has already taken the first step by running playback experiments with goats, pigs, and horses using natural positive and negative sounds. Preliminarily, she discovered that the animals perceive the difference in the middle of the several playback treatments. She emphasized that horses produce two partially independent fundamental frequencies in their whinnies: a phenomenon called biphonation. The lowest one, F0, indicates arousal, while the other, G0, indicates valence. Considering this rare acoustic feature in mammal vocalizations, Briefer pursues the specific function of each of these two frequencies. She expects to contribute very soon with new insights on horses' emotional life.

Part B Human Bioacoustics

Chapter 5

Nonverbal acoustic communication from a psychoethological perspective

Emma Otta[7]

Abstract

The focus of this chapter is on acoustic nonverbal communication and the expression of emotions from a psychoethological perspective. Emotionally modulated speech, indicated by paralinguistic cues – such as speaking rate, tone of voice, and intonation contour – and vocal expressions – such as sobbing, screaming, and laughing – convey information about their sender, whether intentionally or unintentionally. Some paralinguistic cues and vocal cues may be used by the receiver to quickly infer the internal state of the sender, his/her intentions and his/her ensuing behavior, thus influencing the regulation of the interpersonal interaction. In this chapter, I will begin by presenting a definition of emotion and a historical contextualization. The chapter presents evidence of interspecific universals in emotional vocalizations. It distinguishes dimensional and categorical approaches to acoustically transmitted emotions. Evidence for both the universality and the cultural specificity of the detection of emotions from speech and human vocalizations will be presented. Unanswered questions and emerging topics will be pointed out throughout the chapter.

Keywords: Emotions, display, intentionality, paralinguistics, vocalizations.

Defining emotion

I will start by defining emotion from a functional perspective; this perspective reflects the "wisdom of ages" (Lazarus & Lazarus, 1994), as emotion has evolved as an adaptive mechanism through the process of evolution. I will use de Waal's definition, which, in my opinion, is especially appropriate in articulating proximate and distal levels of explanation (de Waal, 2011, p. 194):

[7] Department of Experimental Psychology of the Psychology Institute, University of São Paulo, SP, Brazil. emmaotta@usp.br

> "An emotion is a temporary state brought about by biologically relevant external stimuli, whether aversive or attractive. The emotion is marked by specific changes in the organism's body and mind – brain, hormones, muscles, viscera, heart, etcetera. Which emotion is triggered is often predictable by the situation in which the organism finds itself, and can further be inferred from behavioral changes and evolved communication signals. There exists no one-on-one relation between an emotion and ensuing behavior, however. Emotions combine with individual experience and cognitive assessment of the situation to prepare the organism for an optimal response. (…) Organisms have been selected to enter a particular bodily and mental state under particular circumstances: those who did furthered their interest better than those who did not."

Emotions are triggered by significant life events on the basis of an individual's needs and well-being. According to the component process model (CPM) of emotion (Scherer, 2009), five components are involved: (i) *evaluation of the event* – appraisal of the situation; (ii) *physiological activation* – preparing the body for action; (ii) *expressive movements* – vocal expressions, body postures, and facial expressions; (iii) *sense of purpose* – motivational state directed towards a goal; and (iv) *feelings* – subjective experience (Figure 5.1).

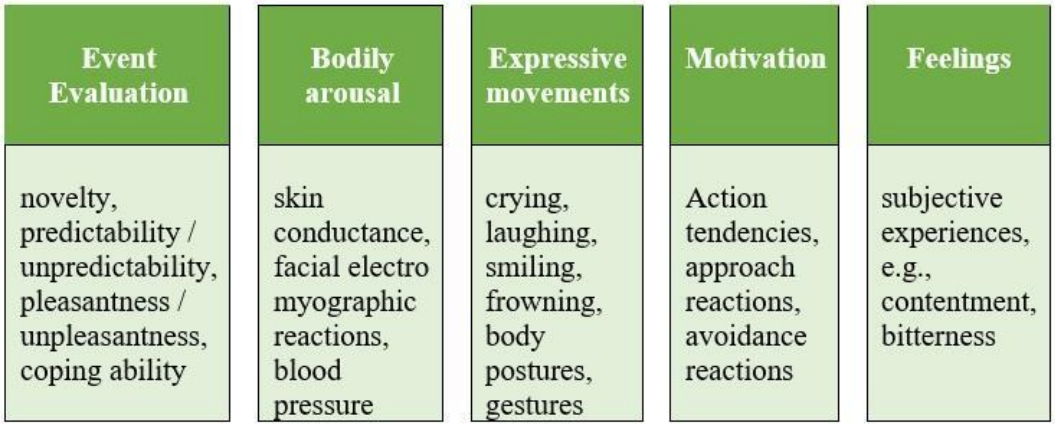


Figure 5.1. The five components of emotion (Source: after Sander et al., 2018).

Historical contextualization

Charles Darwin (1872) wrote about the expression of emotions in his book *The Expression of the Emotions in Man and Animals*. He enunciated the principle of

antithesis according to which certain mental states trigger a strong and involuntary tendency to perform certain movements. The induction of an opposite mental state triggers an equally strong and involuntary tendency to perform antagonistic movements. He illustrated the principle of antithesis with a dog in a threatening body posture, in which the animal appears larger and stronger, and in an appeasing body posture, in which it appears smaller and weaker (Figure 5.2).

> "When a dog approaches a strange dog or man in a savage or hostile frame of mind he walks upright and very stiffly; his head is slightly raised (…); the tail is held erect and quite rigid; the hairs bristle, especially along the neck and back; the pricked ears are directed forward, and the eyes have a fixed stare. (…) These actions follow from the dog's intention to attack his enemy. (…) As he prepares to spring with a savage growl on his enemy, the canine teeth are uncovered. ..." (Darwin, 1872, pp. 50-51)

Darwin then invites us to imagine that the dog suddenly notices that the individual he is approaching is not a stranger but his master. In this case, the dog's behavior changes immediately to the opposite.

> "Instead of walking upright, the body sinks downwards or even crouches, and is thrown into flexuous movements; his tail, instead of being held stiff and upright, is lowered and wagged from side to side; his hair instantly becomes smooth; his ears are depressed and drawn backwards, but not closely to the head; and his lips hang loosely. (…) the eyelids become elongated, and the eyes no longer appear round and staring." (p. 51)

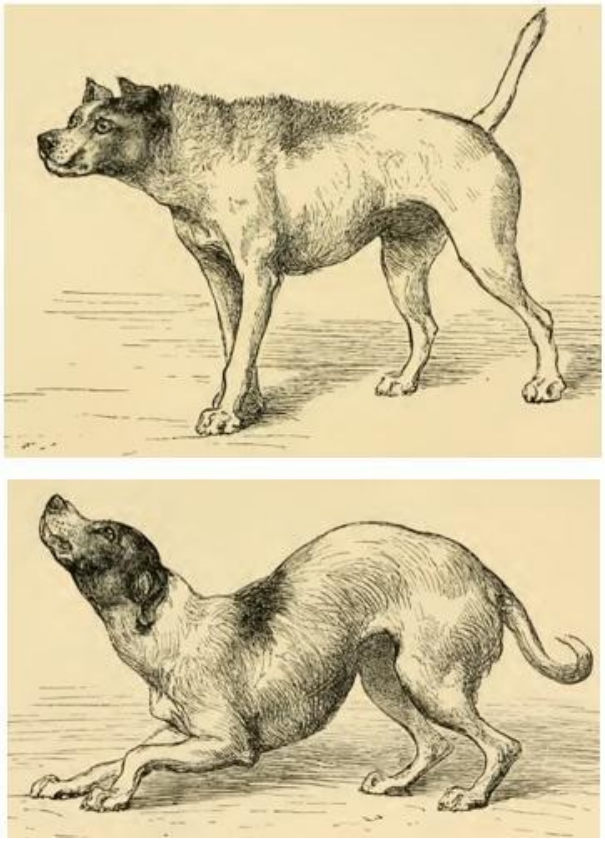


Figure 5.2. Principle of antithesis illustrated by a dog in threatening and appeasing postures (Source: Darwin, 1872, pp. 52-53). http://darwin-online.org.uk/converted/pdf/1897_Expression_F1152.pdf

Approximately 100 years after *The Expression of the Emotions in Man and Animals*, the subject was revisited by Morton (1977, 1983), who compared the physical structure of vocalizations used in the communication of mammals and birds at short distances and the underlying motivation. Animals indicate their mood through vocal sounds. Based on Darwin's antithesis principle, Morton argued that natural selection results in a structural convergence of sounds used in "hostile" and "friendly" contexts. An individual who tries to drive another away uses harsh (broadband), low-frequency sounds. If the receiver does not move away, the interaction may escalate into attack, indicating the sender's aggressive disposition. We can speculate that evolution may have favored roaring in aggressive animals because rough low-frequency vocalizations make the vocalizing individual appear larger and more threatening in the face of a rival (Morton, 1983). In threatening contexts, animals use this type of vocalization, while in friendly contexts, they use relatively tonal, high-frequency sounds. Sound characteristics (harsh quality, tonal quality, and sound frequency [pitch]) interact: 1)

the lower the frequency, the more hostile the sender, and the higher the frequency, the friendlier or more fearful the sender; 2) the greater the harshness, the more hostile the sender, and the more pure and tone-like, the friendlier or more fearful the sender; and 3) a decrease in frequency indicates hostility, and an increase in frequency indicates fear.

In line with Morton (1977, 1983), Ohala (1983) also proposed an innately specified "frequency code", relating the primary meaning of "large vocalizer" and the secondary meanings "dominant, aggressive, and threatening" to low acoustic frequency and the primary meaning of "small vocalizer" and the secondary meanings "subordinate, submissive, non-threatening, and desirous of the receiver's goodwill" to high acoustic frequency. Chuenwattanapranithi et al. (2008) obtained support for the size code hypothesis of emotional speech for two emotions – anger and happiness. Human listeners judged the body size and emotion of the speaker. Thai listeners heard speech sounds produced with a lengthened vocal tract, lowered F0, and roughened voice quality as spoken by an angry individual, and speech sounds produced with a shortened vocal tract, raised F0, and tone-like voice quality as spoken by a happy person.

Inferences about mood in animals can be made based on observations of sender and receiver behavior.

> "This is not to say that animals feel angry when they growl, for we have no idea whether our feelings are the same as the motivation in animals. But by observing what happens when an animal growls (or squeals), we can use words such as "aggressive" or "fearful" to describe whether it will probably attack or flee when it growls or squeals." (Morton, 1983, p. 345)

Returning to Darwin's example to illustrate the principle of antithesis, Morton (1983) added acoustic communication to visual communication.

> "Your pet dog is sleeping on the front porch. As you approach, Fido wakes up and begins barking. The bark means that Fido has perceived something of interest to him but the stimulus is too far away for him to make a "decision." Should he attack or be friendly? When you get closer, or yell his name, he changes from barking to whines, sleeks his fur, and wags his tail at a low angle. On the other hand, if the mailman had elicited the barks, Fido might begin to growl as he approached. It is clear from Fido's actions what moods he exhibited through his vocalizations. (…)

> When a dog growls, it also makes itself visually larger by erecting its fur; when it whines, it sleeks its fur and hunches down to look smaller." (p. 347)

There is redundancy in communication through the auditory and visual channels, and this redundancy may increase the odds of communicating effectively. In humans, the same general relationship exists between the physical structures of sounds and the underlying motivation (hostility or appeasement). One person can say "Go away!" in different ways to another, but when he/she feels truly angry, his/her feelings may be expressed as "growling". Intonation adds information to the content of his/her speech, making it more emphatic. A low or falling voice expresses aggressiveness and assertiveness, while a high or rising voice expresses friendly intentions (Morton, 1983).

Darwin's Musical Protolanguage Hypothesis

Darwin developed a model of language evolution, known as the musical protolanguage hypothesis. Figure 5.3 shows a schematic outline of this hypothesis. In chapter 4 of *The Expression of the Emotions in Man and Animals*, Darwin addressed sound emission and its use as a means of expressing courtship and rivalry. He hypothesized that our ancestors used vocal utterances to express emotions before they had acquired the power to articulate speech in human evolution. In Chapter 3 of *The Descent of Man, and Selection in Relation to Sex*, he proposed a model of language evolution based on a protolanguage that was more musical than linguistic and focused on sexual selection as the underlying mechanism of this evolution.

Darwin (1871, p. 87) stated:

> "I cannot doubt that language owes its origin to the imitation and modification of various natural sounds, the voices of other animals, and man's own instinctive cries, aided by signs and gestures. When we treat of sexual selection we shall see that primeval man, or rather some early progenitor of man, probably first used his voice in producing true musical cadences, that is in singing, as do some of the gibbon-apes at the present day; and we may conclude from a widely spread analogy, that this power would have been especially exerted during the courtship of the sexes, would have expressed various emotions, such as love, jealousy, triumph, and would have served as a challenge to rivals."

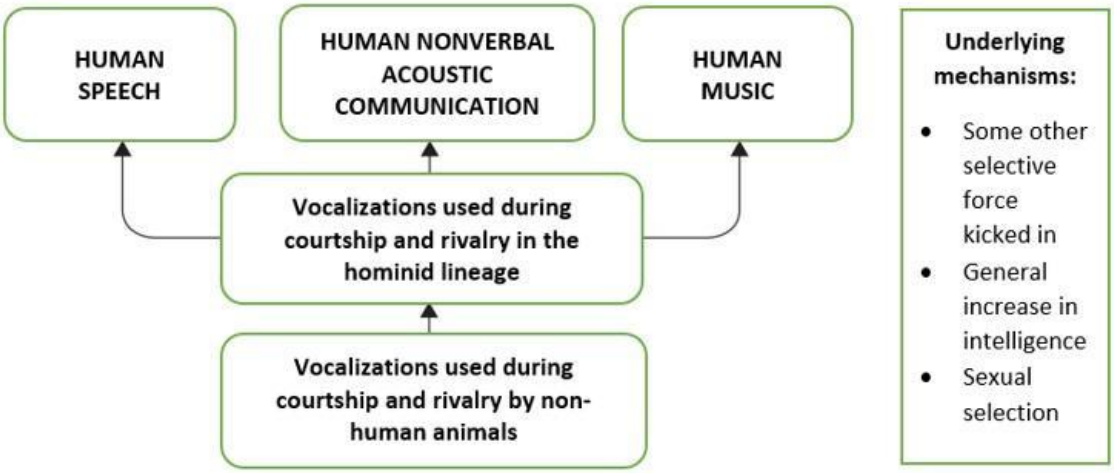


Figure 5.3. Schematic outline of Darwin's musical protolanguage hypothesis (Based on Zimmerman, Leliveld & Schehka, 2013 and Fitch, 2013***).***

Fitch (2010, 2013) summarized and updated this model. The first step was a general increase in intelligence in the hominid lineage. Before vocalizations were used meaningfully, they were used, "so to speak", aesthetically, to fulfil many of the same functions that modern humans use music today (courtship, bonding, territorial advertisement and defense, competitive displays, etc.). (…) at—a later stage (presumably during the evolution of meaningful language) some other selective force kicked in, so that language became equally (or better) expressed in females, and was pushed to develop early" Fitch (2013, p. 494, 497).

In his book *The Singing Neanderthals*, Mithen (2005) also proposed a proto-music/language model. He created the acronym *Hmmmm* to name his model for the early hominid communication system: **h**olistic utterances with their own meaning (not constituted by segmented elements or words), **m**anipulative (with the ability to change the affective states and behavior of others), **m**ultimodal (employing both sounds and movements), **m**usical (rhythmic and melodic), and **m**imetic (uses sound symbolism and gesture).

In search of interspecific universals in emotional vocalizations

Based on Darwin's (1871, 1872) hypothesis that the vocal expression of emotion has ancient roots and Morton's (1977) model on the association of motivation and structural rules of vocalizations, researchers began to search for interspecific

universals in emotional vocalizations. Fillipi et al. (2017) conducted a study in which they asked human participants to evaluate the emotional content of the vocalizations of nine species across three taxonomic classes: Amphibia, Mammalia, and Reptilia (including Aves). They found that speakers of different languages (native speakers of English, German and Mandarin, N=25 in each group) were able to identify increased levels of activation in the vocalizations of all species represented. We conclude from this study that humans are able to identify the emotional content in both conspecific and heterospecific vocalizations, suggesting a biological basis. Basic mechanisms underlying emotion perception in vocalizations may have appeared early during phylogenesis and have been evolutionarily conserved across species.

Cross-species perception of emotion from vocal and visual cues has also been investigated with dogs. In a research study for her master's thesis developed under my supervision as part of the Project Anthrozoo USP at the Postgraduate Program in Experimental Psychology at the University of São Paulo, with the co-supervision of Daniels Mills, Natalia Albuquerque (2013) used an intermodal preferential looking paradigm. Domestic dogs (*Canis familiaris*) were presented with human faces or dog faces with different emotional valences (happy/playful versus angry/aggressive). While the stimuli were projected onto two screens, a single sound was played. The sound could be a dog's bark, a human voice with either positive or negative valence or a control sound (neutral). We found that dogs looked significantly longer at both conspecific and heterospecific faces whose expression corresponded to the valence of the vocalization, leading to the conclusion that dogs have the mental prototypes for positive versus negative categorization of affect and that they are able to integrate acoustic and facial emotional information (Albuquerque et al., 2016). Additionally, the dogs seemed to have a functional understanding of emotional expressions. When they looked at angry human faces, they were more likely to mouth-lick than when they looked at happy human faces (Albuquerque et al., 2018). Mouth-licking may serve as an appeasement signal in dog-human communication (Firnkes, Bartels, Bidoli & Erhard, 2017). It is at the lowest step of Shepherd's ladder of distress signals. Shepherd (2002, 2009) attributes to this behavior the function of a calming signal that defuses conflict and restores harmony in a social interaction. Dogs are the oldest domestic animals, having lived with humans for approximately 10,000 years (Larson, Karlson, & Perri, 2012). It might have been particularly advantageous for them to recognize the emotions of humans and to react with appropriate behaviors during the process of

domestication, in which they may have evolved the ability to read human communication cues.

Dimensional and Categorical Approaches to Acoustically Conveyed Emotions

The chapter proceeds by addressing acoustically transmitted emotions in humans. Emotionally modulated speech and vocal expressions are evaluated in terms of either dimensional or categorical approaches. Self-report instruments linked to these approaches will be presented.

At the beginning of the twentieth century, Wilhelm Wundt (1905), the father of experimental psychology, first proposed a dimensional approach according to which emotions are characterized by their place in a three-dimensional space made up of "arousal-calm", "pleasure-displeasure", and "relaxation-tension". At the end of the twentieth century, Russel (1980) revisited the subject and proposed a circumplex model to express the structure of affect as evaluated by self-report. In this model, affective concepts fell in the circle in the following order: pleasure (0°), excitement (45°), arousal (90°), distress (135°), displeasure (180°), depression (225°), sleepiness (270°), and relaxation (315°). Bradley and Lang (1994) proposed the self-assessment manikin (SAM) as an easy-to-use nonverbal method for evaluating emotional reactions to a wide variety of stimuli, including sounds (e.g., International Affective Digitized Sound [IADS] System; Bradley & Lang, 2007), in terms of pleasure, arousal, and dominance. The pleasure scale shows a smiling figure at one end and a frowning figure at the other. The arousal scale shows a sleepy figure at one end and a wide-eyed figure at the other. The dominance scale shows a small figure at one and a large figure at the other. There was a 9-point rating scale for each dimension: research participants were asked to choose one of five figures in each scale or to place a mark between any two figures. The original paper-and-pencil self-report version evolved into a digital slider version used in smartphones and tablets (Betella & Verschure, 2016). Below each slider are two mirrored isosceles triangles that provide a visual cue for intensity.

Another approach to the study of acoustically conveyed emotions is the categorical emotional approach. At the end of the nineteenth century, Darwin (1872) first proposed a discrete emotions perspective with a main focus on the face, suggesting that facial expressions of emotion are universal. In contrast to Wundt (1905), he considered emotions distinct entities or modules, such as happiness, sadness, fear, anger, disgust, and surprise. Darwin's theory was extended by Ekman

(2009) and Panksepp (1998), assuming that basic emotions have unique characteristics that distinguish them from one another in important ways (behavioral and physiological reactions driven by specific neural reaction systems). These basic emotions are affect programs, phylogenetically evolved adaptation patterns activated by relevant eliciting events. In line with the predictions of the basic emotion theorists, humans assign facial and vocal expressions of emotion to discrete emotion categories with high accuracy. The Product Emotion Measurement Instrument (PrEmo) is an instrument based on the categorical emotional approach (Desmet, 2019). It is a pictorial self-report instrument in which a character expresses 14 different emotions by his/her expressions, body, and voice. Half of the emotions are positive (joy, hope, pride, admiration, satisfaction, fascination and attraction), and half are negative (sadness, fear, shame, contempt, dissatisfaction, boredom and disgust). In addition to basic emotions, this instrument includes social emotions that presuppose the ability to think about emotions and behavior from the point of view of another individual.

Mendl, Burman and Paul (2010) combined dimensional and categorical approaches for the study of emotions. Figure 5.4 schematically represents emotions in a two-dimensional space summarizing a dimensional approach (valence and activation) with an approach of discrete emotions (e.g., happiness, sadness, and fear). The positive emotions are placed in quadrants Q1 and Q2, and the negative emotions are placed in quadrants Q3 and Q4. The Q3-Q1 arrow represents the motivational system of reward acquisition related to increased fitness. The Q4-Q2 arrow represents the motivational system of punishment avoidance triggered by the perception of danger or threat.

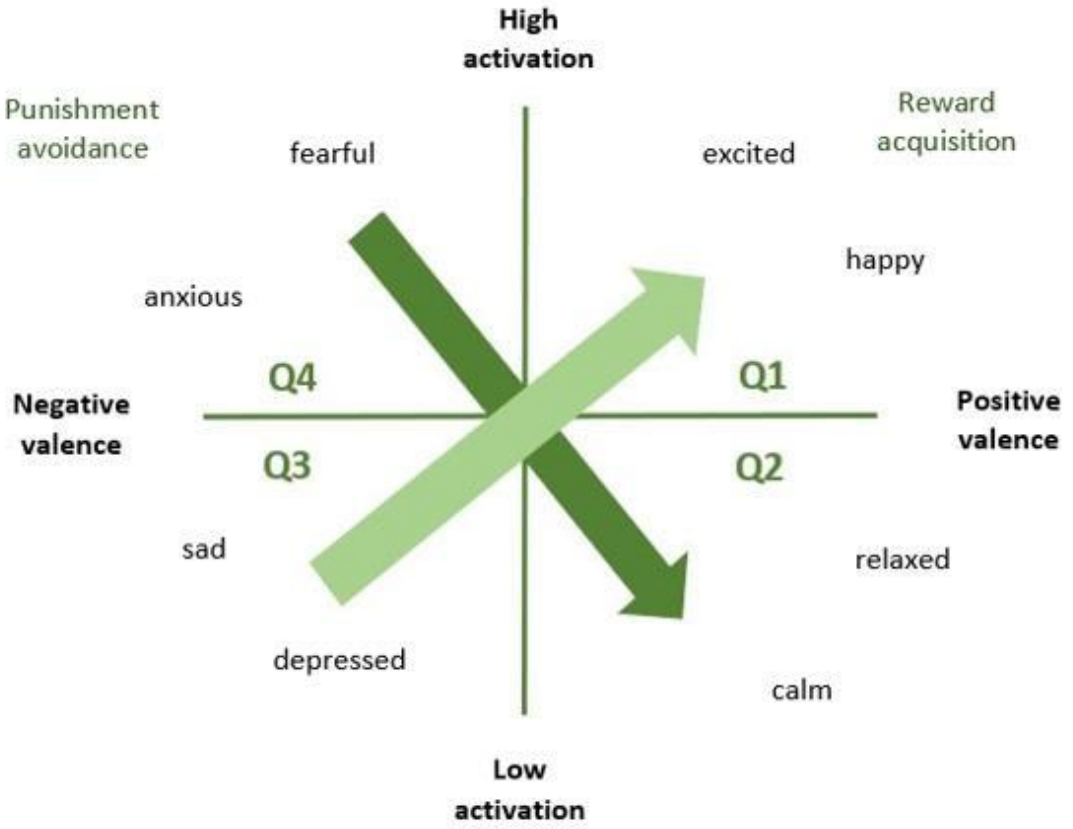


Figure 4. Discrete emotions located in a two-dimensional space of valence and activation (adapted from Mendl et al., 2010).

Stimulus materials

Researchers have developed a variety of stimulus materials that will be illustrated here. Affective prosody recognition tasks are used for research purposes and for clinical practice purposes. Individuals listen to neutral sentences (e.g., "*The girl went to the market"*) spoken in affective tones of voice (happy, sad, angry, etc.) and are asked to name the emotional prosody. The Brazilian version of the Florida Affect Battery (Bowers, Blonder & Heilman, 1999) was adapted from the English original by two researchers from the University of Brasilia (Costa-Vieira & Souza, 2014). In the naming of emotional prosody tasks, there are 20 trials with four repetitions of each of five affects. The Florida Affect Battery also includes tasks in which there is conflict between the semantic content and adequate voice intonation. An example is the sentence "*All the puppies are dead*" said in a happy tone of voice. To correctly evaluate

the affective tone of voice of the speaker, the listener must disregard the content of the message. The Florida Affect Battery also includes tasks in which there is conflict between the semantic context and adequate voice intonation. Additionally, the battery includes cross-modal facial prosody tasks in which individuals are required to match the affect conveyed by a prosodic stimulus with a corresponding facial expression or vice versa.

With the purpose of studying real emotional expressions, Silva, Barbosa and Abelin (2016) selected speech samples of Brazilian Portuguese from the documentary *Jogo de Cena* (*Playing*; Coutinho, 2007). The director had placed an announcement in a newspaper inviting women to tell their stories in his documentary, sharing their joys and sorrows. The film alternates between these women and actresses. There are few studies using real-life voice records, and this study is interesting for this reason.

Further materials used for research focus on vocalizations instead of speech. The Montreal Affective Voices (Belin, Fillion-Bilodeau & Gosselin, 2008) consists of 90 vocalizations corresponding to the emotions of anger, disgust, fear, pain, sadness, surprise, joy and pleasure. Sauter et al. (2010) also created a set of nonverbal vocalizations of negative and positive emotions, such as laughter, an angry growl, retching sounds, screams of fear, moans of sexual pleasure, and sighs of relief.

Detection of emotions from speech and human vocalizations

This section of the chapter will present research findings on the ability of listeners to identify emotions from speakers' voices. Comparisons have been made between accuracy in the recognition of emotion in the face and in the voice. While joy can be almost perfectly identified from facial expressions, listeners have difficulty recognizing this emotion unequivocally in the voice (Scherer, 2003). Anger and sadness were best recognized in the voice, followed by fear, whereas disgust was identified just above chance level.

Comparisons are also made across languages and cultures. In Silva, Barbosa and Abelin (2016), Brazilian and Swedish listeners evaluated authentic speech samples extracted from the Brazilian documentary *Jogo de Cena* and from Swedish television and interview programs. The Swedish and Brazilian listeners evaluated both corpora similarly, leading to the conclusion that the listeners' native language did not influence their perception of the emotions expressed by the speakers.

A meta-analysis conducted by Bąk (2016) showed that research on emotional prosody evaluated from speech is focused on English: 79% of the participants were Germanic language speakers, with a predominance of English; 15% were Japanese speakers; and 6% were speakers of other languages, including Portuguese, Spanish, Arabic, Hindi and Hinba. This meta-analysis showed that humans infer affective states from the emotional prosody of speech in different cultures, although the listeners cannot understand the words and sentences voiced by the speakers (Pell, Monetta, Ekmann & Kotz, 2009). The recognition of basic emotions (e.g., joy, sadness) was superior to chance, but there was an own-language advantage in comparison to foreign languages.

A study on intercultural recognition of basic emotions through nonverbal vocalizations comparing representatives of maximally different populations in terms of language and culture was conducted by Sauter, Eisner, Ekman and Scott (2010). This research team, integrated by a proponent of the "Big Six" basic emotions, compared Himba people from Namibia, who live by herding without contact with Western culture, with English-speaking Europeans. Emotional vocalizations corresponding to the basic emotions of joy, fear, anger, sadness, disgust and surprise communicated the same emotional states regardless of culture, leading to the conclusion that they are universals shared by all humans. In this study, however, the researchers also found a significant interaction between the culture of the listener and the emitter of vocalization, showing that each group performed better in relation to the stimuli produced by members of its own culture. Some social emotions (e.g., pride) were recognized only within culture. Negative emotions were recognized between cultures, but several positive emotions were communicated by culturally specific signals. The researchers concluded that affiliative social signals were shared mainly with in-group members.

In addition to self-report measures, psychophysiological measures have been used to investigate the processing of emotional prosody, including event-related potentials (ERPs), which identify specific brain activity by means of EEG in the presence of speech or vocalization samples. Basic emotions are differentiated by means of the P200 component captured by electrodes located in frontal-central positions. The differentiation between emotional speech and neutral speech occurs in an initial time window of 170 to 230 ms after the onset of stimulus (Paulmann, Bleichner, & Kotz, 2013; Schirmer, Simpson, & Escoffier, 2007). Through ERPs the

differential reactions of humans to fear vocalizations compared to control sounds 150 msec after the onset of the stimulus were demonstrated (Sauter & Eimer, 2010). Rapid detection of affective signals can be important. If conspecific others in the surroundings are frightened, staying alert and preparing for imminent danger may have survival advantages.

Concluding remarks

In this chapter, several topics related to acoustic nonverbal communication and the encoding and decoding of emotions have been addressed from a psychoethological perspective. Although scientists still struggle to define emotions, an operational definition was proposed, taking into account proximate and ultimate levels of causation. Research perspectives were presented based on the categorical approach to emotion, which goes back to Charles Darwin, pioneer in the study of emotions, and the dimensional approach to emotion, which goes back to Wundt, founder of the first psychology laboratory. A review was conducted of research findings of universals in nonhuman emotional vocalizations and cross-cultural recognition of basic emotions among humans. The present chapter shows evidence of interspecific universals and cross-cultural recognition of emotions and of an in-group advantage in the understanding of emotion. A universal acoustic affect program (Sauter et al., 2010) seems to coexist with culturally specific affect programs (Elfenbein & Ambady, 2003).

Acknowledgments

The author acknowledges grant no. 304740/2017-9 of Conselho Nacional de Desenvolvimento Científico e Tecnológico (The Brazilian National Council for Scientific and Technological Development – CNPq) and no. 2014/50282-5 of São Paulo Research Foundation (FAPESP) and Natura Cosméticos S.A. The author would also like to thank Patricia Monticelli for her comments and suggestions on initial versions of the text, which improved its quality.

References

Albuquerque, N. S. (2013). *Reconhecimento de emoções em cães domésticos (Canis familiaris): percepção de pistas faciais e auditivas na comunicação intra e interespecífica*. [Emotion recognition in domestic dogs (*Canis familiaris*):

Perception of facial and auditory cues in intra and interspecific communication] Instituto de Psicologia da USP: Dissertação de Mestrado.

Albuquerque, N., Guo, K., Wilkinson, A., Savalli, C., Otta, E., & Mills, D. (2016). Dogs recognize dog and human emotions. *Biology letters*, *12*(1), 20150883. https://doi.org/10.1098/rsbl.2015.0883

Albuquerque, N., Guo, K., Wilkinson, A., Resende, B. & Mills, D. S. (2018). Mouth-licking by dogs as a response to emotional stimuli. *Behavioural Processes*, *146*, 42-45. http://dx.doi.org/10.1016/j.beproc.2017.11.006

Bąk, H. (2016). The state of emotional prosody research—a meta-analysis. *In Emotional Prosody Processing for Non-Native English Speakers* (pp. 79-115). Springer, Cham.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: a validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40(2)*, 531-539. https://doi.org/10.3758/BRM.40.2.531

Betella, A. & Verschure, P. F. M. J. (2016) The Affective Slider: A Digital Self-Assessment Scale for the Measurement of Human Emotions. *PLoS ONE, 11* (2): e0148037. https://doi.org/10.1371/journal.pone.0148037

Bowers, D., Blonder, L. X., & Heilman, K. M. (1999). Florida affect battery, a manual. Centre for Neuropsychological Studies. *Cognitive Science Laboratory, University of Florida*, 3-19.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49-59. https://doi.org/10.1016/0005-7916(94)90063-9

Costa-Vieira, H. A., & Souza, W. C. D. (2014). O reconhecimento de expressões faciais e prosódia emocional: Investigação preliminar em uma amostra brasileira jovem. *Estudos de Psicologia (Natal)*, *19*(2), 119-127. http://dx.doi.org/10.1590/S1413-294X2014000200004

Chuenwattanapranithi, S., Xu, Y., Thipakorn, B., & Maneewongvatana, S. (2008). Encoding emotions in speech with the size code. *Phonetica*, *65*(4), 210-230. https://doi.org/10.1159/000192793

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*, 1st ed.; London, UK: John Murray.

Darwin, C. (1965). *The expression of the emotions in man and animals.* Chicago: University of Chicago Press. (Original work published 1872)

Desmet, P. M. A. (2019). PrEmo card set: Male version. Delft, Delft University of Technology. ISBN: 978-94-6384-076-7.

De Waal, F. B. (2011). What is an animal emotion? *Annals of the New York Academy of Sciences*, *1224*(1), 191-206. https://doi.org/10.1111/j.1749-6632.2010.05912.x

Ekman, P. (2009). Darwin's contributions to our understanding of emotional expressions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3449-3451. https://doi.org/10.1098/rstb.2009.0189

Elfenbein, H. A., & Ambady, N. (2003). Universals and cultural differences in recognizing emotions. *Current Directions in Psychological Science*, *12*(5), 159-164. https://doi.org/10.1111/1467-8721.01252

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., ... & Güntürkün, O. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1859), 20170990. https://doi.org/10.1098/rspb.2017.0990

Firnkes, A., Bartels, A., Bidoli, E., & Erhard, M. (2017). Appeasement signals used by dogs during dog–human communication. *Journal of Veterinary Behavior*, *19*, 35-44. http://dx.doi.org/10.1016/j.jveb.2016.12.012

Fitch, W. T. (2010). *The evolution of language*. Cambridge: Cambridge University Press.

Fitch, W. T. (2013). Musical Protolanguage: Darwin's Theory of Language Evolution Revisited. In Bolhuis, J.J. & Everaert, M., (Eds.) *Birdsong, Speech, and Language: Exploring the Evolution of Mind and Brain* (pp. 489–503). Cambridge, MA: The MIT Press.

Larson, G., Karlson, E. K., and Perri, A (2012) Rethinking dog domestication by integrating genetics, archeology, and biogeography. *PNAS, 109*(23), 8878-8883. https://doi.org/10.1073/pnas.1203005109

Lazarus, R. & B. Lazarus. 1994. *Passion and Reason*. New York: Oxford University Press.

Mendl, M., Burman, O. H., & Paul, E. S. (2010). An integrative and functional framework for the study of animal emotion and mood. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1696), 2895-2904. https://doi.org/10.1098/rspb.2010.0303

Mithen, S. (2005). *The Singing Neanderthals: The Origins of Music, Language, Mind, and Body*. London: Weidenfeld & Nicholson.

Morton, E. S. (1977) On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *American Naturalist, 111* (981), 855– 869. https://doi.org/10.1086/283219

Morton, E. S. (1983). Animal communication: What do Animals say? *The American Biology Teacher*, *45*(6), 343-348. https://doi.org/10.2307/4447717

Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F_0 of voice. *Phonetica*, *41*(1), 1-16. https://doi.org/10.1159/000261706

Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions*. London, England: Oxford University Press.

Paulmann, S., Bleichner, M., & Kotz, S. A. (2013). Valence, arousal, and task effects in emotional prosody processing. *Frontiers in Psychology*, *4*, 345. https://dx.doi.org/10.3389%2Ffpsyg.2013.00345

Pell, M. D., Monetta, L., Paulmann, S. & Kotz, S. A. (2009) Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*, 107–120. https://doi.org/10.1007/s10919-008-0065-7

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. https://doi.org/10.1037/h0077714

Sander, D., Grandjean, D., & Scherer, K. R. (2018). An appraisal-driven componential approach to the emotional brain. *Emotion Review*, *10*(3), 219-231. https://doi.org/10.1177/1754073918765653

Sauter, D. A., & Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, *22*(3), 474-481. https://doi.org/10.1162/jocn.2009.21215

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408-2412. https://doi.org/10.1073/pnas.0908239106

Schirmer, A., Simpson, E., & Escoffier, N. (2007). Listen up! Processing of intensity change differs for vocal and nonvocal sounds. *Brain Research*, *1176*, 103–112. https://doi.org/10.1016/j.brainres.2007.08.008

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1-2), 227-256. https://doi.org/10.1016/S0167-6393(02)00084-5

Scherer, K. R. (2009) The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*(7), 1307-1351. https://doi.org/10.1080/02699930902928969

Shepherd K. (2002). Development of behavior, social behavior and communication in dogs. In Horwitz D, Mills D, and Heath S. (Eds.). BSAVA *Manual of Canine and Feline Behaviour Medicine (pp. 8–20)*. Gloucester, UK: British Small Animal Veterinary Association.

Shepherd K. (2009). Behavioural medicine as an integral part of veterinary practice. In Horwitz D. and Mills D. (Eds.) BSAVA Manual of Canine and Feline Behaviour (pp 10–23), 2nd. ed.https://doi.org/10.22233/9781905319879.2

Silva, W. D., Barbosa, P. A., & Abelin, Å. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with Brazilian and Swedish listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *32*(2), 449-480. https://doi.org/10.1590/0102-445003263701432483

Wundt, Wilheim (1905). *Grundriss der Psychologie*. [*Outlines of psychology*]. Leipzig: Wilheim Engelmann.

Zimmermann, E., Leliveld, L., & Schehka, S. (2013). Toward the evolutionary roots of affective prosody in human acoustic communication: a comparative approach to mammalian voices. Eckart Altenmuller, Sabine Schmidt, Elke Zimmermann & Eckart Altenmüller (Eds.) *Evolution of emotional communication: from sounds in nonhuman mammals to speech and music in man* (pp. 116-132). Oxford: Oxford University Press.

Peer Commentary

By Sylvia Corte

Emma Otta is a Professor at the Department of Experimental Psychology, Institute of Psychology (IPUSP), of the University of São Paulo, São Paulo, SP, Brazil. I will comment on some of the aspects discussed by Professor Otta on a topic sometimes forgotten but with profound implications for acoustic communication and its emotional components, acoustic nonverbal communication. She approaches the topic from a psychoethological perspective, explaining emotions from a functional perspective and giving evidence of interspecific universals in the interpretation of emotions, for instance, the sounds emitted during laughter and crying in human babies. Otta refers to some research that showed that babies laughed in some situations that also triggered crying. However, this is a developmental process, where laughing substitutes crying. In addition, the stimuli that triggered laughing became less physical, such as tickling, and more cognitive during development. From an evolutionary perspective, there is a discussion about the contexts in which smiling and laughing appear. For example, smiling occurs in more conflictive situations and laughing in rougher and tumble play situations. Finally, we could study the acoustic manifestations of crying, laughing, and smiling from a longitudinal perspective in children.

Otta commented about the Twin Panel's interest in studying these manifestations in their participants. They are now studying acoustic communication in twins (see Claudio Possani chapter). They recorded twins' voices individually in the lab before COVID-19 and compiled the voices of one hundred pairs, which are currently being decoded using the PRAAT software. The Panel now is also studying children. It will be very interesting to study crying in twins and the development of their nonverbal vocalizations (crying and laughing) and language.

I wanted to inquire about the famous "Neanderthal flute," discovered in a cave in Slovenia, to be exposed to another of the topics exposed during the talk. One study dismissed the artifact as nothing more than a bone that had been chewed on by hyenas, while others argue it was a musical instrument. Could early humans, like the Neanderthal, have such sophisticated ways of communication?

Otta addresses archaeological evidence of wind instruments from the Upper Paleolithic (between 12,000 and 50,000 years ago) in association with modern humans (*Homo sapiens*). Neanderthals, the closest relatives of modern humans, became extinct around 40,000 years ago. A juvenile bear femur with two complete holes from the Middle Paleolithic was found in a cave in Slovenia in 1995. This piece of bone divides opinions. Some experts believe it is a flute, the oldest musical instrument, while others believe it is only a chewed carnivore bone, a pseudo-artifact. Computed microtomography studies show that a Neanderthal-made artifact cannot be ruled out. Although we lack a thorough understanding of Neanderthal behavior, some anthropologists believe they may have been intelligent, self-aware individuals using a primatological analogy. Chimpanzees, for example, drum on hollow trees and have a preference for some music over silence. One could surmise that Neanderthals may also have expressed something similar. Perhaps the Neanderthal flute was a flute and not a chewed carnivore bone.

Next, Professor Otta added some comments regarding another complex issue: how and why human language skills differ from our hominid ancestors and other living hominid species. How language evolved has been debated since Darwin. There are perceptual and cognitive abilities underlying language comprehension and production shared with other perceptual and cognitive processes present in nonhuman animals. However, there are grammar and syntax components that differentiate human language from animal communication systems. Human language refers to external things in the world and objects and events distant in place and time employing arbitrary symbols based on rules for combining these elements. According to Herbert Terrace in his article Noam Chomsky, all animals, ourselves included, communicate, but only humans use language as we do. Nonverbal communication appears earlier in nonhuman development and is also present in nonverbal individuals. Very interesting research had been carried out on children with an underdeveloped brain system. In other words, they have no cortex and communicate nonverbally. The researchers discuss their expectations that these individuals will not communicate. Nevertheless, they have a very rich communication system, characterized by smiling and gazing, for example. So, the role of nonverbal and verbal communication in humans can be studied separately.

Non-verbal and verbal communication in humans is a case of multimodal communication that can be exploited in some situations. We can study how they are

related and how we can learn from this. We can examine this multimodal combination in recent research (Albuquerque, 2013) on dogs: does a dog understand that an angry vocalization accompanies an angry face? The situation is more complex when we are dealing with humans. Professor Otta considered that both animals and humans make sense of the variety of sights, sounds, and affective states. They need to coordinate this input and make associations between one sense (e.g., sight) and another (e.g., sound). In the case of dogs, their behavior suggested that they could relate what they saw, heard, and felt. For example, they licked the nose more frequently when observing a negative facial expression than with a positive valence, especially in response to humans. It was an appeasement gesture, and researchers can hypothesize that it was a reaction to a threat exhibited in a prone conflict situation.

Dogs are very sensitive to human behavior and friendly or threatening communication signals. It is the result of the domestication process that probably involved relatively intense selection for tameness. However, nonverbal and verbal communication is a case of multimodal communication, and the combination of these various dimensions is a very interesting subject.

Chapter 6

Physiology of Vocal Production

Domingos H. Tsuji[8]

Abstract

One of the greatest skills acquired by the human species during its evolutionary development was the ability to formulate complex ideas and communicate them to their peers through speech. This can be defined as a set of vocal sounds used for oral communication between human beings and consists of a set of words that are formed by different phonemes which, in turn, are composed of vowels and consonants. Here, I present the sound production and articulatory apparatus in terms of their morphology and functionality.

Keywords: Human voice, source-filter theory; sound producing, speech,

In an overview, the production of human voice used in speech depends on a central neurological control that, through its peripheral nerves, controls the human speaking apparatus, which consists of several organs and anatomical structures (Figure 6.1). The perfectly coordinated functioning of these anatomical structures guarantees speech production, which, at the peripheral level, depends on exhalation, production of the fundamental sound of the larynx (the vocal fold vibration sound), sound resonance, and articulation (Isshiki, 1989). The purpose of this presentation is to provide basic background information for understanding each of these steps.

[8] Ophthalmology and Otorhinolaryngology Department, Faculty of Medicine of the University of Sao Paulo, SP, Brazil. domingostsuji@terra.com.br

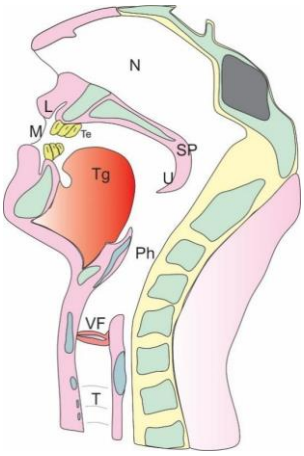


Figure 6.1. Human speaking apparatus.

Exhalation: the primary energy source

Exhalation produces a flow of air molecules that constitute the energetic force for voice production. The physiological mechanisms in this step of vocal production are quite complex; however, it can be said that they depend on adequate motor control of the structures involved in breathing. These structures are the trachea, lungs, bronchi, bronchioles, and alveoli contained in the rib cage or chest cavity. The latter space is delimited by the sternum bone at the front, costal vertebrae, diaphragm, and spine. The main muscles that directly control chest cavity volume are the external and internal intercostal muscles and the diaphragm, which are essential to control inspiration and expiration of pulmonary air (Hansen, 2019). When modifying the volume of the abdominal cavity, the abdominal muscles indirectly influence chest volume and are essential in vocal emission and support for more complex vocal activities such as singing.

The respiratory function as a whole, especially respiratory mechanics, depends on individual anatomophysiological characteristics, such as chest cavity dimensions and the anatomical structures involved, tissue elastic properties, lung volume, respiratory muscle performance and the central nervous system, which coordinates and controls the entire functional process. Thus, it can be considered that everyone has a unique respiratory capacity.

Production of the fundamental sound of the larynx: vocal fold vibration

Anatomically, the larynx is located in the anterior region of the neck (Figure 6.2), contiguous to the trachea, consisting of bone and cartilaginous skeleton, muscles and ligament structures and covered internally by respiratory mucosa. The main cartilages of the larynx are epiglottis, thyroid, cricoid (unpaired) and arytenoid (paired) (Figure 6.3), connected to each other by membranes, ligaments, joints, and muscles. Internally, some membranes and ligaments are covered by mucous tissue on each side, the vocal fold and vestibular fold (false vocal fold), and between them the laryngeal ventricle (sinus of Morgagni). In addition, structures such as the epiglottis, aryepiglottic folds on each side, the arytenoid region, interarytenoid fold, anterior commissure and posterior glottis wall are identified (Figure 6.4).

Figure 6.2. Position of the larynx in the anterior cervical region, connected caudally to the trachea.

With respect to the position of the vocal folds, the laryngeal cavity is divided into three regions or levels: 1 – the glottis, which corresponds to the space delimited by the vocal folds, anterior and posterior commissures; 2 - the supraglottis, corresponding to the region above the vocal folds; 3 – the subglottis or infraglottis, located below the vocal folds, extending to the lower border of the cricoid cartilage (Imamura, Tsuji & Sennes, 2002) (Figure 6.5).

In 1974, Hirano was one of the first to describe the histological structure of the human vocal fold and correlate it with the physiology of vocal production (Hirano, 1975). Based on the layered structure consisting of epithelium, connective tissue and

vocal muscle, he developed the *cover-body theory* for vocal fold vibration, according to which the soft mobile cover, consisting of the mucous membrane, can vibrate over the more rigid and stationary body, composed of the vocal ligament and vocal muscle. The main intrinsic muscles of the larynx are the cricothyroid, lateral and cricoarytenoids, thyroarytenoid and interarytenoid (Figure 6.6).

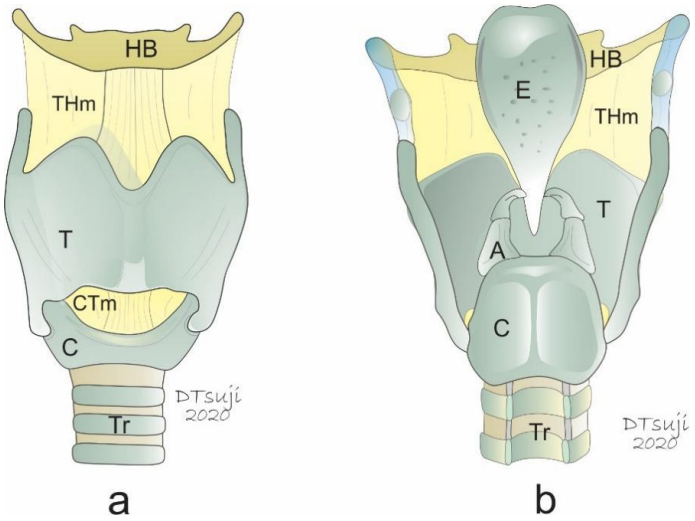


Figure 6.3. Laryngeal skeleton. (a) anterior view; (b) posterior view. HB- Hyoid bone; THm – Thyrohyoid membrane; T- Thyroid cartilage; CTm – Cricothyroid membrane; C- Cricoid cartilage; Tr – Trachea

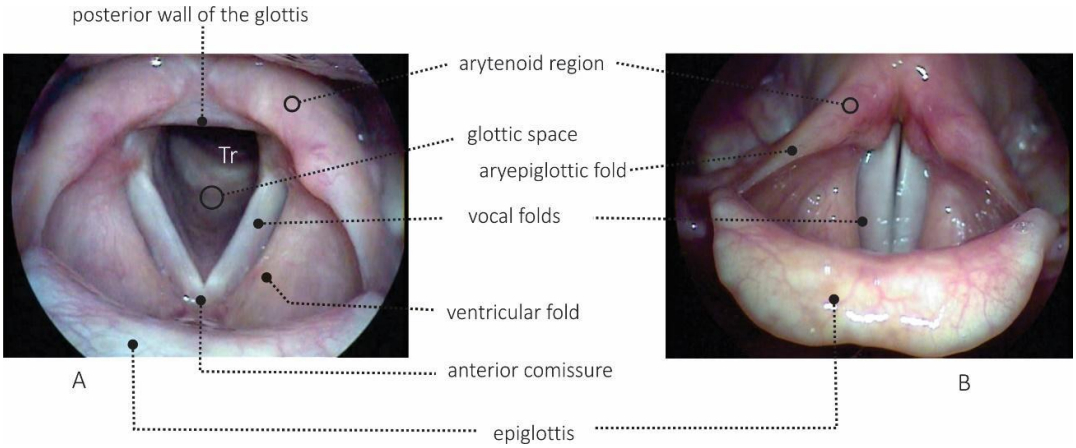


Figure 6.4. Internal view of the larynx. (A) respiratory position of vocal folds; (B) vocal folds in phonatory position.

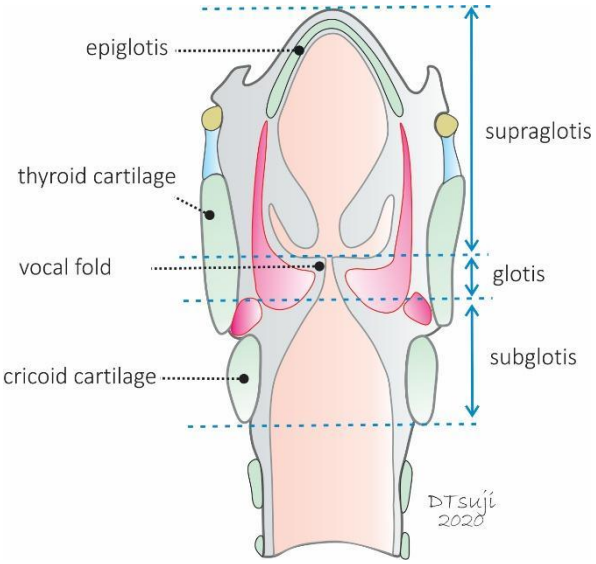


Figure 6.5. Levels of the larynx. The glottis corresponds to the vocal folds level.

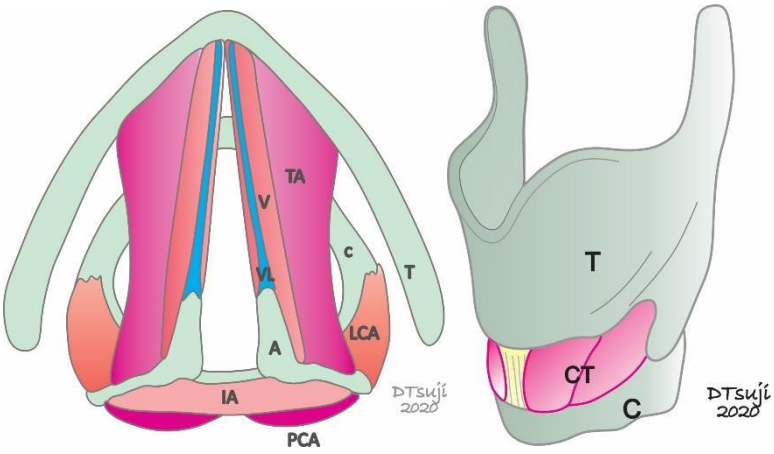


Figure 6.6. Intrinsic muscles and cartilages of the larynx. TA – Thyroarytenoid muscle; V- Vocalis muscle; LCA- Lateral cricoarytenoid muscle; PCA- Posterior cricoarytenoid muscle; IA- Interarytenoid muscle; CT- Cricothyroid muscle; T- Thyroid cartilage; C- Cricoid cartilage; A- Arytenoid cartilage; VL- Vocal ligament.

Under normal conditions, the air inhaled and exhaled by the lungs during breathing experiences practically no resistance to its free transit through the upper airways formed by the larynx, pharynx, oral and nasosinusal cavities. During voice emission, however, the vocal folds assume an adduced phonatory position, closing the glottis, which increases air resistance and raises the air pressure in the subglottic region. The interaction between this pressure, cord structure and cord mechanical

properties produces vocal fold vibration, which transforms pressure energy into sound energy.

The length, thickness, stiffness, and tension of the vocal folds are individual characteristics and can be modified during vocal emission by the intrinsic muscles. Sound properties such as intensity (voice loudness), vocal frequency (voice pitch) and vocal registers depend not only on the anatomical and tissue properties, but also on the functional capacity of these muscles, under neurological command, to modify and modulate the folds' structure, adapting them to a specific vocal task (for example sustained falsetto emission). The muscles' effects on vocal characteristics are presented in Table 6.1.

Table 6.1: Functions of the intrinsic muscles of the larynx

	Muscle	Function
CT	Cricothyroid	Tensor – increases vocal pitch
TA + V	Thyroarytenoid + Vocalis	Adductor, tensor, shorten vocal folds, increases glottic resistance and decreases voice pitch
LCA	Lateral cricoarytenoid	Adductor
PCA	Posterior cricoarytenoid	Abductor
IA	Interarytenoid	Adductor

Source: Tsuji, Watanabe, Imamura & Sennes, in print

The sound produced during vocal fold vibration is the fundamental sound of the larynx. It is a complex sound, consisting of the fundamental frequency (F0) and other harmonic waves of multiple frequencies with decreasing amplitude (Figure 6.7a). This complex sound travels through the vocal tract until it reaches the external environment, where it undergoes modifications to its original characteristics caused by resonance and articulation, allowing the formation of different vowels and consonants, essential in producing the wide range of phonemes present in human speech.

Sound Resonance: vowel production

The pure or simple tone produced by a rigid vibrator such as a tuning fork, corresponds to a single frequency, while the compound or complex sound, such as a guitar string, is formed by a set of simultaneous sound frequencies. Harmonics, or more precisely, harmonic partials, are partials whose frequencies are numerical integer

multiples of the fundamental. With compound sound, the lowest or fundamental frequency (F0) has the highest sound pressure intensity or amplitude. The other sound waves, or harmonic waves (H1, H2, H3…) have progressively lower intensities as the frequency increases. The vocal folds are a malleable vibrating structure that produces a compound sound with harmonic partials when vibrating (Figure 6.7a). This laryngeal or fundamental sound passes through the tubular structure of the vocal tract formed by the pharynx, oral and nasal cavities, and reaches the external environment. During this passage, harmonic waves with different wavelengths undergo the resonance phenomenon, and can be amplified or attenuated, depending on the dimensions of the resonating cavity. This anatomical segment is also known as the phonatory apparatus filter. Because it is an irregularly shaped tubular structure, the vocal tract can be considered a complex open resonator tube, formed by the connection of several tubes with variable calibers and lengths. These dimensions can also be modified in infinite ways by central neurological control during vocal emission. Central neurological control, through peripheral innervation, coordinates the various muscles responsible for mobilizing and modifying the larynx, pharyngeal wall, tongue, oral structures, cheek, and palate configurations.

This complex human ability, capable of modifying vocal tract geometry, allows the fundamental laryngeal sound to be modified by resonance, promoting the amplification and attenuation of different harmonic frequencies. In voice production, the harmonic frequencies amplified by the vocal tract are called harmonic formants and the first three of these - F1, F2 and F3 - define the different vowels (Isshiki, 1989) (Figure 6.7b). Thus, the formants are the characteristic frequency of each vowel. The formant frequencies of vowels /a/, /i/ and /u/ are presented in Table 6.2 (Gonçalves et al., 2009).

The sound resonance phenomenon depends directly on the dimensions and shapes of the resonator tube. Consequently, the resonating characteristics of the vocal tract, composed of different cavities (pharyngeal, oral and nasal), exhibit a supposedly resonance profile, and the coexistence of identical vocal tracts between different individuals is very unlikely, monozygotic twins being an exception, albeit with some caveats.

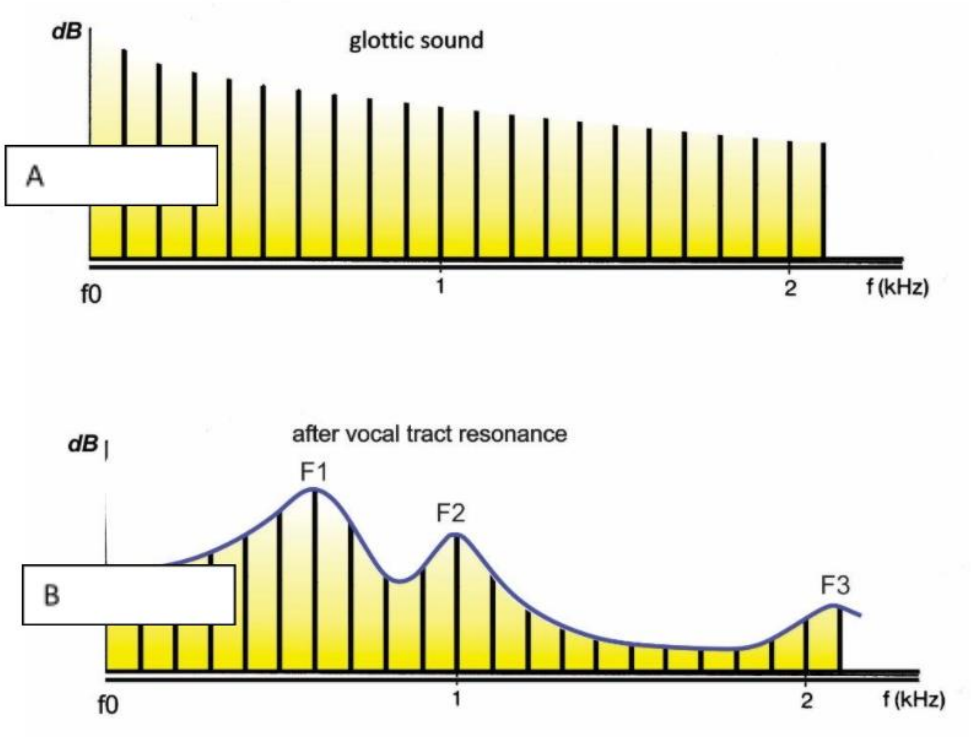


Figure 6.7. Resonance phenomenon. (A) Complex glottic sound with fundamental frequency (F0) and its harmonic frequencies. (B) the same sound spectrum after the influence of vocal tract resonance. The amplified harmonic frequencies (F1, F2 and F3) correspond to the first three formants that characterize vowel production.

Table 6.2. Average harmonic frequency values of the first three formants (F1, F2, F3), in Hz for the vowels /a/, /i/ and /u/.

	Females			**Males**		
Vowels	**F1**	**F2**	**F3**	**F1**	**F2**	**F3**
a	1002.90	1549.95	2959.70	753.87	1278.70	2483.44
i	361.90	2583.89	3378.14	297.80	2150.85	2925.14
u	461.82	763.41	2902.55	345.27	799.51	2351.50

Articulation: consonant production

The laryngeal sound, modified by the resonance phenomenon during its course through the vocal tract, can undergo momentary obstructions in its free flow due to the contact established between structures such as the lips, tongue, soft palate, uvula, teeth,

gingiva, hard palate and posterior pharyngeal wall. Among these, the lips, tongue, soft palate, and uvula are active mobile structures that promote articulatory contact with each other or with other fixed structures. The relationship between active and passive articulators is defined as the point of articulation and defines the characterization of the different consonants.

Depending on the type of occlusive contact obtained (total or partial), complete or partial airflow obstructions may occur. As a result, with the Brazilian Portuguese language as reference, consonants are classified as occlusive or plosive, such as, /p/, /b/, /t/, /d/, /g/, and fricative, such as /f/, /v/, /s/, /z/, /j/. Even among plosive consonants, depending on the lowering of the palate, which releases air through nasal cavities, nasal consonants such as /m/ and /n/ are produced (Seara, Nunes & Volcão, 2011). Differences in spoken language as well as regional and environmental influences can have a major impact on individual articulatory dynamics and promote significant vocal differences between pairs of twins, albeit monozygotic.

Concluding remarks

The human phonatory apparatus, a highly complex system that requires perfect synchronization, is coordinated by the central nervous system, and several organic structures such as the lungs, larynx, pharynx, tongue, and others. The characteristics of the fundamental sound produced in the larynx, as well as the resonance of this sound through the vocal tract, are related to the dimensions, anatomical profiles and dynamic response of the structures involved. Consequently, we can consider that it is almost impossible for identical voices to occur between different individuals, even if they are monozygotic twins. Using various acoustic analysis parameters, San Segundo et al. compared the differences in the voices found between pairs of distinct individuals with those between pairs of voices from the same individual, observing significant differences between the two groups. When confronted with the differences in the voices of pairs of identical twins, she observed that the vocal differences in this group were situated between the differences observed in the other two groups (San Segundo, Tsanas & Gómez-Vilda, 2017). In another study, the same authors demonstrated that the coefficient of similarity of voices decreases when the kinship relationship between pairs of voices declines. This means that the similarity classification, from highest to lowest, was observed in the following scale of values: monozygotic twins, dizygotic

twins, siblings, and reference population (San Segundo & Künzel, 2015). We can therefore infer that, although the pairs of voices of monozygotic twins may be quite similar, there will still be detectable differences in a more detailed vocal analysis.

References

Gonçalves, M. I. R., Pontes, P. A. D. L., Vieira, V. P., Pontes, A. A. D. L., Curcio, D., & Biase, N. G. D. (2009). Transfer function of Brazilian Portuguese oral vowels: a comparative acoustic analysis. *Brazilian journal of otorhinolaryngology*, *75*(5), 680-684. https://doi.org/10.1016/s1808-8694(15)30518-8

Hansen, J. T. (2019). *Netter's Clinical Anatomy*. 4th. ed. (pp.93-155). Philadelphia, PA: Elsevier.

Hirano, M. (1975). Phonosurgery: basic and clinical investigations. *Otol (Fukuoka)*, *21(suppl 1)*, 239-440.

Imamura, R., Tsuji, D. H., & Sennes, L. U. (2002). Fisiologia da laringe. In C. A. H. Campos & C. A. H; H. O. Costa H. O. (Eds.) *Tratado de otorrinolaringologia.* (pp. 743-750). São Paulo: Roca., 743-50.

Isshiki, N. (1989). Surgery to elevate vocal pitch. In N. Isshiki (Ed.), N. *Phonosurgery: Theory and Practice* (pp. 141-155). Tokyo: Springer., Tokyo. p.5-21.

San Segundo, E., & Künzel, H. (2015). Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers. *Loquens*, *2*(2), e021. https://doi.org/10.3989/loquens.2015.021

San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. Forensic *Science International, 270,* 25-38. https://doi.org/10.1016/j.forsciint.2016.11.020

Seara, I. C., Nunes, V. G. & Volcão, C. L. (2011). Fonética e fonologia do português brasileiro: 2º período. Florianópolis: *LLV/CCE/UFSC*.

Tsuji D. H., Watanabe L. M. N., Imamura R., & Sennes, L. U. Anatomía Aplicada a La Función Laríngea. In Kume M, Sulica L. *Patología vocal y Fonocirugía* – El Libro de Los Maters. Mexico: Springer Healthcare. In print.

Peer Commentary

By Lilian C. Luchesi[9]

Professor Domingos Tsuji wrote about anatomic structures, physical movements, and events involved in sound production. There is a relation between vocal tract anatomy and voice production that can explain the individuality of voice. It is unique for each person, with some exceptions on monozygotic twins. What do we know about genetic effects and external influences, like culture, on monozygotic twins' voices and speech?

Professor Tsuji explains that anatomy and genetics affect the form and function of cartilages and muscle on the vocal apparatus. Genetics is an important determinant of each person's anatomical characteristics, and if you have two persons with similar genomes, their vocal apparatus will probably be very similar. Voice production is a mechanical phenomenon in which sound waves go through the vocal tract. The size and extension of our vocal tracts is a mechanical phenomenon too. So if the instrument (our vocal apparatus) producing this mechanical phenomenon is the same, the result will probably be the same.

When comparing a twin pair with the same vocal fold, the same vocal structures, or vocal tract, their sound will probably be very similar. For example, the vocal folds have distinct characteristics. If two vocal folds are exactly equal, the extension or the voice frequency range will probably be equal or very similar. This similarity means they will be similar from the lowest frequency, or lowest pitch, to the highest pitch. Similarly, the intensity range from the lowest intensity to the highest intensity emission will be very alike. These two parameters (pitch and intensity) can be controlled by training, and if you compare, the performance will probably be better in the person more trained. However, basically emission of, for example, /e/ or /a/ or /i/ probably will be very, very similar among twins.

On the other hand, Tsuji reports the uniqueness of the voice: each person, even twin siblings, will have something that differs her from the others. Tsuji compares this

[9] PhD in Psychobiology, University of São Paulo, Ribeirão Preto.

*

individuality with electronic gadgets and says the software inside the brain that controls everything will be different between two persons. To illustrate Tsuji's explanation, Jonsson and colleagues (Jonsson et al., 2021), analyzing genome sequences of monozygotic twins from Iceland, found specific mutations in only one of the siblings in 15% of them, reinforcing that monozygotic is not equal. Two recent studies showing the distinction between monozygotic twins are on vowel formants of Brazilian male pairs (Cavalcanti et al., 2021) and in the vowel filler [e:] in hesitating responses of Spanish pairs (San Segundo et al., 2017). In these two examples above, the twins were raised together.

Otta's research team conducts a case study with twins raised apart in Brazil. The study focuses on a pair of twins separated when they are newborns. One twin was raised in the Northeast of Brazil, and the other in the Southeast. They remained apart during twenty-three years of their life, and now they are together again. In this study, we could see both similarities and differences between them. Although, we are still looking at their voices. Additional information on twin's voice analysis is presented in chapter 12 of this book.

Tsuji says the significant differences among people probably lie in the resonance and not in the phonation of the laryngeal sound. The environment where they were raised, the nationality, or country region significantly influences speech, changing the articulation, intonation, and vowel characteristics. Thus, the environment in which they were raised will be essential, and we expected that siblings would have a high similarity score when they are raised in a very similar environment. If we compare dizygotic twins to non-twins' siblings, dizygotic twins have more similar voices than non-twins' siblings. However, dizygotic twins have lower similar scores than monozygotic twins. Even though they were born on the same day and shared the same environment during their development, they do not share the same genome. Few studies compare MZ twins, DZ twins, and non-twin siblings' voice parameters. Forensic studies using automatic speaker recognition found a decreasing scale in similarity coefficients among MZ > DZ > non-twins > Unrelated Speakers (San Segundo & Künzel, 2015). Although even observing this relation, non-twin siblings can also affect automatic recognition system performance by deteriorating forensic comparison of voice, pointing to the necessity of more investigations on similar-sounding speakers beyond MZ (San Segundo & Yang, 2019).

Comments on the voice production mechanism by humans: robotic voice and singer's voice adjustment

When we are silent during respiration, vocal folds are always opened (Fig 6.4a). The PCA muscle (Posterior cricoarytenoid muscle; see Fig 6.6), which is the only muscle that keeps vocal folds in a lateral position, remains all the time contracted to keep the folds opened, or the glottis opened, for respiration.

The first form of phonation on humans is the sound emission while inspiring. In this phenomenon, instead of the vocal folds vibrating from down to up (from a horizontal to upwards vertical position), movement is downwards from horizontal to vertical down (Figure 6.8). Anyone can produce this kind of sound, but we must remember that vocal folds are not the only sound producer but also the palate, lips, and tongue. All these structures, when constricted, may vibrate during the airflow passage and produce sounds. The main structures in voice production are the vocal folds, but other structures may be involved too in this robotic voice emission.

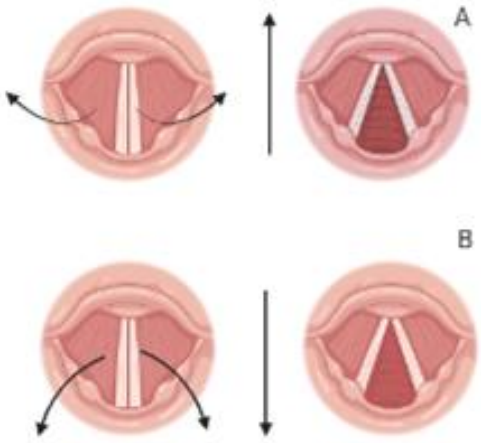


Figure 6.8. Vocal folds movements during silent respiration (A) and respiration with phonation (B). Arrows indicate the fold movement directions. In the silent respiration vocal fold, lateral movement is horizontal upwards oriented, while respiration with phonation is horizontal to down movement in relation to the head position. Created with BioRender.com

Professional singers move the head, shaping the vocal emissions. It is possible that by changing the neck or the head positions, one can stimulate some other muscles around the larynx or even on the neck that can help to give more precision in the voice emission. For example, the format of the vocal tract can be a little bit changed during movements, and probably those highly technically developed singers would have a very sensitive perception of their voice production. So, they use those movements to help them reach the sound they want to produce to that specific interpretation. Sometimes moving the chin-up is better for single high notes. They feel more comfortable producing notes, and the resonance is better for reaching that note with this kind of movement.

To illustrate the relation between postural alignment, vocal resonance, and pitch control, imagine the ears being placed forward the shoulders (the neck thrusting head posture). The anterior neck muscles will be stretched, narrowing the pharynx and negatively impacting the vocal resonance (Wilson Arboleda & Frederick, 2008). When shoulders are rounded forward, inspiration's lung volume decreases and may cause vocal fatigue and difficulty projecting the voice (Wilson Arboleda & Frederick, 2008). Trained singers can control the increase in neck-muscle tension while singing or speaking, preventing the shortening distance between the occiput and seventh cervical vertebra, obtaining a greater resonance, and better breath control (Jones, 1972). The singer's voice emission is boosted in overall aspects, increasing F0 and low-frequency amplitudes when keeping the head extended upward at 15º and 30º positions. In a contrary direction, speaking with the head lowered will affect the low-frequency energy and the gain in Singing Power Ratio (Knight & Austin, 2020) these results were compared to the comfortable habitual head position adopted by each singer.

One more of Tsuji' s comments about the words "vocal cords" and "vocal folds". Indeed, the correct words from the anatomical point of view, the medical point of view, are vocal folds. Because instead of being a chord, a simple chord, it is more like a fold or Portuguese "prega" (pleat), like those on clothes. That is why they are called vocal folds. However, I use "vocal cords" because of some terms used in the medical area. For example, we never say, in Portuguese, "Foldite," we say "cordite" when we want to refer to the vocal fold inflammation, and we never say "preguite." So even in the medical area, in the anatomical scientific papers, we say "Cordite" instead of "Foldite." So, we can keep using this term, in Portuguese, "cordas vocais" (vocal cords), because this term is popular and people can understand it very well. So,

I keep using it, especially with my patients. However, when writing, please write vocal folds.

References

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2021). Acoustic analysis of vowel formant frequencies in genetically-related and non-genetically related speakers with implications for forensic speaker comparison. *PLOS ONE*, *16*(2), e0246645. https://doi.org/10.1371/journal.pone.0246645

Jones, F. P. (1972). Voice Production as a Function of Head Balance in Singers. *The Journal of Psychology*, *82*(2), 209–215. https://doi.org/10.1080/00223980.1972.9923808

Jonsson, H., Magnusdottir, E., Eggertsson, H. P., Stefansson, O. A., Arnadottir, G. A., Eiriksson, O., Zink, F., Helgason, E. A., Jonsdottir, I., Gylfason, A., Jonasdottir, A., Jonasdottir, A., Beyter, D., Steingrimsdottir, T., Norddahl, G. L., Th Magnusson, O., Masson, G., Halldorsson, B. V, Thorsteinsdottir, U., … Stefansson, K. (2021). Differences between germline genomes of monozygotic twins. *Nature Genetics*. https://doi.org/10.1038/s41588-020-00755-1

Knight, E. J., & Austin, S. F. (2020). The Effect of Head Flexion/Extension on Acoustic Measures of Singing Voice Quality. *Journal of Voice*, *34*(6), 964.e11-964.e21. https://doi.org/10.1016/j.jvoice.2019.06.019

San Segundo, E., & Künzel, H. (2015). Automatic speaker recognition of spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers. *Loquens*, *2*(2), e021. https://doi.org/10.3989/loquens.2015.021

San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International*, *270*, 25–38. https://doi.org/10.1016/j.forsciint.2016.11.020

San Segundo, E., & Yang, J. (2019). Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation. *Journal of Phonetics*, *75*, 1–26. https://doi.org/10.1016/j.wocn.2019.04.001

Wilson Arboleda, B. M., & Frederick, A. L. (2008). Considerations for Maintenance of Postural Alignment for Voice Production. *Journal of Voice*, *22*(1), 90–99. https://doi.org/10.1016/j.jvoice.2006.08.001

Chapter 7

Larynx evolution: comparative research with primates and carnivores[10]

By Aline D. Carneiro Gasco and Rogério Grassetto T. Cunha

Bioacousticians have been comparing the anatomy of the larynx and related structures across species for some time already. However, Daniel Bowling paved the way to a novel and more systematic approach including specific measurements and phylogenetic signal analyses. Since plausible arguments have been put forward on the influence of social systems and body sizes on larynx evolution, Daniel Bowling and colleagues focused on comparing larynges of terrestrial and arboreal primates and carnivores, living either alone or in large groups (Bowling et al., 2020). Moreover, the authors assumed that carnivores constitute a special group similar to primates in their diverse social systems, body sizes, and habitats. Furthermore, investigations will continue with Bowling's systematic approach being applied to other taxa.

Their results were quite interesting, with primate larynges proven not only to be larger (relative to body size) but also with larger residual variation. Also, their data points that they have also evolved at a faster rate. Last, changes in larynx size were related to a larger variation in the mean F0 among primates when compare to carnivores, something they suggest is related to a larger prominence of vocal communication among primates. Then, they went on to explore some possible ideas behind these patterns.

Regarding body size, they hypothesize that primate and carnivore larynx sizes would vary according to habitat occupancy. Bowling's social system hypothesis is that group size would proxy social complexity with large-group primates requiring a more robust larynx. So far, both hypotheses have been proved wrong. The differences in group size seemed not to be related to the variety of larynx sizes. Additionally, large-

[10] Daniel Bowling delivered the talk "Rapid evolution of the primate larynx and the Source-Filter Theory", based on the paper *Bowling et al. (2020) Rapid evolution of the primate larynx. PLoS Biol 18(8): e3000764*. This chapter consists of the main ideas of the talk and the following discussion. To see the illustrations, please access https://doi.org/10.1371/journal.pbio.3000764.

group primates varied more than carnivores in terms of larynx size. As sound examples, they point out that the howler monkeys have the largest larynges within primates and do not live in the largest groups, whereas baboons have tiny larynges and live in huge groups. We could add that hylobatids also have very large larynges (both in absolute and relative terms), and also do the Callicebinae, but species from both groups form some of the smaller groups among primates. In short, there are clear differences between primate and carnivore larynx sizes, but there is no simple explanation for the variation, at least based on the chosen index of sociality, as Bowling and colleagues have already learned. He suggested that we could look at relationship quality, for example, or at some of the several indexes of sociality available. In this whole sociality issue, an interesting group to look at are the coatis, a carnivore with large groups, complex sociality and flexible habitat and diet.

Bowling noticed that *Panthera*'s larynx models were some of the largest sized ones in relation to body size, and they also correspond to the roaring feline clade, probably functioning as a body exaggerating message, which might bear some relevance to a sexual selection function. Moving to our species, they did not measure the human larynx in the study, but he expected to position it above the regression line, around the chimpanzee and gorilla larynges on the graph. However, he reminded that since men and women are different from puberty, we must consider that sexual dimorphism affects the larynx size. Thus, men would probably be further above the regression line graph of larynx sizes than women.

Regarding the specific measurements taken in Bowling's systematic approach, they first reduced the data set of ten measurements through a principal component analysis. Their measure was a first approach and roughly indicates how the voice was affected with variable vocal fold lengths, which is reflected in their measurements. Overall, as the larynx becomes larger, the vocal folds lengthen to the extent that they increase the frequency range that can be produced. Perhaps, it sounded like a basic explanation, but those changes are the main blocks to build the understanding around anatomy and voice production. Later on, he suggests that further, steps must be taken in the direction of more sophisticated approaches, for example looking at cartilage morphology. By doing so, he claims we will get some insight on the relation between such finer anatomical differences and the variations in the amplitude of the voice or the shape of the glottal pulse, to cite just some of the possible ways forward.

Conducting Bowling's comparative and systematic approach to vertebrate larynges required a CT scan (computed tomography scan; Figure 7.1) to model the vertebrate larynges. Once vocal tract length measurements would require access to the whole torso, that focused only on excising larynges. The CT scan used the National Museum of Scotland's massive collection of larynges excised from animals that died in European zoos. Before running CT scans, the larynges were blown to moisten the air, thus allowing phonation and vibration.

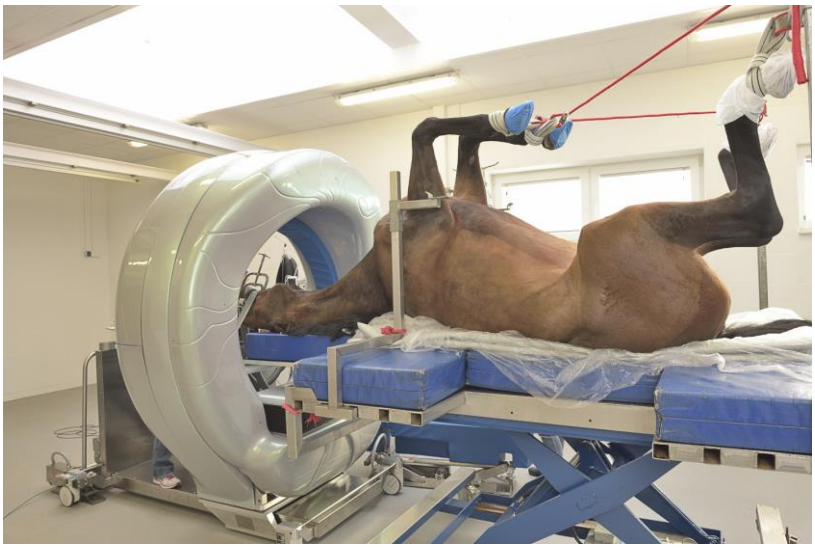


Figure 7.1. The computed tomography scan (CT) creates images up to the level of the thoracic and the limb up of the standing and conscious horses. Font: https://vetgrad.com/.

Reaching out to the end, Bowling briefly tackled the auditory and vocal systems alignment issue across species. What we hear is fine-tuned into the range of what the animals are vocalizing. If one had some proxy for evolution in the auditory and vocal systems, one could look at the different rates at which they evolved. Maybe one of those evolved faster, and then someone could say which system is more likely to be the driver. In terms of brain matters, a growing amount of work on morphology, vocal morphology, and acoustics shows that animals can make complicated use of their vocal systems. Bowling explained that animals could have fine control of laryngeal movement and vocal tract and move their tongue around and make vowel sounds. In short, we do not completely understand what is going on with animal brains. In contrast, the current story about humans is directed to the primary cortical motor descending control that allows us to learn and make different sounds.

Future directions

Future work on the phylogeny of the vertebrate larynx will need to broaden the current National Museum of Scotland's collection to include more mammalian clades. Up to this point, Bowling and collaborators have only looked at primates and carnivores, demonstrating the rapid evolution of primate larynges when compared to their carnivorous counterparts. Furthermore, ongoing collaborative work with Dr. Jacob C. Dunn, director of the Behavioural Ecology Research Group of Anglia Ruskin University (UK), aims to investigate larynx phylogeny to shed light on the evolution of language.

Concluding remarks

We learned that larynx size is strongly related to body length and vocal frequency up to the point, but that the differential flexibility of primate larynx size can potentially affect their vocal communication. Primate larynges are larger, more variable, and evolved more rapidly than their carnivorous counterparts. Primates have turned into a more diverse clade in that matter and display more complex systems than carnivores. Although it seems that primate larynges are less constrained by body size and increase in diversity due, we cannot be sure of the selective pressures responsible for such pattern. Despite adaptations to phonation, which might bear relation to habitat (via the acoustic adaptation hypothesis), social complexity or sexual selection pressures, there are still other major functions for the larynges: the protection of the airway during feeding and the regulation of air supply to the lungs.

Chapter 8

Identifying emotions from voice

Bruna Campos Paula[11]

Abstract

Expression of emotion in vocalizations occurs via modifications of acoustic structure. Vocal expression can be used to infer the emotional state. During speech communication, listeners pay attention to changes in pitch, loudness, rhythm, and voice quality (emotional prosody) to form an impression about the speaker's emotional state in conjunction with linguistic decoding. With respect to acoustic characteristics of vocal expressions in emotion studies, F_0, voice intensity, energy distribution in the spectrum frequency, formant location, and speech rate are frequently used to access the emotional voice.

Keywords: Acoustic characteristics; Human voice; Affective science; Human emotion; Discrete emotions; dimensional approach of emotion.

The emotional expression theory suggests that emotions evolved at different times as an adaptive trait, following the theory of evolution (Hess & Thibault, 2009). Emotional expression can be separated into modules such as anger, fear, happiness, and sadness (Darwin, 1872; Panksepp & Watt, 2011). These modules are the foundation of the theoretical approach of the "basic emotions" concept, which proposes that some aspects of the body and face evolved as an adaptation to mediate specific contexts (Izard, 1992, 2007).

Another perspective is the dimensional approach, which concentrates primarily on one component of emotion, the subjective feeling state, and focuses on identifying emotions based on their placement in a small number of underlying dimensions, such as valence, activation and potency (Laukka, 2004; Laukka, Juslin, & Bresin, 2005). From an evolutionary perspective, emotions can be understood according to the functions they serve (Keltner & Gross, 1999; Cosmides & Tooby, 2000). They evolved

[11] Acoustics and Environment Lab, Mechanical Engineering Department (POLI), University of São Paulo, SP, Brazil. brunacampospaula@gmail.com

to deal with goal-relevant changes in our environment and can be described as relatively brief and intense reactions to these changes (Laukka, 2004).

Psychological theories of human emotion have highlighted the multicomponent nature of emotions, typically including subjective experience and neurophysiological processes, as well as their somatic and endocrine counterparts (Barrett et al., 2007). In the discrete emotion theory, emotion is thought to represent a unique interaction with the environment and its adaptational significance for the individual (Laukka, 2004). A unique cognitive appraisal pattern, physiological activity, action tendency, and expression are related to each discrete emotion (Ekman, 1992; Izard, 1992). Discrete emotion theories explain a limited number of "basic" emotions that have evolved to deal with particularly pertinent life problems: competition and anger; danger and fear; cooperation and happiness; loss and sadness. Studies on the communication of emotions suggest that facial expressions are universally expressed and recognized, and these studies have been the backbone of discrete emotion theories (Ekman, 1992).

Vocal expression of emotions

Emotions are also expressed via the vocal channel through modification of the sound's acoustic structure or specific vocal types (Morton, 1977; Izard, 1977; Anikin et al., 2018). Vocal expression can be used to infer the sender's emotional state (Scherer 2005). The human voice is a diverse and nuanced source of emotional signaling, (e.g., laughter, Bachorowski et al., 2001; teasing, Keltner et al., 2001; motherese, Fernald, 1992; vocal bursts, Simon-Thomas et al, 2009). During speech communication, listeners pay attention to changes in pitch, loudness, rhythm, and voice quality (emotional prosody) to form an impression about the speaker's emotional state in conjunction with linguistic decoding (Wilson & Wharton, 2006).

Most vocal expression studies have used some variant of the "standard content paradigm" to detect emotional speech from real conversations. First, an actor is instructed to read some verbal material aloud while simultaneously portraying particular emotions chosen by the investigator. The emotion portrayals are first recorded and then evaluated in listening experiments to determine whether listeners can decode the intended emotions. The same verbal material is used in portrayals of different emotions, and most typically consist of single words or short phrases. The assumption is that because the verbal material remains the same in the different

portrayals, whatever effects appear in listeners' judgments it should result mainly from the speaker's voice cues.

On the other hand, in the emotional speech method from real conversations, emotions are induced in the speaker using various methods and speech synthesis to create emotional speech stimuli. Both methods have advantages and drawbacks: despite the fact that standard content paradigm ensures control of the verbal material and encoders intention, there are doubts about the validation between posed and natural occurring expressions; using real emotional speech ensures high ecological validity but renders the control of verbal material and encoder intention more difficult.

Acoustic cues of emotions

Linguistic and nonlinguistic information is coded simultaneously in human speech acoustic signals, communicated by the same acoustic voice cues. Furthermore, the voice contains other types of information about the speaker, such as identity, age, sex, and body size (Sell et al., 2010; Smith & Patterson, 2005; González, 2004; Pisanski et al., 2014). Human speech is produced by the speech articulators' continuous movement, such as the tongue, lips, and larynx, which modulate airflow so that speech sounds reach the ears. According to the source-filter model of speech production, vocal acoustics is a combination of an underlying energy source and filtering effects due to pharyngeal, oral, and nasal cavities resonance of the supralaryngeal vocal tract (Fant, 1960). Vibration occurs in the vocal folds, that vibrates the air in the supraglottal vocal tract (pharynx, mouth, and nasal cavity), vocal tract resonances (called formants) are acoustically excited, passing the frequencies in the source waveform that are near the dampened or strengthened energy, but do not totally represent the vocal tract resonances.

With respect to the acoustic characteristics of vocal expressions in emotion studies, F0 (the frequency with which vocal folds open and close across the glottis during phonation), voice intensity, energy distribution in the frequency spectrum (high and low frequencies), formant location, and speech rate are frequently used to access the emotion in voice (Borden & Harris, 1984; Scherer, 1989; Banse & Scherer, 1996; Laukka 2004). Taken together, these aspects constitute prosody. Pittam and Scherer (1993) summarized some of the research evidence, selecting acoustic parameters related to specific emotions:

- Anger, characterized by an increase in mean F0, F0 variability and range, and mean energy;
- Fear, characterized by an increase in mean F0, F0 range, and high-frequency energy;
- Sadness, characterized by a decrease in mean F0, F0 range, downward-directed F0 contours, and mean energy;
- Joy, characterized by increases in mean F0, F0 range and variability, and mean energy.

F0 is subjectively heard as voice pitch, and mainly reflects the differential innervation of the laryngeal muscles and the extent of subglottal pressure (Barbosa and Madureira, 2015). Voice intensity is subjectively heard as vocal loudness and determined by respiratory and phonatory action. Voice quality is subjectively heard as the timbre of the voice, determined by supralaryngeal vocal tract settings and the phonatory mechanisms of the larynx (Barbosa and Madureira *op cit.*). Finally, the temporal aspects of voice concern the temporal sequence of sound production and silence (speech rate).

The basic assumption underlying most work on vocal emotion expression is that there is a set of objectively measurable voice cues that reflect human emotion states. Thus, some researchers have argued that voice cues may reflect only the so-called activation dimension of emotions (Davitz, 1964) or a combination of arousal and valence (Bachorowski, 1999). However, recent research suggests a great deal of acoustic differentiation of emotions in vocal expression (Juslin & Laukka, 2001; Shigeno, 2004; Laukka et al., 2005; Goudbeek & Scherer, 2010; Silva et al., 2016).

Concluding remarks

The literature has shown that combinations of acoustic parameters, such as speech rate and fundamental frequency (F0), cue different emotions. The mean F0 and speech rate are generally higher for emotions associated with high sympathetic arousal such as anger, fear, happiness and feelings of anxiety. However, one of the critical questions about detecting emotion from voice is that a theoretical approach in research is still missing. Studies are needed. The most important aspect in conducting a study on the expression of emotion from voice is choosing which emotional states should be investigated, the quality of speech samples, a multiparameter measure of voice, and physiological and phonatory-articulatory analysis.

References

Anikin, A., Bååth, R., & Persson, T. (2018). Human non-linguistic vocal repertoire: Call types and their meaning. *Journal of Nonverbal Behavior, 42(1)*, 53-80. https://doi.org/10.1007/s10919-017-0267-y

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current Directions in Psychological Science, 8(2)*, 53-57. https://doi.org/10.1111/1467-8721.00013

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*(3), 1581-1597. https://doi.org/10.1121/1.1391244

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70(3)*, 614-636. https://doi.org/10.1037/0022-3514.70.3.614

Barbosa, P. A., & Madureira, S. (2015). *Manual de fonética acústica experimental: aplicações a dados do português*. São Paulo: Cortez.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology, 58*, 373-403. doi: 10.1146/annurev.psych.58.110405.085709

Borden, G., and Harris, K.S. (1984). Speech Science Primer: Physiology, Acoustics, and Perception of Speech, 2nd Edition. Baltimore: Williams and Wilkins.

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In Michael Lewis & Jeannette M. Haviland-Jones (eds.) *Handbook of Emotions*, 2nd ed. pp. 91-115). New York: The Guilford Press.

Darwin C. (1872) The expression of the emotions in man and animals, 1st edition. London: John Murray.

Davitz, J. R. (1964). Auditory correlates of vocal expression of emotional feeling. In Joel R. Davitz (ed.) *The Communication of Emotional Meaning* (pp. 101-112). New York: Mcgraw Hill.

Ekman P (1992) An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169-200. https://doi.org/10.1080/02699939208411068

Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

Fernald, A. (1992). 13. Meaningful melodies in mothers' speech to infants. H. Papoušek,U. Jürgens, and M. Papoušek, (eds.) *Nonverbal vocal communication: Comparative and developmental approaches* (pp. *262-282)*. Cambridger: Cambridge University Press.

González, J. (2004). Formant frequencies and body size of speaker: a weak relationship in adult humans. *Journal of Phonetics*, *32*(2), 277-287. https://doi.org/10.1016/S0095-4470(03)00049-4

Goudbeek, M., & Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America, 128*(3), 1322-1336. https://doi.org/10.1121/1.3466853

Hess U, Thibault P (2009. Darwin and emotion expression. *American Psychologist*, *64*(2), 120-128. http://dx.doi.org/10.1037/a0013386

Izard, C.E. (1977). *Human emotions*. New York: Plenum.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological Review, 99*(3), 561–565. https://doi.org/10.1037/0033-295X.99.3.561

Izard, C. E. (2007). Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on Psychological Science, 2(3),* 260-280. https://doi.org/10.1111/j.1745-6916.2007.00044.x

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381. https://doi.org/10.1037/1528-3542.1.4.381

Keltner, D., & Gross, J. J. (1999). Functional accounts of emotions. *Cognition & Emotion, 13*(5), 467-480. https://doi.org/10.1080/026999399379140

Keltner, D., Capps, L., Kring, A. M., Young, R. C., & Heerey, E. A. (2001). Just teasing: A conceptual analysis and empirical review. *Psychological Bulletin*, *127*(2), 229–248. https://doi.org/10.1037/0033-2909.127.2.229

Laukka, P. (2004). *Vocal expression of emotion: discrete-emotions and dimensional accounts.* Doctoral dissertation, Uppsala University, Acta Universitatis Upsaliensis.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633-653. https://doi.org/10.1080/02699930441000445

Morton E. S. (1977). On the occurrence and significance of motivational-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855-869.

Panksepp, J., & Watt, D. (2011). What is basic about basic emotions? Lasting lessons from affective neuroscience. *Emotion Review, 3(4),* 387-396. https://doi.org/10.1177/1754073911410741

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., Röder, S., Andrews, P. W., ... & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: a meta-analysis. *Animal Behaviour*, *95*, 89-99. https://doi.org/10.1016/j.anbehav.2014.06.011

Pittam, J., & Scherer, K.R. (1993). Vocal expression and communication of emotion. In M. Lewis &. J.M. Haviland (Eds.), Handbook of emotions (pp. 185–197). New York: Guilford Press.

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. In H. Wagner & A. Manstead (Eds.), *Wiley handbooks of psychophysiology. Handbook of social psychophysiology* (p. 165–197). John Wiley & Sons.

Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information, 44(4),* 695-729. https://doi.org/10.1177/0539018405058216

Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., Von Rueden, C., ... & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B: Biological Sciences*, *277(1699)*, 3509-3518. https://doi.org/10.1098/rspb.2010.0769

Shigeno, S. (2004). Recognition of vocal expression of emotion and its acoustic attributes. *Shinrigaku Kenkyu: The Japanese Journal of Psychology*, *74*(6), 540-546. https://doi.org/10.4992/jjpsy.74.540

Silva, W. D., Barbosa, P. A., & Abelin, Å. (2016). Cross-cultural and cross-linguistic perception of authentic emotions through speech: An acoustic-phonetic study with Brazilian and Swedish listeners. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, *32*(2), 449-480. https://doi.org/10.1590/0102-445003263701432483]

Simon-Thomas, E. R., Keltner, D. J., Sauter, D., Sinicropi-Yao, L., & Abramson, A. (2009). The voice conveys specific emotions: Evidence from vocal burst displays. *Emotion*, *9(6)*, 838. https://doi.org/10.1037/a0017810

Smith, D. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, *118(5)*, 3177-3186. https://doi.org/10.1121/1.2047107

Wilson, D., & Wharton, T. (2006). Relevance and prosody. *Journal of Pragmatics*, *38(10)*, 1559-1579. https://doi.org/10.1016/j.pragma.2005.04.012

Peer Commentary

By Plínio A. Barbosa[12]

In her chapter, Bruna Campos Paula guides us to think about the difference between feelings and emotions. She describes feelings as related to thoughts and other cognitive processes and emotions as quick responses to events. Emotions are also expressed in our speech, and this is Paula's interest.

Different emotions are evaluated from voice parameters depending on the researcher's theoretical framework. Two of these are categorical (basic emotions concept) and dimensional theories (valence, activation, and potency). Therefore, when conducting a study on the expression of emotion from voice, Paula suggests that the researcher have that in mind and define the emotional states to be investigated; the quality of the speech samples, a multiparameter measure of voice, and the physiological phonatory-articulatory analysis must also be planned.

In studies on monozygotic twin voices, we expect to find greater similarity in the ways of emotional expression than in dizygotic twin or ordinary siblings' voices. In terms of forensic research, for instance, parameters such as high formant frequencies, as F3 and F4, and temporal variables such as articulation rate and pausing help distinguish speech within twin pairs. Parameters like shimmer and jitter are also important in studying human voice, especially when subject to an emotional charge. Both are aspects of voice quality, where the former is a measure of glottal pulse intensity variation and the latter of glottal pulse period variation. Shimmer and jitter are used to analyze individual vocal differences or the impact of some event on the vocal behavior of a person in a particular situation.

We could also speculate on the impact the study of animal emotions could have on our everyday lives. For example, the results obtained by studying human emotions could be adapted to non-human animal emotions. Probably, the study of facial expressions in animals could be useful for the study of animal communication.

---

[12] Plínio A. Barbosa is Associate Professor at the Department of Linguistics of the Instituto de Estudos da Linguagem at UNICAMP (Brazil), and responsible for the Speech Prosody Studies Group.

In her work, Elodie Briefer illustrated this by using the dimensional approach and focusing on evolutionary aspects (Chapter 4). Her work revealed different kinds of impact. Both emotional arousal and valence can be detected from facial and vocal expressions, and fear inferred from the behavior of animals. Animal welfare can be measured and promoted.

Chapter 9

The use of PRAAT software in acoustic analysis

Plínio A. Barbosa[13]

Abstract

This paper introduces the main features of Praat software for acoustic analyses, going beyond what is suggested by the program association with the phrase "doing phonetics by computer". In fact, annotation, spectral, duration, intensity, and fundamental frequency analyses, mostly used for speech, can be easily adapted to extract parameters from non-human animal vocalizations. In addition to illustrating the Praat features with examples from speech and non-human animal vocalizations, the advantage of scripting is also discussed.

Keywords: acoustic phonetics; acoustic analysis; software

The free Praat software developed by Paul Boersma and David Weenink (2020) is by far the most widely used program for acoustic analysis by phoneticians and speech scientists around the globe. In addition to allowing the user to perform well-known tasks directly related to sound file handling, such as annotation, editing, filtering, and ordinary spectral, durational and intensive analyses, it also allows scripting, enabling the user to perform all tasks automatically through high-level programming. Praat can be downloaded for free at <http://www.praat.org> and is available for most operating systems. The Praat webpage has introductory instructional material for both beginners and advanced users, as well as a list of Frequently-Asked-Questions (FAQs).

[13] Department of Linguistics, Institute for Language Studies, University of Campinas, SP, Brazil. pabarbosa.unicampbr@gmail.com

Provided files are coded in formats such as WAV, AIFF, AIFC, MP3, among others, and any type of signal can be analyzed in Praat. Common examples of non-sound files are speech-related signals such as breath and electroglottographic signals.

When running Praat, two windows open: the Praat objects and Praat picture window. The Praat objects window allows the user to perform numerous tasks and analyses by using a set of dynamic buttons, which makes it possible for the user to conduct all analyses related to a particular object. For example, the button for spectral analysis is available when a sound object is selected, but not when an annotation object is selected (TextGrid in Praat). On the other hand, several tasks can be performed with a TextGrid object, which is not available for a sound object. The concept of dynamic buttons helps users go directly to the relevant analyses they can carry out. In the following sections, some examples of acoustic analyses in Praat will be shown as illustrations of its main features.

Annotation

Figure 9.1 illustrates how speech signals annotation can be performed in Praat by using TextGrid objects that allow several levels of segmentation. In this example, four user-defined interval tiers are shown: phrase tier ("*Em seguida, apareceu um papagaio real*", "*Next, a royal parrot appeared*"). Another type of tier, the point tier, provides a way to fix a position where some types of analyses can be conducted for further verification, such as a Fourier spectrum and LPC analysis for formant value computing, among others.

TextGrid objects can be saved independently from sound files and exported to formats such as XML. Because the user defines the number and nature of the tiers, the Praat annotation system can be easily used to segment non-human animal vocalizations by delimiting intervals of sentences, notes, and phrases, among others.

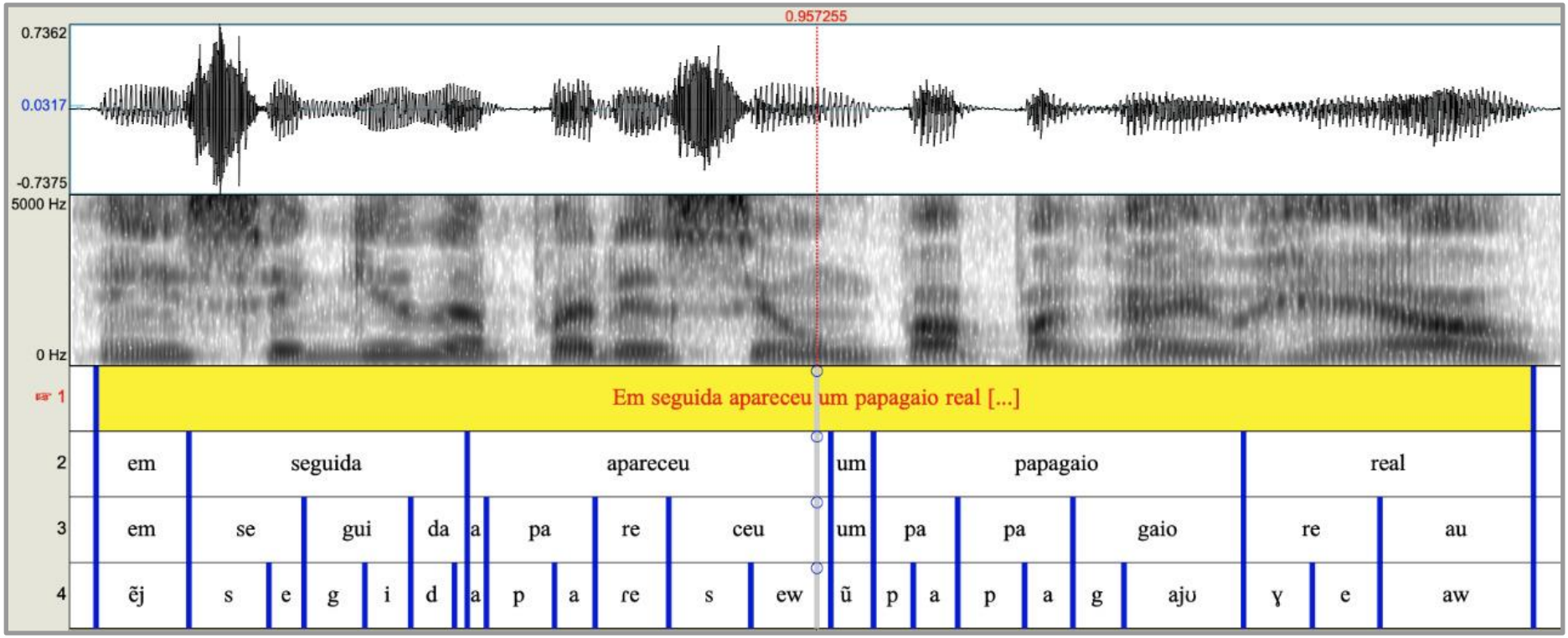


Figure 9.1. Broadband spectrogram (above) and annotation tiers (from top to bottom: phrase, word, syllable, and segment tiers) in Praat.

Spectral analyses

In addition to spectrographic analysis, illustrated in Figure 9.1 above, other types of spectral analyses in Praat are: Linear Predictive Coding (LPC), for computing formant frequencies and bandwidths; Fourier spectrum, for computing the amplitude or phase values of all component frequencies in the signal; and cepstrum, for separating source and filter characteristics and estimating formant frequencies when LPC cannot be used (e.g., when the signal has significant noise or there are bifurcations in the vocal tract). Figure 9.2 illustrates the use of Fourier amplitude spectra for revealing differences in capybara barks in the contexts of feeding behavior (black) and danger (red).

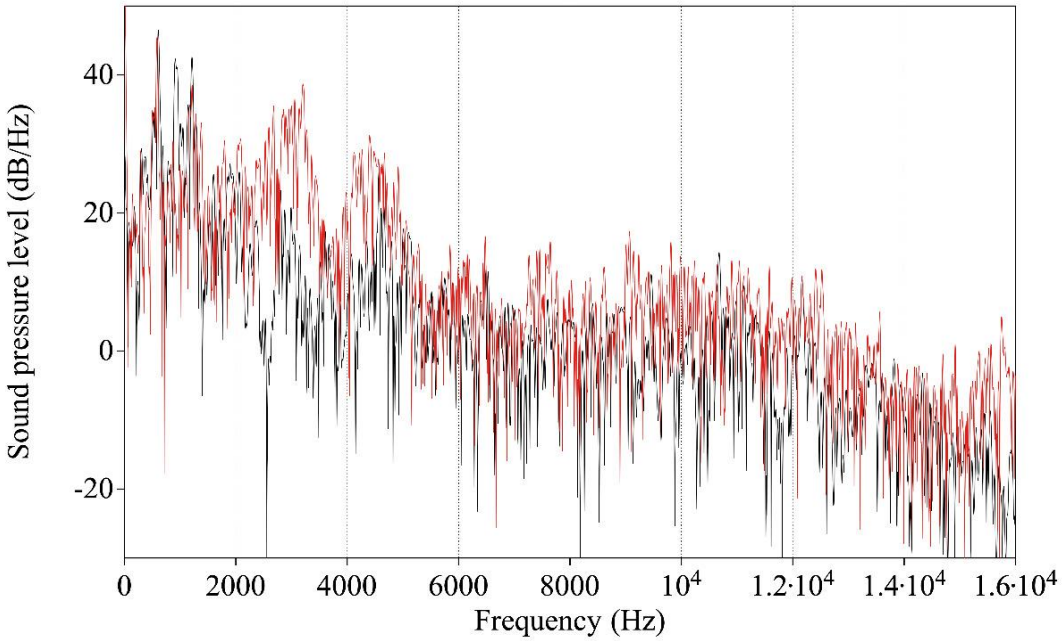


Figure 9.2. The Fourier spectra of two capybara barks from the same individual as an alert for feeding (black) and danger (red).

Quantitative analysis with a number of capybara individuals revealed that when in danger, their barks have a stronger relative amplitude component in the 10-to-15-kHz band compared with the feeding bark.

Duration analysis

In speech, syllable duration is the main parameter for signaling stress in a number of languages, including Brazilian Portuguese, European Spanish and Italian. Based on the study of the variation of normalized syllable duration (Barbosa, 2019) in annotated data, such as those shown in Figure 9.1, it is possible to detect the positions in an utterance where a particular speaker signals stress.

In addition to local measures such as duration, global measures, including speech rate, help understand human behavior in different communicative situations. The same can be applied in non-human animal vocalizations, as illustrated in Figure 9.3.

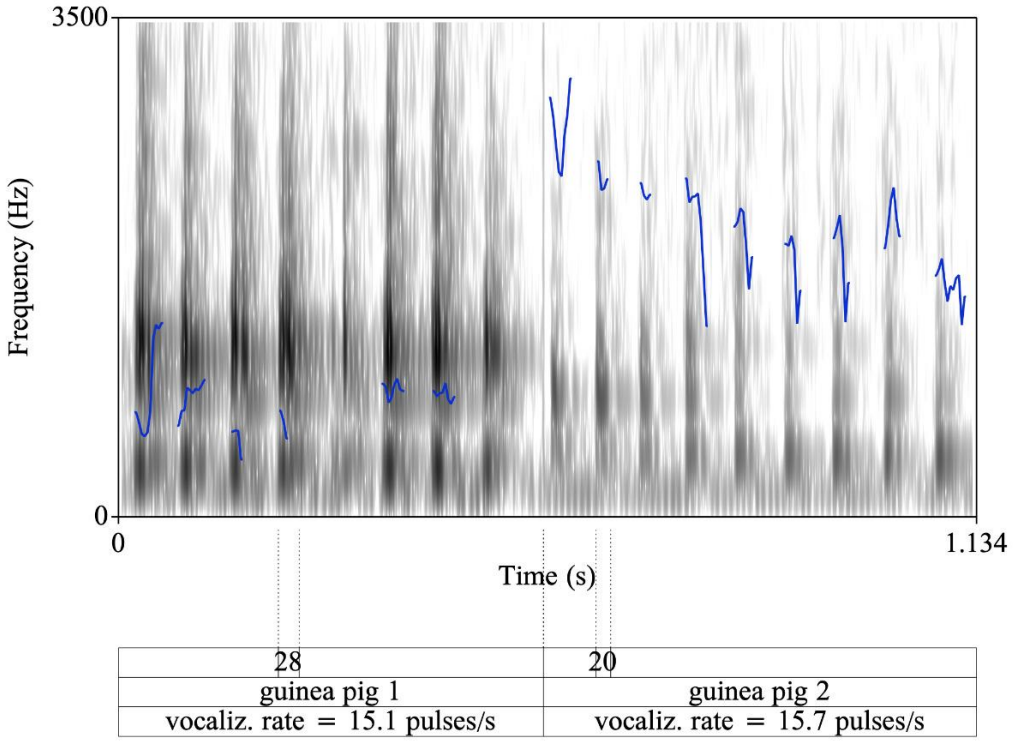


Figure 9.3. Broadband spectrograms of the courtship vocalizations of two guinea pigs with F0 contours superimposed in blue. The vocalization to the left is slower, more intense and has lower F0 than the one on the right.

The broadband spectrograms of the courtship vocalizations of two guinea pigs differ in speech rate (from left to right: 15.1 vs. 15.7 pulses/second), and main-energy chord (note, for some authors: 28 vs. 20 ms). Fundamental frequency is lower for the

2’ guinea pig vocalizations, as discussed below, associated with higher relative intensity (1.5 vs. 0.7 dB). These figures, associated with inferential statistics, contribute to differentiating the two guinea pigs.

Intensity analysis

Praat makes it possible to compute global intensity, but the actual meaning of intensity depends strongly on controlling the distance of the microphone from the sound source. A decibel meter can be used to address this issue and obtain a reliable measure of intensity. As an alternative, relative intensity, a measure of the difference in energy between two band frequencies of a long-term spectrum, is simple to determine in Praat, providing the user with a measure that is not affected by microphone position. It can be computed after the user selects a spectrum object and uses the Query menu to determine the limits of the two frequency bands to be subtracted one from the other. For instance, by subtracting the energy of a low frequency band from that of the total spectrum, a measure correlated to vocal effort can be obtained (Traunmüller & Eriksson, 2000). This also works in vocalizations to detect the strength of a call, as illustrated in the previous section, where the vocalization on the left of the figure is around twice as strong as the one on the right, calculated by computing the difference in energy between the spectrum up to 11 kHz and the band between 0 and 400 Hz.

Fundamental frequency analysis

In speech, fundamental frequency (F0), the acoustic correlate of vocal folds vibration, is the main parameter for signaling pitch and intonation. It plays an important role in the study of prosody, by demonstrating differences in the manner of speaking, which helps distinguish broadcasting from political discourse, for instance. The Praat F0 tracking algorithm is very robust, even in the case of relatively low signal-to-noise ratios, as depicted in Figure 9.4, which shows the F0 contours of the same utterance for the same speaker, with two different levels of additive Gaussian noise expressing approximate signal-to-noise ratios (SNRs) of 16 dB (left) and 8 dB (right).

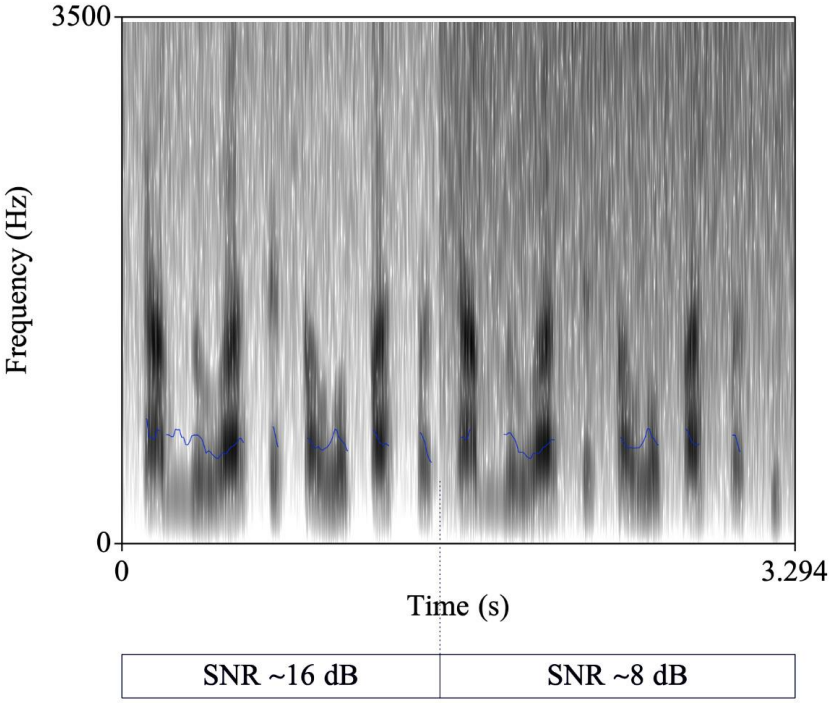


Figure 9.4. F0 contours (blue) in broadband spectrograms for the same utterance from the same speaker for two SNRs: 16 dB (left) and 8 dB (right). See text for details.

There are only a few differences in the traces shown in the figure for the two SNR conditions, which do not produce relevant changes in F0 means and standard deviations. F0 can also be tracked for non-human animal vocalizations, which may help in individual recognition from voice, if combined with relative intensity, duration and spectral analysis, as in the case of Forensic Phonetics (Barbosa et al., 2020).

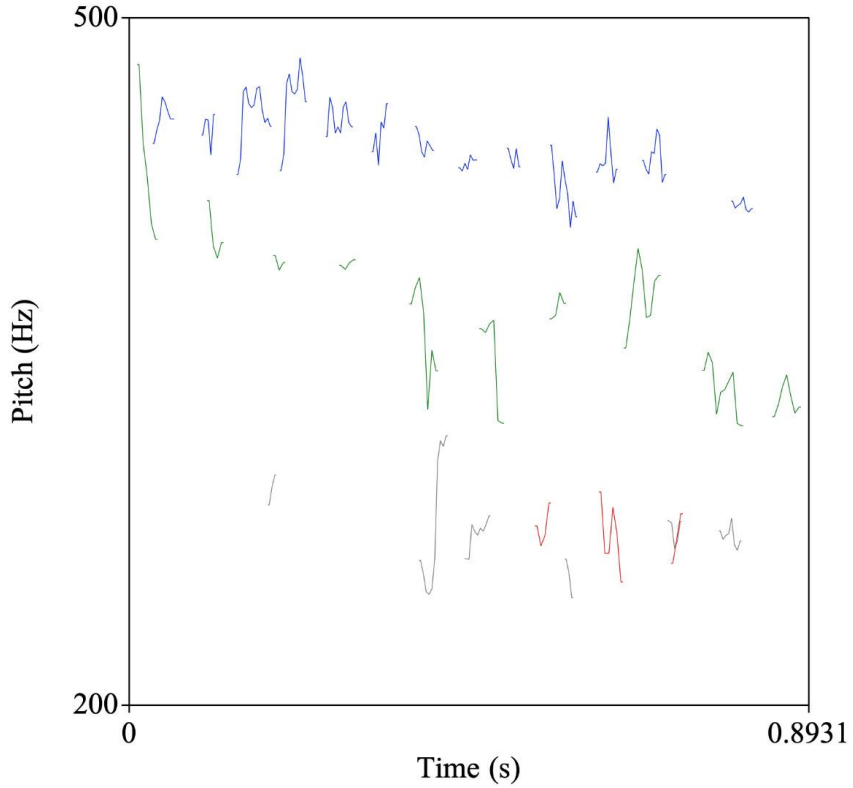


Figure 9.5. Differently colored F0 traces in Hertz for the courtship vocalizations of four different guinea pigs.

Figure 9.5 illustrates clearly differentiated F0 traces of the courtship vocalizations of four different guinea pigs. Despite differences in F0 tracking due to low intensity and noise in the case of the call, whose F0 is in red, the high-pitched individuals are clearly distinguishable from each other and from their low-pitched counterparts. The two high-pitched guinea pigs exhibit F0 medians (and standard deviations) of 369 (33) Hz and 443 (16) Hz against 275 (15) Hz for the low-pitched guinea pigs. This could be used not only for individual recognition purposes, but also to explain female preferences in the case of successful individuals with particular F0 medians and ranges.

Scripting

By far, the main advantage of Praat is the possibility of programming, using the Praat language, which allows any type of algorithm to be implemented, including

perception tests. Communication with the user during execution of a script is also possible in Praat language, which has the advantage of allowing manual corrections for some analyses, when necessary.

Several acoustic analyses can be made completely automatic for sound corpora and are widely used by the speech research community. This is evidenced by the availability of remote repositories with free Praat scripts accompanied by manuals and examples for different types of tasks and experimental situations. Ours can be found at: <https://github.com/pabarbosa/prosody-scripts>.

One of these scripts, the Prosody Descriptor extractor, computes a large set of prosodic parameters, including relative intensity, F0 statistical descriptors, local and global measures of duration for segments and pauses, and voice quality measures, among others. Some of these measures could be adapted to non-human animal vocalizations, making it possible to consider the prosodic features, which are highly relevant for analyzing animal communication because they signal affective states such as emotion and stress.

Acknowledgements

The author is grateful for research grant number 302194/2019-3 awarded by CNPq, and to Patricia Monticelli and her students Bruna Paula and Paula Olívio from the Programa de Pós-Graduação em Psicobiologia, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto, Universidade de São Paulo [Graduate Psychobiology Program of the Faculty of Philosophy, Sciences and Letters of Ribeirão Preto at the University of São Paulo] for their fruitful discussions on non-human animal communication.

References

Barbosa, P. A. (2019). *Prosódia*. São Paulo: Parábola.

Barbosa, P.A., Cazumbá, L.A.F, Constantini, A.C., Machado, A.P., Passetti, R.R., & Sanches, A.P. (2020) [Eds..] *Análise Fonético-Forense: em tarefa de Comparação de Locutor*. Campinas: Millennium Editora.

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.26 Retrieved november 2020: from http://www.praat.org/

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, *107*(6), 3438-3451.https://doi.org/10.1121/1.429414

Peer Commentary

By Patrícia Ferreira Monticelli[14]

The chapter of Professor Plínio Barbosa is fascinating. He delighted us when accepted the invitation to write the book foreword and did it wholeheartedly. I was pleased to chair his session and learn that he is a zoologist at heart. Barbosa collaborates actively for years with the EBAC lab with enthusiasm for non-human animal's vocal production. Now, Barbosa and I gladly share the outcomes of our collaborative work towards advancing animal communication studies with you.

This chapter briefly comments about tools and resources for improving acoustic analyses, such as dealing with compressed audio recordings. For example, the MP3 and OPUS codecs compact speech signals to save space, but acousticians normally avoid these audio recording formats on sound analysis. Barbosa's orientation is that compressed audio can be safely used in sound analysis when we set predefined goals with sound data collection. To illustrate, Barbosa uses the fundamental frequency (F0): the difference between measurements taken on PCM-codified signals and the compressed ones will be only about 4 to 5%. Therefore, the concern should be around sampling rate and signal-to-noise relation.

About the Praat software, it offers broad application in sound analysis either for human and non-human animals. We can easily run basic acoustic analysis in Praat when there is no need to include phonetic analysis. Barbosa wrote Praat scripts (https://github.com/pabarbosa/prosody-scripts) and created a series of short lessons freely accessed on the web at https://youtube.com/playlist?list=PL0d036y-KYM5Q75JCD5-RqUHezbr4gfWO. Thus, Praat tutorial is available on Paul Boersma & David Weenink's website (https://www.fon.hum.uva.nl/praat/manual/Intro.html).

Praat is traditionally used by phoneticians who are not used to the well-known software in the bioacoustics' community, such as Raven Pro (https://ravensoundsoftware.com/software/raven-pro/) and Avisoft SAS Pro (https://www.avisoft.com). Up to date, we cannot find review papers that offer a comparison between Praat and other software for acoustic analysis of animal sounds.

---

[14] Professor head of the Ethology and Bioacoustics Research Laboratory at the University of São Paulo in Ribeirão Preto.

Barbosa and I compared that three software to test whether Praat delivers as fine results as the others. We usually run basic analysis including time-, formant- and F0-related measurements. Praat exceptionally offers jitter and shimmer measurements and speech synthesis that run through scripts commonly available by online users (similarly to the R platform community).

Bioacousticians invest a ton of effort to refine methodologies for acoustic analysis of animal sounds. The terrestrial mammals offer a challenge with their variable non-linear effects and the overlap with background noise below 1kHz. I asked Barbosa to help us with the formants' identification and description. He compared our issue with human babies' cries he studied. Formants are displayed as spectral distances between harmonics on the spectrogram. The lower the frequency, the closer they are, and it will be harder to detect formants between them; it may be impossible if there is background noise. To illustrate, a 500Hz fundamental frequency (F0) will produce a harmonic in 1000Hz; the spectral range between them is only 500Hz. We can use some filtering process, but it will not completely remove noise without harming frequencies of interest.

Barbosa again contributed inspiring ideas to examine prosody in twin cries we are interested in. We should investigate its rhythm and melody if we were to find a difference in twins. Nevertheless, he advises always considering the whole individual, considering unexpected variations among twin pairs, and massively increasing the sound collection.

Lastly, I will briefly describe Barbosa as a detective bringing Forensic Phonetics to police intelligence by comparing two pieces of speech collected from targeted people during criminal investigations. To match convicted people's speech with the speech collected in the crime scene, Barbosa usually prefers searching for acoustic parameters varying more inter- than intra-individuals, like third formant frequency, F0 minimum and mean values. In short, it is possible to catch a person on a lie once the confounding factors of age are gotten rid of; for example, supralaryngeal and laryngeal parameters change substantially until someone reaches adulthood, which influences voice quality.

Chapter 10

Detecting events in acoustic signals

Paulo do Canto Hubert Junior[15]

Abstract

Acoustic event detection is a broad area that is undergoing a new wave of interest due to data availability, the appearance of new methods and algorithms and potential new applications. In this chapter, we analyze the definition of an event and what is meant by event detection; we also propose dividing the problem into three classes: unsupervised, semi-supervised and supervised event detection. We discuss the main challenges of each problem class, and briefly discuss possible approaches for each one.

Keywords: event detection, MFCC, supervised-detection

If you listen to a recording of Johann Sebastian Bach's Cello Suite no. 1 in G Major, you might notice a number of different things. You will perceive every stroke that the cellist makes, every new note emitted, then the sound stops. You might perceive patterns composed of several notes that repeat during the composition, sometimes with slight modifications. You might sense the instant where the mood of the music changes, rising to a climax just before it ends. And, of course, you will most probably perceive when the music begins and ends.

All of these possible perceptions can be broadly described as events that are present in this particular acoustic signal. They differ in nature: some of these events are linked to changes in the frequency distribution of the signal, others in rhythm or intention. In every case, however, there is something that had a distinct, identifiable pattern or form, and then changed, became something else, to change back to what was before, or to something newer still.

This is how we propose to define an acoustic event: some modification in one or many of a signal's features that has a well-defined duration. We are usually interested in events that arise as a result or consequence of a physical phenomenon,

[15] Fundação Getúlio Vargas, Escola de Administração de Empresas de São Paulo (EAESP), EP-USP, paulo.hubert@fgv.br

which would be what we really want to detect. However, in terms of signal processing, an event is a modification in one or many of the signal's features with a well-defined duration, and by event detection we mean the estimation of two instants: the beginning and end of the event.

This definition is not without its shortcomings. For example, we implicitly assume that an event is something that starts abruptly: at time t0 it is absent, at time t0 + epsilon it is present. This is not always the case, however, and we will later discuss how this affects our modelling. We also adopt the name of event detection rather than the more well-known signal detection to highlight an important difference: the signal detection problem, as understood in the engineering sciences, means distinguishing a signal from background noise. Event detection, on the other hand, involves distinguishing a signal among many possible signals and background noise. In this respect, event detection encompasses traditional signal detection problems.

Next, we will focus primarily on the analysis of underwater acoustic signals; in these signals, events are caused by several phenomena: the chanting of fish, the clicking and snapping of claws, the rain, breaking waves, and the presence or passage of vessels driven by humans. These events often overlap, both in time and type of modification they induce to the acoustic signal. This frequent overlapping makes event detection very difficult.

According to our definition of an event, it is clear that to detect an event one must understand what features of the signal are altered, and in what way. This can be determined in a number of ways leading us to a classification of event detection methods, which we will introduce in the next section.

And finally, a disclaimer: in this text, we do not aim to provide a comprehensive theory of event detection. Our goal is to provide the reader with a broad landscape of the types of problems and possible solutions that have emerged.

Classes of event detection problems

We have divided the event detection problem into three classes, depending on what information is available to the analyst. We adopted the nomenclature that is now standard in the field of machine learning, and defined the following three classes of event detection problems:

1. Supervised event detection: in this case, the analyst has access to examples of the signal corresponding to the event of interest. In this situation, algorithms such as neural networks can be used to learn from the examples and determine the main features of the event. This knowledge can be applied to the analysis of new data.
2. Semi-supervised event detection: in this case, the analyst has a mathematical model for the signal corresponding to the presence of the event. This model is usually obtained from theoretical principles and algebraic manipulation. The event detection algorithm in this case usually involves the statistical testing of hypotheses.
3. Unsupervised event detection: in this case, the analyst has no examples or explicit models for the event. The algorithm must then start by segmenting the signal, that is, finding contiguous sections that have stable features different from those of other sections.

Supervised event detection

The current literature on event detection and signal processing (Xia et al, 2019; Dang et al, 2017; Xia et al 2017) shows that this type of problem has raised considerable interest recently. This is partly due to the significant success that algorithms such as neural networks and random forests are experiencing in other fields.

There are numerous challenges involved when applying these algorithms to the problem of acoustic event detection. First and foremost is the database problem: it is well known that for any supervised machine learning algorithm to work efficiently, it must be fed with many examples that are good representations of the event it must learn to detect. In the case of acoustic event detection, it also must incorporate different background noise conditions, distance from the sensors, etc. In addition, all these examples must be annotated, that is, it must be known whether the event is present or absent in each example fed to the algorithm. This means that one must either produce examples in the laboratory (which means that the background conditions might be very different from the real situation in which we want to apply the detection algorithm), or search, listen to and categorize instances of the signal under many different conditions.

Apart from these challenges, there is also the problem of feature engineering and extraction, which involves determining the best way to represent a signal, such

that the presence or absence of the event is readily distinguishable (Kiktova, 2013). One example that illustrates this problem is the use of deep learning methods for acoustic event detection. These methods are highly effective in detecting events on images; here, an event is a given pattern (a dog, cat, house, car) that can appear in different ways and in different positions inside an image. Thus, when applying deep learning methods to acoustic event detection, it is tempting to transform the problem into an image processing problem, for example by first transforming the signal into a spectrogram or an MFCC (Mel Frequency Cepstral Coefficients), two dimensional structures that can be understood as images.

However, standard deep learning architectures are invariable with respect to the position of the event inside the picture; a dog is a dog, whether it appears lying on the floor (bottom of the image) or on top of a closet (top of the image). In a spectrogram or MFCC, however, the same pictorial pattern can have very different meanings if it appears in the low (low frequency events) or top section (high frequency events) of the spectrogram. Currently, the Laboratório de Acústica e Meio Ambiente - LACMAM, at EP-USP [Environment and Acoustics Laboratory at Polytechnic School of the University of São Paulo EP-USP] is conducting research on how to break that symmetry so that the machine learning algorithms can learn that a pattern on a spectrogram can have different meanings depending on their position on the frequency axis.

There are also other difficulties with the choice and extraction of acoustic signal features. Spectrograms and MFCC, for instance, depend on many parameters, such as the time window, time and frequency resolution, windowing, etc. This makes the problem of optimal feature selection very intensive in a computational sense, since a grid search must be performed for all the combinations of these parameters, combined with the network architecture and hyperparameters separately, in order to determine the most accurate detection algorithm.

Semi-supervised event detection

In some situations, the phenomenon that gives rise to the acoustic event we want to detect is well understood in the theoretical sense, and a mathematical model is available that describes what a signal must be like when the event is present. Let’s assume for a moment that we have a model f(t) for the acoustic signal of the event in

the time domain. We define the signal obtained from the sensor as y(t). Now, assume that at any given instant *t* the event *f* might or might not be present. Assume also that the sensor is imperfect, such that there will always be a background noise component, r(t). One can then think of a model with the following form:

$$y(t) = x(t)*z(t) + r(t)$$

Here z is a binary function, which takes the value 1 when the event is occurring at time t, and 0 otherwise. By restricting z to a binary function, we are assuming that the event is either present or not; it can be absent at t0 and present at t0+eps, where eps represents the time resolution of the signal. In principle, this model can be generalized to adopt a nonnegative function z that would allow the event to be present at different, continuous intensities (we could call this z a gain function for the event). The simplest instance of this problem consists of taking a fixed duration section of the signal, assuming that z(t) is constant (either 0 or 1) throughout this entire range, and using some statistical testing procedure to test against.

In Hubert et al (2018) we proposed an algorithm that follows this strategy. The goal of the analysis was to build a boat detection algorithm for underwater acoustic signal data. The model f(t) is obtained from the literature on irradiated ship noise, and the Full Bayesian Significance Test of Pereira and Stern (1999) is the hypothesis testing method adopted. The following has been learned from this study, in terms of semi-supervised event detection: first, the model must be as accurate as possible. A model that is too generic (such as a harmonic model with k harmonics) will inevitably lead to many false positives. There is also a computational challenge, since the statistical testing procedure must be applied ideally to every t during the entire signal duration. Given the typical duration of the event of interest, this problem can be attenuated; nevertheless, the computational cost of the testing procedure cannot be too high, or it will render the detector useless (unless there is no hurry in pointing out that the event was present at some time in the past).

In addition, there is the problem of defining the alternative model, that is, the model for noise term r(t). If detection is performed in a silent environment, where it is very unlikely that an event, different from the one we are trying to detect and from simple noise (for instance Gaussian white noise), will be present, this method will work well. This is the case of traditional signal detection problems that are dealt with in standard signal processing textbooks. However, in more complex environments

such as the sea, the captured signal y(t) will often be formed by the background noise emitted with some other event. For instance, fish choruses, whales chanting, and many other animal vocalizations, but also rain, breaking waves with varying intensity, or other natural or anthropogenic phenomena that can occur in the sea. In this situation, any statistical testing procedure will tend to reject the null hypothesis (signal is noise only), even if f(t) does not accurately describe what is occurring in the signal.

The solutions for the above problems involve obtaining a more accurate or restricted f(t) (in the paper, for example, we chose to adopt a chirp model that would only detect accelerating ships), and collect many different models for other possible events that might be present. From a statistical standpoint, the problem becomes one of model selection (Bretthorst, 1990).

Unsupervised event detection

This type of problem typically appears when a long duration signal is available, with little or no information about the events that may or may not be present at given times. The goals of the analysis are many: to find and extract examples of a specific event; organize the information to allow efficient direct inspection by specialists; discover new or interesting events that were not expected; and describe the environment in terms of what types of events are taking place and when.

It is in this scenario that our definition of an event becomes critical. In unsupervised event detection we must estimate the beginning and end of events about which we have little or no information. Our definition says that an event implies a change in some characteristics of the underlying signal, but if we do not know what characteristics will change, how can we begin to build a detector?

The Acoustics and Environment Laboratory (LACMAM, from the Portuguese acronym) has been developing acoustic sensors and collecting data from many environments for the past 8 years. One of these datasets consists of 10 months of continuous underwater acoustic signals from Parque da Laje, a marine conservation unit off the coast of Brazil. This dataset was built without any particular event in mind. The idea was to explore the underwater soundscape, and later use the data for different purposes (such as extracting examples with the presence of boats, and feeding these examples into a supervised learning algorithm).

Direct inspection of this dataset, however, is costly; spectrograms can be obtained to guide the analyst to potentially interesting sections of the signal that can then be listened to. Adopting this strategy to analyze this many spectrograms visually is not a very efficient way to process the data.

With this problem in mind, in Hubert et al (2019) we propose a signal segmentation algorithm that searches for changes in total signal power. From a statistical standpoint, this means we break the signals into sections with different variances. Given our definition of an event and the discussion above, a change in variance is the most general way in which we believe a signal can change. Of course, this is not necessarily the case; the signal power frequency distribution may change, for instance, without changing the total power. However, for transient events that are taking place in a noisy and rich environment, we believe this will seldom, if ever, be the case. In addition, whenever an animal starts vocalizing, or a boat's engine is turned on, as with many other possible events found in the underwater soundscape, the total power of the signal will change because these events involve a new source, with a power source of its own.

Thus, our model assumes that a sudden change in signal variance occurs at time t0. By using Bayesian (probabilistic) methods, we derive a posterior model for t0 and obtain the maximum posterior (MAP) estimate for t0 in a given signal. Next, we apply a statistical testing procedure (again the Full Bayesian Significance Test) to test the hypothesis that the variances of the two sections are indeed different. If the testing procedure rejects the null hypothesis of equality of variances, the algorithm then proceeds recursively to each of the segments: obtaining the MAP estimate for the variance changepoint, breaking the section into two segments, and then testing the equality of variances hypothesis. The algorithm stops when there are no two contiguous segments with different variances.

This segmentation algorithm can be used as a first step for the unsupervised event detection problem. Given its generality (searching for changepoints for signal variance), it is able to efficiently detect segments of the signal where something new started to take place. In a second step, the segments can be processed and features that are believed to be useful representations extracted.

The third step in an unsupervised event detection procedure would then be to cluster these signals based on the extracted features. Thus, from a continuous long duration, we end up with a few clusters of similar segments (according to the chosen

features). These clusters can then be inspected and labeled, and used for instance to feed supervised learning algorithms. The clusters are also interesting in themselves, since they allow for a concise description of the soundscape depicted in the original, long-duration signal.

Concluding remarks

In this chapter, we aimed at providing a broad vision of the acoustic event detection problem. We categorized this problem into three classes: supervised, semi-supervised, and unsupervised event detection problems. We believe that this division helps us organize the possible solutions and analysis methods and better understand their applications and limitations.

It is often the case, however, that one will encounter two, or even all three forms of the problem in the same research project and using the same data. This is what occurred in LACMAM, when we were given the problem of building a boat detector from underwater acoustic data. We wanted to apply a supervised algorithm; however, we would have needed examples to train the algorithm. We then developed unsupervised methods that could guide the exploration of our sample data and help us find sections of the signal that contained examples of boats passing by. These methods helped us obtain annotated samples for the supervised event detector design.

We also experimented with the semi-supervised setting, obtaining from the engineering literature a closed form model for the irradiated noise of a ship's engine, and applying the traditional solution method of testing the hypothesis of noise only against the hypothesis of noise plus signal. We found that of the three acoustic event detection problems, this can be the most challenging, especially when detection must occur in a rich environment with many concurrent events that might have similar features to the one we are interested in.

Further research on these problems includes modifying machine learning methods to the acoustic event detection problem; investigation of efficient and informative features to be extracted from the signal to facilitate event detection; design of novel segmentation and clustering methods to use in the unsupervised event detection problem.

Acknowledgments

The author is grateful to the University of São Paulo, SHELL Brazil (subsidiary company of Royal Dutch Shell) and FAPESP, through the “Research Centre for Gas Innovation" (RCGI) hosted by the University of São Paulo” (FAPESP Grant Proc. 2014/50279-4).

References

Bretthorst, G. L. (1990). Bayesian analysis. II. Signal detection and model selection. *Journal of Magnetic Resonance (1969)*, *88*(3), 552-570. https://doi.org/10.1016/0022-2364(90)90288-K.

Dang, A., Vu, T. H., & Wang, J. C. (2017, December). A survey of deep learning for polyphonic sound event detection. In *2017 International Conference on Orange Technologies (ICOT)* (pp. 75-78). IEEE. https://doi.org/10.1109/ICOT.2017.8336092

Hubert, P.; Killick, R.; Chung, A.; Padovese, L. (2019). A Bayesian binary algorithm for root mean squared-based acoustic signal segmentation. *The Journal of the Acoustical Society of America, 146(3), 1799-1807.* https://doi.org/10.1121/1.5126522

Hubert P., Stern J.M., Padovese L. (2018). Full Bayesian Approach for Signal Detection with An Application to Boat Detection on Underwater Soundscape Data. In: Polpo A., Stern J., Louzada F., Izbicki R., Takada H. (eds) *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*. Maxent 2017. Springer Proceedings in Mathematics & Statistics, vol. 239. Cham: Springer . https://doi.org/10.1007/978-3-319-91143-4_19

Kiktova, E., Lojka, M., Pleva, M., Juhar, J., & Cizmar, A. (2013, June). Comparison of different feature types for acoustic event detection system. In *International Conference on Multimedia Communications, Services and Security* (pp. 288-297). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-38559-9_25

Pereira, C.A.B. Stern, J.M. (1999). Evidence and credibility: full Bayesian significance test for precise hypotheses. *Entropy, 1(4),* 99-110. https://doi.org/10.3390/e1040099

Xia, X., Togneri, R., Sohel, F., & Huang, D. (2017, July). Random forest classification based acoustic event detection. In *2017 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 163-168). IEEE.doi: https://doi.org/10.1109/ICME.2017.8019452

Xia, X., Togneri, R., Sohel, F. et al. (2019). A Survey: Neural Network-Based Deep Learning for Acoustic Event Detection *Circuits, Systems, and Signal Processing*, *38*(8), 3433-3453. https://doi.org/10.1007/s00034-019-01094-1

Peer Commentary

By Arnaldo Candido Junior[16]

The chapter of Herbert Jr. focused on Machine Learning techniques, including supervised, semi-supervised, and unsupervised methods for detecting sound events in maritime environments. Maritime sounds were captured by a submerged hydrophone created by their research group. An event could be, as he exemplified in his text, a ship passing by.

The main purpose of the research was to detect illegal fishing based on the sounds made by the vessels. The proposed methods make it possible to distinguish between two different events occurring concurrently, as long as the two events do not start simultaneously.

Supervised learning is not recommended in small and variable sample data studies. For this reason, the proposed methods are hybrid and not purely supervised. It is also challenging to separate two different events occurring concurrently. It will work only if the amplitudes and frequencies from these events differ. Alternatively, alternatively, if the sounds caused by these events do not follow the same statistical distribution. It is also recommended, in this case, that a larger time window be used with this type of analysis.

The main purpose of the research was to detect the signal in time and not space. However, the method can be improved to detect location. Furthermore, the presented techniques are useful for working with automatic accounting monitoring in biodiversity-rich environments like tropical forests. Thus, acoustic tomography can also work as a possibility for detecting the origin of signals.

[16] Professor of the Federal Technological University of Paraná, Brazil.

Chapter 11

Automated classification of cry melody in infants

Silvia Orlandi, Claudia Manfredi, Andrea Bandini[17]

Abstract

Melodic features characterize infant crying. To date, several studies have extracted melodic parameters using qualitative and quantitative methods to investigate the significance of cry melody as an instrument to obtain information related to the health conditions of infants. The development of automated techniques based on artificial intelligence algorithms is necessary to classify the melodic contour and understand if cry analysis can be used as a marker of neurological and neurodevelopmental conditions in infancy. Here, we describe the state of the art of melody classification approaches in infant crying. Barriers and facilitators of artificial intelligence techniques for the analysis of melodic features are discussed. In addition, an example of deep learning architecture to categorize melodic features based on a synthetic dataset of infant cries is presented. The optimization of automated cry melody techniques can be performed only with a huge amount of data that are not always available. Additional studies conducted in ambulatory and home environments are needed to evaluate the applicability of cry analysis software tools for early diagnosis and assessment of health status and pathological conditions. Furthermore, the automated melody classification can be the key to describing linguistic precursors for early diagnosis of speech and language disorders.

Keywords: Infant cry, Cry melody, F0 contour, Automated recognition, Acoustical analysis

Infant cry analysis is a non-invasive technique that can support clinicians in the early detection of neurodevelopmental disorders in children (Orlandi et al., 2012, 2017; Esposito et al., 2017). Although acoustic features of cry can be perceptually identified by parents and clinicians, it is not easy to detect the cause and assess the health conditions by only listening to a cry sound. Listening to an audio signal and visually inspecting its spectrogram to identify acoustic characteristics and categorizing

[17] Silvia Orlandi https://orcid.org/0000-0003-2733-8450
Claudia Manfredi https://orcid.org/0000-0001-6364-9753
Andrea Bandini https://orcid.org/0000-0002-3509-2887

the cry units (CRUs) can be a challenging task without an automatic method. Moreover, the perceptual analysis requires highly trained clinicians (Manfredi et al., 2018, 2019).

Automated analysis of infant cry has made great strides in recent years arousing the interest of a growing number of researchers working in clinical, engineering, and computer science fields (Jeyaraman et al., 2018). For example, automated cry analysis may facilitate the detection and assessment of neurological conditions, such as autism spectrum disorders (Orlandi et al., 2012) and hearing conditions (Várallyay, 2004). Moreover, several research studies identified differences between preterm and term infants using cry detection methods based on signal processing and artificial intelligence techniques (Orlandi et al., 2012, 2015; Shinya et al., 2017; Oller et al., 2019).

There are several metrics and parameters that can be extracted from a cry recording. Besides the number and the duration of a CRU, energy and spectrographic features, such as fundamental frequency (F0) and resonance frequencies have been used to detect differences among the types of cry sounds. In the past twenty-five years, several studies analyzed the melodic contour of CRUs (Schönweiler et al., 1996; Várallyay, 2004, 2007; Várallyay et al., 2007; Mampe et al., 2009; Orlandi et al., 2017; Wermke et al., 2017; Manfredi et al., 2018, 2019; Prochnow et al., 2019; Armbrüster et al., 2020). It has been demonstrated that melody plays an important role in the diagnosis of respiratory distress syndrome (Matikolaie et al., 2020) and neurologic disorders (Várallyay et al., 2007). The melodic contour or cry melody is represented by the F0 waveform in the time domain, which was introduced for the first time by Schönweiler et al. (1996) who identified four melody shapes as falling, rising, rising-falling, and flat or plateau. However, later studies identified additional melodic shapes. Várallyay et al. (2007) categorized up to 77 different shapes but only 20 shapes represented 95% of the CRUs analyzed.

A recent study by Armbrüster et al. (2020) showed how regular melodic intervals seem to characterize the cry of typically developing infants, suggesting that melody contour analysis may be a potential marker of voice control in infancy. Moreover, the melodic analysis of neonatal cry represents an effective instrument to discriminate language characteristics (Wermke et al., 2017; Manfredi et al., 2019; Prochnow et al., 2019). Furthermore, the melody of spontaneous crying increases in its complexity over the first months of life (Wermke & Mende, 2016) from a single-

arc structure towards its complex-arc counterparts. As such, cry melody analysis can be used to understand the stages of vocal development towards language. For this reason, several research studies suggested that cry melody may be an indicative parameter of delays in early language development (Wermke et al. 2007, 2011; Shinya et al. 2017).

This chapter discusses the application of automated detection methods to identify the melody contour of infant crying. An overview of the current state of the art based on artificial intelligence techniques is provided to support researchers in the development of novel approaches for the detection of melodic shapes of infant crying. Lastly, we present the first results of deep learning approaches applied on a synthetic dataset of infant cries to classify CRUs based on their melodic shapes.

Materials and Methods

Although several studies have been conducted to identify cry characteristics and melodic features to facilitate non-invasive screening and assessment in newborns and infants, there is a lack of information related to barriers and limitations of automated approaches that can support the use of cry analysis as a neonatal screening technique. In the following paragraphs, we summarize the different ways artificial intelligence techniques have been used for the detection and classification of melodic patterns of infant cry. We also discuss the main challenges that must be overcome in order to develop efficient and accurate methods based on machine/deep learning techniques.

Melodic patterns

A cry recording usually contains several CRUs. Each CRU has a well-defined F0 trend defined by its values in the time domain. The time variation of the F0 values in each CRU represents the cry melody and can be characterized by sharp melodic patterns that can be clustered into different groups. However, consensus on the terminology used to define different aspects of crying and its melody has yet to be reached. Melodic features analyzed in previous articles were referred to as F0 fluctuations (Asthana et al., 2015) or prosodic features (Rodriguez & Caluya, 2019; Ji et al., 2019, 2020) in infant crying. Qualitative studies based on the visual inspection of spectrograms allowed identifying the representative melodic shapes of crying.

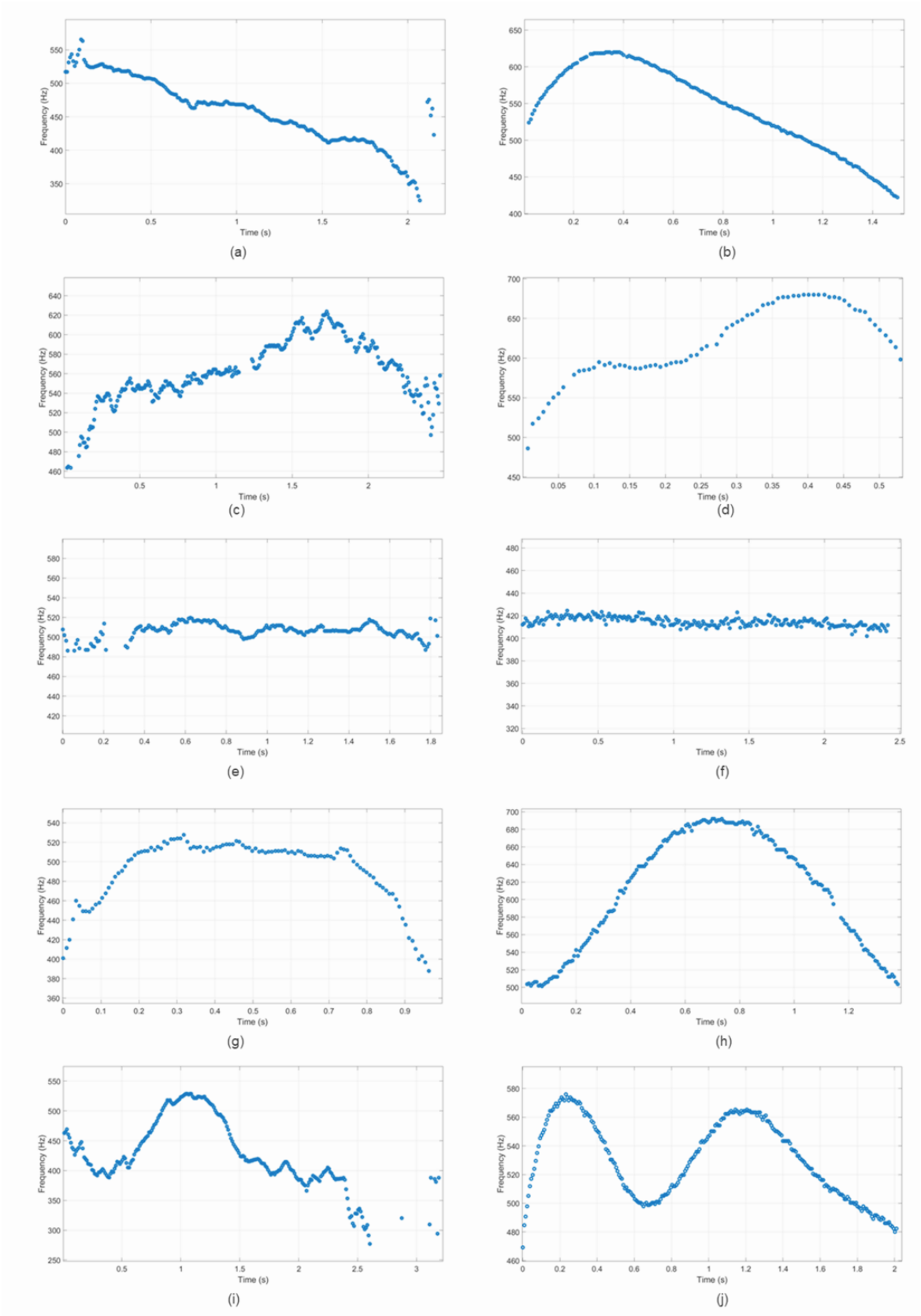


Figure 11.1. Example of melody contours of real (left column) and synthetic (right column) cry units: falling shapes (a and b); rising shapes (c and d); plateau shapes (e and f); symmetric shapes (g and h); and complex shapes (i and j).

Four typical melodic contours have been observed in newborns (Wermke et al., 2009, 2011), namely the symmetrical pattern (i.e., single-arch melody shape almost symmetric with respect to its midpoint), the rising pattern (i.e., single-arch melody shape with a slow F0 increase followed by a rapid decrease, and asymmetric shape skewed to the right), the falling pattern (i.e., single-arch melody shape with a rapid F0 increase followed by a slow decrease, and asymmetric shape skewed to the left), and the plateau (i.e., melody profile with an almost constant frequency). A fifth shape was also identified and denominated as complex (double-arch melody shape) (Wermke et al., 2007; Diaz et al., 2012). Melodic patterns of synthetic and real CRUs are shown in Figure 11.1.

Automated detection of melodic and prosodic features

The first study on the automatic detection of cry melody contour was published in 2009 by Várallyay and colleagues (Várallyay et al., 2009). They identified 39 melodic shape categories using an automated implementation of the Five Line Method (Várallyay, 2004, 2007) based on the representation of the melodic contour as a line on sheet music. Rising, falling, plateau, symmetric, and complex shapes were identified using a threshold applied to the signal energy (Díaz et al., 2012). Then in 2016, three of the basic melodic shapes (i.e., rising, falling, and plateau) were detected using derivatives of the F0 contour (Oren et al., 2016). Automated methods for melody detection have been tested on synthetic and real cry datasets (Orlandi et al., 2017; Manfredi et al., 2018, 2019) identifying falling, rising, plateau, symmetric, and complex shapes using BioVoice software (Morelli et al., in press).

To date, this free software tool is the only one available for performing a fully automated analysis of cry audio recordings extracting information about melody shapes. BioVoice can also support the perceptual analysis of the cry melody, providing the representation of the F0 in the time domain for each cry unit of an audio recording. BioVoice can identify 12 melodic shapes, namely: Falling, Rising, Symmetric, Plateau, Low-Up, Up-Low, Frequency Step, Double, Complex, Undefined, Not-a-Cry, and Other. These shapes are a subset of the shapes presented by Várallyay (2004, 2007). A metric to measure the complexity of cry melody has been defined as the melody complexity index (MCI) by Wermke and colleagues (Wermke et al., 2007). The MCI is the ratio between the number of cries consisting of multiple-arc melodies (MA) and the sum of MA and the number of cries characterized by single-arc melodies.

The MCI is usually applied for intergroup statistical comparisons (e.g., preterm and term comparison, types of crying, health conditions, etc.). Cry melody has also been described using “tilt features”, which were used to parameterize the melody features capturing the F0 contour variations. The amplitudes of the F0 contour are classified as descending and ascending and, along with the lengths of the descending and ascending portions of the contour, are used to determine the tilt features as described in Matikolaie et al. (2020).

Artificial intelligence for melody detection

Most of the research studies on cry melody conducted to date detected melodic features applying signal processing techniques and visual inspection of the spectrograms. The spectrogram and F0 are usually estimated using software tools (Shinya et al., 2017; Prochnow et al., 2019) or in-house scripts (Asthana et al., 2015; Sharma & Mittal, 2017). Six articles described automated signal processing techniques that allow classifying melodic shapes using signal processing techniques and statistical metrics (Várallyay, 2009; Diaz et al., 2012; Oren et al., 2016; Orlandi et al., 2017; Manfredi et al., 2018, 2019). Three studies applied machine learning techniques to distinguish infant crying (i.e., sleep, feeding, pain, discomfort cries) based on F0 fluctuations and melodic/ prosodic features (Osmani et al., 2017; Rodriguez et al. 2019; Matikolaie et al. 2020). Support Vector Machine (SVM), Bagging and Boosted Trees, as well as Decision Tree classifiers, were applied by Osmani et al. (2017) to evaluate infant physiological states such as hunger pain or discomfort. Rodriguez et al. (2019) focused on similar objectives using different algorithms, such as Decision Tree (J48), Neural Network, and Support Vector Machine achieving classification accuracies of up to 83.87%. Matikolaie and colleagues (2020) applied SVM classifiers based on frequency features along with tilt features to detect infants with respiratory distress syndrome obtaining an accuracy rate of 73.80%. Lastly, only one study recently applied deep learning approaches using deep neural networks to detect asphyxia from prosodic features, achieving classification accuracies of up to 96.74% (Ji et al. 2019).

New horizons in artificial intelligence: barriers and facilitators of the automated analysis of infant cry melody

The detection and classification of cry melody through artificial intelligence approaches requires large amounts of data, as well as a uniform and standardized categorization of the cry samples into specific clusters (i.e., falling, raising, etc.). Unfortunately, only one cry recording dataset is available upon request, the Baby Chillanto Database. This database was collected by the National Institute of Astrophysics and Optical Electronics, CONACYT Mexico (Reyes-Galaviz & Reyes-Garcia, 2004; Rosales-Prez, 2015).

Most of the articles reported in the previous paragraphs used the Baby Chillanto Database for algorithm performance evaluation. Moreover, there are no cry datasets with melody ground truth, such as with melody annotation performed with perceptual analysis. Only a very small dataset of 20 synthetic cries with melody annotation is available upon request (Orlandi et al., 2017), but its size is insufficient for training machine and deep learning models. To date, the paucity of infant cry datasets available online, and the lack of guidelines and protocols to conduct perceptual analysis and visual inspection make the application of artificial intelligence techniques based on melodic features a challenging field. In fact, considering the avant-garde of the current machine learning and deep learning algorithms, a rigorous categorization of the melodic shapes would allow not only to identify the different melodic patterns automatically, but also to study the prosodic sequence of a series of cries, providing information on the linguistic characteristics or health conditions of infants. For this reason, neonatal cry and sound experts should work together to build multidisciplinary collaborations to categorize and generate guidelines for visual inspection of melody cry. Rigorous protocols for melody shape categorization will allow researchers to label their datasets using standardized guidelines, in order to unify the efforts towards improving the automated analysis of cry melody via machine and deep learning algorithms. These methods have the advantage of being fast and very efficient if fed with accurate training data, but they require large datasets. Future research studies should focus on building annotated cry corpora available online or upon request.

Another limitation of the automated methods is related to the quality of the audio recordings, which should be collected in quiet environments. Recording infant

cries with good sound quality can be challenging. In cases of a limited amount of good cry samples, a valid alternative is the generation of synthesized cry data (Orlandi et al., 2017; Manfredi et al., 2018) that can be combined with real cry samples for expanding existing datasets or used as a standardized reference to test the performance of novel automated cry melody algorithms.

Deep learning classification of synthetically generated cry units

To illustrate the potential of deep learning techniques in the automated classification of newborn cry melody shapes, we implemented two types of deep neural network architectures on synthetically generated cry units: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNNs are a class of deep neural networks developed for image recognition and classification tasks that base their functioning on convolutional kernels for extracting spatial features from an image in a hierarchical manner (i.e., simpler patterns of image features are combined together to obtain more complex patterns) (LeCun et al., 2015; Krizhevsky et al., 2017). LSTMs fall within the class of recurrent neural networks (RNNs) that are capable of learning long-term dependencies in sequences of data (Hochreiter & Schmidhuber, 1997). LSTMs have been used in speech and language applications such as speech recognition and machine translations (Graves et al., 2013; Bahdanau et al., 2014), although combinations of CNN and LSTM architectures have also been implemented for video recognition tasks (Donahue et al. 2015).

After generating the synthetic data (see next paragraph for further details), three approaches were compared. A rule-based approach (called “Baseline approach" here) was implemented using BioVoice, as proposed by Manfredi et al. (2018), which determines the melodic shape according to the best fit with pre-defined curves. A CNN-based architecture with two convolutional layers (with 8 and 16 filters, respectively) and two fully-connected layers (with 256 and 5 units, respectively) was implemented to classify the 5 melodic shapes from the spectrogram generated with the Short-Time Fourier Transform. Lastly, an RNN-based model composed of an LSTM layer with 256 units followed by two fully-connected layers (with 256 and 5 units, respectively) was implemented to recognize the 5 melodic shapes based on the temporal F0 values of the cry unit. The F0 was estimated using BioVoice F0 estimation (Morelli et al. in press).

Synthetic dataset

A dataset of 10,000 synthesized cry units was generated. This dataset included 2,000 cry units for each one of the 5 basic melodic shapes: complex, falling, plateau, rising, and symmetric. The dataset was generated using the Newborn Cry Synthesizer proposed in Orlandi et al. (2017). This tool synthesizes newborn cry signals with different melody shapes. Given that the F0 of typical development newborns usually varies between 200 Hz and 800 Hz, we synthesized the cry units with F0 values within this range. Moreover, each cry unit was generated by randomly varying the parameters reported in Table 11.1 within predefined ranges, so that each synthesized cry unit had a unique combination of these parameters.

Table 11.1. Parameters and ranges of values used to generate 10,000 synthetic cry units.

	Duration [s]	F0* [Hz]	F1 [Hz]	F2 [Hz]	F3 [Hz]	F0 noise Std [Hz]	Noise amplitude Std
Lower limit	0.5	500	1100	2500	5400	1	0
Upper limit	2.5	700	1300	2800	5600	8	0.0005

*Measured in its max limit (F0 max)

Results and Discussion

The dataset was randomly split into training-set (6000 cry units, 1200 per class), validation-set (2000 cry units, 400 per class), and test-set (2000 cry units, 400 per class). Performance of the three methods was compared based on the results obtained on the test-set and were evaluated using the following metrics: accuracy (i.e., the number of correctly classified CRUs divided by the total number of CRUs), precision (for each shape is the number of true positives divided by the sum of true positives and false positives), recall (for each shape is the number of true positives

divided by the sum of true positives and false negatives) and f1-score (harmonic mean of precision and recall).

CNN- and LSTM-based methods outperformed the baseline approach, whereas the LSTM-based performed slightly better than its CNN-based counterpart. Classification performance is reported in Table 11.2.

Table 11.2. Cry melody recognition performance on a synthetic dataset.

	Accuracy	Precision	Recall	F1-Score
Baseline approach	0.8955	0.9151	0.8955	0.8998
CNN-based	0.9895	0.9896	0.9895	0.9895
LSTM-based	0.9915	0.9915	0.9915	0.9915

This chapter describes the state of the art of melody classification approaches in infant crying. A brief state-of-the-art review was presented identifying barriers and facilitators of artificial techniques for melodic feature detection and classification. Artificial intelligence approaches can be performed only with a huge amount of data, which are not always available. Due to the current limitations in data availability, future studies could use synthetically generated cry datasets for training and testing artificial intelligence algorithms. Our preliminary results show that deep learning approaches can be used to recognize melodic shapes in synthetic data with 99% accuracy. Synthesized datasets may also be used in combination with real data for building approaches able to recognize the cry melody from audio recordings automatically. The proposed architecture should be tested on real infant cry datasets to prove the ability of these techniques to recognize different melodic shapes. Our findings pave the way to building cry melody models that can be used to describe linguistic precursors for early diagnosis of speech and language disorders. In fact, deep learning algorithms will be able to classify the melodic patterns of crying faster,

providing information on the sequences of different CUs describing cry prosodic features.

Furthermore, future studies performed in ambulatory and home environments are needed to evaluate the applicability of cry analysis for early diagnosis and assessment of health conditions. Shareable datasets of synthetic and real cry corpora should be published to foster the development of novel approaches for cry melody classification.

References

Armbrüster, L., Mende, W., Gelbrich, G., Wermke, P., Götz, R., & Wermke, K. (2020). Musical intervals in infants' spontaneous crying over the first 4 Months of life, Folia Phoniatrica *et Logopaedica, 1-12.* i*ttps://doi.org/10.1159/000510622* F

Asthana, S., Varma, N., & Mittal, V. K. (2015, February). An investigation into classification of infant cries using modified signal processing methods. In *2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN)* (pp. 679-684). IEEE. https://doi.org/10.1109/SPIN.2015.7095282

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv*:1409.0473.

Díaz, M. A. R., García, C. A. R., Robles, L. C. A., Altamirano, J. E. X., & Mendoza, A. V. (2012). Automatic infant cry analysis for the identification of qualitative features to help opportune diagnosis. *Biomedical Signal Processing and Control*, *7*(1), 43-49. http://dx.doi.org/10.1016/j.bspc.2011.06.011

Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2625-2634). https://doi.org/10.1109/TPAMI.2016.2599174

Esposito, G., Hiroi, N., & Scattoni, M. L. (2017). Cry, baby, cry: expression of distress as a biomarker and modulator in autism spectrum disorder. *I*nternational Journal of Neuropsychopharmacology, *20(6)*, 498-503. https://dx.doi.org/10.1093/ijnp/pyx014

Graves, A., Jaitly, N., & Mohamed, A. R. (2013, December). Hybrid speech recognition with deep bidirectional LSTM. In 2013 IEEE workshop on automatic speech recognition and understanding (pp. 273-278). IEEE. https://doi.org/10.1109/ASRU.2013.6707742

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9(*8), 1735-1780. https://doi.org/doi.org/10.1162/neco.1997.9.8.1735

Jeyaraman, S., Muthusamy, H., Khairunizam, W., Jeyaraman, S., Nadarajaw, T., Yaacob, S., & Nisha, S. (2018). A review: survey on automatic infant cry analysis and classification. *Health Technology, 8(5)*, 391-404.

Ji, C., Xiao, X., Basodi, S., & Pan, Y. (2019, July). Deep Learning for Asphyxiated Infant Cry Classification Based on Acoustic Features and Weighted Prosodic Features. In *2019 Int Confon Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 1233-1240). IEEE. Retrieved May 21, 2020, from: https://ieeexplore.ieee.org/document/8875427

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84-90. https://doi.org/10.1145/3065386

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521(7553)*, 436-444. https://doi.org/10.1038/nature14539

Mampe, B., Friederici, A. D., Christophe, A., & Wermke, K. (2009). Newborns' cry melody is shaped by their native language. *Current Biology*, *19*(23), 1994-1997. https://doi.org/10.1016/j.cub.2009.09.064

Manfredi, C., Bandini, A., Melino, D., Viellevoye, R., Kalenga, M., & Orlandi, S. (2018). Automated detection and classification of basic shapes of newborn cry melody. *Biomedical Signal Processing and Control*, *45*, 174-181. http://dx.doi.org/10.1016/j.bspc.2018.05.033

Manfredi, C., Viellevoye, R., Orlandi, S., Torres-García, A., Pieraccini, G., & Reyes-García, C. A. (2019). Automated analysis of newborn cry: relationships between melodic shapes and native language. *Biomedical Signal Process Control*, *53*, 101561. https://doi.org/10.1016/j.bspc.2019.101561

Matikolaie, F. S., & Tadj, C. (2020). On the use of long-term features in a newborn cry diagnostic system. *Biomedical Signal Process Control, 59*, 101889. doi:https://doi.org/10.1016/j.bspc.2020.101889

Morelli, M. S., Orlandi S., Manfredi C. (in press). BioVoice: a Multipurpose Tool for Voice Analysis, *Biomedical Signal Process Control*.

Oller, D. K., Caskey, M., Yoo, H., Bene, E. R., Jhang, Y., Lee, C. C., ... & Vohr, B. (2019). Preterm and full term infant vocalization and the origin of language. *Scientific Reports*, *9*(1), 1-10. https://doi.org/10.1038/s41598-019-51352-0

Oren, A., Matzliach, A., Cohen, R., & Friedman, H. (2016, November). Cry-based detection of developmental disorders in infants. In *2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE)* (pp. 1-5). IEEE. https://dx.doi.org/10.1109/ICSEE.2016.7806073.

Orlandi, S., Bandini, A., Fiaschi, F. F., & Manfredi, C. (2017). Testing software tools for newborn cry analysis using synthetic signals. *Biomed Signal Process Control*, *37*, 16-22.:http://dx.doi.org/10.1016/j.bspc.2016.12.012.

Orlandi, S., Bocchi, L., Donzelli, G., & Manfredi, C. (2012). Central blood oxygen saturation vs crying in preterm newborns. *Biomedical Signal Process Control*, *7(1)*, 88-92.

Orlandi, S., Garcia, C. A. R., Bandini, A., Donzelli, G., & Manfredi, C. (2016). Application of pattern recognition techniques to the classification of full-term and preterm infant cry. *Journal of Voice, 30(6),* 656-663. https://doi.org/10.1016/j.jvoice.2015.08.007

Orlandi, S., Manfredi, C., Bocchi, L., & Scattoni, M. L. (2012, August). Automatic newborn cry analysis: a non-invasive tool to help autism early diagnosis. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2953-2956). IEEE.IEEE. doi:https://doi.org/10.1109/EMBC.2012.6346583

Osmani, A., Hamidi, M., & Chibani, A. (2017, November). Machine Learning Approach for Infant Cry Interpretation. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 182-186). IEEE.https://doi.org/10.1109/ICTAI.2017.00038

Prochnow, A., Erlandsson, S., Hesse, V., & Wermke, K. (2019). Does a 'musical' mother tongue influence cry melodies? A comparative study of Swedish and German newborns. *Musicae Scientiae, 23(2),* 143-156. https://doi.org/10.1177/1029864917733035

Reyes-Galaviz, O. F., & Reyes-Garcia, C. A. (2004). A system for the processing of infant cry to recognize pathologies in recently born babies with neural networks. In the 9th *Conference Speech and Computer*. *SPECOM-2004*, 552-557.

Rodriguez, R. L., & Caluya, S. S. (2018). Infants cry classification of physiological state using cepstral and prosodic acoustic features. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, *10*(2-3), 193-196.

Rosales-Pérez, A., Reyes-García, C. A., Gonzalez, J. A., Reyes-Galaviz, O. F., Escalante, H. J., & Orlandi, S. (2015). Classifying infant cry patterns by the Genetic Selection of a Fuzzy Model. *Biomedical Signal Process Control*, *17*, 38-46. https://doi.org/10.1016/j.bspc.2014.10.002

Sharma, S., & Mittal, V. K. (2017, October). A qualitative assessment of different sound types of an infant cry. In *2017 4th IEEE Uttar Pradesh Section International Conference Electrical, Computer and Electronics (UPCON)* (pp. 532-537). IEEE. https://doi.org/ 10.1109/UPCON.2017.8251106

Schönweiler, R., Kaese, S., Möller, S., Rinscheid, A., & Ptok, M. (1996). Neuronal networks and self-organizing maps: new computer techniques in the acoustic evaluation of the infant cry. *International Journal of Pediatric Otorhinolaryngology, 38(1),* 1-11. https://doi.org/10.1016/S0165-5876(96)01389-4

Shinya, Y., Kawai, M., Niwa, F., Imafuku, M., & Myowa, M. (2017). Fundamental frequency variation of neonatal spontaneous crying predicts language acquisition in preterm and term infants. *Frontiers in Psychology*, *8*, 2195. doi:https://doi.org/10.3389/fpsyg.2017.02195

Várallyay Jr, G. (2004). Infant cry analyzer system for hearing disorder detection, Periodica Politechnica, TU Timişoara, *Transactions on Automatic Control and Computer Sciences*, 49, 57–60.

Várallyay Jr, G. (2007). The melody of crying. *International Journal of Pediatric Otorhinolaryngology*, *71(11)*, 1699-1708.

Várallyay Jr, G., Benyó, Z., & Illényi, A. (2007, February). The development of the melody of the infant cry to detect disorders during infancy. In *Proc. IASTED International Conference on Biomedical Engineering (BioMED 2007), Innsbruck, Austria, February* (pp. 14-16).

Várallyay, G., Illényi, A., & Benyó, Z. (2009, December). Melody analysis of the newborn infant cries. In *Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)* (pp. 7-10).

Wermke, K., Birr, M., Voelter, C., Shehata, W. D., Jurkutat, A., Wermke, P., & Stellzig, A. E. (2011). Cry melody in 2-month-old infants with and without clefts. *The Cleft Palate-Craniofacial Journal*, *48(3)*, 321-330. https://doi.org/10.1597%2F09-055

Wermke, K., Leising, D., & Stellzig-Eisenhauer, A. (2007). Relation of melody complexity in infants' cries to language outcome in the second year of life: A longitudinal study. *Clinical Linguistics & Phonetics*, *21(11-12)*, 961-973. https://doi.org/10.1080/02699200701659243

Wermke, K., & Mende, W. (2009). Musical elements in human infants' cries: in the beginning is the melody. *Musicae Scientiae*, *13*(2_suppl), 151-175.

Wermke K, Mende W. (2016). From Melodious Cries to Articulated Sounds: Melody at the Root of Language Acquisition. In: *M.C. Fonseca-Mora and M.Gant (eds.) Melodies, Rhythm and Cognition in Foreign Language Learning*. Lady Stephenson Library. Newcastle upon Tyne: Cambridge Scholars Publishing.

Wermke, K., Ruan, Y., Feng, Y., Dobnig, D., Stephan, S., Wermke, P., ... & Shu, H. (2017). Fundamental frequency variation in crying of Mandarin and German neonates. *Journal of Voice*, *31(2)*, 255-e25. https://doi.org/10.1016/j.jvoice.2016.06.009

Peer Commentary

By Regis Rossi A. Faria and Bruna Lima Ferreira

Silvia Orlandi has conducted exceptional interdisciplinary research work in bioacoustics, employing signal processing for sound analysis and accumulating a large body of knowledge on the characteristics of babies' vocal emissions and the detection process of different types of cries (pain or feeding) based on their melodies. She is now with the Bloorview Research Institute (Holland Bloorview Kids Rehabilitation Hospital) in Toronto, Canada.

Analyzing 100 ten-days-old babies' cries, Orlandi and her team used F0 and the resonance frequencies and obtained contrasting results: F0 should increase, according to the literature, but decrease. Thus, three months old babies have higher frequency cries, suggesting a developmental delay in their very young babies. She tested herself to recognize one baby among others on 12 recordings using the cry melody among other features and suggested that it be considered in future studies. There is also a gap in the normative ranges for typically developing children.

She has been working on a comparison between preterm and full-term babies. The preterm (those born before 37 weeks of gestational age) do not have their vocal tract ready, while the full terms have specific patterns. Therefore, she expects that there are ontogenetic changes in intonation in the prosodic features. It is like a boy experiencing voice change during puberty and even at early development. Nevertheless, in the newborns she studied, cry changes much faster; every week is different.

A distinctive product of Orlandi research supervised by Claudia Manfredi is the BioVoice software (see at https://github.com/ClaudiaManfredi/BioVoice). The software classifies melodic shapes and obtains derivatives, thresholds, and other voice features. It uses a long-short term memory (LSTM) classifier, explores convolutional neural nets (CNN), and generates synthetic datasets based on the interesting sound, in their cases, newborn cry (Orlandi, 2017). The lecture she aired on November 19, 2020, is available online at https://youtu.be/vyIi9s_LPp0.

Chapter 12

Zygosity diagnosis in adult twins based on voice resemblance

Claudio Possani[18]

Abstract

The focus of this chapter is to describe the state of art of the studies on the similarity of twin voices in different aspects such as phonetic parameters (fundamental frequencies and others), recognition of speakers by a jury of listeners and forensic aspects. Studies on the subject of twin voices are new but there is an increasing number of reports and papers attempting to understand this issue. The methods and goals are quite different, as we will analyze in this text. We will also describe a collection of twin voices recorded at the University of São Paulo by the Painel USP de Gêmeos (USP Twin Panel Project) and ask a number of questions about the possibility of detecting zygosity in an adult pair of twins by voice. This approach seems to be unprecedented. It is a multidisciplinary project conducted by a group of researchers from different departments at the University of São Paulo and is based in the Institute of Psychology.

Keywords: Automatic recognition; Euclidean distance; forensic data; Gaussian Mixture Models; speech.

Acoustic analysis of twins' voices

A new area of studies in twins is the acoustic analysis of recorded voices. Different kinds of recordings and different software have been used to understand voice similarity in monozygotic (MZ) and dizygotic twins (DZ) as well as non-related speakers. Fuchs et al. (2000) analyzed seven parameters of vocal performance and three acoustic features. Thirty-one monozygotic twin pairs were compared with 30 control group pairs of non-related persons. The study found more similarity in the MZ group than the non-related group in seven of the ten acoustic characteristics (vocal

[18] Institute of Mathematics and Statistics, University of São Paulo, SP, Brazil.cpossani@ime.usp.br

range, highest and lowest vocal fundamental frequency, fundamental speaking frequency, maximum voice intensity, number of partials, vibrato of intensity).

In a case study with a pair of male monozygotic twins and an age and sex-matched sibling, Whiteside and Rixon (2013) found that the mean fundamental frequency (FO) and sentence duration were more similar in the MZ pair than with the matched sibling. The lowest Euclidean distance values were also recorded between the twin pair, using different phonetic parameters. Significant similarities were observed in all three siblings for the speech tempo and dynamic F0 parameters, suggesting the importance of environmental factors.

Studies on twin voices in Brazil are scarce. Cielo et al. (2012) conducted two studies on two pairs of adult monozygotic twins, one of each sex. The conclusions are similar to those reported in the literature.

All of these studies support the idea that some fundamental voice parameters, such as mean FO, are under considerable genetic influence, while global variation patterns such as temporal patterns, reading style and accent, are shaped more by environmental factors.

Human recognition of identical twins by voice

We will mention here two interesting studies where a jury of selected listeners were asked to identify twin speakers. In the first (Van Gysel et al, 2001), 20 female and 10 male voices of MZ twins were randomly assembled with the voice of a non-related person, creating 30 trios of voices. A jury of ten female Speech and Language Pathology students were asked to detect the twins in the trios, under two conditions: two standard sentences read aloud and a 2.5 second section of sustained /a/. The result was that 82% of the female voices were labelled correctly for the sentences and 63% for the /a/. For the male voices, the numbers were 74 and 52% respectively. The authors performed acoustic analysis of the voices and found a higher correlation in the twin voices.

In another study (Swapna et al, 2013), the sample consisted of 10 monozygotic pairs of twins, 5 females and 5 males. The voices were mixed in such a way that there were some pairs of MZ twin voices and pairs with the voice of the same speaker twice. The sound recorded was a sustained /a/. The jury was composed of 5 Speech Language Pathology students. The result was that 91.6% of the same speaker recordings and 80% of the twin pairs were correctly identified. In both studies, correct identification was

far greater than by chance. It is important to underscore that the number of participants was small.

Automatic recognition of identical twins by voice

The subject of recognition of an individual speaker has been studied for a long time and has several important applications in many aspects of life. It has been used in forensic events as well as security systems. In May, 2017, a British man fooled the security system of one of the world's largest banks by mimicking his twin's voice. For details see Simmons,2017).

An important study in this field was conducted by Kunzel (2011). The author used the Batvox3.13 forensic SPID system to compare the voices of 9 male and 26 female pairs of identical twins. The result was very promising:

"An automatic system for forensic speaker recognition (Batvox 3.1) was used to calculate inter-speaker (non-target), (2) intra-twin pair, and (3) intra-speaker (target) LR distributions. Results show that in certain conditions an automatic Bayesian-based system is capable of distinguishing even the vast majority of very similar sounding voices such as those of identical twins. However, the performance of the system used here was superior for male as compared to female voices." (Künzel, 2011, p. 251)

The author also concluded that "twins cannot generally be considered to be exact copies in terms of voice and speech." (Künzel, 2011, p. 273).

Another important study was conducted by Akin et al. (2016). The cohort consisted of 39 pairs of twins. Three voice records and the images of both ears were obtained from each pair. Two of the voice recordings and the left ear image were used for training purposes, while the third voice recording and the right ear image were used in the test. The goal was to separate one twin from the other, that is, decide which one in the pair the voice and right ear belonged to. The results were good, achieving a 90% correct identification rate.

The forensic point of view

Forensic Phonetics is the study and application of General Phonetics with the goal of contributing to the solution of legal conflicts and speaker identification. In an important paper, Jessen (2008, p. 671) stated that it consists of "the application of the

knowledge, theories and methods of general phonetics to practical tasks that arise out of a context of police work or the presentation of evidence in court". Considering that monozygotic twins share the same genetic material, distinguishing them is a challenge for forensic methodologies.

In this part we will describe three of Eugenia San Segundo's studies with different collaborators. In a 2013 paper, San Segundo and Gómez-Vilda explored phonation similarities in a group of 40 male native Spanish speakers. The distribution was as follows: 7 MZ pairs, 5 DZ pairs, 4 non-twin sibling pairs and 4 pairs of non-related persons. A vector of 65 parameters was created for each speaker, indicated by x_{sji} where s refers to the subject, i the session and j the filler. The parameters of each pair were matched considering the logarithmic likelihood ratio (LLR), as used in forensic voice matching. A Reference Speaker's Model, Γ_R, was used as reference and the logarithmic likelihood between two samples $Z_a = (x_{aji})$ and $Z_b = (x_{bji})$ was defined by

$$\mu_{ab} = log\ log\ \left[\frac{p(Z_b|\Gamma_a)}{\sqrt{p(Z_a|\Gamma_R)p(Z_b|\Gamma_R)}}\right]$$

where the conditional probabilities have been evaluated using Gaussian Mixture Models (Z_a, Z_b, Z_R) as

$p(\Gamma_a) = \Gamma_a(Z_b)$, $p(\Gamma_R) = \Gamma_R(Z_a)$ and $p(\Gamma_R) = \Gamma_R(Z_b)$.

The following conclusions were drawn: The highest LLRs were obtained in intra-speaker (same person) tests, followed by MZ inter-speaker tests and DZ inter-speaker tests. Non-twin siblings had low but above baseline LLRs and non-related speakers showed baseline LLRs. In all categories there were some pairs that had different LLRs than expected, suggesting that behavioral and environmental conditions can play an important role in voice patterns.

In a study with 54 male Spanish speakers, San Segundo and Gómez-Vilda (2014) confirmed the results of the previous work. They analyzed 12 pairs of MZ twins, 5 pairs of DZ, 4 pairs of non-twin siblings (B) and 12 unrelated speakers (US) separated into 6 pairs. Likelihood ratios (LRs) were calculated and the working hypotheses were:

H1) Intra-speaker comparisons would have the highest likelihood ratios.

H2) MZ intra-pair comparisons would yield large LRs, but smaller than in H1).

H3) DZ intra-pairs would have intermediate LRs.

H4) B would have low but above the baseline LRs.

H5) US would yield baseline LRs.

These working hypotheses were expressed by the log-likelihood ratio (LLR) as follows:

H1) LLR > -1

H2) LLR > -1

H3) LLR > -10

H4) LLR > -10

H5) LLR < -10

The general conclusion is that H1 to H5 are confirmed by the data. As an example, in 24 intra-speaker samples they found that H1 was confirmed in 17 pairs of recordings. In 12 MZ pairs, H2 was confirmed 10 times. For DZ twins, 7 out of 10 were successful.

San Segundo et al. (2016) used Euclidean distance (ED) to perform forensic speaker comparisons (FSC) in a set of double sound recordings with the following distribution: 54 same speakers (SS), 54 different speakers (DS) and 12 pairs of monozygotic twins (MZ). They collected both high quality and telephone-filtered recordings. ED was able to identify the SS and DS pairs with no false rejections. Mean ED in the MZ pairs lies between the average ED for SS and DS comparisons, as expected. Some of the MZ pairs exhibited large ED, suggesting that environment could also play an important role in voice characteristics.

USP Twin Panel project

In 2015, Professor Emma Otta from the University of São Paulo created the USP Twin Panel [Painel USP de Gêmeos], a research project focused on twins (Otta et al., 2019). Until then there was no comprehensive study of twins in Brazil. The Panel's mission is to contribute to the understanding of twinship and disseminate information that could be useful to twins and their families. The project involves different research projects and a scientific dissemination program. These studies include well-being, emotions, personality, family relationships and other aspects of interest regarding twins. In this project, a self-report zygosity questionnaire was validated into Brazilian Portuguese using DNA. It consists of four simple questions that indicate if the twins are monozygotic or dizygotic with an accuracy rate of 96%.

Among several of its initiatives, we are interested in the twin database that was collected, especially the voice recordings of 100 pairs of adult twins (around 80% monozygotic). These high-quality recordings were made in a studio.

The project is currently collecting voice recordings of twin children. It is important to underscore that the Panel has a wide range of different data, including images, videos, drawings, and photos of twins that are collaborating with the project. For more details visit the website: https://www.paineluspdegemeos.com.br/

Detecting zygosity through voice resemblance

As we showed in the previous sections of this chapter, there are a large number of studies on twin voices. Reviewing the research literature we point that:

1. Most of these studies were conducted with small samples. Sometimes with fewer than 10 pairs of twins in a group.
2. Many of these studies focused on the resemblance of twin voices. They were conducted to analyze and compare parameters extracted from voice recordings and the discussion surrounded the extent to which monozygotic twin voices are more similar than their dizygotic counterparts and so on.
3. Other studies aimed to discover if one speaker in a pair of twins could be correctly distinguished from the other through their voices. Some studies focused on human detection of voices, although there is an increasing interest in doing so automatically.
4. Other studies on twin voices have a forensic emphasis. They deal with an important issue, namely speaker identification.

We are proposing a new framework for the study of twin voices: is it possible to (automatically) detect zygosity by voice? Let me explain more clearly. Our goal is to create a software to perform the following test: given two voice recordings and knowing that the voices belong to twins, we want to decide (more or less automatically) if the pair of twins are monozygotic or dizygotic.

We are working with two different approaches which are described below[19].

[19] This project is being conducted by Emma Otta, Patricia Ferreira Monticelli, Sandra Maria Aluisio, Bruna Campos, Edresson Casanova, Ricardo Prist, Vinicius Frayze David and the author.

Artificial intelligence approach: we will use our recordings to program a speaker recognition software to differentiate MZ from DZ. The program will be inputted with a number of MZ and DZ voice recordings. It will automatically search for the adequate voice parameters that will lead to the correct identification. In other words, it is a type of reverse cluster analysis. Starting with two clusters, the program will look for the correct parameters.

Voice analysis approach: in this approach, we are trying to create a set of phonetic parameters that could identify zygosity. The idea is to use Euclidean distance or the log-likelihood ratio to measure resemblance. For instance, by following San Segundo's approach described in her 2014 and 2017 papers, we expect that the LLR would be useful in identifying if a pair of twins are mono or dizygotic.

It is important to underscore that all the studies we found in the literature contained a certain number of identification failures. In the traditional zygosity test with questions, there is a certain number of undecidable cases. One of our purposes is to combine the traditional test with the Euclidean distance or LLR to increase the number of correct decisions about zygosity. Another way of addressing this issue is to create a test that would give an answer about zygosity with a certain probability of being correct, depending on the ED or LLR values.

Acknowledgements

The author acknowledges grants no. 2014/50282-5 and 2020/14250-2 of São Paulo Research Foundation (FAPESP) and Natura Cosméticos S.A.

References

Akin, C., Kacar, U., & Kirci, M. (2016). *A multi-biometrics for twins identification based speech and ear.* International Conference on Electronics and Electrical Engineering, Turkey.a*rXiv preprint arXiv:1801.09056.*

Cielo, C. A., Agustini, R., & Finger, L. S. (2012). Características vocais de gêmeos monozigóticos. *Revista CEFAC, 14*(6), 1234-1241. Retrieved May 21, 2021, from: https://arxiv.org/ftp/arxiv/papers/1801/1801.09056.pdf

Fuchs, M., Oeken, J., Hotopp, T., Täschner, R., Hentschel, B., & Behrendt, W. (2000). Similarity of monozygotic twins regarding vocal performance and acoustic markers and possible clinical significance. *HNO, 48*(6), 462-469.

Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass, 2(4),* 671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x

Künzel, H, J. (2011). *Automatic speaker recognition of identical twins. International Journal of Speech, Language & the Law, 17(2), 251–277*.17(2). https://doi.org/10.1558/ijsll.v17i2.251

Otta, E., de Souza Fernandes, E., Bueno, J. A., Dos Santos, K. L., Segal, N. L., Lucci, T. K., ... & Ribeiro, F. J. L. (2019). The University of São Paulo Twin Panel: Current Status and Prospects for Brazilian Twin Studies in Behavioral Research. *Twin Research and Human Genetics*, *22*(6), 467-474. https://doi.org/10.1017/thg.2019.34

San Segundo, E., & Gómez-Vilda, P. (2013) Voice biometrical match of twin and non-twin siblings. In *Proceedings of the* 8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (pp. 253-256). Florence, Italy.

San Segundo, E.. & Gómez-Vilda, P. (2014). Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. *Language and Law, 1(2)*, 22-41.

San Segundo, P., Lopez, A., & Pardalos, P. M. (2016). A new exact maximum clique algorithm for large and massive sparse graphs. *Computers & Operations Research*, *66*, 81-94. https://doi.org/10.1016/j.cor.2015.07.013

San Segundo, E, Tsanas A, & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. *Forensic Science International, 270, 25-38.* https://doi.org/10.1016/j.forsciint.2016.11.020

Simmons, D. (2017). BBC fools HSBC voice recognition security system. Retrieved May 12, 2021, from: https://www.bbc.com/news/technology-39965545 , .

Swapna, S., Anto, S. B., Geethu, K. S. & Balraj, A. (2013.) An investigation into the voice of identical twins. *Otolaryngology Online Journal 3*(2).

Van Gysel, W. D., Vercammen, J., & Debruyne, F. (2001). Voice similarity in identical twins. *Acta oto-rhino-laryngologica Belgica*, *55(1)*, 49-55. Retrieved May 21, 2020, from: https://europepmc.org/article/med/11256192

Whiteside, S. P., & Rixon, E. (2013). Speech tempo and fundamental frequency patterns: a case study of male monozygotic twins and an age-and sex-matched sibling. *Logopedics Phoniatrics Vocology*, *38*(4), 173-181. https://doi.org/10.3109/14015439.2012.742562

Peer Commentary

by Vinicius Frayze David[20]

The presentation of professor Claudio Possani, "Zygosity diagnosis in adult twins by voice resemblance," was described in the previous chapter. He argued that we use too much information to identify a person in real life, not only vision, voices, or head shape; we use information gestaltically. However, it is not easy to understand how to include it in an automatic identification system. He cited a very successful study that used the image of ears to identify a person to talk about his study with twin siblings' differentiation; if they include more information than just voices, he expects to define zygosity easier.

The subject is very new, and they are still not sure exactly what they should be looking for and how. Voice is more straightforward in Possani's viewpoint; there is a long tradition of studying voices and speaker recognition, and the voice analysis framework is very well established. The novelty is using it with twins.

In his attempt to connect twin studies and voice recognition, he aims to conduct even sophisticated tests using simple cell phones that people have in their hands almost all the time. However, most researchers that work with voice recognition prefer using high-quality recordings. Furthermore, people's speech can be affected by environmental noise (Södersten et al., 2005); according to Lavan et al. (2016, p. 1604), "identity-related information from familiar and unfamiliar voices is affected by naturally occurring vocal flexibility and variability, introduced by different types of vocalizations and levels of volitional control during production." Therefore, even when using the voices of very familiar people, it makes much more sense to start with a more controlled situation.

Going back to 1985, Possani remembered Doddington saying, "Fifteen years ago when I first became involved in speech technology, I was frankly not very optimistic about the prospects for commercial application of automatic speech recognition and speaker recognition technology." In the same way, Possani found at least two automatic recognition techniques in his review and assumes that we will also

[20] Vinicius Frayze David – Master's degree in Psychobiology from the Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto da USP (FFCLRP), PhD researcher at the Instituto de Psicologia, Universidade de São Paulo (IPUSP), E-mail: vfdavid@usp.br

use more naturalistic approaches soon. San Segundo et al. (2017) used both controlled and telephone-filtered recordings when establishing similarities using Euclidean distances.

Humans are very good at using multiple information to identify objects and other humans, even if they are partially redundant, which seems to be true for other species. For example, Adachi et al. (2007) studied dogs using paired (owners face and voice) and unpaired (owner's voice and a stranger's face or the opposite) stimuli and found that dogs spent more time looking at the faces in the unpaired condition. According to the authors, it indicates that dogs can produce an internal representation of the owner's face and identify when the voice does not belong to their owners while seeing their faces. It also seems to be true for rhesus monkeys. Sliwa et al. (2011) showed that monkeys could match the voices and faces of familiar conspecifics and humans.

Von Kriegstein and Giraud (2006) experimented that associated voices with faces and other arbitrary sensory inputs. They showed that the association of voices and faces improved the participants' capacity to recognize voices in later trials. Thus, the use of multiple stimuli may also be interesting for automatic voice recognition. Another intriguing aspect of multimodal perception in humans is that we can use voices to identify emotions. However, we are much more accurate if we use visual and auditory cues (Bänziger et al., 2009).

According to Ahn and Bae (2018), many experiments prove the similarity among family members such as parents and children, and even husband and wife. Feiser and Kleber (2012), for example, showed that listeners could identify the voices of brothers, discounting by chance. Kushner and Bickley (1995) and San Segundo et al. (2017) found similar results. Rykova and Werner (2019) showed that the similarity of parents' and children's voices could be identified even when children are as young as 12 years old. However, this effect seems to be influenced by the language they speak. Some examples suggest that using previous studies with non-twins may be a good starting point for studying voice similarity in twins. It is important to underscore that most of the experiments described above were conducted with human participants, and Possani intends to use automatic recognition.

There are two directions in voice studies, and that is one of them. Possani is trying to understand suitable parameters and compare them: the pauses, the accent, which of these parameters gives us information about the similarity or zygosity of two

people. However, since he is using deep learning, the parameters should be chosen by machines. He would like to have artificial intelligence machines that could understand voices and then ask the machine how it did it. However, the machine could probably not answer. So, he does not want to abandon the idea of choosing different aspects of the voice and define which one is better to distinguish monozygotic and dizygotic twins. Nevertheless, as he said before, we do not know the good parameters; we must combine them to find out. It is a completely new field, and we are starting from scratch.

It is extremely important to be cautious when investigating a new field. Nevertheless, studies are indicating that using different aspects of speech could be useful in establishing voice similarity. For example, Kushner and Bickley (1995) showed that whole sentences are more easily recognized than reiterant syllables or individual words. The authors also showed that prosody and volume facilitate identification for the listeners.

Hanani et al. (2013) used computational methods to recognize British English speakers' accents. Gumelar et al. (2019) were not interested in voice recognition, but they used prosodic and spectral feature extraction to identify emotions in recorded voices based on deep neural networks. Again, he still thinks that approaching the subject starting with simpler measures is the best strategy, but there is room for many new designs in the near future.

And what about song learning and ontogenetic changes in voices from childhood to adulthood, especially in male voices? Possani talks about "probability" that the speakers' voices are monozygotic twins "with a very high probability." He thinks that even if life histories are very different, we can assume that genetic factors are more important. However, if there are monozygotic twins with a smaller probability, it will be possible to establish the importance of the life context. Almost every DNA test has a probability of being right and can be used in legal issues such as paternity or crime-solving.

Twin studies are interesting for several reasons, but one of them allows us to estimate the proportion of a trait produced by genetic and context variations. Using ACDE models, for example, it is possible to address this question quantitatively. The most challenging thing is understanding the context variables that can influence our traits and how they influence them. Thus, context aspects are a challenge for voice studies as it is for every other subject. Accents have to be investigated, and there is some evidence that age can play a role in voice (Taylor, 2018). Possani is excited about

this project and uses to say that the most delicious part of working with research is the process of looking for answers.

References

Adachi, I., Kuwahata, H., & Fujita, K. (2007). Dogs recall their owner's face upon hearing the owner's voice. *Animal cognition*, *10*(1), 17-21. https//doi.org/10.1007/s10071-006-0025-8

Ahn, I. S., & Bae, M. J. (2018). On a similarity analysis to family voice. *Advanced Science Letters*, *24*(1), 744-746. https://doi.org/10.1166/asl.2018.11805

Bänziger, T., Grandjean, D., & Scherer, K. R. (2009). Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). *Emotion*, *9*(5), 691-704. https://doi.org/10.1037/a0017088

Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. *Proceedings of the IEEE*, *73*(11), 1651-1664.

Feiser, H. S., & Kleber, F. (2012). Voice similarity among brothers: evidence from a perception experiment. In *21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Santander, Spain*.

Gumelar, A. B., Kurniawan, A., Sooai, A. G., Purnomo, M. H., Yuniarno, E. M., Sugiarto, I., ... & Fahrudin, T. M. (2019, August). Human Voice Emotion Identification Using Prosodic and Spectral Feature Extraction Based on Deep Neural Networks. In *2019 IEEE 7th International Conference on Serious Games and Applications for Health (SeGAH)* (pp. 1-8). IEEE. https://doi.org/10.1109/SeGAH.2019.8882461

Hanani, A., Russell, M. J., & Carey, M. J. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. *Computer Speech & Language*, *27*(1), 59-74. https://doi.org/10.1016/j.csl.2012.01.003

Kushner, R. E., & Bickley, C. A. (1995). Analysis and perception of voice similarities among family members. *The Journal of the Acoustical Society of America*, *98*(5), 2936-2936. https://doi.org/10.1121/1.414098

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(12), 1604-1614. https://doi.org/10.1037/xge0000223

Sliwa, J., Duhamel, J. R., Pascalis, O., & Wirth, S. (2011). Spontaneous voice–face identity matching by rhesus monkeys for familiar conspecifics and humans. *Proceedings of the National Academy of Sciences*, *108*(4), 1735-1740. https://doi.org/10.1073/pnas.1008169108

Rykova, E., & Werner, S. (2019). Perceptual and acoustic analysis of voice similarities between parents and young children. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 262-271). Retrieved 2020 from: https://www.aclweb.org/anthology/W19-6127

San Segundo, E., Tsanas, A., & Gómez-Vilda, P. (2017). Euclidean Distances as measures of speaker similarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International*, *270*, 25-38. https://doi.org/10.1016/j.forsciint.2016.11.020

Södersten, M., Ternström, S., & Bohman, M. (2005). Loud speech in realistic environmental noise: phonetogram data, perceptual voice quality, subjective ratings, and gender differences in healthy speakers. *Journal of Voice*, *19*(1), 29-46. https://doi.org/10.1016/j.jvoice.2004.05.002

Taylor, S. M. (2018). *Acoustic Correlates of Aging and Familial Relationship*. [Master`s thesis, Brigham Young University] Retrieved november 2020 from: https://scholarsarchive.byu.edu/etd/7041/

Von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS biology*, *4*(10), e326. https://doi.org/10.1371/journal.pbio.0040326

Chapter 13

Detecting Respiratory Insufficiency by Voice Analysis: The SPIRA Project

Marcelo Finger & Spira project group[21]

Abstract

This paper describes the initial activities of the SPIRA Project, a COVID-19 motivated research effort to design a system for the early prediction of respiratory insufficiency via audio analysis. It describes the research motivation, its organization in research lines, the initial results obtained in those lines and a preview of the future steps in this research project.

Keywords: Biomarker; Convolutional Neural Networks; COVID-19; health monitoring; prosodic; Small Data approach.

The COVID-19 pandemic has forced most of the population of the world's major cities to social distance. Large gatherings of people can facilitate the spread of the virus, especially in hospitals and health centers. Monitoring potential patients remotely, frequently and automatically is the best way to combine compliance with social distancing and patient safety.

According to specialists, one of the most important symptoms of COVID-19 that leads to hospitalization is respiratory insufficiency, a condition that is amplified in the case of the current pandemic due to the frequent occurrence of silent hypoxia, that is, low blood oxygenation without noticeable shortness of breath (Tobin, Laghi, & Jubran 2020).

[21] The SPIRA Project Group consists of Sandra M. Aluísio (ICMC-USP), Evelyn Alves Spazzapan (Fonoaudiologia-UNESP), Larissa C. Berti (Fonoaudiologia-UNESP), Augusto C. Camargo Neto (IME-USP), Arnaldo Candido Jr (CS-UTFPR-Medianeira), Edresson Casanova (ICMC-USP), Flaviane Fernandes-Svartman (FFLCH-USP), Renato Ferreira (IME-USP), Ricardo Fernandes Jr (CS-UTFPR-Medianeira), Marcelo Finger (IME-USP), Alfredo Goldman (IME-USP) , Lucas R. Gris (CS-UTFPR-Medianeira), Pedro Leyton (IME-USP), Anna S. Levin (FM-USP), Marcus Martins (FFLCH-USP), Marcelo Queiroz (IME-USP), J. Henrique Quirino (Beneficência Portuguesa-SP), Beatriz Raposo de Medeiros (FFLCH-USP), Ester C. Sabino (FM-USP), Daniel da Silva (CS-UTFPR-Medianeira).

An automatic system for early detection of respiratory insufficiency via audio analysis meets both health safety needs and medical triage burden relief. We pursue two complementary approaches to develop a detection tool. The first collects large amounts of data from respiratory insufficiency patients and healthy people, and applies artificial intelligence and machine learning techniques to obtain a speech classification system. We call this predictive task the Big Data approach. However, data-intensive approaches are notoriously unclear and do not yield satisfactory explanations of the underlying phenomena present in the audio signals. The descriptive task of providing a detailed description of signal properties pertaining to respiratory insufficiency in voice and speech signals is our second approach, called the Small Data approach.

Our general approach subscribes to the view of speech and voice as biomarkers (Botelho et al., 2019). In this respect, the goals of the SPIRA Project are as follows:

- Creation of a dataset of audios containing speech records of both respiratory insufficiency patients and healthy people who do not require hospitalization. Patient audios initially originated from COVID-19 wards.
- Development of artificial intelligence algorithms and audio processing necessary for the training and execution of classifiers that will screen patient audios (Big Data approach).
- Development of a broad acoustic description (sound signal and speech and voice acoustic) and a linguistic description of respiratory insufficiency by comparing the audio signals of patients and healthy subjects (Small Data approach).
- Implementation of an automatic audio system based on a support audio classifier to assist the patient screening system.

The present chapter is organized as follows. Related work is presented in the first section, dataset construction in the second section, small data description in the third and fourth sections, and big data analysis in the fifth section 5. Brief conclusions are presented in the last section.

Related work

COVID-19 is too recent a disease to be widely covered in the speech processing literature. Even before the outbreak of the pandemic, the literature contained only a few investigations of speech as a biomarker (Botelho et al., 2019; Trancoso et al., 2019; Nevler et al., 2019; Giovanni et al., 2021). In particular, a framework that models speech production subsystems and their neuromotor

coordination as a biomarker of COVID-19 has been proposed (Quatieri, Talkar, & Palmer, 2020).

With respect to detecting signs of COVID-19 in audio recordings, there are several research initiatives on the internet[22], by startups[23] or public entities[24], a few of which have already published initial results. For instance, the COVID-19 Sounds Data Collection Initiative (Tailor, Chauhan, & Mascolo, 2020) aims to detect the presence and severity of COVID-19, and the COUGHVID crowdsourcing dataset to develop a screening tool (Orlandic, Teijeiro, & Atienza, 2020). The following studies aim to diagnose COVID-19 from speech, breathing or coughing sounds.

Unlike our approach, no study aims specifically at respiratory insufficiency or patient triage, but propose to apply some form of artificial intelligence processing. It is not yet known if there is a detectable difference between enacted and spontaneous coughs, since most recordings are obtained from provoked situations. Positive identification of asymptomatic only COVID-19 infected patients was recently reported (Laguarta et al, 2020); identification uses artificial intelligence techniques over provoked cough-sound cell-phone recordings, and does not perform as well on patients with symptoms.

With the explicit goal of applying artificial intelligence to hospital triage using natural language processing, text extraction from radiology reports was developed (Hassanpour et al., 2017), as well as text processing from patient questionnaires (Spasic et al., 2019). These studies apply written language processing for patient screening and treatment selection.

Dataset

The voice samples were collected from two different sources. Initially, we collected audios from patients infected by SARS-CoV-2, in special COVID-19 wards at three different hospitals in São Paulo: two public hospitals affiliated with the

22 "A Respiratory Sound Database for the Development of ...." 17 nov. 2017, https://link.springer.com/chapter/10.1007/978-981-10-7419-6_6. Accessed on 23 October, 2020.

23 [1] "VoiceMed – Save lives and monitor the health of one billion people." https://www.voicemed.io/. Accessed on 23 October, 2020.

24 "Respiratory Sound Database | Kaggle." 29 jan. 2019, https://www.kaggle.com/vbookshelf/respiratory-sound-database. Accessed on 23 October, 2020.

University of São Paulo (Hospital das Clínicas and Hospital Universitário) and a private institution (Beneficência Portuguesa). Voice samples were collected only from patients with blood oxygenation levels (SpO2) below 92%, indicating respiratory insufficiency. In the hospitals, 536 samples were collected from patients of different age groups.

The second source consisted of audio recorded via a web-based application. A system was specifically implemented to collect speech audio donations from healthy volunteers. It allowed us to form a control group. The system's URL[25] was disclosed through the local news and social networking. After blank samples were eliminated, the resulting dataset was composed of more than 6000 voice donors.

An "appendix" of special recordings was created to address the fact that a COVID-19 ward is a noisy environment: we also collected recordings consisting of pure background noise at the ward, without any voice, typically at the start of a collection session. It is important information, as ward noise is very different from the background noise found in the control group, and noise is a data bias that should be controlled during experiments. Since it is difficult to filter this kind of noise in patient audios, which risks deleting important low-intensity cues to respiratory insufficiency, we decided to gather hospital and device sounds and insert this noise in the control group. It helped address overfitting issues during model training. All collected speech audios contain three different types of utterances:

- Utterance 1, a moderately long sentence containing 31 syllables and syntactic/prosodic branching constituents, designed to allow for possible breathing breaks in major syntactic boundaries (e.g., the syntactic boundary between the branching subject and the predicate) while being relatively simple to be spoken, even by low literacy voice donors: *"O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa"* (*"Love of your neighbor helps strengthening the fight against Coronavirus"*);
- Utterance 2, a well-known nursery rhyme for donors with reading difficulties, due to the lack of reading glasses in hospital, or other types of reading impediments: *"Batatinha quando nasce, espalha a rama pelo chão, nenezinho quando dorme põe a mão ao coração"* (*"When*

25 https://spira.ime.usp.br

small potatoes germinate, branches sprout on the ground; when the baby sleeps, its hands lay over the heart ");

- Utterance 3, a widely known song, was spoken: *"Parabéns a você"* (*"Happy birthday to you"*). The melody is the same as the song in English and the lyrics are: *"Parabéns a você, nesta data querida, muitas felicidades, muitos anos de vida"* (*"Happy birthday to you, on this cherished date, lots of happiness, many years of life"*).

Several issues with the original dataset were identified and addressed: class imbalance, consisting of fewer positive (COVID-19 patients) than negative cases (healthy individuals from the control group); sex imbalance, consisting of a greater number of healthy women than men participating in the process (there were also more men in COVID-19 wards than women); age imbalance, consisting of a higher number of older adults in hospital care than young people in our observations; utterance imbalance, as utterance 1 was more common among patients; healthy people typically recorded all the proposed utterances.

We addressed most dataset issues by sample balancing, taking advantage of the greater number of control group samples. Only audios from utterance 1 were selected and the number of samples used in the experiments was balanced by class and sex, but not by age, to avoid drastically reducing the available data.

Other issues led to discarding samples collected from the dataset. In some patient audios, the collector's voice could be heard, mostly assisting low literacy or visually-impaired patients when reading the utterance. Some control group recordings exhibited popping and crackling noise, possibly due to the characteristics of the recording devices.

The most serious issue for bias removal is the presence of ward background noise in patient audios; we observed that it is easier to insert ward noise in the control group than to remove it from the patients' signal. This process will be addressed in the following section.

Signal Description (small data)

The description of speech and voice has been a challenging task for this project's scope, since data collection took place in different environments and different sound capture configurations and equipment. Thus, the aspects selected for the vocal

and phonetic/phonological analysis could produce more reliable measures, which will be discussed in the section below titled Proposed measures: temporal and spectral.

Respiratory sounds and linguistic utterances are both viewed as important to the signal described in this study. As described in the previous section, creating a target sentence to utter was necessary. The length and syntactic/prosodic branching of constituents were controlled in the creation of this target sentence. A nursery rhyme and the spoken version of "Parabéns a você" were also part of the dataset, but were not recorded for both groups (patients and control) and therefore not considered for the analysis.

It is widely known that the human voice is multidimensional, since it involves a coordinated action of respiration, phonation and resonating systems (Kent, 1997; Patel et al., 2018; Asiaee et al., 2020). Any clinical or health condition that interferes with these systems may affect vocal production and vocal quality, voice aspects known as dysphonia. The literature reports that 28.6% of individuals infected with COVID-19 showed symptoms of dysphonia (Lechien et al., 2020). Asiaee et al. (2020) explain that a patient with COVID-19 may exhibit decreased or lack of energy for vocal production, leading to an interruption or change in speech production.

In order to carry out voice and linguistic analysis, we built a reliable subdataset ($n \cong 200$) that would guarantee suitable answers to our questions. In the next subsections, we present the steps to create our questions, outline the analysis model, provide some details about measures and exhibit preliminary results.

Analysis: general proposal

At the beginning of this study, we expected that the two different groups of speakers (patient versus control) would display significant differences, mainly related to the presence of noisy breathing in the patient's utterance as opposed to its absence in control group participants. However, based on the advice of the medical doctors of our project, we had to rethink our expectations, after we realized that a severe respiratory condition would not be manifested until a very advanced stage of COVID-19. However, even before listening to and visualizing the acoustic signals of voice and speech, we raised two more general hypotheses to explore: (i) presence of more pauses and (ii) vocal deviation in the patient's speech.

From these hypotheses, we were able to design a study model that would allow us to treat and analyze the data to answer the following specific questions: (1) Are

patient utterances longer than those of the control participants? (2) Are there more pauses and are they longer in patient utterance? (3) Are these pauses in the same grammatical locations in patient and control group utterances? (4) Is the speech rate (for example, syllables per second) of patients lower than that of control subjects? (5) Is the patients' mean fundamental frequency (F0) significantly different from that of the control subjects? (6) Do patients exhibit vocal deviation when compared to the control subjects?

Analysis model

The following analysis model was outlined:

- Two groups of speakers: control group and patient group.
- One target sentence.
- Measures of voice and speech aspects: duration, F0 contour, F0 mean and voice harmonicity.
- Voice aspects will be described by sex.

The target utterance was utterance 1 previously mentioned: *"O amor ao próximo ajuda a enfrentar o coronavírus com a força que a gente precisa"* (*"Love of your neighbor helps strengthen the fight against Coronavirus"*).

For the proposed model, analyses were carried out in three domains: temporal, prosodic and spectral. For each of these domains, measures were determined using Praat software, version 6.1.20 (Boersma & Weenink, 2020). In the temporal domain, we measured duration to obtain target sentence length and speech rate. In the prosodic domain, since the target sentence was isolated, we were able to describe the F0 contour and relate it to mean F0. In the spectral domain, in addition to mean F0 per participant, voice harmonicity was determined using the CPPS measure (Cepstral Peak Prominence Smoothed). For the spectral measures, sex was considered.

Proposed measures: temporal and spectral

Using Praat's textgrid annotation resource, target sentence boundaries were obtained and visually isolated from the remaining audio file portions. The criteria used to mark sentence boundaries were the waveform first pulse of the first vowel and the last pulse of the last uttered vowel. This boundary labeling was necessary for both linguistic (phonetic and phonological) and spectral analysis (voice).

Two criteria were used to measure the syllables: (i) a phonological criterion, taking into account the ideal number of syllables (31 syllables) and absence of pauses and (ii) a phonetic criterion, marking actually produced syllables and pauses. It is important to underscore that even for this criterion, in order to avoid infinite detailed segmentation, we used the expected realization in which there is resyllabification, as in / aen / of the *"a enfrentar"* portion (translated: *to face*). Thus, we proposed at least two levels of speech rate: one with 31 syllables and the other with around 26. The speech rate was calculated using the syllable/sentence duration ratio (sr=syllable/sentence duration).

In order to extract F0 and harmonicity voice parameters, the audio recordings were edited to minimize external interference at the time of recording and consider only the participants' continuous speech. Thus, we excluded portions of pauses between one vocalization and another as well as portions with device noises or the voice of a health professional, given that they could interfere with the extraction of the acoustic measures of voice. Our choice to obtain a CPPS measure relies on the fact that it shows the extent to which F0 harmonics are individualized and stand out regarding the noise level present in the acoustic signal (Asiaee et al., 2020). It is worth highlighting that in relation to vocal parameters (F0 and CPPS), the two groups were divided by sex, since male and female voices are different when it comes to the mechanics of vocal fold vibration and signal energy.

In order to determine whether the proposed questions and measures were promising, a very small data subset was analyzed in this initial stage of our study. Syllable and pause duration have yet to be extracted, implying we still do not have precise data on speech rate. However, some preliminary results were obtained, since we were able to use what we call first level speech rate. It consisted of dividing the fixed number of syllables (31) by the actual sentence duration. With respect to voice parameters, some initial observations indicated that F0 values seem to differentiate the two male groups and CCP values the two female groups.

Preliminary results

In the temporal domain, results obtained from a data subset (n=100) indicate a slight difference between groups, in both total duration (I) and speech rate (II) - syllables per second. The patient group (PG n=50) has a tendency to produce longer utterances (average=7.87s) and fewer syllables per second (average = 4.23 syl/s),

which may be related to the duration and number of pauses in patient speech, caused by respiratory insufficiency. On the other hand, for the control group (CG, n=50), the average total duration was 5.40s and the speech rate 5.88 syl/s. In addition, the control group had a smaller standard deviation in both conditions (CGsd I= 0.86, II= 0.92; PGsd I = 2.22; II=0.92).

In the spectral domain, preliminary results indicate a slight difference between female groups in F0 standard deviation (SD) and CPPS measures. The female patient group (PG) showed more unstable emission. This might be due to poorer control in sustaining F0 (PG, F0 SD=36.66Hz) compared to the control group (CG, F0 SD=22.35Hz). In addition, the female PG emitted more noise in the vocal signal, when we compared harmonic behavior (PG, CPPS=7.93dB) between groups, considering the sex variable (CG, CPPS=10.127dB). Females also exhibited more vocal deviation than their male counterparts. Males in the patient group had a higher F0 in relation to the controls (PG = 116.5 Hz; CG=133.02 Hz).

These results indicate that differences in temporal, prosodic and spectral domains may be found between groups. For the large dataset to be analyzed, measures need to be automatically extracted (see the next section of this chapter), which may account for temporal and spectral analysis. An intense dialogue with signal processing colleagues is ongoing to solve problems related to the acoustic signal and corresponding linguistic units.

Signal Processing

The signal processing team at the SPIRA project is involved in two tasks, namely the extraction of features that are important for linguistic and vocal signal descriptions (as outlined in the previous section), and the production of alternative representations that are relevant for machine learning and input signal classification (as described in the next section).

Segmentation

Audio signal segmentation is a preliminary step for many subsequent signal processing tasks. The first segmentation level consists of identifying speech utterances and background noise, which may include sound-producing electrical appliances and other voices (an occurrence that affects mainly the recordings of hospitalized patients).

A straightforward, albeit not perfect, approach is to use energy thresholding to identify the segments containing the main speaker. In this method, the level (in dB) of noise floor is estimated for each audio signal from the minimum values of the energy curve, and a threshold above this noise floor is used for the binary classification of each audio frame; subsequent smoothing (e.g., majority vote) can be applied to avoid rapid alternation between speech and noise segments.

Finer levels of segmentation (e.g., phonetic segmentation) can be obtained using other classification strategies, such as voiced/unvoiced classification, phoneme detection, etc.

Noise reduction

Reducing background noise is important for both improving feature extraction and creating alternative representations for machine learning. Concentrating spectral information on the parts that most likely belong to the speech utterance improves the signal-to-noise ratio, stabilizes the estimation of fundamental frequency (F0) and improves peak-to-valley measurements in both spectrum and cepstrum (e.g., CPP), among other benefits.

Noise gating is a well-known technique for noise reduction based on a gaussian representation of the noise spectrum. Using speech/noise segmentation to define a gaussian model of noise allows the training of an adaptive non-linear filter that selectively suppresses or attenuates specific time-frequency components within the signal's spectrogram, which may then be resynthesized as a new noise-reduced audio signal.

Feature extraction and augmented representations

Many audio features are easily obtained from the original signal and metadata such as speech/noise segmentation (Mitrović, Zeppelzauer, & Breiteneder, 2010). These additional metadata are relevant for linguistic and vocal signal description and investigation of discriminating parameters that would allow the identification of speech utterances affected by respiratory insufficiency, as well as for producing augmented representations in the context of machine learning, since metadata that are known to facilitate the classification of affected and healthy individuals would also probably ease the convergence of hyperparameters during training of automatic learning models.

Speech/noise segmentation provides the first source of many relevant features that may be useful to both description and classification. From this simple on-off description of the signal, one can obtain information on the number and duration of continuous speech utterances and interruptions, such as number of noise segments (a rough proxy for respiratory rate), the ratio between continuous speech duration and the signal's total duration, mean and variance of the duration of both continuous speech utterances and interruptions, among others.

Pitch and timbre-related descriptions may be easily obtained from F0 profiles, such as using pYIN (Mauch & Dixon, 2014), spectrograms, cepstral representations including MFCC (Hibare & Vibhute, 2014)) and harmonic representations such as HPCP (Gomez & Herrera, 2004), among others, obtained for the entire signal or for segments with continuous speech utterances. Several statistical measures can be derived from these representations, such as mean/median/std/min/max of F0 and cepstral peak prominence, which are already under investigation, as well as the characterization of voice formants and voiced/unvoiced segmentation.

Annotation transfer between signals

Automatic labeling of signals according to phonetic information, such as syllable or phoneme transcription, is a difficult task prone to a number of errors even with state-of-the-art techniques. An easier alternative is to transfer labels from signals that already received these manual annotations (which is also difficult and time-consuming for humans). This can be done by exploiting the fact that several recordings have the same spoken sentence, and by aligning recordings that receive manual annotations to those that did not.

Dynamic Time Warping is an algorithm for time-aligning two symbolic sequences that use dynamic programming to build a map of timestamp correspondences between the two sequences. It depends heavily on the choice of a representation that produces similar symbolic sequences for two speech utterances of the same sentence, regardless of the speaker's individual timbre characteristics. This requires representations such as cepstral coefficients, which are less sensitive to pitch or energy variations (related to prosody and thus varying significantly between speakers), and more sensitive to the differences between the spectral structure of the phonemes.

Signal Classification (big data)

For machine learning purposes, the dataset was divided into training (292 audios), validation (292) and tests (108), as is usual in statistical learning. We selected audios with the best signal-to-noise ratio for the test set, and the second best audios were used for validation. The aim of this partitioning is to detect training overfitting. The method chosen to classify the input signal was based on artificial neural networks. The MFCC representation is extracted from the audio and then presented to a convolutional neural network. The first step in this process is to preprocess the audios obtained.

In general, the majority of the audios in the dataset were sampled at 48kHz. We pre-processed these files using Torch Audio 0.5.0 as follows: First, for dimensionality reduction reasons, we resampled these audios at 16kHz. Second, we extracted the MFCCs using a 400ms window employing Fast Fourier Transform (FFT) (Brigham & Morrow, 1967), with hop length 160 and 1,200 FFT components, retaining only 40 coefficients. However, before applying the MFCC feature extraction process, we need to address the duration difference observed in our data.

Ward noise is a serious source of bias. In this scenario, a neural network can be biased during training by focusing only on background noise. To address this issue, we injected pure background noise samples obtained from COVID wards into patient and control group audios. A total of 16 1-minute samples were recorded. To avoid bias in this process, we decided to inject noise into all training and validating samples for both patients and the control group. We also injected noise into some testing samples in order to check for model bias. The test and validation sets were created in such a way as to allow overfitting detection, since they are composed mostly of audios with a very limited amount of noise.

Proposed Model

Several models were tested in preliminary experiments and we describe the one that provided the best results. Regarding topology and model parameters, preliminary experimental results showed that CNNs (Convolutional Neural Networks) applied to MFCCs are useful in analyzing this type of problem. Figure 13.1 presents the selected model's main features including layers, filters, kernels, number of neurons and activation functions. The following conventions are adopted in the figure: kernel size

is represented by K; convolutional dilation size (Yu & Koltun, 2015) by D; and fully connected layers by FC. The input size varies according to the experiment. We investigated the use of Mish activation function (Misra, 2019) due to its regularization effects during training, which helps prevent overfitting.

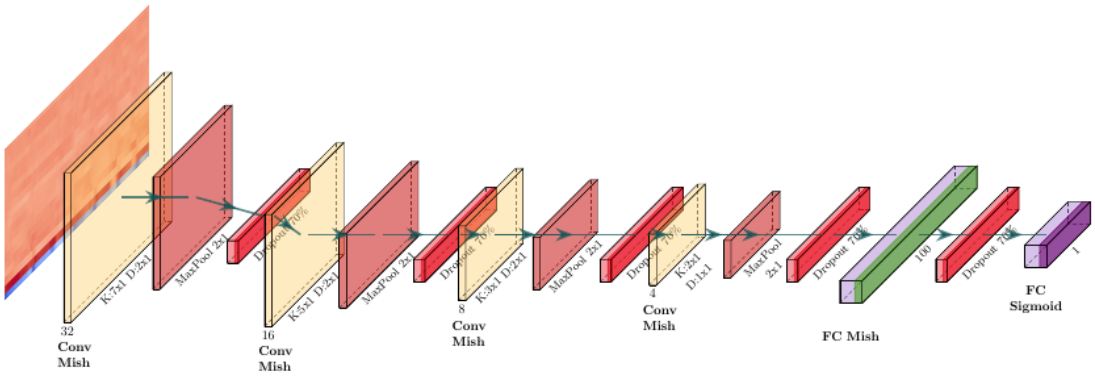


Figure 13.1. CNN topology proposed with four convolutional and two fully connected layers.

We used Binary Cross-Entropy as loss, and Adam optimizer (Kingma & Jimmy, 2014). The initial learning rate was set at 10^{-3}, and the Noam's decay scheme (Vaswani, 2017) was applied every 1,000 steps. For each proposed experiment, we trained the model for 1,000 epochs using a batch size of 30. With respect to regularization, overfitting mitigation is a major concern, given our dataset noise characteristics. As such, several approaches for regularization were applied. In addition to Mish as an activation function, we used three other strategies. First, a global weight decay of 0.01 was applied. Second, a dropout of 0.70 was used in all layers, except the output layer. Finally, after each convolutional layer, we applied group normalization (Wu & He, 2018) to pairs of convolution filters. Thus, the number of groups is half the number of filters.

Our models were implemented using PyTorch 1.5.1. We ran the experiments on a NVIDIA Titan V GPU with 12GB RAM on a server with an Intel Core i7-8700 CPU and 16GB of RAM.

Experiments and Results

Experiments were projected to determine the optimal amount of noise insertion. Note that better results were sometimes obtained without noise in test samples and vice versa. In general, bias is greatly reduced by inserting at least one noise sample into the negative instances. As expected, inserting too much noise decreases model performance. The best overall accuracy was obtained in 3 noise samples, which reached 91% accuracy in the task. The accuracy of each experiment is presented in Figure 13.2, both with and without artificial insertion of ward noise into the test samples.

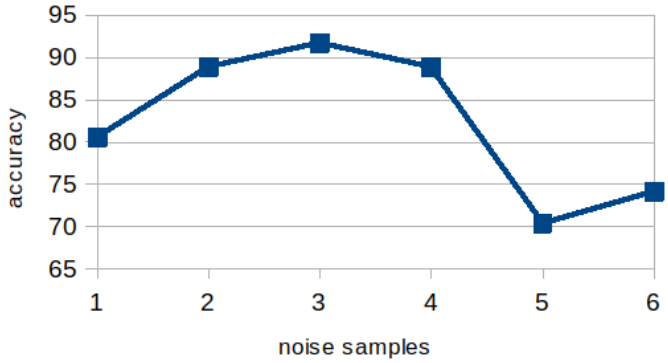


Figure 13.2. Accuracy obtained per number of noise samples inserted in training data

Concluding remarks

The initial results obtained for the SPIRA project seem to validate the original assumption of the project that respiratory insufficiency can be detected to an acceptable level of accuracy from audio signals obtained by remote recordings. Thus, we are encouraged to develop a pre-diagnostic assistance tool to help health professionals in patient triage.

Future work will involve detailed descriptions of the signal properties of patients and non-patients, as well as an extension of the current study to address respiratory insufficiency originating from causes other than COVID-19.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -- Brasil (CAPES) -- Finance Code 001; and Fapesp SPIRA Project 2020/06443-5.

References

Asiaee, M., Vahedian-Azimi, A., Atashi, S. S., Keramatfar, A., & Nourbakhsh, M. (2020). Voice quality evaluation in patients with COVID-19: An acoustic analysis. *Journal of Voice*, https://doi.org/10.1016/j.jvoice.2020.09.024

Boersma, P. & Weenink, D. (2020). Praat: doing phonetics by computer [Computer program]. Version 6.1.26 from http://www.praat.org/

Botelho, M. C., Trancoso, I., Abad, A., & Paiva, T. (2019, May). Speech as a biomarker for obstructive sleep apnea detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5851-5855). IEEE.

Brigham, E. O., & Morrow, R. E. (1967). The fast Fourier transform. IEEE spectrum, 4(12), 63-70. https://doi.org/10.1109/MSPEC.1967.5217220

Giovanni, A., Radulesco, T., Bouchet, G., A Mattei, A. J Révis, J., Bogdanski, E., Michel, J. (2021). Transmission of droplet-conveyed infectious agents such as SARS-CoV-2 by speech and vocal exercises during speech therapy: preliminary experiment concerning airflow velocity. *European Archives of Oto-Rhino-Laryngology, 278(5), 1687-1692*. https://doi.org/10.1007/s00405-020-06200-7

Gómez, E., & Herrera, P. (2004, October). Estimating the tonality of polyphonic audio files: cognitive versus machine learning modelling strategies. In *Proceedings of the International Symposium for Music Information Retrieval (ISMIR*), pp. 92–95.

Hassanpour, S., Langlotz, C. P., Amrhein, T. J., Befera, N. T., & Lungren, M. P. (2017). Performance of a machine learning classifier of knee MRI reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology*, *208*(4), 750-753.

Hibare, R., & Vibhute, A. (2014). Feature extraction techniques in speech processing: A survey. *International Journal of Computer Applications, 107(5)*, 1-8. https://doi.org/10.5120/18744-9997

Kent, R. D. (1997). *The speech sciences*. San Diego: Singular Publishing Group, .

Kingma, D. P. and Ba, J. L. (2014). Adam: A method for stochastic optimization, arXiv:1412.6980, 201

Laguarta, J., Hueto, F., & Subirana, B. (2020). COVID-19 Artificial Intelligence Diagnosis using only Cough Recordings. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 275-281. https://doi.org/10.1109/OJEMB.2020.3026928

Lechien, J. R., Chiesa-Estomba, C. M., Cabaraux, P., Mat, Q., Huet, K., Harmegnies, B., ... & Saussez, S. (2020). Features of mild-to-moderate COVID-19 patients with dysphonia. *Journal of Voice*.https://doi.org/10.1016/j.jvoice.2020.05.012.

Mauch, M., & Dixon, S. (2014, May). pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 659-663). IEEE.

Misra, D. (2019). Mish: A self-regularized non-monotonic neural activation function. arXiv preprint arXiv:1908.08681, 4. https://arxiv.org/pdf/1908.08681.pdf

Mitrović, D., Zeppelzauer, M., & Breiteneder, C. (2010). Features for content-based audio retrieval. In Marvin V. Zelkowitz (ed.) Advances in computers: Improving the Web (Vol. 78, pp. 71-150). London, UK: Academic Press

Nevler, N., Ash, S., Irwin, D. J., Liberman, M., & Grossman, M. (2019). Validated automatic speech biomarkers in primary progressive aphasia. *Annals of Clinical and Translational Neurology*, *6*(1), 4-14.

Orlandic, L., Teijeiro, T., & Atienza, D. (2020). The COUGHVID crowdsourcing dataset: A corpus for the study of large-scale cough analysis algorithms. *arXiv preprint arXiv:2009.11644*.

Patel, R. R., Awan, S. N., Barkmeier-Kraemer, J., Courey, M., Deliyski, D., Eadie, T., ... & Hillman, R. (2018). Recommended protocols for instrumental assessment of voice: American Speech-Language-Hearing Association expert panel to develop a protocol for instrumental assessment of vocal function. *American Journal of Speech-Language Pathology*, *27*(3), 887-905. https://doi.org/10.1044/2018_AJSLP-17-0009

Quatieri, T. F., Talkar, T., & Palmer, J. S. (2020). A framework for biomarkers of COVID-19 based on coordination of speech-production subsystems. *IEEE Open Journal of Engineering in Medicine and Biology*, *1*, 203-206.10.1109/OJEMB.2020.2998051. Retrieved May, 2021 from: https://ieeexplore.ieee.org/abstract/document/9103574

Spasić, I., Owen, D., Smith, A., & Button, K. (2019). KLOSURE: Closing in on open–ended patient questionnaires with text mining. *Journal of Biomedical Semantics*, *10*(1), 1-11. https://doi.org/10.1186/s13326-019-0215-3

Tailor, S. A., Chauhan, J., & Mascolo, C. (2020, September). A first step towards on-device monitoring of body sounds in the wild. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers* (pp. 708-712).https://doi.org/10.1145/3410530.3414440

Tobin, M. J., Laghi, F., & Jubran, A. (2020). Why COVID-19 silent hypoxemia is baffling to physicians. *American Journal of Respiratory and Critical Care Medicine*, *202*(3), 356-360. https://doi.org/10.1164/rccm.202006-2157CP

Trancoso, I., Correia, M. J. R., Teixeira, F., Abad, A., Botelho, M. C. T. , and Raj, B. (2019).. Speech as a (private?) biomarker for speech affecting diseases. In ICIEA 2019 -- The 14th IEEE Conference on Industrial Electronics and Applications. Keynote paper ed. Xi'an, China: IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N, Polosukhi. (2017). Attention is all you need. In: Advances in neural information processing systems. pp. 5998-6008.

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Wu, Y., & He, K. (2018). Group normalization. In: *Proceedings of the European conference on computer vision (ECCV)* (pp. 3-19). https://link.springer.com/chapter/10.1007/978-3-030-01261-8_1

Peer Commentary

by Claudio Possani

Marcelo Finger presented the talk entitled “Detecting Respiratory Insufficiency by Voice Analysis: The SPIRA Project” at the Bioacoustic Meeting Brazil 2020. He is one of the leaders of the Project that involves a large number of participants and institutions. Marcelo is well known in the computational science community for his high-level research.

The SPIRA Project intends to develop an early respiratory problem detection system from human speech recordings. It is hoped that this detection can be established based on low quality sound recordings, such as those obtained by cell phones. The connection with the COVID19 pandemic and its application in this context stands out and confers special importance to all and any advance that the project exhibits in that direction.

The project has 4 main objectives:

1. Create a large database of the voice recordings of healthy people and those with respiratory insufficiency.
2. Develop algorithms based on artificial intelligence and deep learning that classify the recordings and learn to identify the audios of patients with respiratory problems (Big Data Approach)
3. Create a system that describes sound recordings and voices, as well as a linguistic description of the signs of respiratory insufficiency by comparing the recordings of healthy people and those with illnesses, which make it possible to identify individuals with respiratory problems (Small Data Approach)
4. Create an automatic respiratory problem detection system.

Data collection involved unprecedented challenges. A total of 536 recordings of patients with COVID-19 hospitalized at 3 different institutions in the city of São Paulo were initially collected. More than 6000 recordings of healthy people were donated spontaneously to the project. One unexpected difficulty was the fact that a hospital setting where samples from COVID-19 patients were collected is full of background noise from hospital equipment, while the recordings of healthy people

were considered “cleaner”. The best solution was to introduce hospital noise into clean recordings.

Three different recordings were taken from project participants. A sentence containing 31 syllables selected in a structured manner, a well-known nursery rhyme in Brazil (entitled “Batatinha quando nasce”) and the “Happy Birthday to you” song, sung in Portuguese, were used. The first two were used for technical reasons. The researchers’ expectation, based on hard science, was that the vocal recordings of patients would be quite different from those of healthy people. The medical researchers were the most skeptical in this respect and they were right. Thus, researchers began to look for more subtle differences between the recordings of the two groups as the basis for the Small Data Approach strategy. Differences such as the amount and duration of speech pauses were considered. This approach was much more promising.

The initial results of the project are encouraging. Differentiating respiratory problems accurately has been achieved, which validates the initial project proposal. In addition to being very contemporary, this chapter is relevant and illustrates the plasticity that science demands from researchers to revise strategies and approaches that lead to the desired results.

Chapter 14

Deep Learning approaches for Speech Synthesis and Speaker Verification

Edresson Casanova[26], Christopher Shulby[27] and Sandra Maria Aluísio[28]

Abstract

Speech synthesis is the artificial production of human speech, which can be used in applications such as text-to-speech, music generation, navigation systems and accessibility for visually-impaired people. As for the speaker recognition task, we can define it as the process of recognizing the speaker of a speech segment by processing speech signals, which can be broadly classified as speaker identification and verification. This chapter summarizes the Deep Learning practices applied in the field of speech synthesis and speaker verification. Speech synthesis and speaker verification have been widely investigated in speech technology applications, especially due to the popularity of virtual assistants. Considerable research has been conducted and significant progress has been made in the last 5-6 years. As Deep Learning techniques advance in most fields of machine learning, older state-of-the-art methods are also being replaced by Deep Learning methods in both speech synthesis and speaker verification areas. Thus, Deep Learning has apparently become the next generation solution for the synthesis and verification of speakers.

Keywords: Speech Technologies; Speech Synthesis; Speaker Verification; Deep Learning approaches.

Speech synthesis systems, also known as Text-To-Speech (TTS), have received considerable attention in recent years due to the popularization of virtual assistants, such as Amazon Echo (Purington et al., 2017), Google Home (Dempsey, 2017) and Apple Siri (Gruber, 2009). However, according to Tachibana et al. (2017), traditional Speech Synthesis systems are not easy to develop, since they are typically composed of many specific modules, such as a text analyzer, grapheme-to-phoneme

[26] PhD student at the Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil
[27] Defined Crowd, Lisbon, Portugal.
28 Professor at the Institute of Mathematics and Computer Science, University of São Paulo, São Carlos, SP, Brazil

converter, duration estimator, F0 generator, spectrum generator and vocoder. Figure 14.1 presents the main components of a traditional speech synthesis system. In summary, given an input text, the text analyzer module converts dates, currency symbols, abbreviations, acronyms, and numbers into their standard formats to be pronounced or read by the system, i.e., carries out text normalization and tackles problems such as homographs, then with the normalized text, the phonetic analyzer converts the grapheme into phonemes. In turn, the duration estimator estimates the duration of each phoneme. The acoustic model is used to generate acoustic characteristics such as F0 and a spectral envelope that corresponds to linguistic characteristics. Finally, the vocoder converts the spectrum into a waveform (Ze et al., 2013).

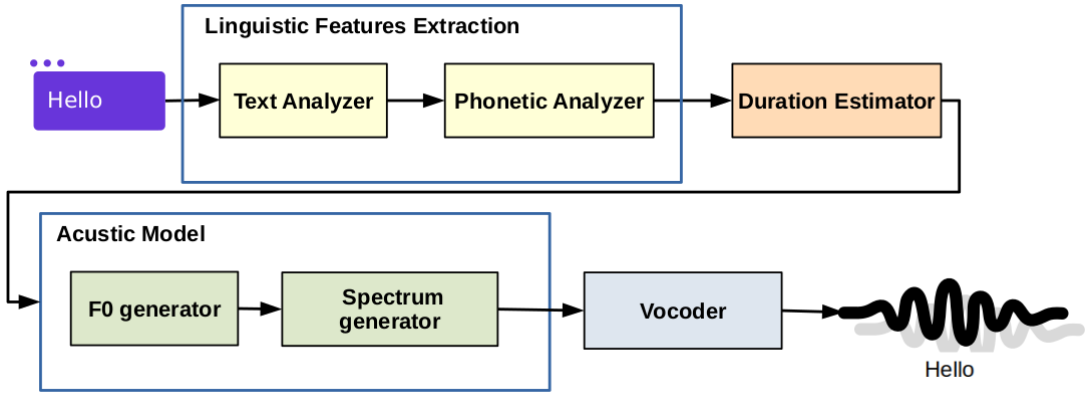


Figure 14.1. The main components of a traditional speech synthesis system.

The advent of Deep Learning (Goodfellow et al., 2016) has made it possible to integrate all processing steps into a single model and connect them directly from the input text to the synthesized audio output, which is known as end-to-end learning. Although neural models are sometimes criticized as being difficult to interpret, several end-to-end trained speech synthesis systems (e.g., Sotelo et al., 2017, Wang, Skerry-Ryan et al., 2017, Shen et al., 2018, Tachibana et al., 2018, Ping et al.,2018, Kim et al.,2020, and Valle et al., 2020) have been able to estimate spectrograms from text entries with promising performances.

Due to the sequential characteristic of text and audio data, the recurring units were the standard building blocks for speech synthesis, as in Tacotron 1 and 2 (Wang,Skerry-Ryan et al., 2017; Shen et al., 2018). In addition, the convolutional

layers showed good performance while reducing computational costs, as observed in the DeepVoice 3 (Ping et al., 2018) and Deep Convolutional Text To Speech (DCTTS) (Tachibana et al., 2018) models. On the other hand, with the recent popularization of Transformers (Vaswani et al., 2017), some transformer-based synthesis models have emerged, such as that proposed by Li et al. (2019), which performed similarly to Tacotron 2 (Shen et al., 2018), and trained 4.25 times faster. Finally, the flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the speech synthesis area, where the Flowtron (Valle et al., 2020) model surpassed the results reported by Tacotron 2 for enabling the manipulation of the latent space, allowing a change in characteristics such as speech speed and prosody. On the other hand, Kim et al. (2020) proposed GlowTTS, whose performance resembled that of the Tacotron 2, synthesizing speech 15.7 times faster.

The advent of Deep Learning has also enabled significant advances in speaker recognition. Speaker Recognition can be divided into three different subtasks: Speaker Verification (SV), Speaker Identification and Speaker Diarization. The objective of SV is to determine if two distinct audios contain the voice of the same speaker. On the other hand, speaker identification seeks to ascertain which speaker produced the voice on the audio file. Finally, Speaker Diarization splits an input audio stream into homogeneous segments according to the speaker's identity. In this study, we will only address Speaker Verification because it can be used in both of the other tasks cited above (Sztahó et al., 2019).

Currently, state-of-the-art (SOTA) Speaker Verification systems (Wang, Wang, Law et al., 2019; Deng et al., 2019; Chung, Huh et al., 2020; Casanova, Candido Junior, Shulby et al., 2020) allow the identification of new speakers without the need to retrain the model. This feature is very useful for different applications, such as meeting loggers, telephone-banking systems (Bowater & Porter, 2001) and automatic question answering (Ferrucci et al., 2010).

The objective of this study was to review the SOTA methods using Deep Learning that are applied in the speech synthesis area, focusing on Sequence-to-Sequence (seq2seq) models and speaker verification tasks. This text is subdivided in Speech datasets (main datasets employed in speech synthesis and speaker verification tasks); Deep Learning for the speech synthesis task; Deep Learning for the speaker verification task and conclusions and reflections.

.

Speech datasets

As with many tasks related to machine learning, the issue of the dataset used is fundamental. The methods developed can be evaluated and compared only if the same test circumstances are used. It is difficult to say whether an approach performs better if it is evaluated on a different dataset (or corpus) (Sztahó et al., 2019). Some datasets are used for speaker recognition and speech synthesis. Section *Speech synthesis datasets* presents the most commonly used datasets for speech synthesis in the English language, as well as the unique dataset publicly available for Brazilian Portuguese. Section *Speaker verification datasets* presents the main datasets used in the training and evaluation of speaker recognition models.

Speech Synthesis datasets

For the speech synthesis task, high quality datasets recorded in controlled environments are required. Since the purpose of speech synthesis is to synthesize high quality voice, if the training dataset contains noise, the model can synthesize it, which is not desired. The most widely used for training single-speaker speech synthesis models is the LJ Speech (Ito, 2017) dataset, which consists of 24 hours of speech by an English-language speaker. On the other hand, for multi-speaker synthesis, the LibriTTS (Zen et al., 2019) and VCTK (Veaux et al., 2016) datasets are the most commonly used. Although the most popular datasets are for English, other languages also have open datasets. With Portuguese, for example, the only publicly available dataset is the TTS-Portuguese Corpus (Casanova, Candido Junior, de Oliveira et al., 2020). Table 14.1 shows the approximate number of hours and total number of speakers of the main publicly available datasets for speech synthesis in English and the only dataset available for Portuguese.

Speaker Verification Datasets

For the SV task, the datasets created for the development of Automatic Speech Recognition (ASR) systems are commonly used due to their characteristics. Unlike speech synthesis datasets, their ASR counterparts generally have several speakers and few samples for each speaker; this feature is desired, since for Speaker Verification we want as many speakers as possible during model training (Sztahó et al., 2019). Thus, the datasets built for ASR models can be used to train and evaluate SV models. However, some datasets are made specifically for Speaker Verification. For example, VoxCeleb 2 (Chung et al., 2018) is currently the largest dataset built for SV. It consists of samples from more than 6,000 speakers downloaded from YouTube. Table 14.2 shows the approximate number of hours and total number of speakers in the main publicly available datasets for ASR and provides information about the VoxCeleb 2 dataset.

Table 14.1. Speech Synthesis datasets

Corpus	**Hours (~)**	**Total Speakers (~)**
LibriTTS (Zen et al. 2019)	586	2,456
M-AILAB	75	2
VCTK (Veaux et al. 2016)	44	109
LJ Speech (Ito 2017)	24	1
TTS-Portuguese Corpus (Casanova, Candido-Jr, de Oliveira, et al. 2020)	10.5	1

Table 14.2. Speaker Verification datasets

Corpus	**Hours (~)**	**Total Speakers (~)**
LibriSpeech (Panayotov et al. 2015)	986	2,848
Common Voice (Ardila et al. 2019)	2,508	58,250
TED-LIUM V3 (Hernandez et al. 2018)	452	2,028
VoxCeleb (J. S. Chung et al. 2018)	2,000	6,112

Sequence-to-Sequence Voice Synthesis Approaches

With the advent of Deep Learning, speech synthesis systems have evolved considerably and are still being studied intensively. Models based on Recurrent Neural Networks such as Tacotron Wang, Wang, Skerry-Ryan et al., 2017), Tacotron 2 (Shen et al., 2018), Deep Voice 1 (Arik, Chrzanowski et al., 2017) and Deep Voice 2 (Arık, Diamos et al., 2017) have gained prominence, but have high computational costs because they use recurring layers. This led to the development of fully convolutional models, such as DCTTS (Tachibana et al., 2018) and Deep Voice 3 (Ping et al., 2018), which sought to reduce the computational costs while maintaining good synthesis quality. On the other hand, more recently with the popularization of the Transformers, new Transformer-based models (Li et al., 2019; Kim et al., 2020) have emerged, and due to the parallelization of this architecture, the models achieved results similar to those of recurrent architectures with lower computing costs. Finally, the flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the synthesis area, allowing the training of simpler models with reduced computing costs. For example, the quality of the GlowTTS (Kim et al., 2020) model is similar to that of the recurrent Tacotron 2 model, but it can synthesize speech 15.7 times faster. The speech synthesis models are trained by receiving a text as input and a spectrogram as an expected output that represents the speech of the respective text input.

The model must learn to generate a spectrogram given the input text; the spectrogram is then transformed into a waveform using a vocoder. Neural vocoders have better quality speech synthesis, while phase reconstruction methods such as Griffin-Lim (GLA) (Griffin & Lim, 1984) and RTISI-LA (Real-Time Iterative Spectrogram Inversion with Look-Ahead) (Zhu et al., 2007) are based on Short Fast Fourier Transform (SFFT) redundancy (Sorensen & Burrus, 1988) and have higher synthesis speed and reduced quality. Figure 14.2 presents a general flow diagram of a TTS system based on Deep Learning. Briefly, given an input text, it is passed to the TTS model, which returns a spectrogram. Finally, this spectrogram is converted into a waveform by the vocoder.

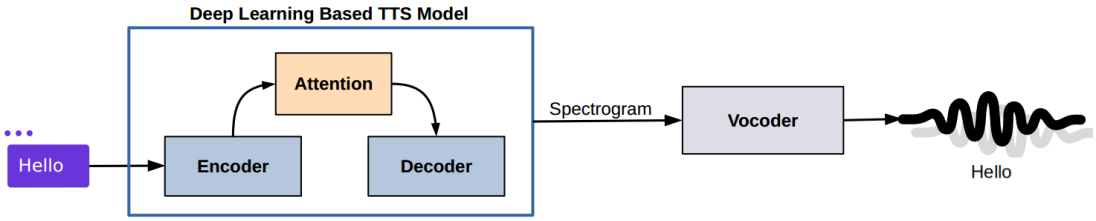


Figure 14.2. General flow diagram of a TTS system based on Deep Learning.

The most popular neural vocoders today are Wavenet (Tamamori et al., 2017), WaveRNN (Kalchbrenner et al., 2018), Waveglow (Prenger et al., 2019), GAN-TTS (Bínkowski et al., 2019), MelGan (Kumar et al., 2019) and more recently WaveGrad (Chen et al., 2020). Each of these vocoders has its advantages; some focus on higher quality and others on faster synthesis. In this study, we will not discuss vocoders, but they play a very important role in speech synthesis, converting a spectrogram into a waveform. In this chapter, we will only focus on models that convert text into spectrograms.

As mentioned above, a large amount of data is required to train speech synthesis models. For the English language, the most popular single speaker dataset for speech synthesis is called LJ Speech (Ito, 2017) and contains 24 hours of speech. On the other hand, in Brazilian Portuguese, the only available dataset is TTS-Portuguese Corpus (Casanova, Candido Junior, de Oliveira, et al., 2020) and contains 10 hours of speech. The speech synthesis models are subjectively evaluated using the Mean Opinion Score (MOS). Ribeiro et al. (2011) proposed a methodology for calculating MOS in speech synthesis and the vast majority of studies follow this technique. To calculate the MOS, the evaluators are asked to assess the naturalness of the statements generated on a five-point scale (from 1 = Bad to 5 = Excellent). Each participant evaluates the audio and the average MOS of the participant is calculated.

Tacotron 1 (Wang, Wang, Skerry-Ryan et al., 2017) was one of the first speech synthesis models to use only neural networks to transform text into a spectrogram. The authors proposed the use of a single deep neural network trained from end-to-end. Tacotron 1 includes an encoder, decoder and post-processing module, in addition to using an attention mechanism (Bahdanau et al., 2014) and convolutional filters, skipping connections (Srivastava et al., 2015) and Gated Recurrent Unit (GRU) neurons (Chung, Gulcehre et al., 2014). Tacotron also uses the Griffin-Lim algorithm

to convert the STFT spectrogram into the waveform (Griffin & Lim, 1984). Simultaneously, the Deep Voice 1 (Arik, Chrzanowski et al., 2017) model also emerged, which uses several neural submodels to synthesize speech into text. The Deep Voice 2 (Arık, Diamos et al., 2017) model was then proposed. This model is based on Deep Voice 1; however, the authors proposed improvements to surpass the results obtained by Tacotron 1. In addition, the authors proposed improvements in Tacotron 1 and changed the Griffin-Lim vocoder in favor of the WaveNet neural vocoder, thereby increasing the quality of the synthesized speech.

On the other hand, Shen et al. (2018) proposed an improvement on the Tacotron 1 model. They simplified the architecture and combined the new model with a modified version of the WaveNet (Tamamori et al., 2017) vocoder. Tacotron 2 is composed of a recurrent network of sequence prediction features that maps the incorporation of characters to Mel spectrograms, followed by a modified WaveNet model acting as a vocoder to synthesize waveforms in the time domain from these spectrograms. They also demonstrated that the use of Mel spectrograms as a conditioning input for WaveNet, instead of linguistic characteristics, allows for a significant reduction in the size of the WaveNet architecture, and consequently faster speech synthesis.

Furthermore, with the popularization of Transformers (Vaswani et al., 2017) in the Natural Language Processing (NLP) area, and the use of several language models such as BERT (Devlin et al., 2018), some transformer-based synthesis models have emerged. We can cite the work proposed by Li et al. (2019) which achieved quality comparable to that of Tacotron 2 (Shen et al., 2018), but trained 4.25 times faster.

Finally, more recent flow-based models (Kingma et al., 2016; Hoogeboom et al., 2019; Durkan et al., 2019) attracted attention in the synthesis area. Valle et al. (2020) proposed the Flowtron model, which reformulates from Tacotron 2 to provide high-quality and significant Mel spectrogram synthesis. Flowtron is optimized to maximize the likelihood of training data, which makes training simple and more stable. It allows the manipulation of several aspects of speech synthesis, such as pitch, tone, speech rate, cadence and accent. It achieved MOS scores slightly higher than those of Tacotron 2 and also allows for speech manipulation. On the other hand, Kim et al. (2020) proposed GlowTTS, whose quality is similar to that of Tacotron 2, but synthesizes speech 15.7 times faster. It uses transformers in its architecture and also

allows one to manipulate the velocity of speech. Both Flowtron and GlowTTS use the Waveglow neural vocoder.

Speaker Verification approaches

In the last decade, the area of speaker recognition has undergone major changes. In the past, speaker identification models could only identify speakers seen during training, and required a reasonable amount of speaker data to be able to learn to identify that speaker. Currently, speaker recognition models are able to identify speakers not seen in training using just a few seconds of the speaker's voice; this is known as the open-set scenario. This advance was possible due to the evolution of the machine learning area and the introduction of new cost functions applied to the training of these models.

Current speaker verification methods are trained using acoustic features, such as Mel-frequency cepstral coefficients (MFCCs) (Davis & Mermelstein, 1990) or Mel spectrograms, as inputs and use speaker IDs to calculate the loss. The models aim to learn a representation (speaker embedding), which is a vector of fixed size, to which the distance of the vectors of two different speakers is the greatest possible, while the distance of vectors of two samples of the same speaker are as close as possible. After training, the distance between these embeddings is usually calculated, allowing speakers to be identified. The performance of SV systems is commonly evaluated by the Equal Error Rate (EER) (Cheng & Wang, 2004). EER is a biometric security system algorithm used to predetermine threshold values due to its false acceptance index and false rejection rate (Cheng & Wang, 2004). EER indicates that the proportion of false acceptances is equal to that of false rejections, and the lower the EER, the more accurate the biometric system (Sztahó et al., 2019).

An SV system can be evaluated in two scenarios. In the closed-set scenario, where samples of speakers seen in the training of the SV model are used, the model recognizes these speakers. In the Open-set scenario, where speaker samples never seen in the training of the model are used, the model does not recognize these speakers. The models usually report only EER results for the Open-set scenario, since the goal of SV systems is to learn to differentiate speakers never seen in training, eliminating the need to retrain the neural model (Casanova, Candido Junior, Shulby et al., 2020).

The first studies to use deep neural networks in speaker recognition in an open-set scenario used speaker embeddings learned via the Softmax loss. Although the Softmax classifier can learn different embeddings for different speakers (Snyder et al. 2017, 2018), it is not non-discriminatory enough (Chung et al., 2020). To overcome this problem, the models trained with Softmax were combined with backends built in Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006) to generate scoring functions (Ramoji et al., 2020; Snyder et al., 2018). On the other hand, Liu et al. (2017) proposed Softmax Angular, where the cosine similarity is used as logit input for the Softmax layer, showing its superiority over Softmax alone. Subsequently, Wang et al. (2018) proposed the use of Additive Margins in Softmax (AM-Softmax) to increase inter-class variance by introducing a cosine margin penalty to the target logit. However, according to Chung, Hu et al. (2020), training with AM-Softmax and AAM-Softmax (Deng et al., 2019) proved to be a challenge, since they are sensitive to scale and margin value in the loss function.

The use of contrastive (Chopra et al., 2005) and triple loss (Schroff et al., 2015; Bredin, 2017) has also achieved promising results in speaker recognition, but these methods require a careful choice of pairs or triplets, which is time-consuming and can interfere with performance (Chung, Hu et al., 2020).

Wang, Wang et al. (2019) proposed the use of prototypical networks (Snell et al., 2017) in speaker recognition. Prototypical networks seek to learn a metric space in which the classification of open sets of speakers can be performed by calculating distances for prototypical representations of each class. Generalized end-to-end loss (GE2E) (Wan et al., 2018) and Prototypical Angular (Chung, Hu et al., 2020) follow the same principle and recently achieved SOTA results in speaker recognition. Chung et al. (2020) compared the different loss functions mentioned above in the training of two convolutional models proposed by the authors. They showed that the Prototypical Angular loss function performs better than the others, demonstrating that it is more suitable for training SV models.

Finally, Casanova, Candido Junior, Shulby, et al. (2020) proposed a new training approach consisting of reconstructing the 1-second pronunciation of the phoneme /a/ in the voice of the speakers. After training, the model is able to approximate the pronunciation of /a/ in the voice of any speaker and an embedding of this reconstruction is extracted from an intermediate layer of the neural network. Given that the reconstruction of /a/ from the same speaker is always closer to their own than

to others, the model is applied in open-set scenarios. In addition, the method surpassed a model trained in a 500x larger dataset with the GE2E loss function. It also surpassed the result of the best model proposed by Chung, Hu et al. (2020) and trained with the Angular Prototypical loss function in one of the four datasets used to compare the models. Therefore, the method requires fewer data points to achieve competitive results.

Concluding remarks

In this chapter, we aimed to list the main Deep Learning approaches applied in the fields of Speech Synthesis and Speaker Verification. In the era of Deep Learning, as in most tasks involving machine learning, significant improvements in performance have been achieved when compared to classic/traditional methods. As Deep Learning techniques advance in most fields of machine learning, older, state-of-the-art methods are also being replaced by those using Deep Learning in both speech synthesis and speaker verification. Thus, Deep Learning has apparently become the next generation solution for speech synthesis and speaker verification (Sztahó et al., 2019). In some cases, Deep Learning opened up new research fronts, allowing us to meet demands that were not previously possible. In addition, speaker verification and speech synthesis systems are still evolving. In the Speech Synthesis field, the current goal is to reduce the computing cost of the models and improve speech manipulation mechanisms, with a view to synthesizing more expressive speech (Valle et al., 2020; Kim et al., 2020). On the other hand, in Speaker Verification, researchers still seek to advance the current results and focus more on new training methods for modeling (Chung, Hu et al., 2020; Casanova, Candido Junior, Shulby et al., 2020).

References

Arik, S. O., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). *Deep voice: Real-time neural text-to-speech.* arXiv preprint arXiv:1702.07825. http://proceedings.mlr.press/v70/arik17a/arik17a.pdf

Arık, S. O., Diamos, G., Gibiansky, A., Miller, J., Peng, K., Ping, W., Raiman, J., & Zhou, Y. (2017). *Deep voice 2: Multi-speaker neural text-to-speech*. arXiv preprint arXiv:1705.08947 https://arxiv.org/pdf/1705.08947.pdf

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate.* arXiv preprint *a*rXiv:1409.0473. https://arxiv.org/pdf/1409.0473.pdf?utm_source=ColumnsChannel

Bínkowski, M., Donahue, J., Dieleman, S., Clark, A., Elsen, E., Casagrande, N., & Simonyan, K. (2019). *High fidelity speech synthesis with adversarial networks*. arXiv preprint arXiv:1909.11646. https://arxiv.org/pdf/1909.11646.pdf

Bowater, R. J., & Porter, L. L. (2001, August 21). Voice recognition of telephone conversations. *Google Patents.* (US Patent 6,278,772).

Bredin, H. (2017). Tristounet: triplet loss for speaker turns embedding. (2017, March) In: *2017 Acoustics, Speech and Signal Processing (ICASSP) International Conference on Acoustics, Speech and Signal Processing (IEEE)* (pp. 5430–5434).

Casanova, E., Candido Junior, A., de Oliveira, F. S., Shulby, C., Teixeira, J. P., Ponti, M. A., & Aluisio, S. M. (2020). *End-to-end speech synthesis applied to Brazilian Portuguese*. arXiv preprint arXiv:2005.05144. https://arxiv.org/pdf/2005.05144.pdf

Casanova, E., Candido Junior, A., Shulby, C., da Silva, H. P., Cordeiro, A. F., Guedes, V. d. O., & Aluisio, S. M. (2020). *Speech2phone: A multilingual and text independent speaker identification model*. arXiv preprint arXiv:2002.11213. https://arxiv.org/pdf/2002.11213.pdf

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., & Chan, W. (2020). *Wavegrad: Estimating gradients for waveform generation*. arXiv preprint arXiv:2009.00713. https://arxiv.org/pdf/2009.00713.pdf

Cheng, J.-M., & Wang, H.-C. (2004, December*).* A method of estimating the equal error rate for automatic speaker verification. In: *International Symposium on Chinese Spoken Language Processing* (pp. 285–288).

Chopra, S., Hadsell, R., & LeCun, Y. (2005, June). Learning a similarity metric discriminatively, with application to face verification. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 1* (pp. 539–546).

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *Empirical evaluation of gated recurrent neural networks on sequence modeling.* https://arxiv.org/pdf/1412.3555.pdf?ref=hackernoon.com

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. S., Choe, S., Ham, C., Jung,S., Lee, B.-J., & Han, I. (2020). In defence of metric learning for speaker recognition. *In Interspeech*. https://arxiv.org/pdf/2003.11982.pdf

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). *Voxceleb2: Deep speaker recognition*. https://arxiv.org/pdf/1806.05622.pdf

Davis, S. B., & Mermelstein, P. (1990). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, *28*(4), 357-366. 10.1109/TASSP.1980.1163420 Retrieved 2020 from: https://ieeexplore.ieee.org/document/1163420

Dempsey, P. (2017). The teardown: Google home personal assistant. *Engineering & Technology*, *12*(3), 80–81. 10.1049/et.2017.0330

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4690–4699). https://openaccess.thecvf.com/content_CVPR_2019/papers/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. https://arxiv.org/pdf/1810.04805.pdf&usg=ALkJrhhzxlCL6yTht2BRmH9atgvKFxHsxQ

Durkan, C., Bekasov, A., Murray, I., & Papamakarios, G. (2019). Neural spline flows. In: *Advances in neural information processing systems* (pp. 7511–7522).

Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, L., Murdock, J. M., Nyberg, E., Prager, J., Schlaefer, N., & Welty, C. (2010). Building watson: An overview of the deepqa project. *AI magazine, 31*(3), 59–79. https://doi.org/10.1609/aimag.v31i3.2303

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. http://www.deeplearningbook.org

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *32*(2), 236–243. 10.1109/TASSP.1984.1164317. Retrieved 2020, from: https://ieeexplore.ieee.org/abstract/document/1164317

Gruber, T. R. (2009). Siri, a virtual personal assistant-bringing intelligence to the interface. In: *Semantic Technologies Conference*.

Hoogeboom, E., Berg, R. V. D., & Welling, M. (2019). Emerging convolutions for generative normalizing flows. *arXiv preprint* arXiv:1901.11137. Retrieved 2020, from: http://proceedings.mlr.press/v97/hoogeboom19a/hoogeboom19a.pdf

Ioffe, S. (2006). Probabilistic linear discriminant analysis. In: *European Conference on Computer Vision* (pp. 531–542). Springer, Berlin, Heidelberg.

Ito, K. (2017). *The lj speech dataset*. Retrieved April, 29 2020 from: https://keithito.com/LJ-Speech-Dataset/

Kalchbrenner, N., Elsen, E., Simonyan, K., Noury, S., Casagrande, N., Lockhart, E., Stimberg, F., Oord, A. van den, Dieleman, S., & Kavukcuoglu, K. (2018). *Efficient neural audio synthesis.* arXiv preprint arXiv:1802.08435. http://proceedings.mlr.press/v80/kalchbrenner18a/kalchbrenner18a.pdf.

Kim, J., Kim, S., Kong, J., & Yoon, S. (2020). *Glow-tts: A generative flow for text-to-speech via monotonic alignment search*. arXiv preprint arXiv:2005.11129. https://arxiv.org/pdf/2005.11129.pdf.

Kingma, D. P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., & Welling, M. (2016). Improved variational inference with inverse autoregressive flow. In: *Advances in neural information processing systems* (pp. 4743–4751).

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., Brebisson, A. de, Bengio, Y., Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. In: *Advances in neural information processing systems* (pp. 14910–14921). Retrieved 2020 from: https://arxiv.org/pdf/1910.06711.pdf

Li, N., Liu, S., Liu, Y., Zhao, S., & Liu, M. (2019). Neural speech synthesis with transformer network. In: *Proceedings of the AAAI Conference on Artificial Intelligence: 33* (pp. 6706–6713).

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 212–220).

Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., & Miller, J. (2018). *Deep voice 3: 2000-speaker neural text-to-speech*. [Conference paper]. Proceedings of the International Conference on Learning Representations - ICLR (pp. 214–217). Retrieved 2, May 2021 from: https://openreview.net/references/pdf?id=SyD5g8sPM

Prenger, R., Valle, R., & Catanzaro, B. (2019). Waveglow: A flow-based generative network for speech synthesis. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3617–3621).

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF" social roles, user satisfaction, and personification of the amazon echo. In: *Proceedings of the 2017 Conference on Human Factors in Computing Systems (CHI)* (pp. 2853–2859).

Ramoji, S., Krishnan V, P., Singh, P., & Ganapathy, S. (2020). Pairwise discriminative neural plda for speaker verification. *arXiv preprint* arXiv:2001.07034. https://arxiv.org/pdf/2001.07034.pdf

Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). Crowdmos: An approach for crowd-sourcing mean opinion score studies. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2416–2419).

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 815–823).

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, Rj., Saurous, R. A., Agiomvrgiannakis, Y., & Wu, Y. (April, 2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 4779–4783). 10.1109/ICASSP.2018.8461368. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In: *Advances in neural information processing systems* (pp. 4077–4087).

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In: *Interspeech* (pp. 999–1003).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018, April). X-vectors: Robust dnn embeddings for speaker recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 5329–5333). 10.1109/ICASSP.2018.8461375. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Sorensen, H. V., & Burrus, C. S. (1988*).* Efficient computation of the short-time fast fourier transform. In: *International Conference on Acoustics, Speech, and Signal Processing* (pp. 1894–1895).

Sotelo, J., Mehri, S., Kumar, K., Santos, J. F., Kastner, K., Courville, A., & Bengio, Y. (2017). Char2wav: End-to-end speech synthesis. *Proceedings of the International conference on learning representations.*

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Training very deep networks. In: *Advances in Neural Information Processing Systems* (pp. 2377–2385).

Sztahó, D., Szasz´ak, G., & Beke, A. (2019). *Deep learning methods in speaker recognition: a review.* arXiv preprint arXiv:1911.06615. https://arxiv.org/ftp/arxiv/papers/1911/1911.06615.pdf.

Tachibana, H., Uenoyama, K., & Aihara, S. (2018). Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4784-4788).

Tamamori, A., Hayashi, T., Kobayashi, K., Takeda, K., & Toda, T. (2017). Speaker-dependent wavenet vocoder. In: *Proceedings of Interspeech* (pp. 1118–1122).

Valle, R., Shih, K., Prenger, R., & Catanzaro, B. (2020). Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. arXiv preprint arXiv:2005.05957. https://arxiv.org/pdf/2005.05957.pdf.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. In: *Advances in neural information processing systems* (pp. 5998–6008).

Veaux, C., Yamagishi, J., MacDonald, K., et al. (2016). Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh*. The Centre for Speech Technology Research (CSTR).

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018, April). Generalized end-to-end loss for speaker verification. *In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4879–4883). 10.1109/ICASSP.2018.8462665. Retrieved 2020, from https://ieeexplore.ieee.org/xpl/conhome/8450881/proceeding

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, *25*(7), 926–930.

Wang, J., Wang, K.-C., Law, M. T., Rudzicz, F., & Brudno, M. (2019). Centroid-based deep metric learning for speaker recognition. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3652–3656).

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A.(2017). *Tacotron: A fully end-to-end text-to-speech synthesis model.* arXiv preprint arXiv:1703.10135. https://arxiv.org/pdf/1703.10135.pdf%EF%BC%89

Ze, H., Senior, A., & Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In: *2013 IEEE. International Conference on Acoustics, Speech and Signal Processing* (pp. 7962–7966).

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., & Wu, Y. (2019). Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint* arXiv:1904.02882. https://arxiv.org/pdf/1904.02882.pdf

Zhu, X., Beauregard, G. T., & Wyse, L. L. (2007). Real-time signal estimation from modified short-time fourier transform magnitude spectra. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1645–1653.

Peer Commentary

by Claudio Possani

This chapter by Edresson Casanova, Christopher Shulby and Sandra Maria Aluísio presents the contents of the first author's lecture at the Bioacoustic Meeting Brazil 2020. Edresson is a young researcher (doctoral student at ICMC/USP/SC) trained in Computer Sciences with a focus on Neural Networks and Deep Learning.

In this chapter the focus is on Speech Synthesis and Speaker Verification. This area has been receiving increasing attention for several years, with the emergence of so-called virtual assistants. The early 21st century saw the birth of methods known as Deep Learning. There was a revolution in the scope and possibilities that emerged.

Speech Synthesis techniques obtain good reproductions of human voices. The voice quality obtained is fundamental. It is important to underscore that the English language has received the largest number of resources and hours of recording and therefore, the best results. The chapter presents the primary models used, recording times and general characteristics of this type of study. The neural network concepts play an important role in this area.

Under the general name of Speaker Recognition, recent decades have seen enormous progress in tasks involving: (SV) Speaker Verification, (SI) Speaker Identification, and (SD) Speaker Diarization. Deep Learning also caused a revolution in the field of studies.

The specific aim of SV is to decide whether two different audio recordings were produced by the same person/speaker. SI attempts to identify the speaker that produced a certain sound recording from previously collected recordings. This is what some bank security systems do. SD splits the audio input stream into homogeneous segments according to the speaker's identity.

In the present chapter the authors address only questions related to Speaker Verification. One of the significant recent advances obtained from Deep Learning techniques are the so-called Open-set scenarios in which the system recognizes a speaker with even just a few seconds of acoustic recording, even if the speaker's recordings were not used by the system in the "learning" phase.

The final part of the chapter includes an original contribution by the first author and collaborators, which consists of the reconstruction of 1 second of the pronunciation of the /a/ phoneme, constant in the speaker's voice and, after a training period, the model is capable of producing the /a/ sound of any speaker, even from a very short recording. This makes it possible to identify the speaker or determine whether two recordings are of the same speaker by comparing the sounds produced. The present chapter is an introduction to this type of study, which is becoming increasingly relevant.

In conclusion

We hope that with ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH we have met the goal of producing research-based educational material, combining multidimensional approaches to the study of human sound production with an evolutionary perspective. That was why we started with sound production in non-human animals: there are anatomic and functional similarities involved in sound production in mammals, that from the most parsimonious view represent a continuity between non-human and human animals. Where there seems to be a novelty in human acoustic communication, it is in fact only a "relative novelty", as our mentor César Ades used to say. With respect to human language, he wrote that "studies on the ability of non-human primates to acquire the use of symbols in their interactions with humans, if they do not prove (and the intention is not to prove at all) that these animals can speak as human beings, show that they have skills that foreshadow the language (Savage-Rumbaugh, Shanker and Taylor, 1998[29])." (Ades, 2009[30], p. 12, our free translation).

In addition to the biological overview, several chapters had an instrumental role, dealing with methodological issues and presenting the state-of-the-art of the analytical frameworks adopted by the scientific fields that study voice and vocalizations. In the "About the contributors" section at the beginning of the book, the reader finds information about the authors and their affiliation. We encourage interested readers to look for them in databases for academic research, such as the https://www.researchgate.net/, https://scholar.google.at/, https://orcid.org and, for Brazilians, http://lattes.cnpq.br/

The book was produced after the scientific online event also titled "Acoustic Communication: An Interdisciplinary Approach". All the talks and generated discussions have been video recorded. The videos can be freely accessed on YouTube.

Finally, we hope this digital open-access material provided by the University of São Paulo (USP) will preserve the exchange of ideas and experiences shared among

[29] Savage-Rumbaugh, E. S., Savage-Rumbaugh, S. G. S. S., Taylor, T. J., & Shanker, S. (1998). *Apes, language, and the human mind*. Oxford University Press.

[30] Ades, C. (2009). Um olhar evolucionista para a psicologia. In Otta, Emma & Maria Emília Yamamoto (eds.). *Psicologia Evolucionista (pp. 10-21)*. Rio de Janeiro: Guanabara Koogan

researchers during the two-days meeting. We gratefully acknowledge the support given by USP's Pro-Rectory of Research and the São Paulo Research Foundation (FAPESP).

Finally, we invite you to listen to Professor Régis Farias' musical composition, presented at the closing of the online event ACOUSTIC COMMUNICATION: AN INTERDISCIPLINARY APPROACH. The participants were delighted to listen to SAPOS (2014, 6'30"). Regarding this composition, he had written to us: The performance expands the natural space and reorders the recorded time of a swampy song, polemicizing the amphibious language with a human examination of its microvilli and rhythms: a dispatch of technology towards an increased perception of the music of other beings. Croaks are orchestrated to talk about the place occupied, the abundance of water and the organized life that exists there. Effects reveal and transmute an aquatic principality, for the introspection of human fantasy.

The online event gave us the opportunity of coming together, strengthening existing connections and creating new ones through science. At the closing ceremony, science met art.

Patrícia Ferreira Monticelli & Emma Otta