2023

Painting: Tomka, Anina (2021). *Songbirds*. Basel, Switzerland.

TOWARDS AUTOMATED QUANTIFICATION OF VOCAL COMMUNICATION DURING SOCIAL BEHAVIORS IN SONGBIRDS

TOMAS TOMKA

DISS. ETH NO. 29067

DISS. ETH NO. 29067

TOWARDS AUTOMATED QUANTIFICATION OF VOCAL COMMUNICATION DURING SOCIAL BEHAVIORS IN SONGBIRDS

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

TOMAS TOMKA

MSc ETH in Biotechnology, ETH Zurich

born on 19.02.1992

citizen of Basel, Switzerland

accepted on the recommendation of

Prof. Dr. Richard Hahnloser

Prof. Dr. Dina Lipkind

Prof. Dr. Benjamin Grewe

2023

“When the pregnant approaches the beautiful it becomes not only gracious but so exhilarate, that it flows over with begetting and bringing forth.”

— Plato, *Symposium*, 206d[1]

“In every piece, there is somewhere a heart. Somewhere some spot. Where the piece is particularly close to things which we cannot tangibly perceive. Beyond God, or whatnot people imagine exists, but we have no real access to. But in music we have. This passage that comes quietly out of nowhere and goes into nowhere. […]. I am waiting for this passage. […]. This passage decides everything. If it does not fit, the whole thing is lost.”

— Patricia Kopatchinskaja[2]

[1] Taken from (Plato, 1925).

[2] Loosely translated quote from (Batthyany, 2012).

Acknowledgements

This dissertation would not have been possible without the support and goodwill of many different people. First, I want to thank my supervisor Prof. Dr. Richard Hahnloser, together with the Institute of Neuroinformatics and the Neuroscience Center Zurich of ETH Zurich and University of Zurich. I thank them for their trust, resources, support, and for giving me the opportunity to conduct my doctoral studies. Additionally, I thank Richard together with Prof. Dr. Dina Lipkind and Prof. Dr. Benjamin Grewe for their guidance as thesis committee members.

Secondly, I am grateful to all newly found colleagues and friends for the time spent. Namely, Linus Rüttimann, Corinna Lorenz, Dr. Anja Zai, Dr. Diana Rodrigues, and Yingqiang Gao, for years of inspiring coffee breaks, social drinking and dining, mutual care, and companionship. Additionally, I thank Anja for her valuable mentorship and feedback.

I am thankful for the expertise and wisdom shared by Dr. Hazem Toutounji, Dr. Mariana da Rocha, Dr. Jörg Rychen, and Dr. Stefan Reimann. Similarly, I thank Roman Doronin, Qi Zhi Winston Tan, and Youssef Farag for conducting stimulating scientific projects under my supervision.

My doctoral studies have relied on experimental data recorded from zebra finches of our lab's colony, caringly managed by Heiko Hörster. I thank him and all people who contributed to the work I present here (detailed in "Source and authorship attribution"). Importantly, I am likewise thankful for the zebra finches' enormous contribution to our scientific understanding of vocal communication.

I am indebted to current and former members of the Hahnloser lab for not only taking over my share of caretaking duties for the lab's bird colony – but, in addition, also making sure I stay well-hydrated during lunch time. It worked with remarkable efficiency; one can tell that you all excel at "Labortierkunde" (LTK)!

Most importantly, I am grateful to my family and closest friends for their unconditional support. Especially, I thank my mother, Anina Tomka, my father, Zdenek Tomka, and my brother, Stefan Tomka, who have immeasurably contributed to me being where I am today. I thank my uncle, Dr. Stephan Parak, for valuable advice and for proofreading the grammar of this dissertation. And I express my gratitude to all other people who took a stand for me on my educational and professional path, allowing me to leverage my potential.

Lastly, I am grateful for nature to have found its way to develop rich and diverse ways of communication among individuals of a given species – facilitating solicited copulations in zebra finch couples, as well as systematic doctoral studies thereof.

Source and authorship attribution

This thesis contains parts of a published scientific article and two manuscripts in preparation. Here, I detail the sources and my personal contributions.

Chapter 1 (with appendices II and III) is adapted from a manuscript in preparation, co-authored with Xinyu Hao, Aoxue Miao, Kanghwi Lee, Dr. Anja. Zai, Dr. Stefan Reimann, and Prof. Dr. Richard Hahnloser. Together we thank Dr. Prof. Dina Lipkind for making her previously published data (subset 1) available for post-hoc analyses. I have contributed to the conceptualization of the study, data annotation and dataset curation, implementation of the retrieval algorithms, data analysis and writing the manuscript. To fit the storyline and style of this thesis, I have made minor adaptations to the text.

Chapter 2 (with appendices IV-VI) has been adapted from a manuscript in preparation, co-authored with Linus Rüttimann, Heiko Hörster, Prof. Dr. Richard Hahnloser, and Dr. Mariana da Rocha. Together, we acknowledge Dr. Homare Yamahachi for experimental, and Dr. Jörg Rychen for technical, and Dr. Anja Zai for statistical assistance/consultancy. I have contributed with vocal annotation, data analysis, as well as drafting, reviewing, and editing of the manuscript. To represent my contribution better, I have shortened the original manuscript content by removing several paragraphs and subsections about the copulation detection algorithm that we have developed. To fit the storyline and style of this thesis, I have made large adaptations to the text.

Chapter 3 is adapted from the published manuscript "A system for controlling vocal communication networks" (Rychen et al., 2021), co-authored with Dr. Jörg Rychen, Dr. Diana Rodrigues, Linus Rüttimann, Dr. Homare Yamahachi, and Prof. Dr. Richard Hahnloser. I have contributed with data analysis, as well as drafting, reviewing, and editing of the manuscript. To represent my contribution better, I have shortened the original manuscript content by removing several paragraphs and subsections about the engineering and technical performance of our system. To fit the storyline and style of this thesis, I have made minor adaptations to the text.

The following general comments apply throughout the thesis. Firstly, the "Abstract" has been adapted from the abstracts of the above-mentioned manuscripts and article, and conversely original parts of this dissertation might be reused in future versions of the manuscripts in preparation. Secondly, in addition to the declarations above, any thesis chapters or sections adapted from articles or manuscripts, as well as figures that have not been produced by me, are marked with a corresponding footnote. Thirdly, I extracted the zebra finch icons used in Figures 2.1, 3.1, and 3.3 from pictures that were kindly provided by Heiko Hörster (taken in the colony of our lab). All experimental procedures used to collect the shown data (except subset 2 in "Chapter 1", which was not recorded in our lab) were approved by the Veterinary Office of the Canton of Zurich. Furthermore, "Appendix I" lists details on central code scripts that I have used to perform analyses. Lastly, whenever I write "I", I refer to contributions or opinions that are my own; and I write "we" when those were made as part of the collaboration with the people listed above.

Table of Contents

Abstract

Vocalizations are produced by highly specialized motor gestures and regulate social interactions in many species. Vocal learners, such as songbirds or humans, acquire their vocal repertoire through cultural transmission with an intriguing computational efficiency. The zebra finch, a highly social songbird, has been a model organism of outstanding importance for our understanding of vocal learning. Primarily reductionist research has generated valuable insights into molecular, neural, and behavioral aspects of vocal learning. But the combinatorial effect of social factors on cultural transmission remains largely unknown. Today, the field is transitioning towards more holistic inquiries at the social level, using big data paradigms to uncover systemic principles. However, multiple challenges need to be solved to enable conclusive longitudinal studies of entire animal groups.

Reliable vocal detection in large-scale sound data has been a longstanding problem and has served as playground for many machine learning efforts, but benchmark animal datasets with labelled vocalization boundaries are scarce. Creation of such datasets requires tedious screening for vocalizations that have been missed with machine-based approaches. The challenge of faithfully annotating vocal data aggravates when studying interactive behaviors, due to overlap of individual vocalizations and noises from animal interactions. Additionally, contextualization of vocal interactions with relatively rare and brief non-vocal events, such as copulations, previously required strenuous and time-consuming inspection of video data. Lastly, correlation-based hypotheses need to be tested for causality, which requires experimental control over individual social interactions. We tackle these challenges in threefold manner.

First, we introduce a benchmark dataset of vocal segments from single zebra finches at different developmental stages. We test how well zebra finch vocalizations can be retrieved as vocal neighbors of each other in spectrographic space, using different distance measures. Interestingly, the Spearman distance outperforms other popular distance measures such as the cosine and Euclidean distances. We find excellent performance for adults (F1 score of 0.93 ± 0.07) using 50 labelled examples (templates), but not for juveniles (F1 score of 0.64 ± 0.18), which produce highly variable vocalizations. For juveniles, the retrieval is improved when searching with equally sized overlapping template slices (F1 score of 0.72 ± 0.10), compared to searches with entire templates. As an addition to a growing array of computational tools for vocal communication research, our vocal retrieval method is useful to proofread human- or computer-annotated datasets.

Secondly, we introduce a dataset of interacting mixed-sex zebra finch couples engaging in copulations. We have found that animal-borne wireless sensors, which have been originally introduced to assign vocalizations to individuals, are highly suitable for automated copulation detection. We have observed that the female radio transmitter's carrier frequency is modulated by the physical mounting of the flying male. Copulation attempts are detected by joint occurrence of this modulation and male wing flaps. Annotating vocal and non-vocal behaviors, we find behavioral signatures signaling solicited copulations roughly 25-30 s in advance: for instance, with frequent female nest/whine calls, or changes in courtship song tempo and composition. Monitoring, or even predicting, copulations based on behavioral signatures could benefit animal caretaking and wildlife conservation programs.

Thirdly, our group has developed a system for real-time control of vocal interactions among separately housed and digitally connected animals. We have characterized vocal interactions between pairs of connected birds by the cross-covariance function and we have shown that birds engaged in reliable vocal interactions constrained by the imposed network topology. Our system and analysis could be applied in the

future to probe detailed causal relationships in vocal interactions among songbird couples or during vocal learning in juvenile birds.

Taken together, our main contribution is to democratize access to large-scale curated zebra finch datasets, which can be used in the future to train machine-based solutions to detect vocalizations or predict reproductive behaviors. Additionally, we provide a computational tool for proofreading existing datasets, and a system to manipulate vocal interactions in real-time. With these efforts, we aim to accelerate systemic insights into the structure, development, and function of vocal expressions – and positively impact human coexistence with animal wildlife.

Zusammenfassung[3]

Vokalisationen werden durch hochspezialisierte motorische Bewegungen erzeugt und regeln soziale Interaktionen bei vielen Arten. Sprech- oder Gesangslerner, wie Singvögel oder Menschen, erwerben ihr vokales Repertoire durch kulturelle Weitergabe mit einer verblüffenden rechnerischen Effizienz. Der Zebrafinke, ein äußerst sozialer Singvogel, ist ein Modellorganismus von herausragender Bedeutung für unser Verständnis des vokalen Lernens. Primär reduktionistische Forschung hat wertvolle Erkenntnisse über molekulare, neuronale und verhaltensbezogene Aspekte des Gesangslernens erbracht. Die kombinatorische Auswirkung sozialer Faktoren auf die kulturelle Weitergabe ist jedoch noch weitgehend unbekannt. Heute geht dieses wissenschaftliche Feld zu ganzheitlicheren Untersuchungen auf sozialer Ebene über, um in grossen Datensätzen systemische Prinzipien aufzudecken. Um aussagekräftige Längsschnittstudien ganzer Tiergruppen zu ermöglichen, müssen jedoch noch zahlreiche Herausforderungen gelöst werden.

Die zuverlässige Erkennung von Vokalisationen in großen Datenmengen ist seit langem ein Problem und diente als Spielwiese für viele Bemühungen im Bereich des maschinellen Lernens; jedoch gibt es nur wenige Referenzdatensätze mit abgegrenzt-annotierten Tierlauten, um maschinelle Lösungen zu testen. Die Erstellung solcher Datensätze erfordert ein mühsames Suchen nach fehlenden Vokalisationen, die bei maschinenbasierten Ansätzen übersehen wurden. Die Herausforderung, vokale Daten zuverlässig zu annotieren, verschärft sich bei der Untersuchung interaktiver Verhaltensweisen, da sich individuelle Vokalisationen überschneiden können oder von Geräuschen, die durch Tierinteraktionen verursacht werden, überdeckt werden. Darüber hinaus erforderte die Kontextualisierung von vokalen Interaktionen mit relativ seltenen und kurzen nicht-vokalen Ereignissen, wie z. B. Kopulationen, bisher eine anstrengende und zeitaufwändige Inspektion von Videodaten. Und schließlich müssen korrelationsbasierte Hypothesen auf Kausalität geprüft werden, was eine experimentelle Kontrolle über einzelne soziale Interaktionen erfordert. Wir gehen diese Herausforderungen auf dreifache Weise an.

Erstens stellen wir einen Referenzdatensatz mit Vokalsegmenten von einzelnen Zebrafinken in verschiedenen Entwicklungsstadien vor. Wir testen, wie gut Zebrafink-Vokalisationen als akustische Nachbarn voneinander im spektrographischen Raum unter Verwendung verschiedener Abstandsmaße wiedergefunden werden können. Interessanterweise übertrifft die Spearman-Distanz andere populäre Distanzmaße, wie die Kosinus-Distanz oder die euklidische Distanz. Wir erhalten ausgezeichnete Resultate für erwachsene Tiere (F1-Wert von 0.93 ± 0.07), indem wir mit 50 annotierten Beispiel-Vokalisationen (Schablonen) suchen, aber nicht für Jungtiere (F1-Wert von 0.64 ± 0.18), die sehr variable Vokalisationen produzieren. Bei Jungtieren verbessern sich die Resultate durch das Suchen mit gleich großen, sich überlappenden Schablonen-Scheiben (F1-Wert von 0.72 ± 0.10) im Vergleich zur Suche mit ganzen Schablonen. Als Ergänzung zu einer wachsenden Anzahl von computergestützten Werkzeugen für die Erforschung der vokalen Kommunikation ist unsere Methode nützlich, um von Menschen oder Computern annotierte Datensätze zu überprüfen.

Zweitens stellen wir einen Datensatz von kopulierenden gemischt-geschlechtlichen Zebrafinkenpaaren vor. Wir haben herausgefunden, dass die von Tieren getragenen Funksensoren, die ursprünglich eingeführt wurden, um Vokalisationen einzelnen Individuen zuzuordnen, sehr gut für die automatische

[3] I have used DeepL (Kutylowski, 2017) for an initial translation of my original "Abstract" written in English. I modified the automatic German translation with manual corrections.

Kopulationserkennung geeignet sind. Wir haben beobachtet, dass die Trägerfrequenz des weiblichen Funksenders durch das physische Besteigen seitens des fliegenden Männchens moduliert wird. Kopulationsversuche werden durch das gemeinsame Auftreten dieser Modulation und der Flügelschläge des Männchens erkannt. Bei der Analyse von vokalem und nicht-vokalem Verhalten finden wir Verhaltenssignaturen, die auf eine einvernehmliche Kopulation etwa 25-30 Sekunden im Voraus hinweisen: z.B. häufige Nest- und Heulrufe der Weibchen oder Veränderungen in Tempo und Zusammensetzung des Balzgesangs. Die Registrierung oder gar die Vorhersage von Kopulationen auf der Grundlage von detektierten Verhaltenssignaturen könnte für Tierpflege- und Wildtierschutzprogramme von Nutzen sein.

Drittens hat unsere Gruppe ein System für die Echtzeitkontrolle von vokalen Interaktionen zwischen getrennt untergebrachten und digital verbundenen Tieren entwickelt. Wir haben vokale Interaktionen zwischen jeglichen zwei Vögeln durch die Kreuzkovarianzfunktion charakterisiert und gezeigt, dass die Vögel zuverlässige vokale Interaktionen durchführen, die durch die vorgegebene Netzwerktopologie eingeschränkt werden. Unser System und unsere Analyse könnten in Zukunft eingesetzt werden, um detaillierte kausale Zusammenhänge bei vokalen Interaktionen zwischen Singvogelpaaren oder während des Gesangslernens bei Jungvögeln zu untersuchen.

Insgesamt besteht unser Hauptbeitrag darin, den Zugang zu großen kuratierten Zebrafink-Datensätzen zu demokratisieren. Diese können in Zukunft verwendet werden, um maschinelle Lösungen zur Erkennung von Vokalisationen oder zur Vorhersage des Fortpflanzungsverhaltens zu trainieren. Darüber hinaus stellen wir ein computergestütztes Werkzeug zum Korrekturlesen bestehender Datensätze und ein System zur Echtzeit-Manipulation von vokalen Interaktionen bereit. Mit diesen Bemühungen wollen wir ganzheitliche Erkenntnisse über die Struktur, Entwicklung und Funktion von Vokalisationen beschleunigen – und das Zusammenleben von Mensch und Tier positiv beeinflussen.

Introduction

Every day, we try to deduce the emotional states of the people we meet. The slightest changes in their posture, their facial expressions, or their voice can give us the hints about their inner state – such as their general level of arousal, positive or negative emotions, motives. Reliable inference of such inner states of our fellow beings comes to our outmost benefit in navigating our social lives. Is it the right time to bring up a fun fact about songbird copulations? Does the person next to me need medical help? Am I at risk of getting physically attacked within the next three seconds? Do I trust someone to be a companion, possibly for the rest of my life? The first premise found in any arbitrary how-to guide on private or professional relationships seems most likely to be: open communication is key. Why infer inner states from a slight twitch in a facial muscle, when we can just tell each other "what's going on"?

Our vocal apparatus has been shaped throughout millions of years for this exact purpose: to express inner states voluntarily, efficiently, and precisely, with highly specialized and delicate vocal gestures. The beauty of human spoken languages is that, in this evolutionary process, they have not been carved into our genome rigidly, but are relearned at each generation, by children that pick up words and syntactical rules from their surroundings. This cultural process is called vocal learning. Practitioners, biologists, and more recently computational scientists have been fascinated by the ease and speed with which an average child learns a vast linguistic repertoire that spans an infinite space of possible expressions. Compared to currently popular machine learning systems, such as deep neural networks, children outstand with their vocal learning efficiency, adaptability, capability to generalize, and with their creativity. The backlog of machine learning systems in these qualities motivates researchers today to pay special attention to natural learning processes.

To me, shedding light on the strategies and mechanisms underlying vocal learning and communication is appealing for two reasons in particular. Firstly, to understand, appreciate, and guide vocal learning and communication – especially in children with difficulties. Secondly, for the potential "inspiration" of machine-based solutions with some of the "intelligence" that natural vocal learners exhibit. For such ends, modern research relies on animal models that exhibit rich vocal learning abilities and are easy to keep in controlled experimental conditions.

The male zebra finch, who courts the female with a complex directed birdsong, has proven to be a model organism of outstanding importance for our understanding vocal learning. Highly controlled experiments of the past have generated valuable reductionist findings about the underlying molecular basis (George et al., 1995; Gurney & Konishi, 1980; Haesler et al., 2007; Warren et al., 2010), neural circuits and dynamics (Brainard & Doupe, 2000; Doupe & Konishi, 1991; Fee & Goldberg, 2011; Gadagkar et al., 2016; Hahnloser et al., 2002; Okubo et al., 2015; Olveczky et al., 2005), and behavioral changes in response to limited live (Kollmorgen et al., 2020) or playback exposure to tutor song (Lipkind & Tchernichovski, 2011; Lipkind et al., 2013, 2017; Ravbar et al., 2012; Tchernichovski et al., 1999, 2001). Today the neuroscience of vocal learning is transitioning from a mainly reductionist to a more naturalistic and systemic approach at the social level: the goal is to probe the interdependence between learning and social dynamics (Rüttimann et al., 2022).

This endeavor comes with numerous challenges. Social dynamics are mediated via multimodal communication channels, such as vocalizations or non-vocal gestures, and they increase in complexity with the number of animals. Big data is therefore needed to resolve them adequately. However, the primary bottleneck is often not the acquisition of such multimodal data streams, but the reliable and standardized annotation of each individual's behaviors. Automated machine-based annotation solutions are needed. The extraction of meaningful *passages*[4] buried within big data is the basis for a sound understanding of complex natural phenomena. This dissertation aims to provide “midwifery equipment” to assist the vocal research field in its transition towards a big-data-driven science that integrates neural, ethological, and social aspects.

In the following, I will review these central concepts to the dissertation: vocalizations as expressions of internal states, the zebra finch as a model for vocal learning and communication, as well as computational approaches and challenges to vocal analysis. I will then outline the approaches taken in this dissertation.

Affective vocal expressions as evolutionary origins of spoken language

In 1872, Charles Darwin hypothesized that vocal signals are an expression of an animals’ internal state and that calling for a potential mate was one of the most primeval evolutionary origins of the voice (Darwin, 1872, 1871). In his works, he takes the reader on an anecdotal journey through the animal kingdom, discussing, for instance, “love-calls” in taxa such as insects, spiders, fish, frogs, and birds – telling of fisherman who imitate mating noises of male umbrina fish to take them without bait (Darwin, 1871). Such sex-specific vocal patterns – predominantly observed in male individuals – have served Darwin as illustrations of his theory of sexual selection. And he has drawn a close connection between vocalizations and emotional states – emphasizing mating calls, but also fearful sounds, in his treatises. On the potential origins of vocal expressions of affective states, Darwin wrote *[pp. 331 of* (Darwin, 1871)*]*:

> *“All the air-breathing Vertebrata necessarily possess an apparatus for inhaling and expelling air, with a pipe capable of being closed at one end. Hence when the primeval members of this class were strongly excited and their muscles violently contracted, purposeless sounds would almost certainly have been produced; and these, if they proved in any way serviceable, might readily have been modified or intensified by the preservation of properly adapted variation.“*

Today, it remains the predominating theory that such affective vocalizations are the most ancestral emotional expressions (Bass et al., 2008; Bryant, 2021). In terms of central vocal control, this origin has been mapped structurally to caudal hindbrain and rostral spinal cord of vocalizing fish (Bass et al., 2008). And, in contrast to Darwin, we have now access to systematic reviews of the vocal correlates of emotions observed in mammals (Briefer, 2012), and specifically in humans (Scherer et al., 2003, 1995), as well as reviews discussing their evolution (Bryant, 2021). In humans, for example, a conserved mapping between vocal features and affective states has been reported across diverse societies (Elfenbein & Ambady, 2002). However, it is evident that there is an enormous evolutionary leap from first vocal expressions of attraction and fear to the wealth of human spoken language.

Two hallmarks of language are its composition from vocal units that carry semantic information and the use of syntax to string these vocal units into sequences (Fitch, 2017; Hauser et al., 2002; Jarvis, 2019). Besides communicating the attraction to mates, animal calls have been shown to communicate meaningful information such as detection of predators or discovery of food in monkeys, dogs, chicken, and songbirds (Dittus, 1984; Fischer, 1998; Gill & Bierema, 2013; Gouzoules et al., 1984; Hauser, 1998; Marler et al., 1986a, 1986b; Seyfarth et al., 1980; Slobodchikoff et al., 1991; Suzuki, 2016; Templeton et al., 2005). The

[4] Reference to the quote from Patricia Kopatchinskaja displayed on page 2 of this dissertation. My interpretation of this quote in light of my dissertation is that an artist tries to express significant parts of her inner life in specific passages of her work. The listener needs to attend to these passages. Only if both are successful, the fog of uncertainty, that separates the listener from the artist, can be lifted – otherwise the message is lost. It is an analogy for vocalizing animals (“artists”) and the conspecifics or scientists (“listeners”) that attend to them.

other feature – syntax – is the ordering of acoustic units according to certain rules. Birdsong has such syntactical rules, which I will discuss in the following section. But other non-human species exhibit rudimentary syntax as well. For instance, serial calls in titi monkeys have been shown to have meaning primarily when occurring in sequence (Robinson, 1979). Taken together, there is evidence for both semantics and syntax in non-human species. But where is the unprecedented wealth of our human languages stemming from?

Humans uniquely combine a multitude of highly developed abilities necessary for spoken language, with complex vocal learning playing a prominent role (Jarvis, 2019). Many semantic and syntactical characteristics of animal communication can be innate and do not require vocal learning, in principle. In fact, complex vocal learning – the formation of auditory template memories and their subsequent vocal imitation – is believed to be a rare trait among vocalizing taxa; thus far only found in humans, cetaceans, pinnipeds, bats, elephants, songbirds, parrots, and hummingbirds (Jarvis, 2019; Petkov & Jarvis, 2012; Tyack, 2020). It has been hypothesized that a primary driver for vocal learning capacities to evolve must have been the selective pressure of female preferences on male vocal abilities (Jarvis, 2004, 2006; Fernando Nottebohm, 1972; Okanoya, 2002).

Taken together, it is thought that the attraction of sexual partners has played a major role throughout the evolution of language – involved in *bringing forth*[5] not only affective vocalizations, but also complex vocal learning (Darwin, 1872, 1871; Jarvis, 2004, 2006; Nottebohm, 1972; Okanoya, 2002). Interestingly, the latter seems to have emerged independently in a few distantly related groups of taxa (Jarvis, 2004). One of these taxa is a highly social songbird: the zebra finch.

The zebra finch as a model for speech learning and communication

In 1954, Desmond Morris introduced the Australian zebra finch[6] as an ideal model organism, which is robust to laboratory handling, habituates quickly, breeds all through the year, and requires no special diet for nestlings (Morris, 1954). Today, this organism (*Taeniopygia guttata*) is subject to multi-million-dollar research projects around the world (Griffith & Buchanan, 2010; Hauber et al., 2021). What else, apart from its low maintenance in the laboratory, has led such massive study of the zebra finch?

A key steppingstone has been the shift in research focus to the male zebra finch song (Immelmann, 1969). This occurred at a time when modern birdsong research has been just pioneered by Mark Konishi, who studied white-crowned sparrows and introduced the concept of juvenile vocal learning as imitation of memorized auditory templates that the juvenile acquires through exposure to song of conspecific adults (Konishi, 1965). In following decades of intensive research, zebra finch birdsong has become an established

[5] Reference to the quote from Plato's *Symposium* displayed on page 2 of this dissertation. The quote has three interpretations through the lens of birdsong research:

(i) The male zebra finch *brings forth* directed song in presence of a "beautiful" female.
(ii) The female *brings forth* solicitation, and potentially offspring, in presence of "beautiful" song.
(iii) Evolution *brings forth* vocal abilities in the mating context, which "beautifully" perpetuates life.

[6] All experimental data presented in this dissertation has been collected from the zebra finch (*Taeniopygia guttata*). For this reason, I will review here primarily zebra finch literature, while recognizing that other songbird species have advanced our scientific understanding of vocal learning significantly as well (Brainard & Doupe, 2002; Doupe & Kuhl, 1999).

animal model of human speech learning and production (Bolhuis et al., 2010; Brainard & Doupe, 2002; Doupe & Kuhl, 1999; Lipkind et al., 2020; Zann, 1996).

Humans and songbirds acquire complex vocabularies and syntax in similar developmental phases through similar mechanisms. Individuals of both taxa exhibit critical periods during development, where exposure to vocalizing conspecific adults is crucial for vocal learning (Bolhuis et al., 2010; Doupe & Kuhl, 1999; Lipkind et al., 2020). Development consists of two, potentially overlapping, phases: a sensory phase, where auditory memories of adult vocalizations are formed, and a sensorimotor phase, where auditory feedback guides vocal imitation of template memories (Bolhuis et al., 2010; Doupe & Kuhl, 1999; Lipkind et al., 2020). In zebra finches, the sensory phase spans roughly the window of 25-65 days-post-hatch (dph), while the sensorimotor phase occurs around 35-90 dph (Brainard & Doupe, 2002; Mooney, 2009). The adult zebra finch courtship song consists of stereotyped syllables that are ordered in a motif with stereotypic syntax. Syntax and phonology of this song are independently learned through complex vocal learning (Lipkind et al., 2013, 2017; Tchernichovski et al., 2001). Syllables in songbirds and humans start as highly unstructured and variable "babbling" sounds and subsequently "crystallize" into distinct vocal categories (Doupe & Kuhl, 1999; Lipkind et al., 2020; Tchernichovski et al., 2001). Regarding syntax learning, a stepwise acquisition of new syllable transitions has been observed in both taxa (Lipkind et al., 2020, 2013). A difference between human language and birdsong, however, is that birdsong is thought to only convey rudimentary semantic information compared to complex semantic constellations in human speech (Lipkind et al., 2020). This intriguing comparative study of vocal development is the primary reason for the high scientific impact of zebra finch research. However, the crosstalk with adjacent disciplines has turned out to be fruitful as well.

Since the mid-1970s, the field has developed a growing focus on the brain, generating fundamental insights into its workings (Brainard & Doupe, 2000; Doupe & Konishi, 1991; Fee & Goldberg, 2011; Gadagkar et al., 2016; Gurney & Konishi, 1980; Hahnloser et al., 2002; Nordeen & Nordeen, 1988; Nottebohm & Arnold, 1976; Okubo et al., 2015; Olveczky et al., 2005). The brain in itself is a matter of public fascination, philosophical controversy, and hotspot of medical attention, which motivates its scientific investigation. Additionally, the productive interaction between neuroscience and artificial intelligence (Macpherson et al., 2021) has incentivized the exploration of the neurocomputational dimension of natural learning, leading, for instance, to applications of reinforcement learning theory to birdsong learning (Doya & Sejnowski, 1995; Lipkind et al., 2017; Toutounji et al., 2022).

Since Immelmann's pioneering work (Immelmann, 1969), the spotlight has been on the zebra finch song, and much less research has focused on the rest of its diverse call repertoire. Although different call types have been described and studied, their classification so far relies on expert knowledge, and the semantics are not inferred from large-scale controlled experiments, but from anecdotal observations in the field (Elie & Theunissen, 2020, 2016; Zann, 1996). Despite this backlog, interesting new findings have been discovered, such as the existence of invariant neural responses for calls of a given category (Elie & Theunissen, 2019) or call signatures that signal the individual's identity (Elie & Theunissen, 2018). Taken together, today's zebra finch research has grown to a diverse field of study, spanning multiple scientific disciplines: vocal development, neurobiology, neurocomputation, and semantics.

A very common scientific strategy in birdsong research thus far, has been the reductionist approach (Brainard & Doupe, 2000; Doupe & Konishi, 1991; Fee & Goldberg, 2011; Gadagkar et al., 2016; George et al., 1995; Gurney & Konishi, 1980; Haesler et al., 2007; Hahnloser et al., 2002; Kollmorgen et al., 2020;

Lipkind & Tchernichovski, 2011; Lipkind et al., 2013, 2017; Okubo et al., 2015; Olveczky et al., 2005; Ravbar et al., 2012; Tchernichovski et al., 1999, 2001; Warren et al., 2010), at the cost of missing phenomena that would emerge from complex interactions in more naturalistic settings. The zebra finch, however, is widely known as a highly social bird and thus non-vocal social interactions might impact vocal learning. From human studies we know that passive exposure to adult speech is not enough for successful vocal development in infants (Goldstein et al., 2003; Goldstein & Schwade, 2008; Kuhl et al., 2003, 2007). And indeed, it has been recently shown in zebra finches that non-vocal feedback signals from females can guide male vocal learning (Carouso-Peck & Goldstein, 2019). Additionally, interactions among juvenile birds can affect song learning as well (Tchernichovski & Nottebohm, 1998; Volman & Khanna, 1995). The combinatorial effects of such social interactions and the resulting impact on song learning remains unknown. The main hurdle for the study of song learning in naturalistic settings lies in resolving and annotating the behaviors of each individual (Rüttimann et al., 2022).

Previous technical approaches and challenges in vocal analysis

Sequences of discrete signals can convey meaning more efficiently than other signals (Shannon et al., 1949) and in nature they do prominently, for example, as genetic code or languages (Hauser et al., 2002; Kershenbaum et al., 2016). Although not all vocal signals occur in immediate sequence and as discrete signals, throughout the animal kingdom they do occur in distinct units (Hauser et al., 2002; Kershenbaum et al., 2016; Lipkind et al., 2020; Marler, 1967). Any vocal analysis is thus first concerned with extracting those units from sound recordings and labelling them (Kershenbaum et al., 2016; Sainburg & Gentner, 2021). This challenge can be divided into two subtasks: finding vocal segments of each individual involved (segmentation), followed by clustering of the extracted vocal segments into vocal categories such as song syllables or call types.

The segmentation problem is easiest for single birds recorded in a well isolated environment. As discussed in the section above, many studies have used such reductionist settings, with a major advantage being that the analysis is simpler, due to a reduced number of factors (such as noises) that need to be considered and controlled. The segmentation can often be performed by simple thresholding of the vocal envelope and assuming that any segment that matches the statistics of the vocalizing individual is a vocal segment (Sainburg & Gentner, 2021; Tchernichovski et al., 2000). However, even in this simplest setting there are problematic cases such as vocalizations that are continuous with non-vocal noises – and we have recently developed a semi-supervised method to segment onsets and offsets more accurately in such cases, by defining them as two neighborhoods in an embedding plane (Lorenz et al., 2022). Once vocal segments are obtained, they might be categorizable into discrete classes.

Clustering of data samples is a standard machine learning task and many semi-supervised or unsupervised procedures have been applied to categorize vocal data. Besides low-dimensional embeddings of vocal segments or trajectories (Lorenz et al., 2022; Sainburg et al., 2020), researchers have explored hierarchical clustering (Burkett et al., 2015), as well as transformers (Morita et al., 2021). A disadvantage is that all these clustering algorithms require either some human interaction or an already segmented dataset (Lorenz et al., 2022; Sainburg & Gentner, 2021).

The modern alternative, that can solve segmentation and clustering in parallel, is the training of neural networks (Cohen et al., 2022; Steinfath et al., 2021). Their disadvantage is that the performance typically depends on the abundance and quality of labelled data. For human speech there are large corpora of labelled speech data available to test state-of-the-art neural architectures (Karita et al., 2019; Nassif et al., 2019).

For the zebra finch, however, high quality labels of vocalizations are rare. Many publicly available datasets contain motifs, single vocalizations, or call sequences without defined vocalization boundaries (Elie & Theunissen, 2020; Goffinet et al., 2021; Pearre, 2017). In fact, to our knowledge, there only exists one public zebra finch dataset consisting of 473 annotated syllables of female-directed song recorded from a single individual (Clemens, 2021; Steinfath et al., 2021). Consequently, given that today's science is moving to big data paradigms, there is a large unmet need for high-quality annotated datasets of model organisms such as the zebra finch.

When pretrained models are used for automated labelling of unseen data, there is a need to validate that these models generalize well to the new task, or else to retrain them with few manually labelled examples. This approach, called transfer learning, has been used in speech recognition – for instance, models trained for the recognition of one language can be used to detect a different language with little or no retraining data (Wang & Zheng, 2015). For the highly stereotyped female-directed zebra finch song syllables mentioned above (Clemens, 2021), the authors report that 48 manually annotated syllable examples were sufficient to retrain their cross-species model to achieve 90% of the performance reached with the complete training dataset (Steinfath et al., 2021). We note, however, that in the real-world scenario it is unknown how well transfer learning has worked, without further efforts. While one can visually detect false positives, it is particularly hard to screen potentially terabytes of data for false negatives (missed detections) – which is one of the challenges tackled in this dissertation.

Everything gets more complicated with the size of the group of vocalizing individuals that is recorded. When attending a crowded cocktail party, we humans are astoundingly good at separating sounds and attending to a single persons' speech. We solve the so called "cocktail party problem" with comparable ease (Bronkhorst, 2015; Haykin & Chen, 2005), but it remains a hard engineering problem. To robustly disentangle the vocalizations in a social group of zebra finches, researchers have therefore devised animal-borne sensors, attached to miniature "backpacks" worn by the zebra finches (Anisimov et al., 2014; Gill et al., 2015; Hoffmann et al., 2019; Rüttimann et al., 2022; Stowell et al., 2016; Ter Maat et al., 2014). In our current setup, we use accelerometer sensors, which record body vibrations as caused, for example, by vocalizations (Anisimov et al., 2014; Rüttimann et al., 2022). First unsupervised machine learning approaches to segment these vocalizations reveal that the challenge is much harder compared to single-bird microphone recordings (Lorenz et al., 2022). To overcome these difficulties, it is imperative to compile gold-standard datasets that can be used to supervise and benchmark annotation systems.

Even when vocal behaviors in an interacting group of animals would be completely resolved, conclusions on causality are difficult. To probe causality, i.e., that a component X is necessary for a phenomenon or function F, one needs ideally proof loss-of-function (without X, no F) and gain-of-function (adding X recovers F). Thus, one needs to manipulate the system. Novel biotechnological tools such as optogenetics allow targeted manipulation of neural dynamics (Ausra et al., 2021; Roberts et al., 2012; Zhao et al., 2019), and the very recent introduction of an immortalized zebra finch cell line opens new future avenues (Biegler et al., 2022). However, to apply such manipulations in highly naturalistic settings and more, to perturb atomic social interactions selectively, is still all up in the air.

Beyond the scientific objectives in academic laboratories, there is a need for automated recognition of animal sounds in wildlife management and conservation (Buxton & Jones, 2012; Digby et al., 2013; Lewis et al., 2021; Marques et al., 2013; Stowell et al., 2016). The hope is that monitoring vocal activity could help programs such as the restoration of nocturnal seabirds on Aleutian Islands after eradication of nonnative predators (Buxton & Jones, 2012). I hope that with our scientific efforts, towards automated quantification of vocal communication, we will ultimately also contribute to the understanding of animal expressions, as well as ameliorate our interactions and coexistence with wildlife.

Approaches taken in this dissertation

In the following three chapters I set forth the approaches we have taken to address some of the needs described in the previous section, namely: the compilation of gold-standard datasets from single-bird and multi-bird zebra finch experiments; tools to screen datasets for (previously missed) vocalizations; real-time manipulation of vocal communication networks achieved with a system developed in our lab.

In "Chapter 1", we introduce a gold-standard dataset of over 53'0000 vocalizations recorded from male zebra finches in isolation at different stages of development, compiled in a group effort of our lab. We plan to make this dataset publicly available upon publication of our work in a peer-reviewed journal; thereby democratizing the ability to train neural networks for automated annotation of zebra finch recordings. We have identified exhaustive nearest neighbor retrieval as a suitable method to proofread annotations or query large dataset for specific vocalizations. We benchmark retrieval on our dataset, to test how well it works on unstructured juvenile subsong versus adult vocalizations.

In "Chapter 2", we introduce another gold-standard dataset of over 54'000 vocalizations recorded from mixed-sex zebra finch couples using animal-borne accelerometers. Again, we plan to publish this data upon appearance of our work in a peer-reviewed journal. Linus Rüttimann and other colleagues have developed an accelerometer-based remote detection of copulation events in these couples. Here, I annotate vocal data stemming from 6-minute-long copulation and control episodes, and investigate the vocal signatures associated with copulations. I contextualize the findings with video-based annotations of non-vocal data, labelled by Dr. Marianna da Rocha.

In "Chapter 3", I evaluate whether a system developed in our lab by Dr. Jörg Rychen and colleagues can indeed control communication networks between digitally connected and separately housed zebra finches in real time. In experiments conducted by Dr. Diana Rodrigues, we have tested different networks connecting three zebra finches by uni- or bidirectional communication links between bird pairs. My contribution was to evaluate whether the type of the link set by the experimenter is indeed reflected in the cross-covariance functions between vocal activity in any two birds – with unidirectional links constraining the directionality of causal relationships.

Taken together, the "midwifery equipment", that we here provide for zebra finch research to transition to a big-data-driven social research field, comes primarily in form of annotated single-bird and multi-bird datasets with a total of over 107'000 vocalizations, which can drive future training of automated labelling systems. Additionally, we contribute by providing a computational tool for proofreading annotations and a system to probe causal relationships in communication networks.

Chapter 1 – Benchmarking nearest neighbor retrieval of zebra finch vocalizations across development[7]

The reliable and reproducible neuroethological study of vocalizations is challenging, especially when longitudinal research requires large datasets (see "Previous technical approaches and challenges in vocal analysis", partially summarized in the following). Challenges of segmenting vocalizations from noise in vocal learners such as the zebra finch arise from the vocal repertoire dramatically changing over the course of development (Kollmorgen et al., 2020; Tchernichovski et al., 2001). Songs start out as unstructured subsong – analogous to human babbling – and gradually differentiates into distinct classes of stereotyped syllables (Lipkind et al., 2020). Zebra finches also produce less stereotyped innate calls (Elie & Theunissen, 2016) whose acoustic features vary depending on behavioral context (Elie & Theunissen, 2016; Perez et al., 2015). To supervise machine-based annotation systems that capture such diverse repertoires, training data of sound recordings with labelled vocalization boundaries are needed. However, although vocal learning and communication in zebra finches have been extensively researched, public data comprising labels of segment boundaries is scarce: only one dataset containing 473 annotated syllable segments from female-directed song of one male bird is publicly available (Clemens, 2021; Steinfath et al., 2021).

To generate high-quality datasets, a particularly tedious task is to eliminate false negatives that have been missed by conventional machine learning methods (Lorenz et al., 2022). False negatives are hard to detect with supervised algorithms, because if they are contained in the training set, then they will likely show up in the test set. There is a need for powerful methods that allow swiftly annotating and proofreading large datasets of vocalizations suitable for training data-hungry deep learning systems. For proofreading, we propose to use nearest neighbor (NN) search algorithms.

NN ranking is a highly successful information retrieval method (Cover & Hart, 1967): it is used in tasks such as tagging images (Guillaumin et al., 2009), recommendation systems (Adeniyi et al., 2016), and for inference in language models (Khandelwal et al., 2019). NN search scalability has improved massively since the popularization of graphics processing units (GPUs) for parallel computing (Garcia et al., 2008) and with the advent of powerful approximate nearest neighbor methods (Andoni & Indyk, 2008; Becker et al., 2016; Indyk & Motwani, 1998; Muja & Lowe, 2009). Neural networks are highly popular alternatives for supervised annotation tasks (Cohen et al., 2022; Steinfath et al., 2021), as well as information retrieval (Guo et al., 2019; Liu et al., 2015; Mitra & Craswell, 2017). One of the advantages of NN ranking over neural networks is that NN retrieval has few parameters and an interpretable distance measure. Importantly, it allows controlled detection of (out-of-distribution) vocalizations that reside outside of the feature subspace spanned by the labelled examples (templates). It is therefore a suitable approach for searching previously missed detections.

Here, we use NN search and human proofreading to create a GS dataset of zebra finch vocalizations that can serve to supervise and test machine-based annotation systems. We then benchmark our NN search method on that dataset to evaluate its performance and identify the ideal distance measure for the task.

[7] The text and figures in this chapter have been adapted from a manuscript in preparation. My personal contributions and changes to the text are detailed in the section "Source and authorship attribution".

To reduce the manual annotation workload and retrieve unlabeled renditions of specific vocalization types, template-based detection of vocalizations has been previously tested (Anderson et al., 1996; Brooker et al., 2020). Anderson and colleagues applied a dynamic time warping algorithm to find the optimal path traversing continuous template frames and minimizing their distance to the input frames from the search space (Anderson et al., 1996). Brooker and colleagues benchmarked commercially available song detection software such as “monitoR” (Katz et al., 2016), which uses a Pearson correlation as a similarity measure, separately and in an ensemble approach (Brooker et al., 2020). However, the sample size was either limited to single birds and unique similarity measures (Anderson et al., 1996), or certain vocalization types were excluded from the analysis (Brooker et al., 2020). This might have been motivated by the fact that the computational cost grows linearly with number of templates and length of the test recordings. Today, state-of-the-art GPUs greatly facilitate it to benchmark several similarity measures on the retrieval of complete repertoires of several individuals at different developmental stages.

Due to the inherent lack of categorical structure in developing vocal repertoires (Lipkind et al., 2020), we benchmark NN search on the vocal segmentation task only, which is to determine for each time point (e.g., 4-ms sound interval) in a sound recording whether it contains a vocalization or not. Our dataset is divided into two subsets: adult (subset 1) and juvenile (subset 2) male zebra finch vocalizations. In our WHOLE approach, we use the entire templates for NN retrieval, whereas in the PART approach, we use fixed windows cut from the templates. The PART approach allows search for conserved vocalization subparts and has the benefit that all NN have the same dimensionality.

Methods

Sound recordings and spectrograms

We used datasets from four adult and four juvenile male zebra finches (each of the latter was recorded at three different ages, see “Table 1.1” for details). Recording was triggered by vocalizations (or other sounds); thus, recordings are unevenly spaced in time depending on the activity of the bird. Each recording contains vocalizations with some silence before and after the vocalizations. All adult birds (subset 1) were raised in the animal facility of the University of Zurich. During recording, birds were housed in single cages in custom made soundproof recording chambers equipped with a wall microphone (Audio-Technica Pro42), and a loudspeaker. The day/night cycle was 14/10 h. Vocalizations were saved using custom song-recording software (Labview, National Instruments Inc.). Sounds were recorded with a wall microphone and digitized at 32 kHz. We analyzed data recorded on days on which birds had already spent the previous three days in their stable recording environment. Juvenile individuals (subset 2) have been randomly sampled from serial tutoring experiments published by Lipkind and colleagues (Lipkind et al., 2017). In these experiments, the juvenile has been exposed to a second tutor song, starting around 55 to 65 days post-hatch, after successful learning of a first song. The day when the tutoring playback has been switched to the second song, we select for our dataset and refer to as the baseline (BL). We additionally added samples from 10 days (-10BL) and 20 days (-20BL) before the playback switch.

We computed spectrogram columns $Y_t \in \mathbb{N}^b$ by Fourier transforming data segments $X_\tau \in \mathbb{R}^b$ of b = 512 samples:

$$Y_t = \text{int8}(\ln(|\text{FFT}(X_\tau \Omega_{\tau-t})|) \cdot 128/\beta),$$

where Ω is a hamming window of length b, and β = 6.54 (or 4.93 in subset 2) is a parameter controlling the dynamic range. The hop size between adjacent Fourier segments is 128 samples. For distance computations, we removed low frequencies (0-688 Hz in adults and 0-947 Hz in juveniles) due to large background noises in these ranges.

Generation of gold-standard annotations

For a fraction of spectrograms of each day-long set, vocal segments (not further classified into vocalization types) have been annotated by human experts with high accuracy, resulting in gold-standard (GS) annotations. Human annotators used the semi-supervised segmentation method from (Lorenz et al., 2022) and exhaustive NN search to reduce the workload. A detailed annotation protocol is provided in the "Appendix II".

Nearest neighbor vocalization retrieval using gold-standard templates

The simplest approach to template-based vocalization retrieval is to take a single template of duration τ and compute a spectrographic distance to any potential candidate. Candidates are defined as any spectrogram window of the same duration τ that can be sampled from the search space (Figure 1.1). To reduce computational cost, we restrict the search to non-silent periods (using a threshold on the root-mean-squared audio signal). The best candidate vocalization can be identified by its minimal distance to the template.

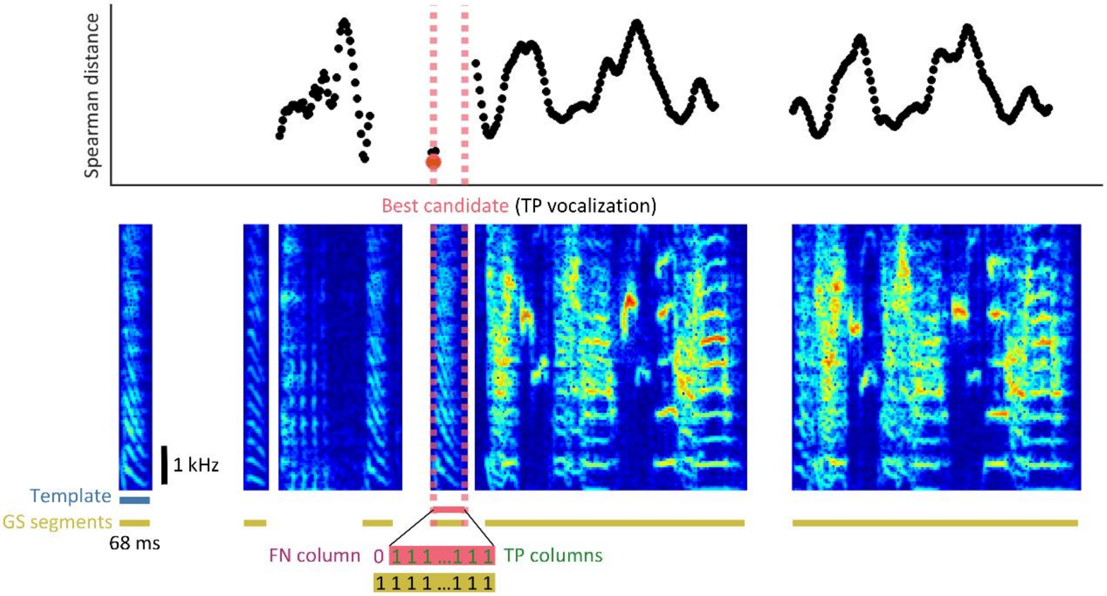


Figure 1.1: Template-based search of vocalizations (WHOLE approach). For an exemplary template drawn from our gold-standard (GS) dataset, we compute the spectrographic distance to any potential candidate (Spearman distance plotted at the candidate onsets in the top panel) of the same duration in the remaining search space (which excludes silent periods). The best candidate vocalization minimizes the computed distance. For the evaluation of retrieval performance, we identify the confusion matrix between vocally labelled spectrogram columns of the retrieved set compared to the GS set (using label "1" for vocal, and label "0" for non-vocal columns). In this example, all columns within the best candidate are true positives (TP), but the corresponding GS segment has its onset one column earlier – which results in a false-positive (FP) column label. Since this deviation is within a reasonable tolerance (≤ 5 columns or 20ms), we regard this segment as a TP vocalization.

In the following, we retrieved vocalizations based on a repertoire sample of n templates of duration τ_i with $i = 1, \dots, n$, randomly chosen from a total of N GS-labelled vocalizations of a given bird and age. In the WHOLE approach, we computed the spectrographic distances of any possible template-candidate pair.

These distances populated a matrix D, with elements $D_{i,j}$ representing distances between the i-th template and potential candidates with onsets at position j in the search space. Since longer templates fit less often in the search space, and therefore have fewer potential candidates, D is sparsely populated in general. A challenge arises from having a template set with different durations τ_i, when distances are biased by the duration (or dimensionality) of the template.

To address this fact, we tested different normalization strategies. Besides not normalizing, we explored dividing distances $D_{i,j}$ by τ_i, $\sqrt{\tau_i}$, or min-max normalizing them for each template separately as

$$D_{i,j}^{\mathrm{norm}} = \frac{D_{i,j} - \min\limits_{k} D_{i,k}}{\max\limits_{k} D_{i,k} - \min\limits_{k} D_{i,k}}.$$

The latter is making the simplifying assumption that all templates are expected to occur at least once in the dataset, since each template will be forced to yield distances with range [0,1], including the distance 0 (best possible candidate) and 1 (worst possible candidate).

After we had computed all distances, we started an iterative retrieval process of valid candidates that currently minimize the computed distance. After each retrieval iteration, we discarded candidates that overlap with or are adjacent to the retrieved candidate – reducing the search space. By iteratively retrieving the top M nearest neighbors in this manner, we followed an ultra-greedy procedure of iteratively selecting with every candidate our overall best guess in terms of its proximity to a template.

Vocalization retrieval using template slices

In the PART approach, we circumvent any duration-induced distance bias by slicing each template into overlapping slices of w spectrogram columns (Figure 1.2). We obtained in total $n_w = \sum_i \mathrm{floor}(\frac{\tau_i}{w})$ template slices. We have chosen the parameter w to be shorter than a typical template duration. To any i-th template with $\tau_i < w$, we append a trailing zero-pad to reach a total length of w. The best candidate slice was again chosen by its minimal distance to any template slice. The retrieved vocalization was chosen to be of the same duration as the template from which the matched slice originated, with one exception: the candidate vocalization was cropped if it extended into silent periods.

Spectrographic distance measures

We tested the Euclidean, cosine, Jaccard, and Spearman distances using the built-in MATLAB function pdist2. Additionally, for the WHOLE approach, we evaluated earth mover's distances (EMD) measuring sound-probability-transport along a single spectrogram axis: either summing EMD distances row-wise (EMDr, transport along the temporal axis) or summing column-wise (EMDc, transport along spectral axis).

Performance evaluation

To evaluate the performance of our approaches for a given bird and age, we continued retrieval until we retrieved a set of $N - n$ candidates. In the ideal case with perfect retrieval, the union set of templates and retrieved candidates would be identical to our GS set. Once we had completed this retrieval process, we evaluated the segmentation results with two scores: one on the binary labels of spectrogram columns indicating vocal activity, and one on the obtained segments. For the column-wise assessment, we first computed the confusion matrix of these binary labels (see Figure 1.1 for an example of true-positive and false-negative labels). From this confusion matrix we computed the F1 score, which is defined as the harmonic mean of precision and recall. For the segment-wise assessment, we have defined a vocalization score (VocScore) as the F1 score of detected vocal segments. A segment is considered a true-positive (TP) vocalization if both, its predicted onset and offset, are within a temporal tolerance ε around the gold-standard values. This tolerance reflects the fact that even experts disagree on precise segment boundaries. Here, we have chosen a generous tolerance of $\epsilon = 5$ spectrogram columns.

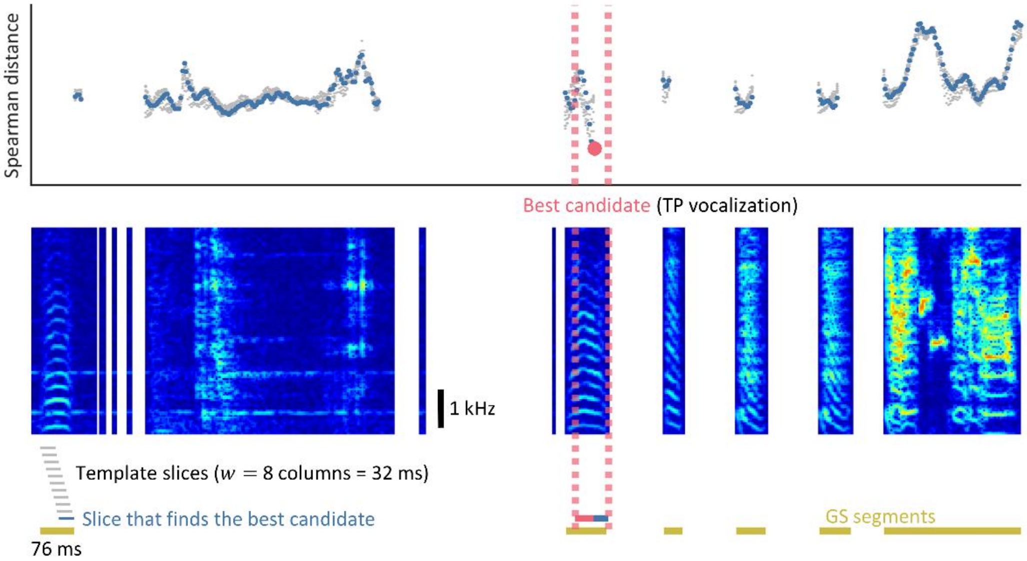


Figure 1.2: Search of vocalizations using template slices (PART approach). We chopped an exemplary template drawn from our gold-standard (GS) dataset into overlapping slices of width w, for which we again computed the spectrographic distance to any potential candidate slice in the search space (Spearman distance shown in top panel; blue for the slice that finds the best candidate; grey for the other ten template slices). The best candidate slice minimizes the computed distance (red dot in top panel). Once this slice was identified, we retrieved the best candidate (delimited by dashed red lines), by using the templates original duration (here 76 ms). This candidate is true-positive, because its relative onset (+5 columns) and offset (+1 column) to the corresponding GS segment are both within the accepted tolerance of up to 5 columns.

Results

A gold-standard (GS) dataset of juvenile and adult vocal segments

We release our GS dataset containing a total of 53’326 vocalizations in annotated recordings of 370 min total time, which we sampled from day-long recordings of zebra finches at different developmental stages (Table 1.1). To allow rigorous comparisons of vocalizations across species, individuals, and developmental stages, the human-reliant annotation process ideally follows stringent conventions. We publish guidelines that specify two decision boundaries involved in segmentation: the decision whether there is a silent period between two sounds (Figure AII.1), and the distinction of vocal from non-vocal sounds (Figure AII.2-AII.3).

Even two human experts draw segment boundaries differently. To compare different expert annotations in absence of aligned conventions, we share an additional set of labels annotated by a second expert for two adult and two juvenile datasets. We quantified expert disagreement with our performance scores on the annotations of expert 2 (using the GS data as a reference): While the F1 score was generally high across both subsets (0.981 ± 0.014), the VocScore fluctuated more (0.923 ± 0.046). For example, the adult bird g19o3 is a case where two vocal sounds are very close by, resulting in a low VocScore (F1 score: 0.975, VocScore: 0.8831), while g19o10 has distinct segments, and therefore most disagreements are within the VocScore tolerance of 5 columns (F1 score: 0.9918, VocScore: 0.9983).

Table 1.1: Dataset overview. The age of the birds is specified in days-post-hatch (dph). There are 3 samples for each juvenile bird in the dataset, chosen at specific days relative to the day of switch between playback of two tutor songs (see "Sound recordings and spectrograms"). The last four columns specify how many minutes of the day-long recordings have been annotated, the number of annotated vocalizations, the fraction of time with vocal activity in annotated recordings ("label imbalance"; perfect balance corresponds to 0.5), and the range of vocalization durations, respectively.

Developmental stage	Bird name	Sex	Hatch date	Age (dph)	Annotated (min)	Number of vocalizations	Label imbalance (vocal/total columns)	Vocalization duration range (ms)
Adult (subset 1)	g17y2	male	14.4.2015	197	84.34	10050	0.4714	20-656
	g4p5	male	28.12.2012	115	104.18	26045	0.5155	16-300
	g19o3	male	13.11.2015	154	7.72	2045	0.4238	20-240
	g19o10	male	08.11.2015	198	7.68	1998	0.548	28-400
Juvenile (subset 2)	R3406	male	29.11.2011	35	1.27	139	0.22	20-357
				45	8.28	243	0.0486	9-377
				55	39.42	2281	0.1077	12-372
	R3428	male	16.12.2011	39	7.30	1316	0.2931	15-514
				49	6.86	780	0.2496	12-418
				59	52.19	4026	0.1862	23-435
	R3549	male	17.02.2012	43	7.33	781	0.2411	15-581
				53	9.02	929	0.2209	15-438
				63	10.52	1068	0.2372	12-343
	R3625	male	13.04.2012	45	11.67	728	0.1216	26-372
				55	7.23	534	0.1363	12-418
				65	4.71	362	0.1575	15-293
All					**370**	**53326**		**9-656**

Performance of nearest neighbor retrieval

We tested our two template-based vocal retrieval approaches on our GS dataset using various distance measures and normalization strategies (Figure 1.3), obtaining excellent WHOLE results for adults (F1 score of 0.93 ± 0.07), but not for juveniles (F1 score of 0.64 ± 0.18). Performance for juveniles was improved in the PART approach (F1 score of 0.72 ± 0.10). Generally, as we retrieved $N - n$ candidates, the precision in labelling spectrogram columns decreased slowly, reflecting the intuition that nearer neighbors are better candidates, while the candidate's distance increased monotonically per definition (Figures 1.3a, AIII.1, and

AIII.2). Qualitatively, we observe that birds with better precision tend to have a critical point in retrieval progression (close to the point where we stop retrieval), where the distance accelerates with each new retrieval (Figures AIII.1-AIII.2). This acceleration is usually accompanied by a drop in the current retrieval precision (Figures AIII.1-AIII.2), and may serve as a natural criterion for stopping unsupervised retrieval.

We found that the Spearman distance outperforms other measures – especially in juvenile birds (Figure 1.3b-e). It is followed by the cosine and Jaccard distances, which are computationally less expensive. The worst tested measure is the Euclidean distance. For the WHOLE approach, we explored Earth mover's distances allowing transport along a single spectrogram axis, testing whether discounting similarity of slightly distorted vocal renditions is beneficial. We found that they perform poorly (Figure 1.3b); with allowing "column-wise" transport along the frequency axis (EMDc) yielding better results than allowing it "row-wise" along the temporal axis within the duration of the template (EMDr). Taken together, we found correlation-based measures (Spearman and cosine distances) to be superior for vocal NN retrieval.

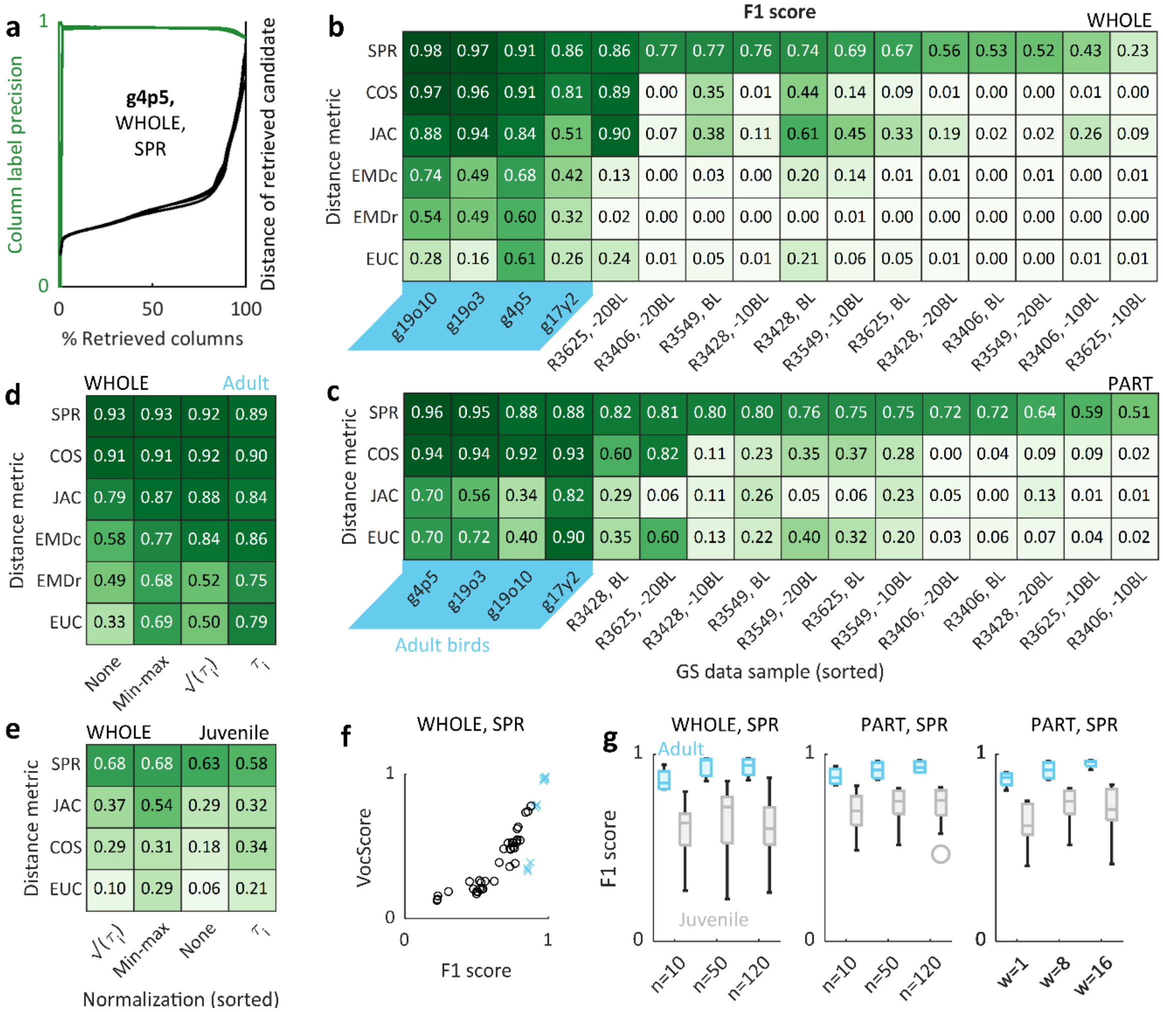
We normalized distances in the WHOLE approach with four different strategies. For adults, not normalizing was among the best strategies for the Spearman distance, while being worst for Earth mover's, Jaccard, and Euclidean distances (Figure 1.3d). As expected, these latter distances benefit from division by the template duration, because they scale with the dimensions of the compared vectors. The template-wise min-max normalization is a good alternative, working well across distance measures and GS data subsets (Figure 1.3d-e). Taken together, NN search yielded best results using the PART approach for juvenile data and the unnormalized WHOLE approach for adults.

During development or even during a single day, zebra finches can join or separate adjacent vocal elements (Figure AII.2). The VocScore will be very sensitive to any segmentation error occurring in between such elements: e.g., if a gap that is present in the GS data has not been inferred in the test set, we report a long false-positive (FP) and two short false-negative (FN) vocalizations. The VocScore generally correlates with the F1 score (Figure 1.3f). The F1 scores were often variable across individuals at low values (juvenile birds, Figure 1.3g). At high values (for adult birds) they were sensitive to the number of templates n and (for the PART approach) the slice width w (Figure 1.3g).

Next, we wondered whether there are some detrimental templates that confuse the retrieval process and could be filtered out by an expert before starting the search (Figures 1.4, and AIII.3-AIII.4). To investigate this possibility, we examined three exemplary birds, an adult and two juveniles, more closely[8]. We found that retrieval rates of 50 different templates varied strongly in all three showcased birds (Figure 1.4a-c), as well as in all other birds of the GS set, for both, the WHOLE (Figure AIII.3) and the PART approach (Figure AIII.4). In the juvenile birds, there were a few templates that yielded excessively low retrieval precision (high fraction of FP detections; Figure 1.4a-f). For one search replicate per exemplary bird, we plotted the retrieved candidates of the worst three templates (Figure 1.4g-i). Detrimental templates had either background noises (Figure 1.4e, templates "1" and "2"), very faint harmonic extensions (Figure 1.4e, template "3"), or they were very short (Figure 1.4f, templates "1" and "2"). This latter case was highly

[8] We showcase only three birds, because detailed spectrographic examination and simulation of many replicates (required for statistical analysis in Figure 1.4j-l) is time consuming. Note, that the retrieval histogram of R3428 (Figure 1.4b), seems qualitatively different compared to three other replicates shown in Figure AIII.3 (having three outstandingly bad templates), but is not an outlier in term of overall performance (Figure 1.4k) – which is why I kept it to illustrate bad templates. The example for R3549 in contrast, is an outlier in terms of performance (Figure 1.4l) – which is why I consciously picked it, to illustrate an extreme case.

detrimental – not so much because it caused low precision (e.g., by retrieving a noise sound, template “1”), but by lowering recall through retrieval of vocalization onsets of longer vocalizations (template “2”).



b — F1 score, WHOLE

	g19o10	g19o3	g4p5	g17y2	R3625, -20BL	R3406, -20BL	R3549, BL	R3428, -10BL	R3428, BL	R3549, -10BL	R3625, BL	R3428, -20BL	R3406, BL	R3549, -20BL	R3406, -10BL	R3625, -10BL
SPR	0.98	0.97	0.91	0.86	0.86	0.77	0.77	0.76	0.74	0.69	0.67	0.56	0.53	0.52	0.43	0.23
COS	0.97	0.96	0.91	0.81	0.89	0.00	0.35	0.01	0.44	0.14	0.09	0.01	0.00	0.00	0.01	0.00
JAC	0.88	0.94	0.84	0.51	0.90	0.07	0.38	0.11	0.61	0.45	0.33	0.19	0.02	0.02	0.26	0.09
EMDc	0.74	0.49	0.68	0.42	0.13	0.00	0.03	0.00	0.20	0.14	0.01	0.01	0.00	0.01	0.00	0.01
EMDr	0.54	0.49	0.60	0.32	0.02	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
EUC	0.28	0.16	0.61	0.26	0.24	0.01	0.05	0.01	0.21	0.06	0.05	0.01	0.00	0.00	0.01	0.01

c — PART

	g4p5	g19o3	g19o10	g17y2	R3428, BL	R3625, -20BL	R3428, -10BL	R3549, BL	R3549, -20BL	R3625, BL	R3549, -10BL	R3406, -20BL	R3406, BL	R3428, -20BL	R3625, -10BL	R3406, -10BL
SPR	0.96	0.95	0.88	0.88	0.82	0.81	0.80	0.80	0.76	0.75	0.75	0.72	0.72	0.64	0.59	0.51
COS	0.94	0.94	0.92	0.93	0.60	0.82	0.11	0.23	0.35	0.37	0.28	0.00	0.04	0.09	0.09	0.02
JAC	0.70	0.56	0.34	0.82	0.29	0.06	0.11	0.26	0.05	0.06	0.23	0.05	0.00	0.13	0.01	0.01
EUC	0.70	0.72	0.40	0.90	0.35	0.60	0.13	0.22	0.40	0.32	0.20	0.03	0.06	0.07	0.04	0.02

d — WHOLE, Adult

	None	Min-max	$\sqrt(\tau_i)$	τ_i
SPR	0.93	0.93	0.92	0.89
COS	0.91	0.91	0.92	0.90
JAC	0.79	0.87	0.88	0.84
EMDc	0.58	0.77	0.84	0.86
EMDr	0.49	0.68	0.52	0.75
EUC	0.33	0.69	0.50	0.79

e — WHOLE, Juvenile

	$\sqrt(\tau_i)$	Min-max	None	τ_i
SPR	0.68	0.68	0.63	0.58
JAC	0.37	0.54	0.29	0.32
COS	0.29	0.31	0.18	0.34
EUC	0.10	0.29	0.06	0.21

Figure 1.3: Nearest neighbor retrieval performance for various distance measures and normalization strategies. (a) As columns of $N - n$ candidates (n = 50 templates) were retrieved (x-axis) for the exemplary bird g4p5, the column-based precision (green) slowly declined (after fluctuations in the labelling of the very first columns), while the distances computed for the current candidate increased. Results are shown for 3 replicates; because replicates behaved highly similar, their curves may overlay each other. **(b,c)** Mean F1 scores (3 replicates) across the dataset for different distance measures, using 50 templates in the unnormalized WHOLE (b) or PART (c) approach (slice window w=8 columns). The tables have been sorted along the rows and columns that contain the entry with the best performance. Abbreviations: SPR=”Spearman”, JAC=”Jaccard”, COS=”Cosine”, EMDc=”column-wise Earth mover’s distance”, EMDr=”row-wise Earth mover’s distance”, EUC=”Euclidean”. **(d,e)** Sorted tables of mean F1 scores (as in b,c) for different normalization strategies used in the WHOLE approach for pooled adult (c) or juvenile (d) replicates, using 50 templates (see “Nearest neighbor vocalization retrieval using gold-standard templates”). **(f)** The relationship between F1 score and VocScore for adult (crosses) and juvenile (circles) birds, computed for the Spearman distance used in the WHOLE approach across 3 replicates per sample. **(g)** Sensitivity analysis for number of templates n and slice window w for the Spearman distance.

Removing the worst three templates (searching with 47 templates only) did not increase performance in the adult (Figure 1.4j) but did so in the juvenile (Figure 1.4k-l). This indicates that NN search can be improved by selecting representative and clean templates.

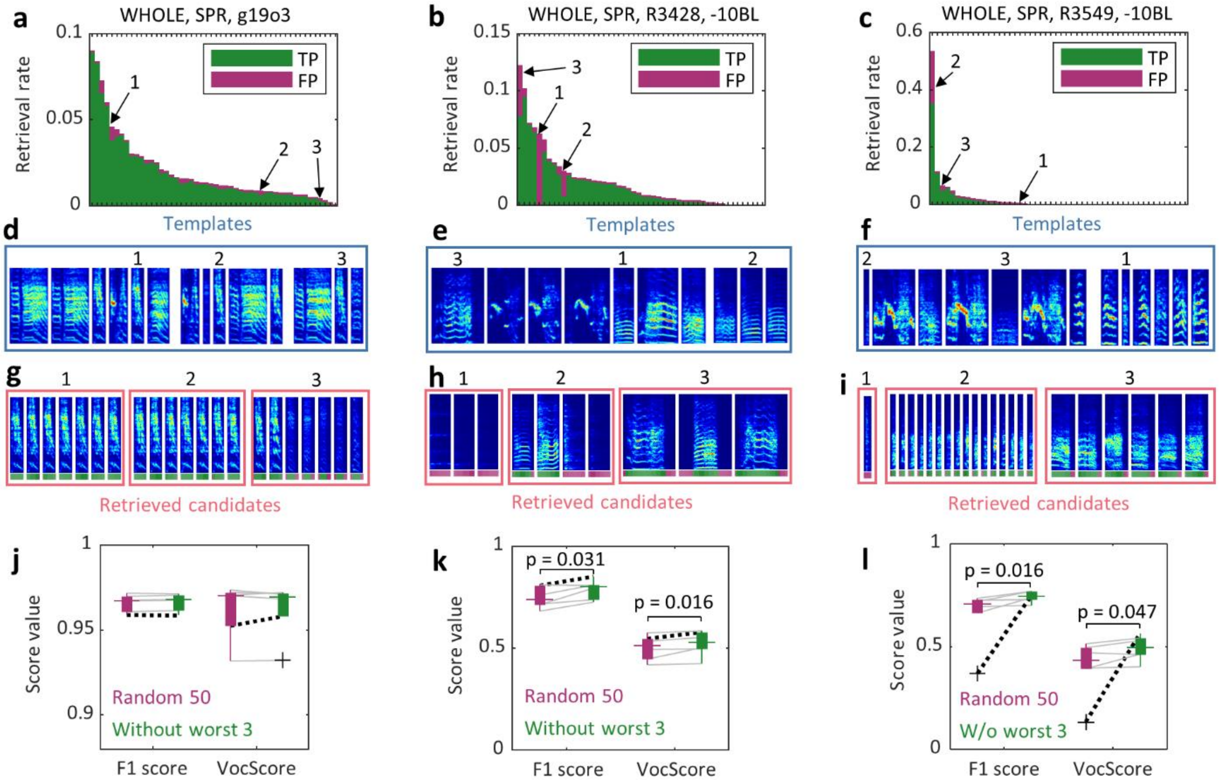


Figure 1.4: Noisy or outlier templates are detrimental for retrieval performance in exemplary juveniles. (a-c) For the WHOLE approach with the Spearman (SPR) distance, we examine how retrieval rates are distributed in an exemplary adult (a) and two juvenile (b,c) birds. For this end, we sort the 50 templates by retrieval rate (summed TP and FP retrieval, top panels). The worst 3 templates in terms of retrieval precision are labelled with numbers 1-3 ("1" being the worst template). **(d-f)** For each example bird, out of the 50 templates, we plot several spectrogram examples, including the worst 3 templates. **(g-i)** For each example bird, we plot several candidates retrieved by the worst 3 templates. **(j-l)** For each example bird, we simulate n=6 retrieval search replicates, and use boxplots to show the effect on performance scores when removing the worst 3 templates (green box) from the initial set of random 50 templates (purple box). A significant increase in performance is observed for the juvenile birds ($p<0.05$, one-sided paired-sample Wilcoxon signed rank test). The performance change for the replicate showcased in (a-i) is highlighted (black dotted line; grey lines are used for the remaining 5 replicates).

Discussion

To accelerate the comparative large-scale study of animal vocalizations there are currently several unmet needs faced by researchers of this field: (i) the availability of gold-standard benchmark datasets of vocal segments reflecting the diverse repertoire of an individual across development, (ii) that are of enough volume to train supervised automated annotation algorithms, and (iii) adhere to standardized annotation conventions, as well as (iv) methods to systematically screen for false negatives when proofreading annotations. Here, we contribute to mitigate these needs: by sharing a zebra finch dataset with over 53'000 vocalizations recorded across different developmental stages, proofreading it with exhaustive NN search, and characterizing the retrieval difficulty across the birds' development.

We have illustrated our decision boundaries and the difficulties in manual annotation (see "Appendix II"). In summary, we advocate for the definition of vocal segments as tightly restricted intervals of continuous vocal activity. In a first line of comparative research, these segments should be defined independently from functional considerations, which are often unknown and require carefully controlled experiments. This is especially true for rich and diverse repertoires as in the case of the zebra finch. Across development or even across a single day, the zebra finch might join adjacent vocal elements or separate them with arbitrarily short silent pause (Figure AII.1). To highlight the importance of shared labelling conventions among human experts, we demonstrated that in their absence experts can draw very different segmentation boundaries (VocScore of down to 0.88).

To characterize template-based retrieval across our dataset, we deployed two search approaches: based on whole templates or template slices (Figures 1.1-1.2). For both approaches we tested several commonly used distance measures (Figure 1.3). We found that the Spearman distance outperforms other measures – especially so in the juvenile data samples. Together with the well-performing cosine distance, it is correlation-based, invariant to global changes in the power (or loudness), and works well with templates of different durations, since correlations between two vectors do not scale with the vector dimension. In contrast to the cosine distance, which captures linear relationships, the Spearman distance can capture other (non-linear) monotonic relations as well (Kaufman & Rousseeuw, 1990; Spearman, 1906). Recently, it has been shown to be successful in other applications such as spam email detection (Sharma & Suryawanshi, 2016) or indoor localization based on received Wi-Fi signal strength (Xie et al., 2016).

The Euclidean metric, often the first choice when comparing songbird vocalizations (Anderson et al., 1996; Kollmorgen et al., 2020; Sainburg et al., 2020; Tchernichovski et al., 2000), exhibited the overall worst performance. By its nature, it is not discounting any translations or distortions in the spectrographic space: a candidate that is ten times louder will have a large distance to its template. On the other extreme, the EMD distances we used, measuring sound-probability-transport along a single spectrogram axis, have not been successful: it might be that mapping powers to the probability simplex combined with allowing lateral transport is taking the abstraction from the original signal too far.

Taken together, we conclude that suitable distance measures for NN search of vocalizations discount for certain transformations of the template in spectrographic space. We think that discounted transformations, such as varied loudness, ideally reflect natural axes of variance in the animals' repertoire. Such discounts would therefore optimally allow out-of-distribution detection – such as proofreading a large dataset for missed detections. Examining three birds more closely (Figure 1.4), whenever possible, we recommend to select templates that do not have large background noises, very short durations, or outlier features for initial search. Instead of biasing results by excluding such templates totally, it might be a good strategy to do a two-stage search: first with stereotyped templates, then with apparent outliers.

For the best performing Spearman distance, the juvenile data was better retrieved using the PART approach rather than the WHOLE approach. One possible rationale is that early vocalizations might feature more conserved (useful template slices) and more variable parts. Both our retrieval approaches however suffer from inflexibility of segment duration: the retrieved set of candidates will only exhibit durations observed in the template set. We can imagine two machine-based solutions for this limitation. One possible approach is the fine-tuning of durations using dynamic time warping in the temporal neighborhood of retrieved candidates. Another approach could be to modify the PART approach to follow a water-shed strategy: conserved slices serve as seeds, and from these seeds the vocalizations could be elongated using a threshold on the distance profile.

Taken together, based on our results we recommend using the Spearman distance for highest performance, using the PART approach for juvenile data (low stereotypy) and the unnormalized WHOLE approach for adults (high stereotypy). We recommend that commonly used analysis methods that use Euclidean distance (Kollmorgen et al., 2020; Tchernichovski et al., 2000) or detect linear relationships (Katz et al., 2016), consider the Spearman distance as an option for future applications.

Where does our approach fit into the landscape of tools available for vocal annotation? Deep neural networks are the state-of-the-art to learn birdsong segmentations from gold-standard labels (Cohen et al., 2022; Steinfath et al., 2021). In comparison, our approach is ideally suited for proofreading existing datasets, because it can control out-of-distribution detection with a well-defined and interpretable distance measure. Additionally, template-based retrieval can be performed with as little as one positive example, making large data accessible to specialized queries. However, a disadvantage compared to neural networks, is that the cost scales with the number of labelled examples.

In our view, the comparative study of animal vocalization of the future will employ a multitude of computational tools assisting the researcher to increase scientific reliability and reproducibility. While the human experts are "in the loop" it is critical to define standardized annotation conventions that can be generalized to any dataset without prior knowledge on the functional role of vocalizations. NN search can then assist to proofread expert-labelled annotations. Iteratively, deep neural networks can be trained, and annotations can be proofread to capture more complete corpora of vocal data. Finally, additional annotation layers can be added based on functional or structural assessments, which may depend on whether the vocal units are assessed in the domain of perception (receiver) or production (sender).

Data availability

We will release our dataset (Table 1.1) upon publication of our work in a peer-reviewed journal.

Funding

This study was partly funded by the Swiss National Science Foundation (Grant 31003A_182638; and the NCCR Evolving Language, Agreement No. 51NF40_180888) and by the China Scholarship Council (Grant No. 202006250099).

Chapter 2 – Behavioral signatures of copulations in freely behaving zebra finches[9]

The human species, as many other species, is sustained by means of sexual reproduction. In mammals and birds, for example, this reproduction is achieved through copulations that mediate the internal fertilization of female egg cells. The choice of partners, participating together in this act, drives sexual selection of traits and behaviors that reinforce mate attraction (Darwin, 1871). This has led to the evolution of complex sexual ornaments such as birdsong (Jarvis, 2004, 2006; Nottebohm, 1972; Okanoya, 2002). Birdsong is a vital research field due to its structural properties and cultural transmission that parallel human speech (Doupe & Kuhl, 1999; Lipkind et al., 2020). Big data research, wildlife management, and animal care taking would greatly benefit from monitoring copulations, or predictive sexual behaviors such as birdsong, in freely behaving animals. In endangered populations, for instance, these key events can decide the fate of entire populations. Unfortunately, in-action monitoring of such events has been hampered by technical challenges in systematic copulation detection and the resolution of signature behaviors.

In zebra finches, a model organism for sexual and vocal behavior, copulations have been hard to study in the past, mainly because of their brevity of 1-2 s. Their detection has required strenuous visual inspection of live experiments or movies, which is extremely labor intense. Consequently, courtship behavior has been typically reported by experts based on experience (Morris, 1954; Zann, 1996), studied without physical contact (Avey et al., 2005; MacDougall-Shackleton et al., 1998), ignoring non-vocal behaviors (Bischof et al., 1981; Elie et al., 2010; Sossinka & Böhner, 1980), with rare or not reported copulatory activity in exclusive, usually 15-minute-long, mixed-sex encounters (Arnold, 1975; Gill et al., 2015; Goodson et al., 2009; Harding et al., 1983), or by hormonally stimulating the female to be sexually receptive (Bharati & Goodson, 2006). These studies have provided particularly valuable insights into hormonal (Arnold, 1975; Harding et al., 1983) and dopaminergic (Bharati & Goodson, 2006; Goodson et al., 2009) control of courtship, as well as the modulation of gene expression through sensory sexual stimuli (Avey et al., 2005). However, to our knowledge, nobody has studied zebra finch behavior as a function of temporal proximity to copulation – the critical act through which evolution is thought to have selected impressive traits, such as culturally learned vocal abilities (Jarvis, 2004, 2006; Nottebohm, 1972; Okanoya, 2002).

Vocal communication is an important part of the behavioral repertoire of songbirds, including courtship. Individuals of many songbird species utter thousands of calls and songs per day, and to understand the social dynamics within songbird groups it is essential to know "who says what". This is usually not possible using stationary microphones, because these do not allow discrimination between the vocalizations of multiple interacting birds. To tackle this problem, sensor nodes have been developed that can be attached to songbirds and that allow selectively recording the vocalizations of the bird that carries the sensor node on its back (Anisimov et al., 2014; Gill et al., 2015; Hoffmann et al., 2019; Rüttimann et al., 2022; Stowell, Gill, et al., 2016; Ter Maat et al., 2014). However, these animal-borne sensors do also record non-vocal vibrations and disturbances (Anisimov et al., 2014).

[9] The text and figures in this chapter have been adapted from a manuscript in preparation. My personal contributions and changes to the text are detailed in the thesis section "Source and authorship attribution".

We find that a useful byproduct of using this technology is the unprecedented access to zebra finch copulation attempts. During a copulation, males typically flap their wings to hover above the female's back, which leads to a particular signature on the on-bird sensors, so called accelerometers, which we use. Furthermore, the proximity of the male's body to the female's wireless sensor node leads to changes in the inductive properties of the female's antenna, which results in a modulation of the carrier frequency of its transmitter. Based on these observations, we have developed the idea to detect copulations by the coincidence of audio-traces of flapping wings in males and stereotyped modulations in the radio carrier frequency of female sensors[10]. We have tested this idea in groups of mixed-sex zebra finch pairs, housed together over a period of 9-14 days, using our group's previously developed experimental setup (Rüttimann et al., 2022). We have found that our method of detecting copulation attempts reduces the time spent on visual inspection of movies by a factor of more than 1000.

We set out to investigate the vocal and non-vocal signatures in automatically detected and manually verified episodes of solicited copulation attempts (SCA). Typically, copulation events in zebra finches start with a courtship display by the male, which includes female-directed song and dance (Morris, 1954). If interested, the female will respond with a tail-quiver as a solicitation display, and then the male will hop onto her back while flapping his wings to achieve cloacal contact. The exclusively male song is typically composed of a few introductory notes, followed by stereotyped song motifs, consisting of several syllables, which are often repeated within song bouts. It has been shown that female-directed song bouts have more preceding introductory notes and contain more motifs that are slightly shorter (Sossinka & Böhner, 1980). Calls produced by both sexes are assumed to play a role in pair formation and maintenance as well, which is currently not well understood (Elie et al., 2010; Gill et al., 2015). Call categorization into subtypes has been described phenomenologically, based on expert knowledge synthesized from field and laboratory studies (Elie & Theunissen, 2016; Zann, 1996). However, it has been noted to be difficult to cleanly separate vocalizations into distinct clusters based on this nomenclature (Elie & Theunissen, 2016; Elie et al., 2010). Taken together, the repertoire of zebra finch courtship and solicitation is diverse and has been studied to largely different degrees.

Here, we examine these diverse behavioral signatures and share novel insights into their differential regulation in proximity to copulations. We have compiled an annotated vocal dataset based on accelerometer recordings, complemented by a dataset of non-vocal behaviors labelled on video recordings. We uncover that copulations are signaled by elevated behavioral rates roughly 25-30 s in advance. Additionally, we show that song composition, song tempo, and call durations are varied around copulations. Interestingly, "nest/whine" call durations covary between both sexes around copulations, potentially indicating synchronized sexual arousal.

In the following, I will emphasize vocal annotation and analysis, which is my primary contribution to our collaborative efforts. I will describe the compilation of labelled datasets and analysis of conditional

[10] The development and evaluation of remote copulation detection is the primary contribution of Linus Rüttimann and has been excluded from this chapter, to better represent my own contribution (see "Source and authorship attribution"). Note, that although this detection methodology is not reported here, Dr. Marianna da Rocha has visually verified that our annotated SCA occur, and only occur, at the specified times in the presented dataset – this verification forms the basis for our results and conclusions.

behaviors, given previously detected copulation attempts thanks to the primary efforts of Linus Rüttimann. I will then present and discuss the rate-based and feature-based signatures that we have found.

Methods

Spectrograms

To annotate and illustrate vocal behaviors, we transform accelerometer signals into spectrograms. The raw signals were sampled at a rate of 24'000 Hz (first replicate) or 24'414 Hz. We removed low-frequency noise by applying a high-pass filter with 300 Hz cutoff. We computed spectrogram columns $Y_t \in N^b$ by Fourier transforming data segments $X_\tau \in R^b$ of $b = 384$ samples:

$$Y_t = \text{int8}(\ln(\alpha|\text{FFT}(X_\tau \Omega_{\tau-t}|) \cdot 128/\beta),$$

where Ω is a hamming window of length b. To scale the power range, we used scaling factor $\alpha = 5$ and dynamic range $\beta = 4.93$. The hop size between adjacent Fourier segments is 96 samples.

Annotation of vocal and non-vocal behaviors

From all detected and visually verified copulation attempts (CA) across six replicates of mixed-sex experiments as described in (Rüttimann et al., 2022), we excluded one CA due to failure of the accelerometers. For each of the remaining 57 CA, we selected the interval I = [-5 min, 1 min] relative to the CA onset and annotated vocal exchanges in all 7-minute-long accelerometer recordings that intersect with I. Data from the interval I around a given CA is referred to as a copulation episode (COP). We randomly drew control episodes (CTRL) defined by intervals I relative to random time points (RTP) uniformly chosen within the remaining data of a given experiment.

For vocal annotation of accelerometer recordings, we used the semi-automatic tools developed in (Lorenz et al., 2022) and "Chapter 1"; however, we had to manually adjust labels to achieve gold-standard quality. A detailed vocal annotation protocol is provided in "Appendix IV", where we first specify segmentation into vocal units (Figure AIV.1) and then categorization into vocalization types (Figures AIV.2-AIV.5). The most complete description and dataset of the zebra finch call repertoire has been published by Elie and Theunissen (Elie & Theunissen, 2016, 2018, 2020). We base our expert-based categorical classification on their nomenclature, unfortunately noting inconsistencies with other zebra finch literature (Gill et al., 2015; Zann, 1996). As others (Elie & Theunissen, 2016; Elie et al., 2010), we often could not separate all categories into distinct clusters in spectrographic space (Figure AIV.2). In our data, we observe "nest/whine" and "tet" call clusters for both sexes, as well as exclusively male clusters corresponding to song syllables, introductory notes, distance calls, and "wsst" calls (Figure AIV.5).

The experiment "b8p2male-b10o15female" has been recorded in an older setup (with a smaller chamber, recording single-angle videos). We excluded its video data from the dataset because it did not allow faithful annotation of non-vocal behaviors. The behavioral annotation of COP / CTRL video episodes in the remaining 5 experiments has been conducted by Dr. Mariana da Rocha, and a detailed definition of these 20 behaviors is provided in "Appendix V". A copulation has been categorized as "solicited" only if the female displayed a tail-quiver shortly before the copulation, which was the case in 55 of 57 CA.

Exclusion criteria for behavioral analyses

From the annotated dataset we exclude two CA from further analyses, because the mounting was very brief and not solicited by the female, resulting in a non-stereotypic copulation signature. Additionally, during inspection of the video data, we found an undetected mounting attempt within a CTRL episode, occurring inside the nest. This mounting attempt was unlikely successful and has not been detected because the male did not flap his wings; however, we found a call signature characteristic for copulations. We decided to exclude this episode as well. We proceeded analyzing the remaining 55 COP episodes exhibiting solicited copulation attempts (SCA) together with the remaining 56 CTRL episodes.

Within these episodes, there are periods in the dataset (4.6 ± 6.6 % for vocal data and 3.8 ± 6.2 % for video data, across experiment replicates), where annotation of a particular behavior is impossible: either due to intersection with the beginning or end of day-long recording session, due to accelerometer failures, or an animal sleeping (we decided to focus on the awake state). We ignore these periods by representing them with "not a number" (NaN) values.

Behavioral rate analysis

We aimed to express behavioral signatures as a function of the temporal proximity to copulations. For this end, we compute the conditional rate function (CRF) of a vocal or non-vocal behavior B relative to the reference event (SCA or RTP) for a given episode as follows. We first extracted all onset times of the chosen behavior. We then computed the density of onset times in units of Hz as a function of the time lag τ to the reference event. The CRF is given by this density that we smoothed with a Gaussian of length of 20 s and standard deviation 4 s.

Because our data is hierarchically structured with unequal sample sizes across bird couples, we used hierarchical bootstrapping (Carpenter et al., 2003; Saravanan et al., 2020) to compare statistical groups (COP and CTRL episodes). First, we hierarchically bootstrapped the dataset by resampling it n=1000 times (sampling with replacement from experiment replicates and their associated episode replicates). We then computed the mean CRFs $\mathrm{r}_{B,\mathrm{G,i}}(\tau)$ of behavior B across episodes of a given statistical group G for each of these datasets ($i = 1, \dots, n$). From the bootstrapped distribution of $\mathrm{r}_{B,\mathrm{COP,i}}(\tau)$, we computed again the mean $\overline{r}_{B,\mathrm{COP}}(\tau)$. We define the mean normalized CRF as

$$\overline{n}_{B,\mathrm{COP}}(\tau) = \frac{\overline{r}_{B,\mathrm{COP}}(\tau) - CL_{B,\,\mathrm{CTRL}}(\tau)}{CU_{B,\,\mathrm{CTRL}}(\tau) - CL_{B,\,\mathrm{CTRL}}(\tau)},$$

where the upper and lower confidence interval bounds $CU_{B,\,\mathrm{CTRL}}(\tau)$ and $CL_{B,\,\mathrm{CTRL}}(\tau)$ are estimated using the 0.5% and 99.5% quantiles bootstrapped from shuffled CTRL episodes, respectively. We shuffled the data within each CTRL episode using a random circular shift in every resampled dataset. We did this to artificially increase sample size of sparse events along the time axis. In this way, $\overline{n}_{B,\mathrm{COP}}(\tau)$ is expected to lie between zero and one 99% of the time, if the (onset) density of behavior B is not related with copulation events. In other words, values of the normalized CRF above one or below zero are considered significant deviations from baseline, i.e., behaviors that occur excessively often during COP episodes or that are excessively suppressed during such episodes.

Vocal feature statistics

We annotated motifs and bouts computationally. Motifs are defined by the stereotyped sequence of consecutive syllables. Some motifs can get interrupted, usually between two syllables, we excluded these examples from further analysis. "Complete motifs" were those that exceeded a duration threshold defined for each male separately. Following others, we defined bouts (initially named "strophes" by (Bischof et al., 1981; Sossinka & Böhner, 1980)) as a sequence of motifs separated by less than 2 s (Sossinka & Böhner, 1980; Jarvis et al. 1998). Since durations of complete motifs vary across birds, we normalized their durations for each male by the mean of median durations computed over CTRL episodes (one median per episode). In the following, we compared features of complete motifs, bouts, or vocalizations across statistical groups (e.g., <50s per-SCA vs CTRL, or DIR vs UNDIR, Figure **2.1**2.2).

Due to relatively small numbers of copulations per animal pair, we performed statistical analysis only across all pairs. As for behavioral rate analysis, we used hierarchical bootstrapping to compare a vocal feature V in two statistical groups G_1 (e.g., COP) and G_2 (e.g., CTRL). For a given statistical group G, we first construct a hierarchically bootstrapped distribution of n=1000 episode-averaged values $\mathrm{x}_{V,\mathrm{G},\mathrm{i}}$ of vocal feature V (number of introductory notes, motif duration, motifs per bout, call duration).

To visualize the bootstrapped distribution of averages, we use a violin plot (Hintze & Nelson, 1998). To test significant difference of these distributions for G_1 and G_2, we evaluate whether the proportion r of bootstrapped datasets for which $\mathrm{x}_{V,G_2,\mathrm{i}} - \mathrm{x}_{V,G_1,\mathrm{i}} \leq 0$ satisfies either $r < \frac{\alpha}{2}$ or $r > 1 - \frac{\alpha}{2}$, with significance level $\alpha = 0.05$. In other words, if one of these conditions for r is satisfied, we conclude that V significantly depends on the group for which it is measured.

Results

A behavioral dataset of mixed-sex zebra finch pairs in copulatory and non-copulatory contexts

Manual labelling of accelerometer recordings resulted in a dataset of 54’148 vocalizations annotated across six experiment replicates (Table 2.1, examples shown in Figure 2.1a-b). Our dataset reveals novel findings on the categorical structure of zebra finch calls, such as the fact that two short “nest” calls can get arbitrarily close, transitioning gradually to “tet” calls (Figure AIV.1). In none of the 14 birds we were able to separate “nest” call clusters from “whine” calls, which is why we annotated both with the broader category label “nest/whine” (see “Appendix IV”). Our vocal dataset is complemented with manually annotated video recordings comprising 4074 non-vocal behaviors (Table 2.2, examples shown in Figure 2.1a,c).

Rate-based and feature-based behavioral signatures of copulations

Examining our dataset, we first probed whether behavioral rates change with proximity to the copulations. For vocal behaviors (Figure 2.1d, left side), we found highly elevated singing rates (mean normalized CRF > 1) on average around 27 s before the copulation, rapidly declining to CTRL-associated levels after the copulations. In contrast, “nest/whine” call rates were more symmetrically elevated around copulations (average interval: [-27 s, 30 s] in females, [-13 s, 24 s] in males). Interestingly, the structural difference of “tet” versus “nest/whine” calls is reflected in a different CRF, with “tets” being elevated only prior to copulations. The rates of three highlighted non-vocal behaviors (Figure 2.1d, right side), were highly elevated roughly < 25 s prior to copulations (the CRFs of the remaining behaviors are included in “Appendix VI”). We observe occasional significance at a much lower level on longer time scales; it might be that the burst-like and diverse nature of zebra finch behavior leads to spurious significance that would vanish for larger sample sizes. However, some long-range effects, such as a slightly elevate male “tet” call rate seem to be persistent across the time axis.

Knowing that song composition and tempo reportedly vary in female-directed behavior (Sossinka & Böhner, 1980), and do so increasingly with the quality of female stimuli (Bischof et al., 1981), we wondered whether a similar effect is observed prior to copulations. Indeed, we find that the songs occurring less than 50 s prior to copulations show the same differences to CTRL songs, as have been reported previously for directed versus undirected songs (Figure 2.2a-c). When labelling video recordings, we

Table 2.1: Vocal dataset overview. "Vocal density" is defined as the fraction of time with vocal activity in annotated recordings, averaged across both birds (fifth column). Note, that only a fraction of all annotated vocalizations resides within the COP / CTRL episodes (last column).

Experiment replicate	**Vocally annotated (h)**	**Number of female vocalizations**	**Number of male vocalizations**	**Vocal density**	**Number of COP / CTRL episodes[11]**	**Fraction of vocalizations within COP and CTRL episodes**
b8p2male-b10o15female	2.74	6076	4586	0.049	14/14	1
CopExpBP03	6.58	6446	5476	0.027	15/15	0.4942
CopExpBP04	3.16	4020	6706	0.052	7/7	0.5264
CopExpBP06	3.51	5226	4887	0.041	9/9	0.5701
CopExpBP07	1.86	2622	2598	0.048	4/4	0.5569
CopExpBP08[12]	1.53 (3.58)	2105 (3818)	1687	0.043	8/8	1 (0.5104)
All	**19.39**	**28208**	**25940**		**57**	

Table 2.2: Non-vocal dataset overview. Behaviors are assigned to a specific bird (female or male), except for cloacal contact, which is a shared behavior.[13]

Experiment	**Male behaviors**	**Female behaviors**	**Cloacal contact (shared behavior)**
b8p2male-b10o15female	654	612	10
CopExpBP03	522	259	10
CopExpBP04	566	381	6
CopExpBP06	221	183	4
CopExpBP07	389	253	4
CopExpBP08	654	612	10
All	**2352**	**1688**	**34**

annotated directed (DIR) and undirected (UNDIR) songs manually (see "Appendix V"); this is a hard task, because directedness has to be judged based on the relative posture of the birds, and additionally, singing has to be distinguished by ear. The differences between songs with an onset in DIR versus UNDIR video-segments have showed a less pronounced effect in the indices for female-directedness: although all indices show the expected trend on average, only one is significant (Figure 2.2a-c).

Since the main focus in zebra finch research has been on the male song and much less on the calls, we decided to extend our feature-based analysis to call durations. We analyzed call durations in 25 s intervals

[11] From these episodes, two COP and one CTRL episode have been excluded from further analyses (see "Annotation of vocal and non-vocal behaviors")

[12] Due to a technical failure, from the originally annotated 3.58 h (denoted in parentheses), we have lost the annotations of 1925 male vocalizations that occurred outside of the COP / CTRL episodes. The numbers that are provided without parentheses correspond COP / CTRL episodes (1.51 h).

[13] Data has been annotated by Dr. Mariana da Rocha.

– either before copulations, for "tet" calls, or around copulations, for "nest/whine" calls (distinguished based on the shape of call rate curves in Figure 2.1d). As shown in Figure 2.2d-e, we found that call durations

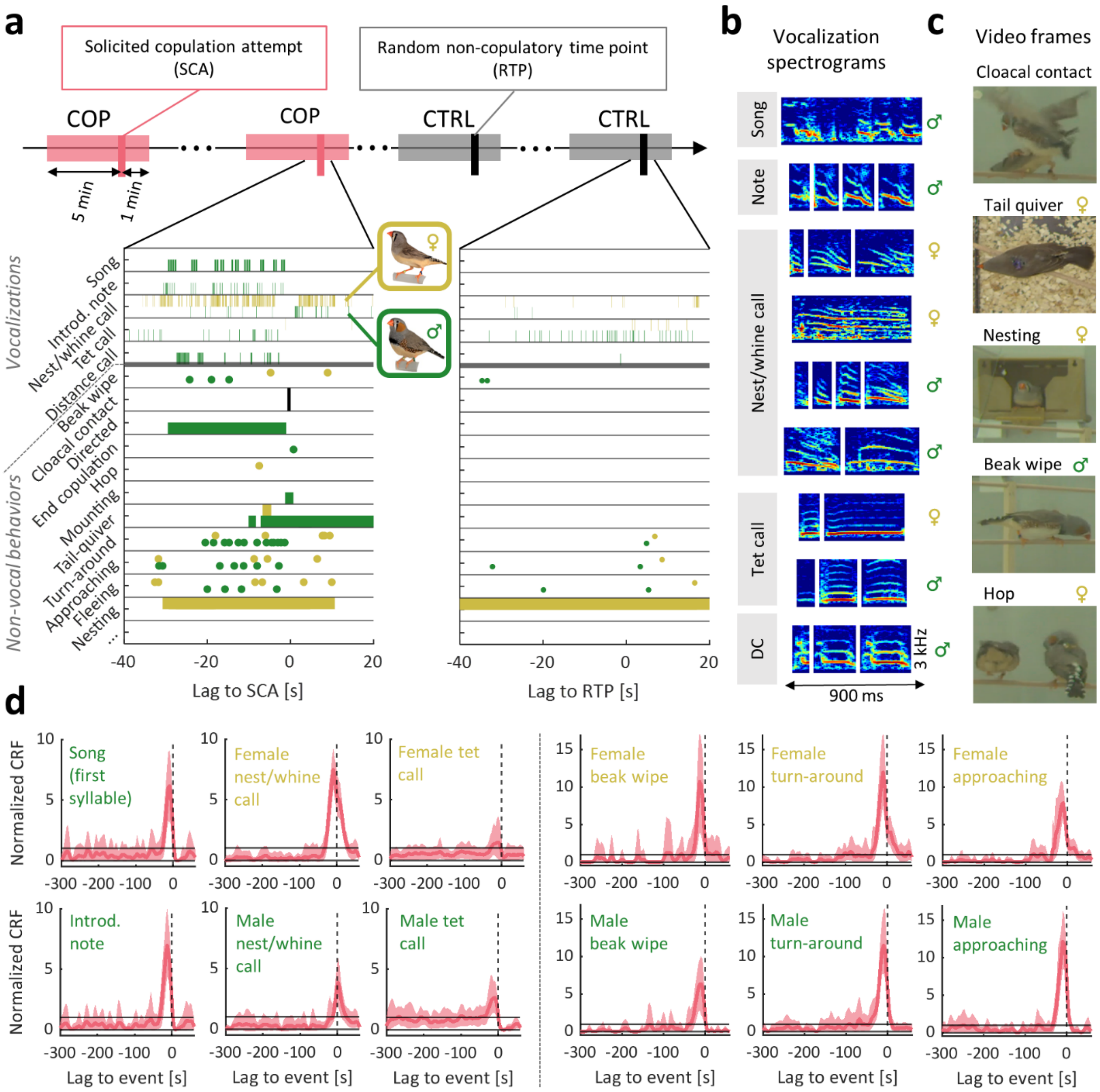


Figure 2.1: Copulations are characterized by a categorical signature of vocal and non-vocal behaviors. (a) Behavioral annotation of copulation (COP) and (CTRL) episodes collected from an exemplary bird couple. **(b-c)** Accelerometer spectrograms of vocalizations (b) and video frames of a few non-vocal behaviors (c) annotated for the exemplary bird couple. DC: distance call. **(d)** Normalized mean conditional rate functions (CRF, thick red lines) with 99% confidence intervals (red shaded areas) of selected vocal and non-vocal behaviors across n=55 solicited COP samples collected from n=6 (vocal) and n=5 (non-vocal) experiment replicates. To allow for comparison between COP and CTRL samples, bootstrapped COP rates were normalized using a 99% confidence interval computed from bootstrapped and shuffled CTRL rates from n=56 control samples (black horizontal lines; see "Behavioral rate analysis"), which sets the threshold for statistical significance at a fixed unit distance along the y axis.

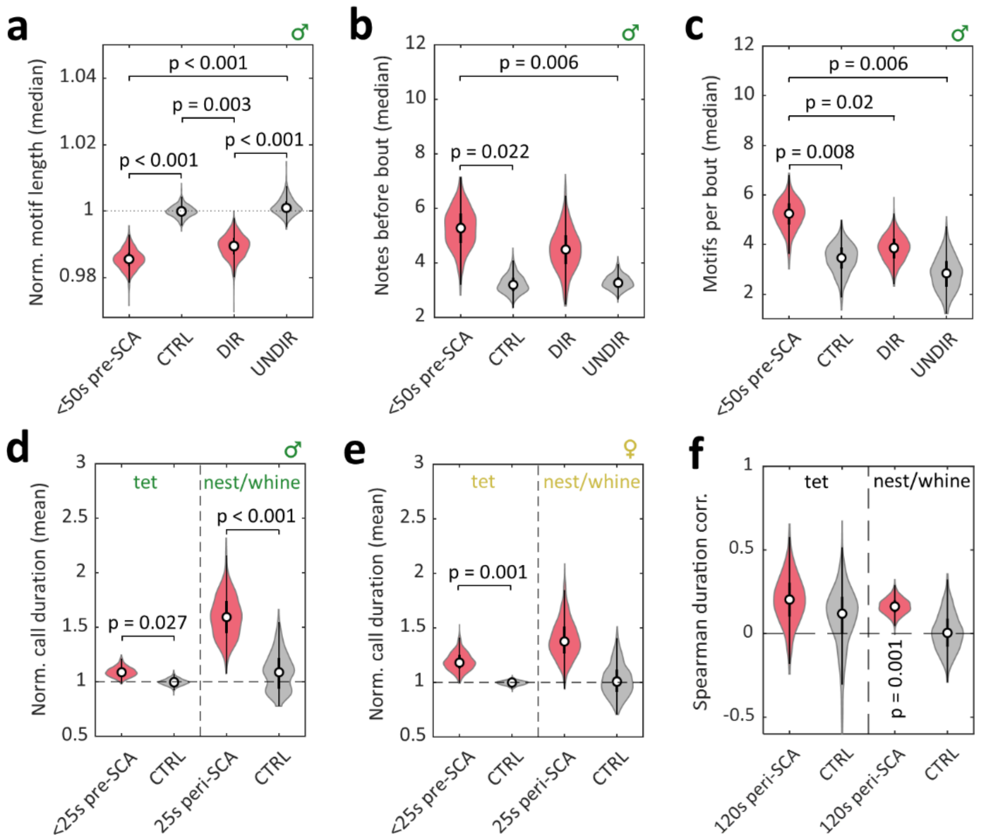


Figure 2.2: Variation of song composition, song tempo, and call durations around copulations. (a-c) Hierarchically bootstrapped statistics of median song features computed for each episode separately: median lengths of complete motifs normalized for each male to the means of CTRL episode medians (a), median number of introductory notes before a bout (b), and median number motifs per bout (c). Medians are compared in these categories: in the 50 s window before a SCA ("<50s pre-SCA"); in CTRL episodes; for motifs/bouts with an onset within video-annotated directed song ("DIR") or undirected song ("UNDIR"). See the section "Vocal feature statistics" for a definition of the features, as well as details on bootstrapping and hypothesis testing. **(d-e)** Similarly as in (a-c), we hierarchically bootstrapped median "tet" and "nest/whine" call durations, for male (d) and female birds (e), normalized to "tet" durations in CTRL episodes. For "tets" we have chosen a 25 s window before SCA ("<25s pre-SCA"), and for "nest/whines" a 25 s window around SCA ("25s-peri-SCA"). **(f)** Similarly as in (a-e), we hierarchically bootstrapped Spearman correlations between durations of same-type, different-sex calls as follows: from a sequence of calls of a given type, we selected pairs of consecutive calls where a female precedes a male call by up to 2 s; we then computed the correlation between female and male durations of these selected call pairs. P-values calculated from the bootstrapped distributions reflect a test for nonzero correlation.

are elevated in these intervals compared to CTRL episodes, except for female "nest/whine" calls, which failed the hypothesis test by a small margin (p=0.058, see "Vocal feature statistics" for our hypothesis testing method). Seeing the large between-episode variation, we wondered whether the mixed-sex couples covary call durations on a short timescale (with an inter-call-interval of up to 2 s) across a longer timescale (in a 2 min window) around copulations, as they upregulate and downregulate elevated call durations. We found a small but highly significant Spearman correlation between successive female and male "nest/whine" calls (Figure 2.2f). In other words, around copulations, male "nest/whine" calls share a monotonic duration relationship with temporally proximal female "nest/whine" calls.

Discussion

Copulations are key events that can decide about the survival of entire animal populations and the traits of future generations. We provide the tools to detect songbird copulations automatically and remotely with on-bird vibration sensors, and we show that solicited copulations are signaled on vibration and video data 25-30 s in advance (Figure 2.1d). We will release our manually labelled dataset of vocal and non-vocal behaviors, which we expect to be useful to train machine-based recognition systems. Remote detection and prediction of copulations have several use cases across research, animal caretaking, or wildlife management.

Copulations come in different variants and their frequency might be used as a readout of animal wellbeing. The success of the copulation depends on whether prolonged cloacal contact is achieved or not[14]. Copulations without female solicitation, where the male forces himself, often by holding the female by her head feathers, have also been observed[15] (Birkhead et al., 1988, 1989). It would be of great interest to further develop our copulation detection approach to enable the automated distinction of these different variations of copulation events, as they likely entail distinct consequences for reproductive success. Our remote detection system allows to probe for factors that influence copulatory or reproductive success, by targeted experimental interventions. A further application we can envision for automated copulation detectors, is to measure copulation frequency to evaluate the wellbeing of experimental animals, and thus help with efforts concerning animal welfare. Validity of such a readout is supported by the evidence that copulation frequency is significantly reduced in zebra finches treated with stress-inducing corticosterone (Scalera & Tomaszycki, 2018). An even broader potential impact of our work lies in the study of vocal copulation signatures.

We expect that the tools presented here will open new avenues of investigation into the functions of vocal communication and the implications for reproductive success. An increasing number of studies have used animal-borne sound recorders to investigate vocal behavior in various species, such as whales and seals (Johnson et al., 2009), bats (Greif & Yovel, 2019), chipmunks (Couchoux et al., 2015), and songbirds (Gill et al., 2015). These studies try to better understand the role of animal vocal communication for alarming, food source advertisement, social learning, territory defense, pair bonding, mate attraction, and offspring care. Investigating how vocal communication can successfully elicit mate attraction and stimulation has important implications for songbird conservation, as it can help inform captive breeding efforts for species reintroduction programs (Lewis et al., 2021). Having the knowledge to support such programs is of high importance given the recent reports of widespread songbird population declines (Bairlein, 2016; Rosenberg et al., 2019). Beyond these application-oriented paths, novel insights into vocal encoding of information can be gained.

With over 54’000 labelled zebra finch vocalizations, our vocal dataset is of unprecedented volume, allowing detailed investigation into the categorical nature of zebra finch vocalizations. A fundamental concept in the study of semantics is the distinction of graded versus discrete signals (Marler, 1967). Here, we report intermediate forms between previously distinguished call types (“Appendix IV”), which would indicate

[14] Dr. Mariana da Rocha has annotated cloacal contact based on video recordings. Future work could investigate whether there are specific behavioral signatures for different degrees of cloacal contact.

[15] We observed two unsolicited mountings, which exhibited atypical vocal copulation signatures. Due to the low sample size (n=2), we excluded them from our analyses.

structural gradedness. This structural gradedness has to be distinguished from functional gradedness; zebra finch calls of a given type have previously been shown to elicit invariant neural responses in some auditory areas, although to different degrees (Elie & Theunissen, 2015, 2019). Structural call categorization is challenging, even more so from accelerometer data. Compared to wall microphones, the accelerometer signal is attenuated for higher frequencies, influenced by body movements, and dependent on tight skin contact (Rüttimann et al., 2022). Semi-supervised approaches on accelerometer recordings perform poorly (Lorenz et al., 2022) and consequently manual annotation is needed. We encourage to investigate the structural nature of zebra finch call repertoires in more detail, for instance, with methods such as fuzzy clustering (Cusano et al., 2021), or the Cuzick-Edwards test statistic (Cuzick & Edwards, 1990). The latter is a suitable statistical method to quantify the degree of categorical overlap. Here we consult two-dimensional embeddings to categorize calls, which leads us to distinguishing fewer categories as others (Elie & Theunissen, 2016; Gill et al., 2015; Zann, 1996). We find that the main two call types observed in our experiments across both sexes, "tet" and "nest/whine" calls, show a different copulation signature (Figure 2.1d). Thus, their structural difference is mirrored in differential regulation, with "nest/whines" being upregulated both, before and after of copulation, but "tets" are elevated only in advance.

To probe the functional role of calls conclusively requires more controlled experimental settings, but nevertheless many researchers have been tempted to hypothesize on "what animals say". In 1872, Charles Darwin wrote a treatise on "*The expression of emotion in man and animals*", in which he carefully examined plethora of anecdotal evidence on vocal expression in light of his theory of evolution (Darwin, 1872). The fundamental idea has been that affective vocalizations express inner states such as joy or fear. Recently, it has been shown that stress can be transferred among zebra finch mates by means of vocal communication (Perez et al., 2015). Vocalizations could therefore be used not only to express inner states, but to create resonance between inner states of bonded animals. Interestingly, we report that the durations of "tet" and "nest/whine" calls are simultaneously upregulated and that the "nest/whine" durations covary between birds around copulations (Figure 2.2d-f). This could potentially hint at a synchronized inner state, such as arousal. How mates orchestrate courtship and what role the different behaviors play remains an exciting research topic – even more when neural dimensions are taken into account.

Songbirds are important model species for the study of the neural basis of vocal learning. Vocal learning, the ability to imitate conspecific vocalizations, with many parallels shared between birdsong learning and human speech acquisition (Doupe & Kuhl, 1999; Lipkind et al., 2020). Birdsong has been shown to differ in tempo and composition when directed towards a female (Sossinka & Böhner, 1980). We have annotated directedness manually from video data, which is non-trivial, compared to previous experiments where the experimenter takes control over presence of female stimuli. We find that song markers of female-directedness are more pronounced prior to copulations compared all "directed" songs as judged from video data. This is in line with the finding that these markers depend on the quality of female stimuli (Bischof et al., 1981). We hope that more quantitative measurements of directedness in naturalistic environments can be developed in the future, leveraging recent advances in computer vision that allow posture tracking (Graving et al., 2019; Mathis et al., 2018). Our dataset is ideally suited to supervise automated behavioral tracking, which would open the doors for real-time copulation predictions.

Birdsong is acquired through self-reinforced learning, however, the important role of reinforcement in adult songbird's vocal learning has so far mainly been studied through the use of aversive stimuli (Andalman & Fee, 2009; Charlesworth et al., 2011, 2012; Tian & Brainard, 2017; Tumer & Brainard, 2007; Warren et al., 2011). Being able to connect adult vocal learning to a potent positive reinforcer, such as successfully eliciting solicitation or copulation in a mate, could lead to new avenues of investigation for birdsong neuroscience. Performance-contingent playback of female copulation-related behaviors could be potentially used to test whether the changes in female-directed song can be overridden through reinforcement.

Taken together, we have developed the first automated detector for copulation events that can be applied in complex and naturalistic environments. We annotated vocal and non-vocal behaviors, and characterized copulation specific behavioral signatures, announcing solicited male mountings roughly 25-30 s in advance. We envision that our work will have numerous benefits: wildlife monitoring and management of endangered species, as well as providing new insights into vocal communication, the implications of vocalizations to an animal's fitness, and the origins of language.

Data availability

We will release our dataset upon publication of our work in a peer-reviewed journal.

Funding

Financial support was provided by the Swiss National Science Foundation (Grant 31003A_182638; and the NCCR Evolving Language, Agreement No. 51NF40_180888).

Chapter 3 – A system for controlling vocal communication networks[16]

The exchange of information using sequences of vocally produced acoustic elements is widespread among animal species. Studies of animal communication remain challenging because the meaning of vocal signals depends not just on their sound features, but also on the behavioral state of animals and the environmental context (Ciaburri & Williams, 2019; Ljubičić et al., 2016; Vignal et al., 2004). As a result, the complexity of the vocal dynamics grows rapidly with group size, making it difficult to detect and assign the information conveyed to conclude causality.

A simple technique to study animal communication in a controlled setting is video and audio playback (Böhner, 1983; Burt et al., 2001, 2007; Evans & Marler, 1991; James et al., 2019; Ljubičić et al., 2016; Perez et al., 2015). Even simple playback systems can mimic a conspecific or heterospecific individual to some degree: Male zebra finches and Bengalese finches sing directed song to video presentations of female conspecifics (Ikebuchi & Okanoya, 1999), female zebra finches perform courtship displays to videos of male conspecifics (Swaddle et al., 2006), and videos of "audience" hens potentiate alarm calls when produced in the presence of a predator model (Evans & Marler, 1991).

Moreover, modern playback systems can interact with animals in a feedback loop (King, 2015). These interactive playback systems (IPS), also referred to as virtual social environments (that simulate social environments) impose artificial exchanges between an animal and a robot or a computer. In songbirds, IPS have been extensively used to study social interactions and influences during developmental song learning (Ljubičić et al., 2016; Perez et al., 2015).

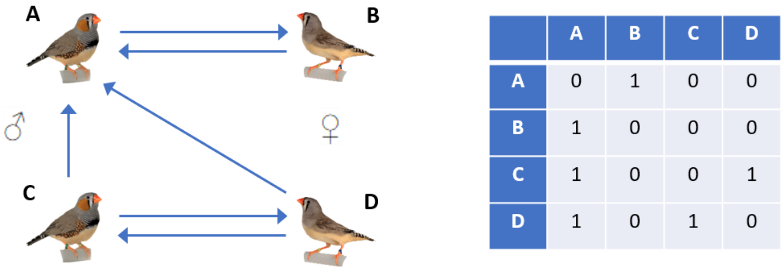
In some cases, no qualitative difference was found in the response to live versus video stimuli (Elie et al., 2010; Swaddle et al., 2006; Takahashi et al., 2017; Vignal et al., 2004). However, in other studies, an attenuated (Evans & Marler, 1991) or enhanced response to video stimuli was reported (Ikebuchi & Okanoya, 1999), suggesting that interactions among animals may exhibit dynamics that are hard to mimic using pure sound and video playback. For example, juvenile zebra finches learn better from live tutors than from interactive vocal playback (Derégnaucourt, 2011), indicating that some aspects of natural communication are hard to mimic using playback.

We propose a new approach to studies of vocal communication in a naturalistic setting, which consists of connecting live animals via programmable auditory channels. The system we present allows flexible control of the communication network among up to four animals housed in separate, electronically connected sound-isolation chambers. To offer controllability of the auditory scene akin to playback systems, the auditory link between any pair of animals can be programmatically enabled or blocked in each direction independently (Figure 3.1).

The main technical challenge inherent to such a communication system is to prevent transitive sound propagation in serially connected chambers. For example, in an asymmetric network C → A ↔ B in which animal B shall hear animal A but not animal C (Figure 3.1), sound leakage from C to B must be prevented using a dedicated sound gating mechanism. Another challenge is to prevent acoustic feedback instabilities,

[16] The text and figures in this chapter have been adapted from a published manuscript (Rychen et al., 2021). My personal contributions and changes to the text are detailed in the thesis section "Source and authorship attribution".

which can occur in closed microphone-loudspeaker loops when the closed-loop gain is higher than 1 at any frequency.



	A	B	C	D
A	0	1	0	0
B	1	0	0	0
C	1	0	0	1
D	1	0	1	0

Figure 3.1: Controlling the topology of communication networks in zebra finches[17]. Left: Schematic of a specific communication network among 4 zebra finches. In this example, the communication links within two male-female couples are symmetric, but only male A can hear the other couple. In other words, there are links from birds C and D to bird A but there is no link in the reverse direction. Right: This network can be represented as a binary 4-by-4 connection matrix in which the diagonal elements are zero and six off-diagonal elements are one.

We addressed these challenges with a least mean square (LMS) echo attenuation filter and a dynamic squelch. The echo attenuation filter subtracts out a large fraction of the microphone signal elicited by the loudspeaker in the same chamber and the dynamic squelch prevents transitive sound propagation in linked chambers. Furthermore, the squelch suppresses the playback of microphone noise when the associated animal is silent. These technical aspects are detailed in the published manuscript (Rychen et al., 2021). In this thesis chapter, I present data from applications in adult male zebra finches, demonstrating reliable vocal interactions constrained by the imposed network structure.

Methods

Animals and experiments

Zebra finches (*Taeniopygia guttata*) bred and raised in our colony (University of Zurich / ETHZ) were kept on a 14/10 h light/dark daily cycle, with food and water provided ad libitum. All experimental procedures were approved by the Cantonal Veterinary Office of the Canton of Zurich, Switzerland (license numbers ZH207/2013 and ZH077/17). All methods were carried out in accordance with relevant guidelines and regulations (Swiss Animal Welfare Act and Ordinance, TSchG, TSchV, TVV).

In all experiments, the LMS filter was trained each day right before the starting of the first recording session. Birds had unlimited access to food, water, cuttlefish bone, sand bath, water bath, millet, and three perches. Before recording any data, we provided birds with 5 days habituation time in the setup, 1 hour on the first day, and 1 additional hour each day until a maximum of 4 hours was reached. After the habituation period, interaction channels were engaged during experiment sessions in the range from approximately 30 min to around 2.5 h, depending on the birds' vocal activity. For the remainder of the day, the birds were housed in a large social cage.

In the experiment shown in this thesis, male zebra finches could move freely inside the recording chamber (60x60x60 cm^3) that was equipped with a swing, except for one replicate, where birds were housed in

[17] This figure has been jointly produced with Dr. Jörg Rychen (who drafted the figure).

39x23x39 cm^3 plexiglass cages. Following the 5-day habituation period, birds were placed into the setup for up to 4 h/day. We noticed that under these more transient housing conditions, vocalization rates tended to be smaller than in the 24 h/day setting. To incentivize birds to vocalize in asymmetric cyclic networks, we played a female or male call roughly every 15-30 s to the top bird (T). To minimize interference, in case of ongoing vocal interactions, playback was automatically delayed by 3.5 s.

Cross-covariance analysis

We characterized vocal interactions between pairs of connected birds by the cross-covariance (CCV) function

$$CCV_{A,B}(\tau) = \frac{1}{T}\int_0^T (\delta_A(t) - \overline{\delta_A})(\delta_B(t+\tau) - \overline{\delta_B})dt$$

of their mean-subtracted vocalization onset trains δ_A, δ_B and where T denotes the duration of the session. We computed CCV functions up to a maximum lag of 2 s and smoothed them with a 300-ms Gaussian filter with standard deviation of 60 ms.

To assess the significance of CCV peaks, we shuffled the data using circular shifts during intervals of vocal activity. To identify these intervals, we first grouped call onsets of the responding bird into time intervals such that consecutive call onsets separated by less than 500 ms were grouped in the same interval. In case an interval was less than 2 s long, we extended it, to make the minimum interval duration 2 s. This grouping procedure was either running forward in time starting with the session beginning, or backward in time starting with the session end, with equal probability. On average, this grouping procedure resulted in 258 ± 350 intervals per hour.

Within an interval, we circularly shifted the onsets by a common amount that we uniformly sampled in [0 t_i], where t_i is the interval duration. By repeating this random circular shifting procedure n=200 times, we obtained a distribution of shuffled CCV functions. Significant CCV peaks had to exceed the standard deviation of this distribution by a factor of 3, corresponding to a p-value of roughly 0.01.

To compare CCV functions in a common plot, we normalized them as

$$CCV_{\text{norm}}(\tau) = \frac{CCV(\tau) - CI_{\text{lower}}(\tau)}{CI_{\text{upper}}(\tau) - CI_{\text{lower}}(\tau)},$$

where the upper and lower confidence interval bounds CI_{upper} and CI_{lower} lied 3 standard deviations away from the mean of our random shuffle predictor.

Results

Communication networks constrain vocal interactions

We tested whether birds engaged in reliable vocal interactions constrained by the network topology. To this end, we imposed on the vocal interactions two distinct networks, either a symmetric hierarchical network, or an asymmetric ring network that we judged was sufficiently different from the hierarchical network to observe an effect of topology. In the hierarchical network, the male T (top) could interact with the two other males L (left) and R (right) and the males could each interact with T but not with each other. This hierarchical network models a sort of anti-eavesdropping situation in which T can simultaneously hear both other males, but not in the context of ongoing communication, which normally sets the stage for eavesdropping. The ring network models a middleman communication situation, in which message passing between any pair of birds is direct in one direction but indirect in the other.

In one experiment among three males, we found that switches between the hierarchical and the cyclic networks triggered strong changes in vocal interactions (Figure 3.2). Adding a feedback connection to a unidirectionally connected bird pair, from R ↔ T (cyclic) to R ↔ T (hierarchical) could result in rapid and vigorous vocal responses in R right after the first call in T that was audible to R (Figure 3.2a). Conversely, when switching from hierarchical to cyclic, at the most extreme case of two birds at the bottom of the hierarchy (L and R), one bird switched from not responding to a single call of a given type when not connected (hierarchical) to responding to virtually every single call of that type when the cyclic connection appeared (Figure 3.2b) demonstrating that animals can react dramatically to imposed network changes.

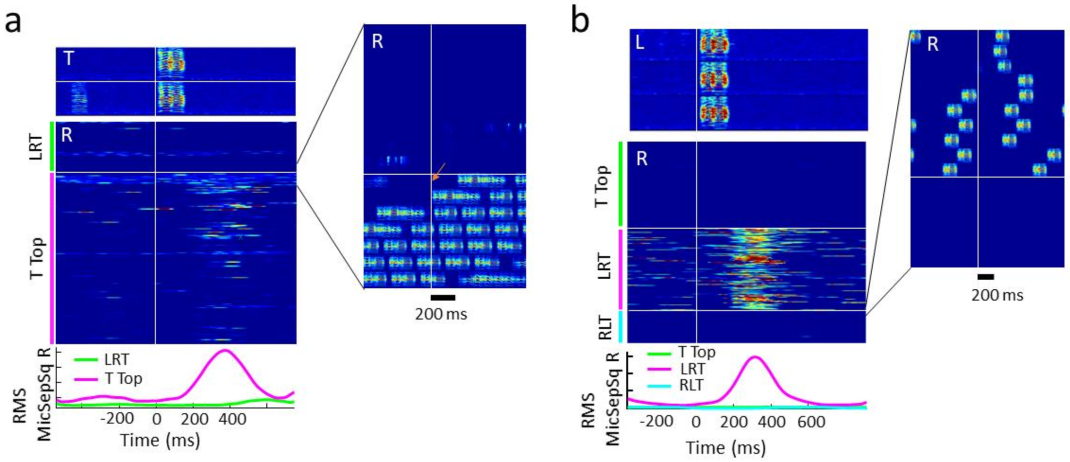


Figure 3.2: The communication network constrains vocal interactions[18]. Vocal interactions can change very sensitively in response to switches between communication networks, shown here for hierarchical and cyclic networks. (a) Switch from L → R → T → L (cyclic, LRT) to L ↔ T ↔ R (hierarchical, T Top) leads to an initial string of call response in R to the first audible call in T (orange arrow in inset). Shown are spectrograms of two example calls in T (top) and several hundred responses in R depicted as a root-mean-square (RMS) stack plot (middle), aligned to the onset of the calls in T when T's calls are not audible (LRT, top) and when they are (T Top, bottom). The calling times in T run from top to bottom. The bottom curves represent the RMS MicSepSqR curves averaged over all calls in T in both conditions. R responds with a latency in the range 300-500 ms, with waning reliability. The inset on the right depicts spectrograms of responses in R right before the network switch (top) and right after the switch (bottom), aligned to the onsets of T's calls (vertical white line). Both T and R produce dense strings of calls, which leads to multiple depictions of a given call in R in subsequent rows. (b) Switch from L ↔ T ↔ R (hierarchical, T Top) to L → R → T → L (cyclic, LRT) uncovers vigorous responses in R to calls in L (example spectrograms on top). However, L does not respond to R when the cyclic network changes direction to R → L → T → R (cyclic, RLT). The inset on the right illustrates the immediate silence following the switch to the second cyclic network.

[18] This figure has been produced by Prof. Dr. Richard Hahnloser.

We quantified the reliability of call-call interactions in pairs of birds in terms of the cross-covariance (CCV) function (see section "Cross-covariance analysis"). We found that a connection from bird A to bird B typically entailed the presence of reliable vocal responses in B to calls in A: the CCV function often peaked above a shuffle predictor (corresponding to p<0.01, see section "Cross-covariance analysis"). As expected, when connections were unidirectional, the CCV functions displayed at most a single peak at a positive time lag (Figure 3.3a,c), in agreement with the causality imposed by the network.

Pairs of disconnected birds can be prevented from hearing each other and from direct vocal interactions by appropriate separation and sound isolation of recording chambers. Nevertheless, calls in non-connected birds could be correlated as shown in Figure 3.3b-c, in which two birds L and R at the bottom of a hierarchical network exhibited a CCV peak near a zero-time lag, indicating that both birds tended to respond

to the same calls in T. Such observation illustrates the well-known fact that correlation does not imply causation, because correlations can arise from a common cause, i.e., bird T at the top of the hierarchy. We observed such non-causal correlations in 2/3 non-connected bird pairs at the bottom of the hierarchy.

The same was not observed in bidirectionally coupled bird pairs (L ↔ T and R ↔ T). In 4/6 of such pairs, we observed two significant CCV peaks: one at a negative time lag (bird T responds) and one at a positive time lag (bird T is responded to). Such symmetric interactions are characteristic of turn-taking, which is typical in many species including zebra finches (Elie et al., 2010; Hoffmann et al., 2019; Okobi et al., 2019; Takahashi et al., 2013). Moreover, in 5/6 bidirectionally connected bird pairs, the birds on top of the hierarchy were less responsive (average normalized CCV peak 1.56) than the lower birds (average peak 3.31), suggesting that a larger social network entails less reliable communication.

Discussion

Using standard off-the-shelf components, we implemented a digital system for controlling the vocal communication network among a small group of animals. The system yields high-quality recordings of each animal's vocalizations, provided the animals are separately housed in acoustically distinct environments.

To test the system's capability to control communication networks, we restricted vocal exchanges to diverse sub-networks and thereby regulated the social complexity among animals. The communication networks we imposed were sufficient to enable non-trivial vocal exchanges that were not merely reflexive but reflected birds' personalities or states (Figure 3.2), and ranks in the group (Figure 3.3). As such, there are many possible uses for our system when applied to three or more birds. For example, our system could complement observational approaches using small backpack recorders attached to animals (Anisimov et al., 2014; Gill et al., 2015; Stidsholt et al., 2019). That is, our system can help to overcome a shortcoming of observational studies, which can merely yield hypotheses about the "meanings" of certain types of vocal interactions but are not amenable to selective testing of these hypotheses because vocal exchanges among animals are virtually impossible to manipulate without a dedicated communication system. Thus, when a certain meaning has been hypothesized from observation in freely interacting animals, it would be reassuring to infer the same meaning in loss-of-function (removed connection) and gain-of-function (e.g., playback) experiments implemented with our system.

There are several limitations of our system, which could be addressed in future extensions. For example, it is currently not possible to manipulate sound direction because we use only one loudspeaker per chamber. Birds can estimate sound source direction from interaural time differences (ITDs) and interaural level

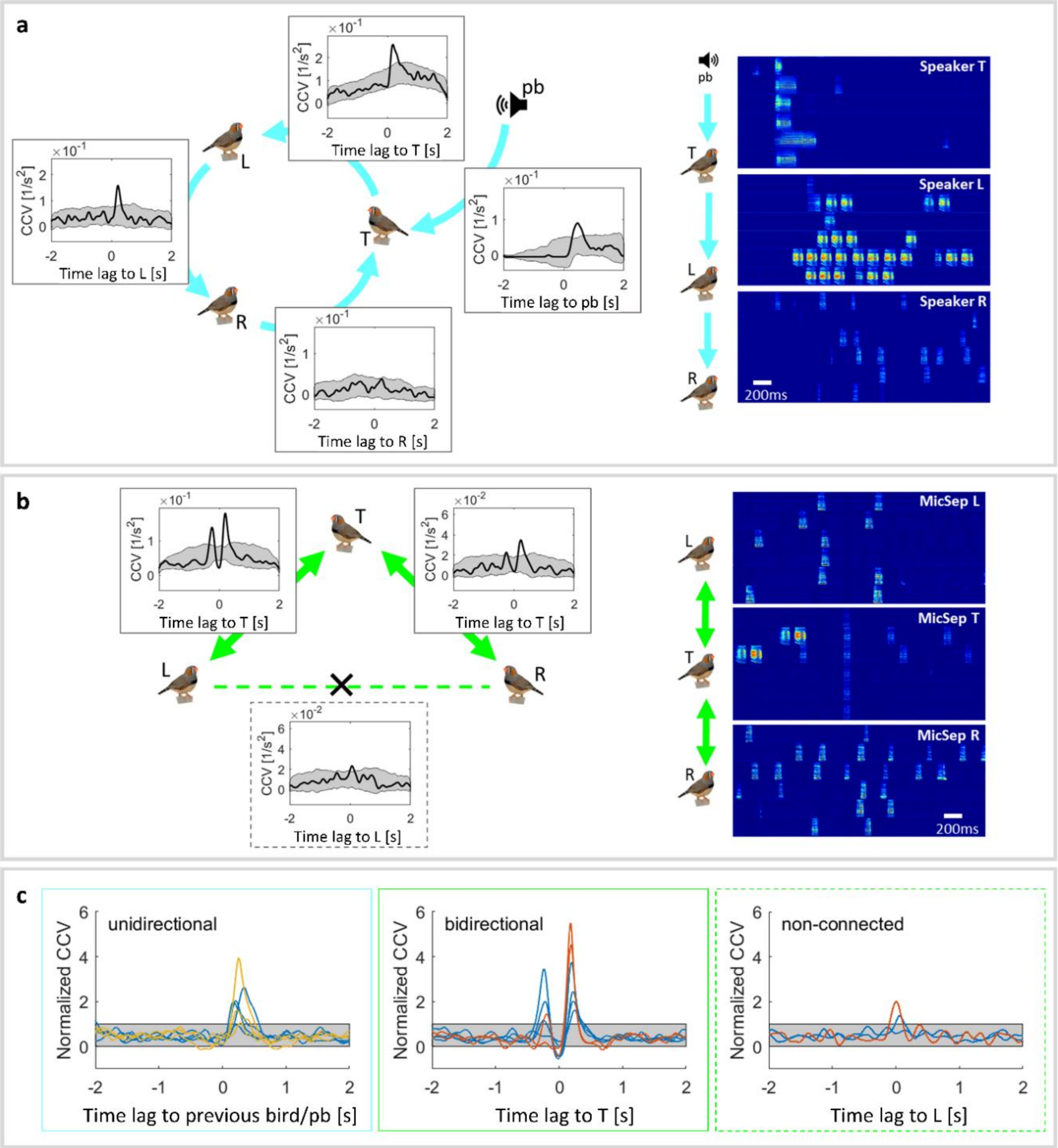


Figure 3.3. The structure of vocal interactions mirrors that of the imposed network. (a) In an asymmetric network, the interactions tend to be asymmetric, and (b) in a symmetric network, they tend to be symmetric. (a) In pairs of asymmetrically connected birds, cross-covariance (CCV) functions (black lines, see section “Cross-covariance analysis”) indicate unidirectional vocal exchanges revealed by unimodal peaks. Stacks of example spectrograms are shown (right) with the auditory stimulus (pb) presented in chamber T (top), T’s response broadcast to bird L (middle), and L’s response broadcast to R (bottom); corresponding rows in the 3 sub-panels are from simultaneous recordings. (b) In a symmetric hierarchical network, CCV functions reveal bidirectional vocal interactions. (a,b) The gray areas represent 3 standard deviations of a random shuffle predictor (see section “Cross-covariance analysis”). (c) Normalized CCV functions in unidirectional (left), bidirectional (middle), and non-connected (right) bird pairs, n=3 bird groups (blue in all 3 subpanels, orange in the middle and right subpanel, and yellow in the left subpanel) across a total of 5 different network configurations (two asymmetric, three symmetric). To allow for comparison among experiments, CCV

functions were normalized by the shuffle predictor (gray areas), which sets the threshold for statistical significance at a fixed unit distance along the y axis.

differences (ILDs). We could manipulate these cues to some degree by using a distinct speaker for each link in the network, in which case, in a network of 4 birds, we would need up to 12 speakers, 3 in each chamber. Accordingly, we would need to calculate up to 12 LMS filters in total, which would mildly increase the complexity of our hardware and software architecture.

Although we digitized only the acoustic communication mode, it is a simple matter to digitize the visual communication channel using cameras and computer screens. Advances in generative modeling of animal imagery (Brock et al., 2018; Goodfellow et al., 2014) could open the door to countless possibilities such as artificial visual societies. In combination, combined audio-visual communication systems could provide a means to play evolutionary games.

Because we make use of a powerful FPGA, additional signal processing is possible to enhance the function of the system. For example, we could add routines for real-time detection of a certain syllable (Pearre et al., 2017) and computation of its pitch. Such processing is required in operant conditioning experiments in which birds adapt the pitch of their syllables (Tumer & Brainard, 2007). In our context, selective pitch estimation would allow us to study the role of pitch and its adaptation in a social context. Even a vocoder could be implemented that shifts the pitch in real-time (Sober & Brainard, 2009), which would allow studying the effect of pitch variability on the receiver bird.

The system as described is laid out for the hearing range of zebra finches. By using different microphones and loudspeakers, the signal range could be expanded. As a result, many species could be studied that vocalize in the ultrasonic range, such as bats (Chaverri et al., 2018), rodents (Heckman et al., 2017), and frogs (Shen & Xu, 2016). In terms of signal processing, the ultrasonic range is more challenging to work with because the sampling rate must be higher. Also conceivable are extensions to underwater environments. For example, interactions among cetaceans could be experimentally examined by keeping animals in separate pools. Such a setup has been proposed as enrichment for captive cetaceans (Law & Kitchener, 2017). The squelch could play an important role in such an application because playback experiments have shown that cetaceans react to even soft noises (Smith, 1965). In the free-range and under-water setting, echo cancelation filters may need to be much longer (because sounds propagate much further in water), which should be well possible with our chosen system architecture.

Last but not least, instead of merely switching a binary connection matrix, the connection links could be more finely manipulated using a gain and a delay, with the result of simulating virtual distances between animals. Because acoustic communication evolved to be useful over large distances and without visual contact, experimental manipulation of virtual distance can be useful (Mouterde et al., 2014; Theunissen et al., 2013). Furthermore, adding noise to the communication would allow exploring the strategies employed by animals to cope with adverse environments. For example, the Lombard effect and its neural underpinnings are still debated (Shen & Xu, 2016). Also, a further important field of research in acoustic communication is the concept of turn-taking (Hoffmann et al., 2019; Pika et al., 2018), which could be dissected in detail using the described system.

Funding

This work was supported by the NCCR "Evolving Language" and by project Grants 31003A_156976/1 and 31003A_182638 from the Swiss National Science Foundation. Dr. Diana Rodrigues was supported by a fellowship from Fundação Ciência e Tecnologia [SFRH/BD/52196/2013] (through GABBA Ph.D. program).

Conclusions and outlook

Together with my colleagues, I have set out to assist birdsong research in its transition from reductionist to holistic and big-data-driven approaches at the social level. In the following, I will integrate our main findings, outline limitations and potential future avenues, and discuss the broader societal impact our work could have.

Towards democratized development of automated vocal detectors and gold-standard vocal datasets

To find our way through big data we need to structure it effectively, tagging important *passages*, and annotating it with meaningful labels. Cost-effective annotation of big data requires machine-based solutions that minimize the time that an expert needs to curate a dataset. To conduct longitudinal studies in songbird research, ideally, we would have access to fully automated detectors of animal vocalizations. However, solutions of this kind, before being applied, typically need to be supervised with large amounts of human labelled examples.

In the case of the zebra finch, our model organism for vocal learning, there is only one public dataset with labelled vocal segments available (Clemens, 2021; Steinfath et al., 2021). It consists of female-directed (highly stereotyped) song from a single male individual (Steinfath et al., 2021). However, our long-term aim has been to study vocal learning in juvenile birds, as well as interactions in groups of animals. For this end, we needed to create our own gold-standard datasets, which we have introduced in this dissertation.

Our first gold-standard dataset ("Chapter 1") consists of over 53'0000 vocalizations recorded from male zebra finches in isolation at different stages of development. Our second gold-standard dataset is recorded from mixed-sex zebra finch couples equipped with animal-borne accelerometers, which we have used to assign over 54'000 vocalizations to the birds that uttered them ("Chapter 2"). Labor-intense generation of these datasets has not only drawn our attention to the fascinating biological intricacy of zebra finch communication, but also prompted us to carefully scrutinize and declare annotation conventions.

The most impactful biological research often integrates observations from diverse species to identify conserved principles, exemplified prominently by the work of Charles Darwin (Darwin, 1859, 1871, 1872). To allow modern statistical analysis of multi-species data, the definition and use of standardized annotation conventions is imperative. However, publicly available zebra finch datasets that contain motifs, single vocalizations, or call sequences without defined vocalization boundaries (Elie & Theunissen, 2020; Goffinet et al., 2021; Pearre, 2017), as well as the one that contains annotated syllables (Clemens, 2021; Steinfath et al., 2021), do not systematically illustrate the decisions taken by the annotating expert.

In our own comparatively large-scale datasets, we have been exposed to limiting cases, where the choice of the label is non-trivial, even for an expert. In fact, we report that without aligned annotation conventions, two experts can produce largely deviating annotations, e.g., by grouping adjacent vocal sounds as one or two vocalizations ("Chapter 1"). In the appendices II and IV, we illustrate our decision boundaries for labelling microphone recordings of single birds and accelerometer recordings in multi-bird experiments. When first inspecting juvenile subsong, I have been personally tempted to take the developmental endpoint – the adult song syllables – as a reference, to find out which sequential subsong notes "belong together". I quickly realized that such a convention is impractical when the endpoint is unknown and could bias analysis by imposing the expert's interpretation. Although today it is regarded as ideal to base segmentations of animal sound recordings on the functional roles of the vocal signals (Kershenbaum et al., 2016; Sainburg

& Gentner, 2021; Suzuki et al., 2006), these roles are often hard to estimate experimentally, are therefore unknown in general, and could additionally depend on the behavioral context. One of our main suggestions is therefore, to base any ground segmentation on the acoustic structure only – separating two vocal sounds into two segments when there is a detectable silence between them. Once this agnostic segmentation is provided, additional (potentially sparse) annotation layers that are based on inferred functions, can be added.

As we plan to make our datasets publicly available upon publication of our work in peer-reviewed journals, we hope to contribute to the democratization of training and testing novel automated detectors of zebra finch vocalizations. Having access to labelled microphone and accelerometer data will allow developers to tailor and benchmark their own solutions for their own purposes. In our group, we[19] currently develop a lightweight network to detect zebra finch vocalizations, using the single-bird microphone dataset introduced in "Chapter 1".

Since it cannot be known how well an automated vocal detector will work on unseen data, such as data obtained from a longitudinal zebra finch study in a novel naturalistic environment, we envision proofreading tools to be an essential component of future songbird research. Here, we have proposed nearest neighbor retrieval for proofreading because it is predestined for controlled out-of-distribution detection of vocalizations that lie "in the neighborhood" of already annotated examples. We tested different commonly used distance measures and found the Spearman distance to perform best. Taking 50 labelled examples, we find that retrieval performance is much worse for juveniles (F1 score of 0.64 ± 0.18) than for adults (F1 score of 0.93 ± 0.07). Juvenile vocal retrieval is moderately improved when searching with equally sized overlapping template slices (F1 score of 0.72 ± 0.10) instead of whole templates. A potential reason for this improvement is that early vocalizations might feature more conserved (useful template slices) and more variable parts (which confuse retrieval with whole templates). Critically reflected, our results imply that proofreading juvenile datasets might be more strenuous compared to adult datasets, since the true-positive candidates will likely have more dispersed nearest neighbor rank values, instead of being aggregated at the top of the nearest neighbor ranking. Additionally, retrieved candidates might still need manual corrections of the segment boundaries[20]. Nevertheless, we expect that our approach can significantly lower the cost to obtain a gold-standard dataset. Our method allows to iteratively train automated detectors, such as available deep neural networks (Cohen et al., 2022; Steinfath et al., 2021), and proofread annotations to capture more complete corpora of vocal data.

How can proofreading be further improved? Two approaches come to mind: either we find a better distance measure, or we improve the data representation. We have noted in "Chapter 1" that an ideal measure would discount for transformations between the template and its candidates, if and only if these occur along axes of natural variation, such as loudness. In other words, a much louder version of a call should still be detected, although the Euclidean distance is relatively large. The Spearman distance, the best distance in our tests, indeed discounts changes in loudness because it compares ranks of features, instead of their absolute values. For the second approach, improving the representation of the data, I have sketched a

[19] Co-authored with joint first authors Xinyu Hao and Kanghwi Lee, as well as Linus Rüttimann, Aoxue Miao, Nianlong Gu, Dr. Vivi Nastase, and Prof. Dr. Richard Hahnloser

[20] This could be improved for example by post-processing the segment durations using dynamic time warping.

proposal together with my colleague Yingqiang Gao[21]: we have proposed to enrich the single channel vocal data with information from multimodal data streams (e.g., data recorded in the experiments of "Chapter 2"), by learning a shared embedding with transformer architectures (Dosovitskiy et al., 2020; Vaswani et al., 2017; Yu et al., 2020). This revolutionary method has been developed for natural language processing, where text consisting of recurring discrete words is the input data. The models learn an embedding – a space where similar units of input sequences (such as synonymous words from textual data), when be mapped into (being represented as dense vectors), will lie closely together. In our case, we want to annotate two-dimensional spectrograms and video frames. Interestingly, the initial transformer architecture has recently been adapted for image captioning using multi-view image input (Yu et al., 2020). This work is the closest to our need that I could find. One advantage of transformers is that they can be pre-trained in self-supervised manner, e.g., by predicting masked data from the context, assuming that the input sequence is non-random and features recur in similar contexts (Dosovitskiy et al., 2020; Vaswani et al., 2017). This is especially useful when annotated (or "captioned") data is scarce. One disadvantage of this approach is its large computational cost, which could be mitigated using reduced dimensionality of input features. It would be interesting to adapt these approaches for multimodal behavioral captioning. The resulting embeddings could be used, for example, to screen for multimodal nearest neighbors.

Towards live monitoring and prediction of reproductive behaviors

Charles Darwin deceased 140 years, but his theory of evolution has prevailed. It states that naturally occurring variants in individual traits are preserved if beneficial either for survival or attraction of sexual partners (Darwin, 1859, 1871). Successful copulations are a primary way to pass such traits onto the next generation in mammals and birds[22]. Mediating sexual reproduction, these key events therefore decide the fate of entire populations. In our work presented in "Chapter 2", we found vocal and non-vocal behavioral signatures that signal copulations of mixed-sex zebra finch couples 25-30 seconds in advance. These signatures include elevated singing and calling rates, frequent non-vocal behaviors such as beak wipes and approaches, changed song composition and tempo as expected for female-directed singing (Sossinka & Böhner, 1980), as well as elevated call durations.

Detecting copulations based on such signatures has multiple applications. Firstly, it could inform wildlife management and conservation (Buxton & Jones, 2012; Digby et al., 2013; Lewis et al., 2021; Marques et al., 2013; Stowell et al., 2016). With detailed knowledge of vocal and reproductive activity, such programs could better protect endangered populations, for instance, by intervening when detection rates drop acutely due to natural or human-caused perturbations. Secondly, reproductive activity has been shown to decrease with stress-inducing corticosterone treatment (Scalera & Tomaszycki, 2018), and courtship behavior could therefore contribute to monitor stress levels of animals in captivity, noninvasively.

Our approach opens several avenues for technical improvement. Our copulation detection method, developed primarily by Linus Rüttimann, leverages the animal-borne accelerometers that we attach to the birds to distinguish their vocalizations. In contrast to our human approach, it has been shown that zebra finches can identify each other based solely on vocal signatures (Elie & Theunissen, 2018). In principle, with the improvement of our recording devices and computational tools, it could be that we too can assign

[21] During the stimulating course "My thesis and beyond: Developing an interdisciplinary research idea", taught by Dr. Elizabeth Amadei, with inputs from Dr. Mariana da Rocha and Dr. Prof. Richard Hahnloser.

[22] They mediate genetic or epigenetic inheritance, but complex interactions with environmental factors exist (e.g., birdsong is an excellent example for cultural transmission of traits).

vocal behaviors without these sensors in the future, using little or no training data for a given individual. It might be possible to then detect copulations purely from remote audio and video recordings. This would be attractive for any real-world use case of our method. Another avenue is not only the monitoring, but the prediction of (successful or unsuccessful) copulation attempts, and the investigation of the factors that promote them.

Prediction of behavior is a powerful capability. When automated behavioral detectors are fast and accurate, they allow for real-time interventions, which can manipulate behavioral outcomes, such as the choice of mating partners. In a broader societal context, behavioral surveillance and prediction are hot topics (Liang et al., 2018; Richards, 2013; Zuboff, 2015). Such technology can be used in the interest of an individual, or against it. I therefore encourage todays and future generations to use such technology moderately, transparently, and wisely. I have exemplified use cases that I judge beneficial for our society, such as reducing harmful perturbations to endangered species, by monitoring their vocal expressions.

Towards understanding the structure, development, and function of animal vocal expressions

One of the few advantages of curation-intense generation of gold-standard biological datasets is in-depth exposure to their richness in structural complexity and variation. For anyone who will use our annotations to compute their own summary statistics, I recommend to first get a grasp of this richness (some of which I have tried to capture in appendices II and IV) by visual or auditory inspection. My own inspection has led us to report novel structural findings on the zebra finch repertoire (“Chapter 1”and “Chapter 2”), for example, that two “nest” calls can get arbitrarily close and thereby transition to “tet” calls (“Appendix IV”). To categorize calls in “Chapter 2”, we consulted two-dimensional spectrogram embeddings. We have found that “nest” and “whine” calls populated the same cluster in all birds, exhibiting structural gradedness on a “nest/whine” continuum.

This vocal categorization could be improved in several ways. Firstly, we only used accelerometer data to save time, and have consulted microphone recordings only in exceptional cases, e.g., when high syllable notes were poorly visible on the accelerometers. Using a multimodal approach could reduce mislabeling, which might occur, for example, when accelerometers have perturbed skin contact due to body movements (Rüttimann et al., 2022). Secondly, methods such as fuzzy clustering (Cusano et al., 2021), or the Cuzick-Edwards test statistic (Cuzick & Edwards, 1990) could be used to quantify separation between vocal categories. Lastly, although we advocate to always first use structural properties for baseline annotation (appendices II and IV), functional assessments can increase our confidence in categorical labeling.

Indeed, we have found that our structural categorization is mirrored in functional differences: while “tet” calling rates are only elevated prior to copulations, “nest/whine” rates are elevated more symmetrically around copulations. Both of these shared call types have elevated durations prior to copulations (“tets”), or around copulations (“nest/whines”)[23]. The “nest/whine” durations covary significantly between sexes only around copulations, but not in control episodes. A difficult open question remains: What is the exact function of these call types?

I have introduced the distinction of innate affective expressions of internal states and complex learned vocalizations earlier (see “Introduction”). The zebra finch exhibits both traits, with birdsong being a well-

[23] Note that the long calls of the “nest/whine” continuum correspond to the “whine”-extremum (“Appendix IV”).

known culturally learned sexual display, used “to impress”. Much less systematic knowledge exists on call functions. It has been shown by others that zebra finches can vocally induce mirrored physiological states in their mates (Perez et al., 2015). Together with the “nest/whine” call duration covariance that we observe around copulations, this could indicate that states such as sexual arousal are expressed and synchronized by vocal means, if it is “the right time” to engage in reproductive behaviors. However, such hypotheses need careful testing.

To probe causal relationships, one needs to manipulate a system, showing loss-of-function and gain-of-function dependent on the controlled presence of the putative cause. The control of atomic social interactions in freely behaving animals is currently not feasible, to our judgement. To probe the causality in vocal interactions without physical contact, we have developed a system to manipulate these interactions among separately housed, but digitally connected animals (Rychen et al., 2021), discussed in “Chapter 3”. Compared to playback experiments of the past (Böhner, 1983; Burt et al., 2001, 2007; Evans & Marler, 1991; James et al., 2019; Ljubičić et al., 2016; Perez et al., 2015), our system allows perturbations of atomic interactions within multi-animal networks with otherwise naturalistic vocal exchanges. It can be therefore used to evaluate hypotheses, such as the one introduced above, in greater depth: Is it sufficient to exchange naturally occurring female “nest/whine” calls with longer versions, to induce a physiological response in certain out of several conspecifics? These conspecifics could be multiple competing males of varying hierarchical rank, or males with different mating partners. Extrapolating current technological developments (Ausra et al., 2021; Biegler et al., 2022; Roberts et al., 2012; Zhao et al., 2019), our system could be complemented in the future with minimally invasive methods to selectively control specific perceptual or behavioral aspects of an otherwise freely behaving individual.

A particularly interesting question raised by our work and approachable with perturbation experiments is whether the renditions of female behaviors prior to copulations could be used as positive reinforcers for adult vocal learning, which has been mainly studied with aversive stimuli (Andalman & Fee, 2009; Charlesworth et al., 2011, 2012; Tian & Brainard, 2017; Tumer & Brainard, 2007; Warren et al., 2011). More generally, one could test the existence of causal relationships between changes in male birdsong composition and female courtship displays prior to copulations.

A typical question when transitioning from reductionist to holistic system descriptions is the following: how is information encoded (reductionist part) and which global patterns emerge from local interactions (holistic part)? While I have emphasized the transition of birdsong research to holistic descriptions[24], this does not imply that the reductionist part is solved. In principle, vocal information can be encoded in many ways. In “Chapter 2” we examined behavioral rates, features of single vocalizations, or song composition. As for single vocalizations, a classical distinction is their classification in discrete and graded signals (Marler, 1967). In my opinion, it could be that there is no clear dichotomy between such vocalizations. As in human language (Scherer et al., 2003; Scherer, 1995), it could be that a vocalization has an explicit meaning that could be discrete (or graded, in principle), but multiple implicit messages, such as information about the discrete sender identity (Elie & Theunissen, 2018) or graded internal states (Perez et al., 2012, 2015). Extrapolating this idea, call duration modulation around copulations could be an implicit message, while an unknown explicit message would contribute to the differential shape of the CRF curves of “tet”

[24] In my opinion, neuroscience research in general is moving slowly in this direction at the level of neuronal populations and cortical networks.

and “nest/whine” calls. This is another hypothesis that would need to be tested with experimental perturbations. Finally, our datasets are very rich, and consequently we could not exhaustively mine for more possible patterns, such as stereotyped complex behavioral sequences of a single individual or bird pairs[25].

The zebra finch has gained its scientific impact primarily for being a model organism for vocal learning. Our research will hopefully contribute to shed light onto how this small and highly social bird learns complex songs from its conspecifics – with ease and efficiency. It has been shown that, additionally to tutor song exposure, juvenile learning can be guided by non-vocal feedback signals from females (Carouso-Peck & Goldstein, 2019) and affected by interactions among juvenile birds (Tchernichovski & Nottebohm, 1998; Volman & Khanna, 1995). Diverse social interactions are therefore expected to shape vocal learning, but the mechanisms governing their interplay remain uncovered. I envision that our efforts will help to understand learning as a phenomenon that emerges within a social system from underlying atomic social interactions, which shape neural substrates.

[25] This idea has been formulated together with the former Master student Roman Doronin.

Appendix I – Readme on central MATLAB scripts used in this dissertation

Scripts for "Chapter 1"

Path to parent folder[26]: cmatlab\DivPrograms\Individual\TomasTomka\bruteforce

`TT_NNsearchBruteForceWholeMultiDist.m`

Description: This script is testing the WHOLE approach across multiple template set sizes, multiple gold-standard data samples, multiple distance measures, multiple distance normalization strategies, and multiple replicates. The data is loaded from flatclust[27] "Archives" called `Flat`. These are the main steps of the script:

1. For each replicate, templates are extracted from `Flat.X`, and the GS labels of the remaining search space are stored in `labels_true`. The variables `LIndices` and `LIdx` are initialized and later used to map enumerated candidates to the correct label indices[28].
2. The search space is further reduced by excluding silent periods based on the "stencil" information stored in `Flat.Z`.
3. Distances between templates and candidates are computed (using the in-built function pdist2 or custom distance measures). The shortest template has the largest number of potential candidates, and thus defines the dimension of the matrix D with elements $D_{i,j}$ representing distances between the i-th template and potential candidates with onsets at position j in the search space. Longer templates fit less often in the search space, and therefore have fewer potential candidates, which is why D is sparsely populated, in general.
4. After distance computation and normalization, candidates with minimal distance to their template are retrieved iteratively, while avoiding overlaps or immediate adjacency with previously retrieved vocalizations.
5. After retrieval has terminated (the sum of the template set size and retrieved set size equals the total number of GS segments), the performance is evaluated[29]. The results are stored in the output structure `BFS`.

[26] Cmatlab is our lab's current code repository.

[27] Flatclust is a customized MATLAB software that has been developed in the Hahnloser lab. It has been documented elsewhere and therefore its architecture is not detailed here.

[28] This part of the code can be excluded in future applications when no GS data is available, and retrieval is applied to unlabeled data.

[29] Footnote 30 applies here too.

Input parameters:

Parameter name	*Description*
`ns`	Array of template set sizes to be iterated over.
`Nreps`	Number of retrieval replicates.
`archives_name`	Name of the flatclust "Archive collection" (collection of annotated spectrogram-sets).
`distNs`	Cell array of distance measures (specified as strings) to be iterated over.
`norms`	Cell array of normalization strategies (strings) to be iterated over (for each computed distance matrix).
`birds`	Cell array of structure arrays (one per bird) that specify which data to use. The entry `birds{i}.name` contains the name of the i-th bird folder, and `birds{i}.arch` contains a list of its flatclust archives, to be taken from the specified `archives_name`.
`freqs`	Array of frequency bins (enumerated along the y-axis of a spectrogram), which are to be considered for distance computation (default: [12, 13, …, 128], excluding low frequency noises).
`store_spec`	Boolean that controls whether extracted vocalization spectrograms are stored (1) or not (0).
`nscore`	Number of times that the performance is evaluated during the retrieval progression.
`tolerance`	Temporal tolerance (in spectrogram bins) of the VocScore.
`newrun`	Boolean specifying whether an old run is continued (0), for example when adding a new bird to the BFS structure, or a new run is initiated (1).

Output: The structure `BFS` stores the main parameters (for reproducibility) in separate fields, as well as the results in the six-dimensional cell array `BFS.Z`. These six dimensions are: the bird identity, the archive identity, the distance measure, the normalization strategy, the number of templates, and the replicate identity. Each entry `X = BFS.Z{a,b,c,d,e,f}`, with any iterated indices `a-f`, is a structure with these fields:

Field name	*Description*
`T`	Cell array of template spectrograms.
`VVrTi`	Array of template identities responsible for the retrieval of a given column (appending values as retrieval progresses).
`VVr`	Array of true labels of retrieved columns (appending values as retrieval progresses).
`VVrIdx`	Array of column indices of retrieved columns (within their files; appending values as retrieval progresses).
`VVrDAT`	Array of file identifiers of retrieved columns (appending values as retrieval progresses).
`Dr`	Array of distance values of the candidate-template pair that lead to the retrieval of a given column (appending values as retrieval progresses).
`SyllScore`	Array with the third row containing the VocScore values as retrieval progresses (last value taken as the final performance). The first two rows store the corresponding precision and recall values.
`F1Score`	Array with the third row containing the F1 score values as retrieval progresses (last value taken as the final performance). The first two rows store the corresponding precision and recall values.
`cost`	Time cost for computing the distance matrix D.

TT_NNsearchBruteForcePartMultiDist.m

Description: This script is testing the PART approach across multiple template set sizes, multiple gold-standard data samples, multiple distance measures, multiple template slice widths, and multiple replicates. The script works similarly to TT_NNsearchBruteForceWholeMultiDist.m with these adaptations:

1. Per default, templates which are shorter than the specified slice width are zero-padded (at their end).
2. The distance computation is simplified since all template slices are of equal size, resulting in a densely populated distance matrix D.
3. The retrieval procedure is more elaborate, since the original template duration needs to be considered when retrieving a vocalization based on slice distances. When this original duration protrudes out of the search space (into a silent period), we crop the retrieved candidate using the "stencil" information.
4. Again, candidates are retrieved iteratively, while avoiding overlaps or immediate adjacency with previously retrieved vocalizations, per default. However, we added an option "elongate", to allow new candidate slices to extend previously retrieved vocalizations. Since this option did not result in higher performances, we excluded it from further research.

Input parameters: Same as for TT_NNsearchBruteForceWholeMultiDist.m, except the norms parameter is removed, and these additional parameters need to be specified:

Parameter name	*Description*
Ws	Template slice width (in spectrogram columns).
zeropad	Boolean that specifies whether short templates are zero-padded (1) or not (0).
elongate	Boolean that specifies whether retrieved vocalizations can be subsequently elongated (1) or not (0).

Output: The structure BFS stores the main parameters (for reproducibility) in separate fields, and the results in a seven-dimensional cell array BFS.Z. These seven dimensions are: the bird identity, the archive identity, the elongation option, the distance measure, the template slice width, the number of templates, and the replicate identity. The fields of BFS.Z are the same as in TT_NNsearchBruteForceWholeMultiDist.m.

Peripheral scripts

To generate the figures of "Chapter 1", I have used these additional scripts:

Figure	*Scripts*
1.1	`TT_BFS_introfig.m`
1.2	`TT_BFS_introfig_part.m`
1.3	`TT_BFS_cumsumvoc0.m` (a) `TT_BFS_heatmaps0.m` (b) `TT_BFS_heatmaps.m` (c) `TT_BFS_heatmaps_normalization.m` (d-e) `TT_BFS_score_relationship.m` (f) `TT_BFS_sensitivity.m` (g)
1.4	`TT_BFS_goodandbad.m` `TT_BFS_goodandbad_plot.m`

Additionally, the live-script `inspect_annotated_spectrograms_isobird.mlx` shows a light-weight way of inspecting annotations of an exemplary file on the raw audio signal, as well as spectrograms. Please note that we have used `imagesc` for plotting and annotation, which places the axes origin at (0.5,0.5). When using different languages or plotting functions, the location of the labels might need to be adjusted for this fact.

Scripts for "Chapter 2"

Path to parent folder: cmatlab\DivPrograms\Individual\TomasTomka\backpack

`TT_extract_bp_experiment_from_flatclust.m`

Description: For a given mixed-sex experiment, this script extracts annotated events from flatclust "Archives" (vocal data), and video data annotated by Dr. Marianna da Rocha. The output structure `Exp` is a compact representation of the behavioral data and is used to compute more derived statistics.

Input parameters:

Parameter name	*Description*
`in`	Path to the experimental data (previously "bird folder").
`boris_cop`	Path to video-based annotations of copulation episodes.
`boris_ctrl`	Path to video-based annotations of control episodes.
`bird_channels`	Integer value array specifying the channel numbers that correspond to backpacks.
`is_old_bp`	Boolean specifying whether the experiment has been recorded in the old (1) or new "birdpark" (0).
`from_scratch`	Boolean specifying whether the `Exp` variable should be generated from scratch (1) or adapted from previous runs (0).
`get_boris`	Boolean that controls whether video-based annotations should get loaded.

Output: The structure Exp stores experimental details in these fields:

Field name	*Description*
ID	Experiment instance name (previously “bird name”).
Archives_folder	The name of the archives folder.
Archives_list	The names of the archive(s).
bird_list	The identities of the birds that participated in the experiment.
parameters	Structure containing recording and spectrogram parameters.
sample_info	General information about the samples taken in this experiment.
bird_info	General information about the participating birds.
cluster_names	List of annotated vocal clusters.
state_event_names	List of video-annotated state events.
point_event_names	List of video-annotated point events.
samples	Cell array containing data for each sample.
bird_umap	Cell array containing UMAP coordinates of the vocalizations of each bird.

TT_compute_Exp_readouts2.m

Description: This script computes rate curves for vocal and non-vocal behaviors, for each sample of every specified experiment.

Input parameters:

Parameter name	*Description*
Exps	List of stored Exp variable identifiers to be iterated over.
clusters_f	Array of female vocal cluster numbers to iterate over.
clusters_m	Array of male vocal cluster numbers to iterate over.
secs_per_sample	Originally defined episode duration (pad shorter episodes with NaN values).
tref	Originally defined time of the reference event within an episode in seconds (align short episodes, such that the reference event is located at this time).
plot_samples	Resolution of the rate curves (number of data points per curve).
sr_nonvoc	Scan rate used for non-vocal annotations.
filterwin	Gaussian filter window in seconds.
filtersigma	Standard deviation of Gaussian filter in seconds.

Output: The structure Exp is extended with:

Field name	*Description*
Exp.readouts.nonvocal_rates	Cell array with a structure per bird storing its non-vocal rates.
Exp.readouts.vocal_rates	Cell array with a structure per bird storing its vocal rates.

TT_copul_annotate_motifs_and_bouts.m

Description: This script is automatically annotating male song motifs and bouts.

Input parameters:

Parameter name	*Description*
`Exps`	List of stored `Exp` variable identifiers to be iterated over.
`do_plot`	Option to generate some plots for a sanity check.
`th_bout`	Silence duration threshold used to assign motifs into bouts.

Output: The structure of the n-th sample `Exp.sample{n}` is extended with:

Field name	*Description*
`Exp.samples{n}.motifs`	Structure storing motif annotations.
`Exp.samples{n}.bouts`	Structure storing bout annotations.

`TT_assign_dir_undir.m`

Description: This script annotates male song motifs and bouts as directed or undirected.

Input parameters:

Parameter name	*Description*
`Exps`	List of stored `Exp` variable identifiers to be iterated over.

Output: Stored labels of motifs and bouts in the `Exp` structure as directed or undirected.

Peripheral scripts

To generate the figures of "Chapter 2", I have used these additional scripts:

Figure	*Scripts*
2.1	`TT_plot_copul_vocbehav_example.m` (a) `TT_BackpackVocPoster` (b) `TT_plot_copul_videoframes_example` (c) `TT_plot_voc_rate_curves.m` (d)
2.2	`TT_copul_plot_motif_length.m` (a) `TT_copul_plot_inotes.m` (b) `TT_copul_plot_motifs_per_bout.m` (c) `TT_copul_plot_nest_durs.m` (d-e) `TT_copul_plot_calldur_corr.m` (f)

Additionally, the live-script `inspect_annotated_spectrograms_copul.mlx` shows a light-weight way of inspecting annotations of an exemplary file on the raw audio signal, as well as spectrograms. Please note that we have used `imagesc` for spectrogram plotting and annotation, which places the axes origin at (0.5,0.5). When using different languages or plotting functions, the location of the labels might need to be adjusted for this fact.

Scripts for "Chapter 3"

Path to parent folder: cmatlab\DivPrograms\Individual\TomasTomka\skype

`TT_vocal_xcorr_full_newexp2.m`[30]

Description: This function computes the cross-covariance (CCV) function of the mean-subtracted onset trains of two sets of vocal annotations, as well CCV functions of randomly shuffled data. It uses the in-built function `xcorr`.

Input arguments:

Argument name	*Description*
`F1`	Flat variable containing the first annotated set.
`F2`	Flat variable containing the second annotated set.
`X1_elems`	Element (annotated segment) identifiers of the first set.
`X2_elems`	Element (annotated segment) identifiers of the second set.
`maxlag`	Maximal lag in audio samples to compute the cross-covariance for (argument for `xcorr`).
`norm`	Normalization option (argument for `xcorr`).
`test`	Choice of the shuffling method for onset trains: 'shift_bout_X2' (default; circular shift within each bout of the second set), 'shift_X2' (circular shift of the second set across the analyzed period), 'shuffle_X2' (random permutation of the second set).
`nsample`	Number of shuffled replicates.
`total_time`	Total time of the analyzed experimental period.
`start`	Time of the day in hours when the analyzed period starts.

Output variables:

Variable name	*Description*
`CC`	CCV function of the original annotated sets.
`CC_test`	Two-dimensional array of CCV functions of shuffled data (replicates are stored in separate columns).

`TT_skype_xcorr_full_opencomm_newexp2.m`[31]

Description: This script extracts specified elements of any two annotated sets within a communication network, uses `TT_vocal_xcorr_full_newexp2.m` to compute the original and shuffled CCV functions, applies a Gaussian filter, and plots the results. It requires that the annotated sets of the network are loaded in the `Flats` variable.

[30] I have used other variants of this function; this version is the most generic one.

[31] I have used variants of this script for different experiments (using different channels and tag conventions); this one is the most recent example.

Input parameters:

Parameter name	*Description*
`sessions`	Identifier of the experimental session (which network has been imposed).
`vocal1`	List of annotation tags (strings) to be looped over for the first set.
`vocal2`	List of annotation tags (strings) to be looped over for the second set.
`clust`	Structure that specifies which clusters to consider for the analysis (per annotation tag).
`maxlag_sec`	Maximal lag in seconds to compute the cross-covariance for (argument for `xcorr`).
`norm`	Normalization option (argument for `xcorr`).
`test`	Choice of the shuffling method for onset trains: 'shift_bout_X2' (default; circular shift within each bout of the second set), 'shift_X2' (circular shift of the second set across the analyzed period), 'shuffle_X2' (random permutation of the second set).
`nsample`	Number of shuffled replicates.
`win`	Gaussian filter window in audio samples.
`sigma`	Standard deviation of Gaussian filter in audio samples.
`day`	Tag that specifies the analyzed day.

Output: Panels as shown in Figure 3.3a-b (on the left).

`TT_skype_summary_xcorr.m`

Description: This script produces Figure 3.3c, based on previously computed CCV functions.

Input parameters:

Parameter name	*Description*
`maxlag_sec`	Maximal lag in seconds to compute the cross-covariance for (argument for `xcorr`).
`sr`	Audio scan rate.
`win`	Gaussian filter window in audio samples.
`sigma`	Standard deviation of Gaussian filter in audio samples.

Output: Figure 3.3c.

Appendix II – Protocol for manually annotating microphone recordings of single birds (“Chapter 1”)

Vocal signals tend to arise from distinct acoustic units, which is a characteristic shared across the polymorphic landscape of vocalizing species (Hauser et al., 2002; Kershenbaum et al., 2016). Animal studies in monkeys, dogs, chicken, and songbirds have shown that animal calls can be used to communicate semantic meaningful information such as detection of predators, discovery of food, or attraction of mates (Dittus, 1984; Fischer, 1998; Gill & Bierema, 2013; Gouzoules et al., 1984; Hauser, 1998; Marler et al., 1986a, 1986b; Seyfarth et al., 1980; Slobodchikoff et al., 1991; Suzuki, 2016; Zuberbühler et al., 1999). Nevertheless, the functions of animal vocalizations are generally unknown for most calls and species (Kershenbaum et al., 2016; Sainburg & Gentner, 2021). To advance our understanding of vocal communication in animals, we need to study large and well-annotated datasets. Here we address the problem of how to segment audio recordings of a given species. The segmentation problem is to distinguish the times at which an animal vocalizes from the times at which it does not.

One of the simplest methods of segmenting vocalizations from continuous recordings is to consider sound amplitude and to define as vocalizations all sounds that are above a given threshold. However, this procedure will misclassify certain noises as vocalizations, which is why more refined approaches are needed that potentially make use of the statistics of the individual (Tchernichovski et al., 2000). In the extreme case, we need to inspect every single potential vocalization and decide based on expert knowledge where to cut the dividing line between vocalization and noise.

To standardize the segmentation task, we have created this set of guidelines based on two decisions boundaries for a vocalization:

a) The decision whether there is a silent period between two sounds, which we take by inspecting spectrograms (Figure AII.1, left).
b) The decision whether a sound is vocal or non-vocal (Figure AII.1, right; Figure AII.2-AII.3).

Birds, especially when young, tend to vary the gaps between vocalizations. An example is shown in Figure 1 (yellow dotted box): This sequence of three vocal elements looks like a precursor of syllable C that the juvenile tries to imitate, but they appear with sufficiently large gaps, which is why we sometimes classify them as 3 distinct syllables. Thus, for a) we infer a gap where we can visually detect one, irrespective of other singing attempts in the animal.

The second decision boundary (b) is harder to define universally from single-microphone recordings; ideally, we would like to have simultaneous recordings from the trachea to measure sounds and air flow there. In practice, it is a human expert, who judges whether a sound is vocal or non-vocal by listening to examples and inspecting the corresponding spectrograms. Again, this task is relatively simple for highly stereotyped vocalizations, but more difficult for faint, short and variable vocalizations in juveniles (Figure AII.1, right; Figure AII.1, left, Figure AII.3).

A special case consists of faint sounds (usually at around 6 kHz) that frequently occur after (or, less frequently, before) vocalizations (Figure AII.2, left). We consider them to be inhalation sounds (Goller & Daley, 2001; Riede et al., 2013; Tchernichovski et al., 2000) and exclude them from the vocal dataset (default setting).

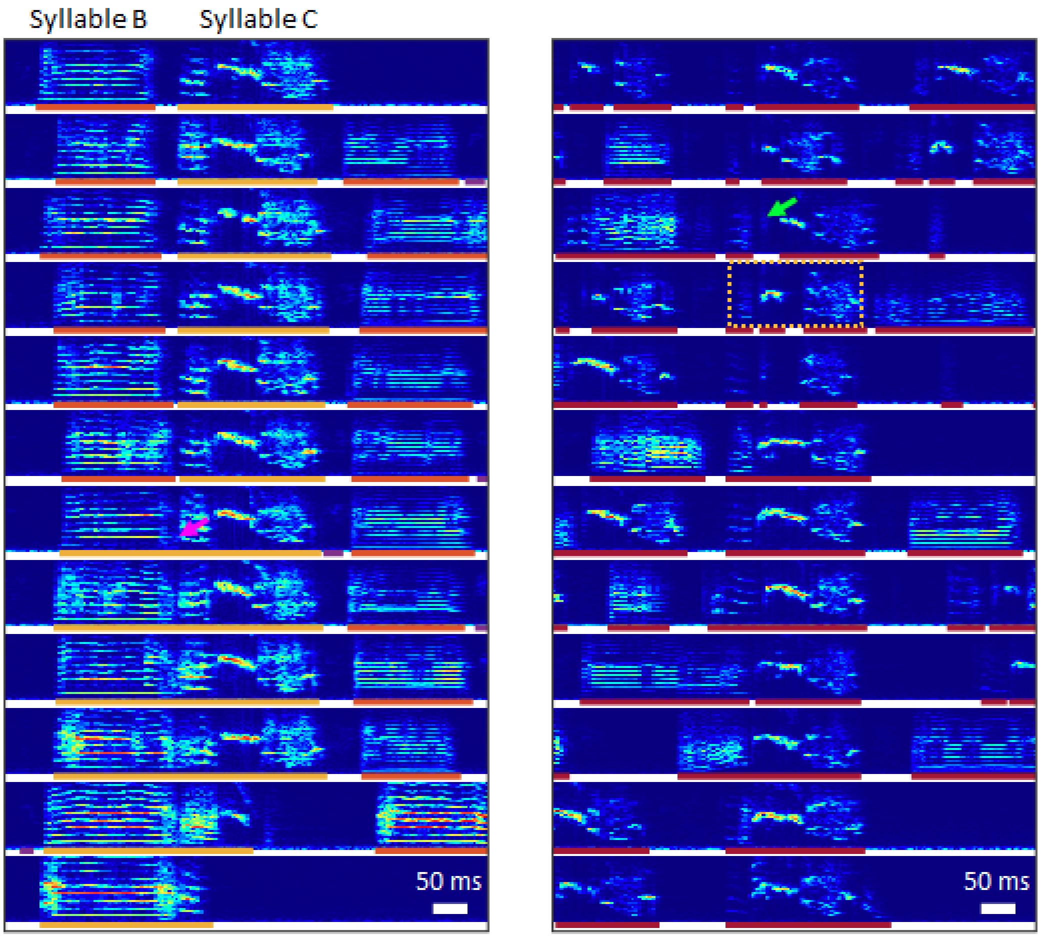


Figure AII.1: Definition of vocal segments as continuous intervals of vocal activity. (left) Zebra finch song examples at 59 day-post-hatch, aligned to notes that resemble the beginning of syllable C. At this stage, syllable C is surrounded by clear gaps most of the time (top 6 examples). However, in a minority of cases, no silent gap is visible between the preceding syllable B and the first note of syllable C (bottom 6 examples, boundary case indicated with magenta arrow). Gold-standard segmentation labels of syllable-C-notes (yellow) and of other vocalizations (orange, purple) are indicated by bars below the spectrograms. **(right)** Vocalizations recorded at 49 day-post-hatch (red bars), aligned to examples that resemble syllable C. Short noisy sounds within syllable precursors (green arrow) have not been classified as vocal activity based on isolated visual inspection, but likely would be, if the context would be taken into account. The yellow dotted box marks three vocal elements that could potentially be interpreted as a unitary precursor of syllable C, if the developmental endpoint were to be taken into account. Bars as on the left.

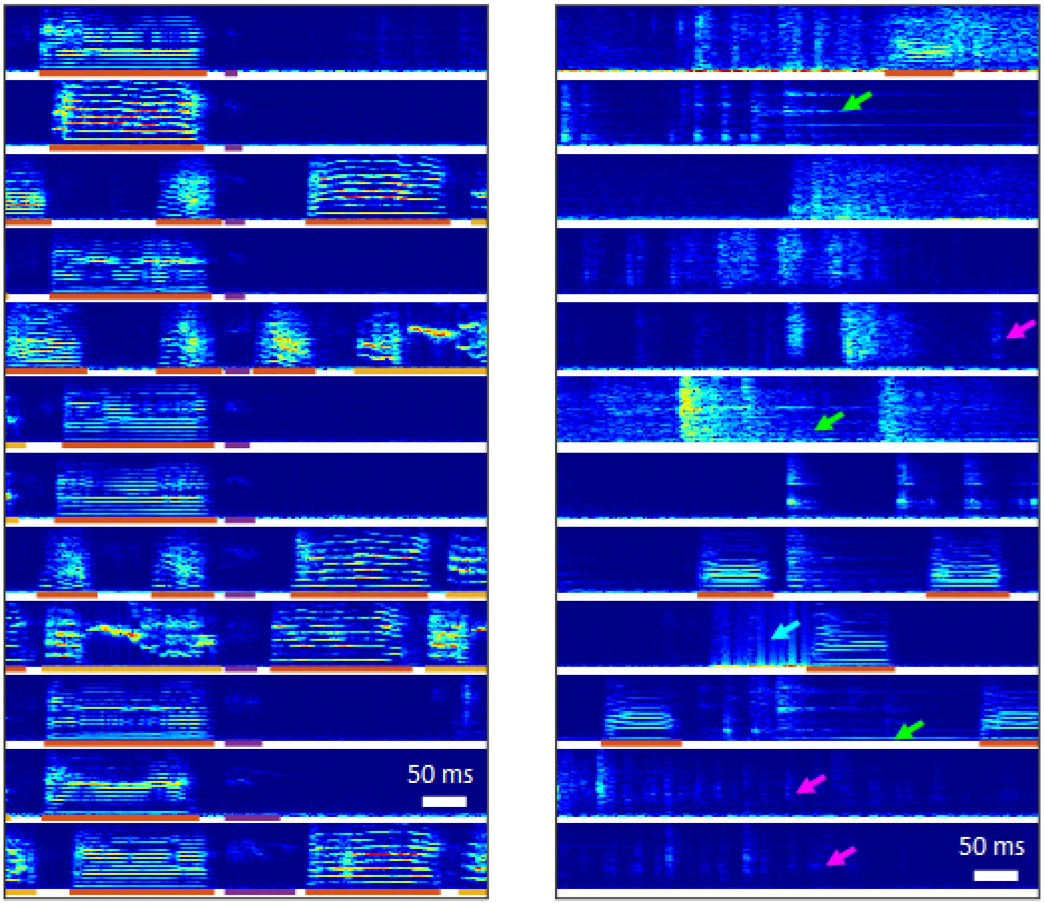


Figure AII.2: Decision-boundary between vocal and non-vocal sounds. **(left)** Spectrogram examples of putative inhalation sounds (indicated with purple bars) observed in a zebra finch at 59 day-post-hatch (excluded in the gold standard by default). **(right)** Examples of non-vocal noises which may include prominent tones (green arrows), wide-band noise (blue arrows), or very faint signals (magenta arrows).

The examples we provided illustrate our decision boundaries and the difficulties with segmentation approaches. In summary, we advocate the definition of vocal segments as tightly restricted intervals of continuous vocal activity. These segments should be defined independently from functional considerations. How to extract functional units from vocal segments is an open question, the answer may depend on whether the vocal units are assessed in the domain of perception (receiver) or production (sender). Still, it is regarded as ideal to validate chosen segmentations based on the functional roles of the vocal signals (Kershenbaum et al., 2016; Sainburg & Gentner, 2021; Suzuki et al., 2006). However, recent work in songbirds suggests that "syllables may not be perceptual units for songbirds as opposed to common assumption" (Mizuhara & Okanoya, 2020).

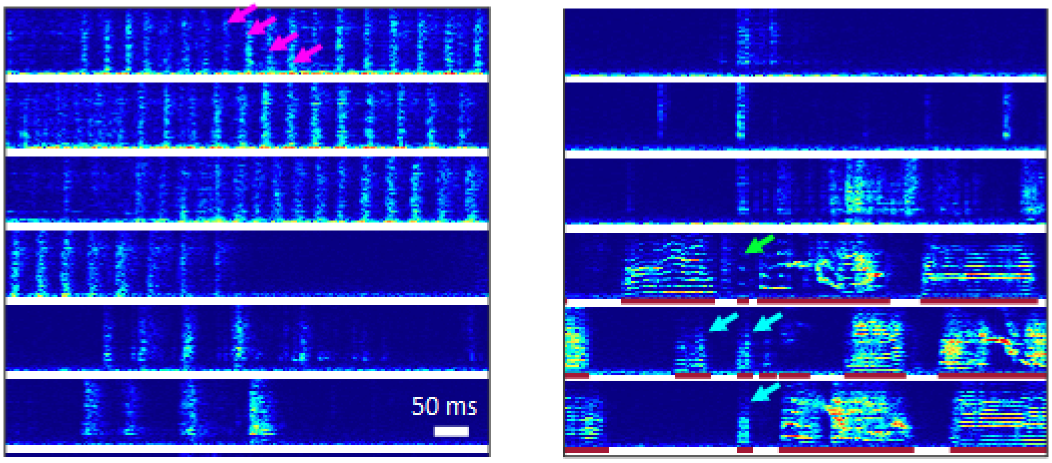


Figure AII.3: Detailed decision-boundary between vocal sounds and wing flaps. Spectrogram examples short noises. Wing flaps are easy to detect on spectrograms when occurring in serial repetition (i.e., when the bird is flying; magenta arrows). For short sounds, indicators of vocal activity can be harmonics (green arrow) or a strong skew in the spectral density towards certain frequencies (low frequency sounds indicated with blue arrows).

Appendix III – Extended set of retrieval analyses (“Chapter 1”)

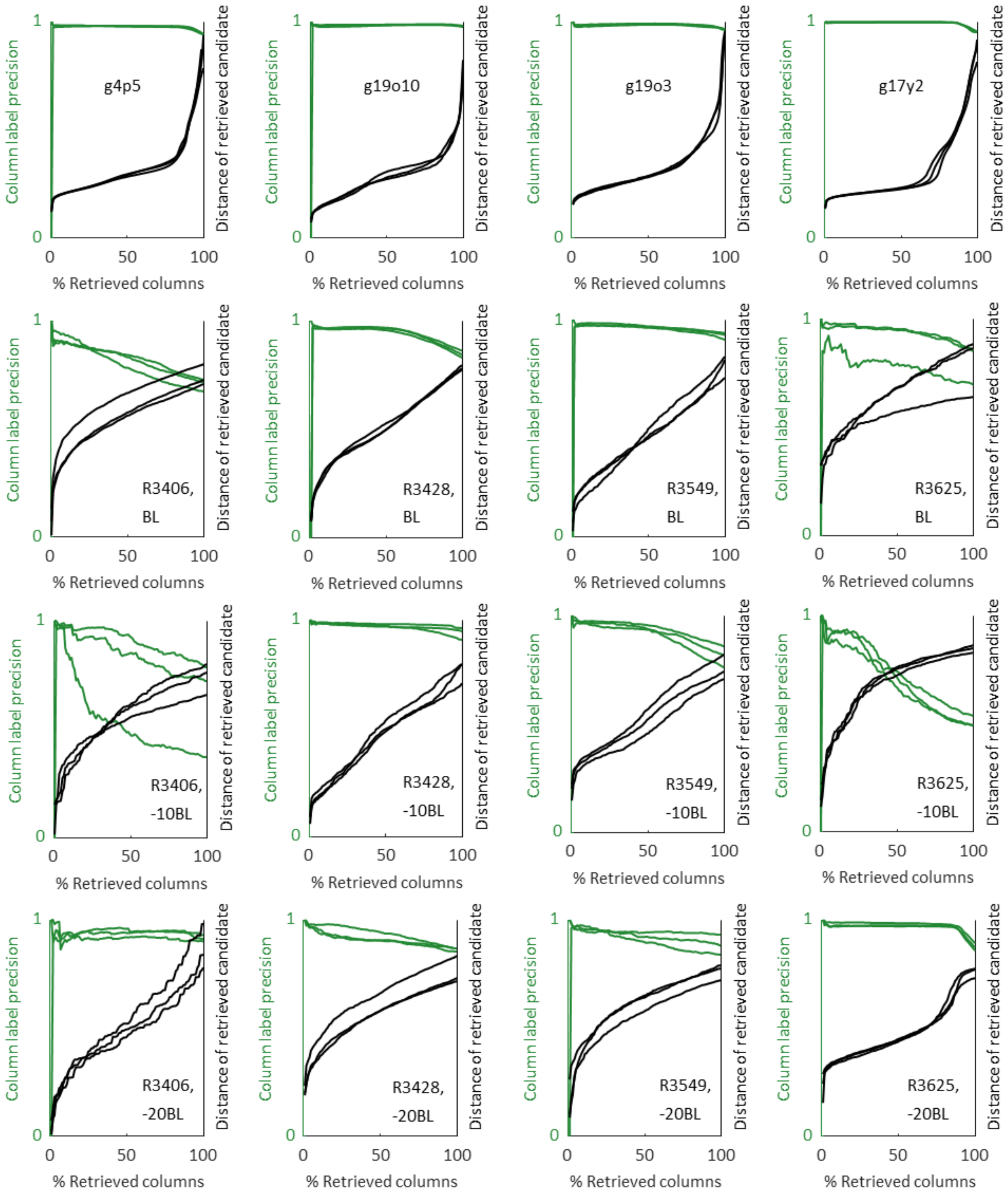


Figure AIII.1: Extended set of precision and distance curves as a function of retrieval progression, using the WHOLE approach with the Spearman distance and 50 templates (replicated for all birds). The top row shows adult birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure 1.3a for a detailed description.

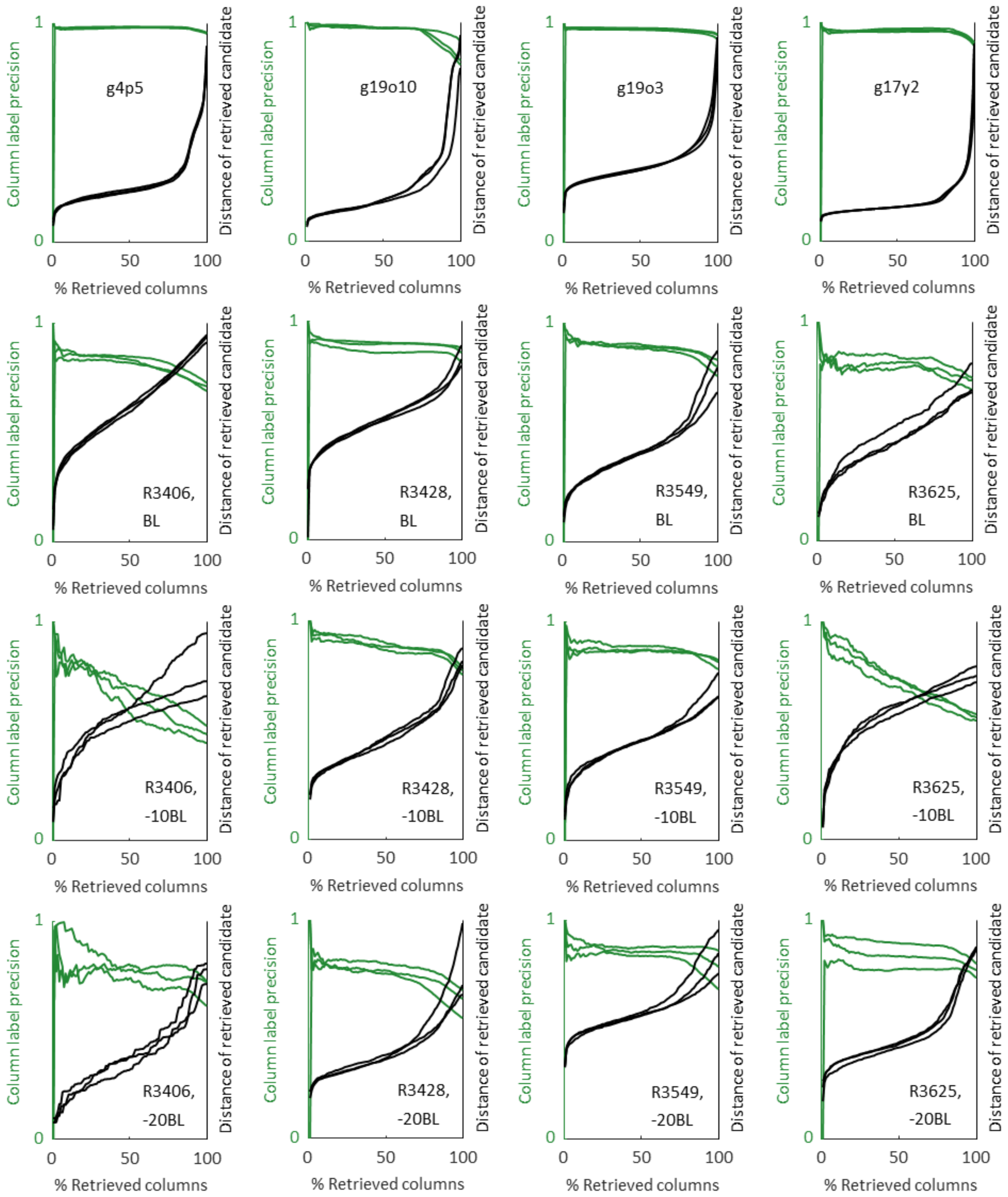


Figure AIII.2: Extended set of precision and distance curves as a function of retrieval progression, using the PART approach with the Spearman distance and 50 templates (replicated for all birds). The top row shows adult birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure 1.3a for a detailed description.

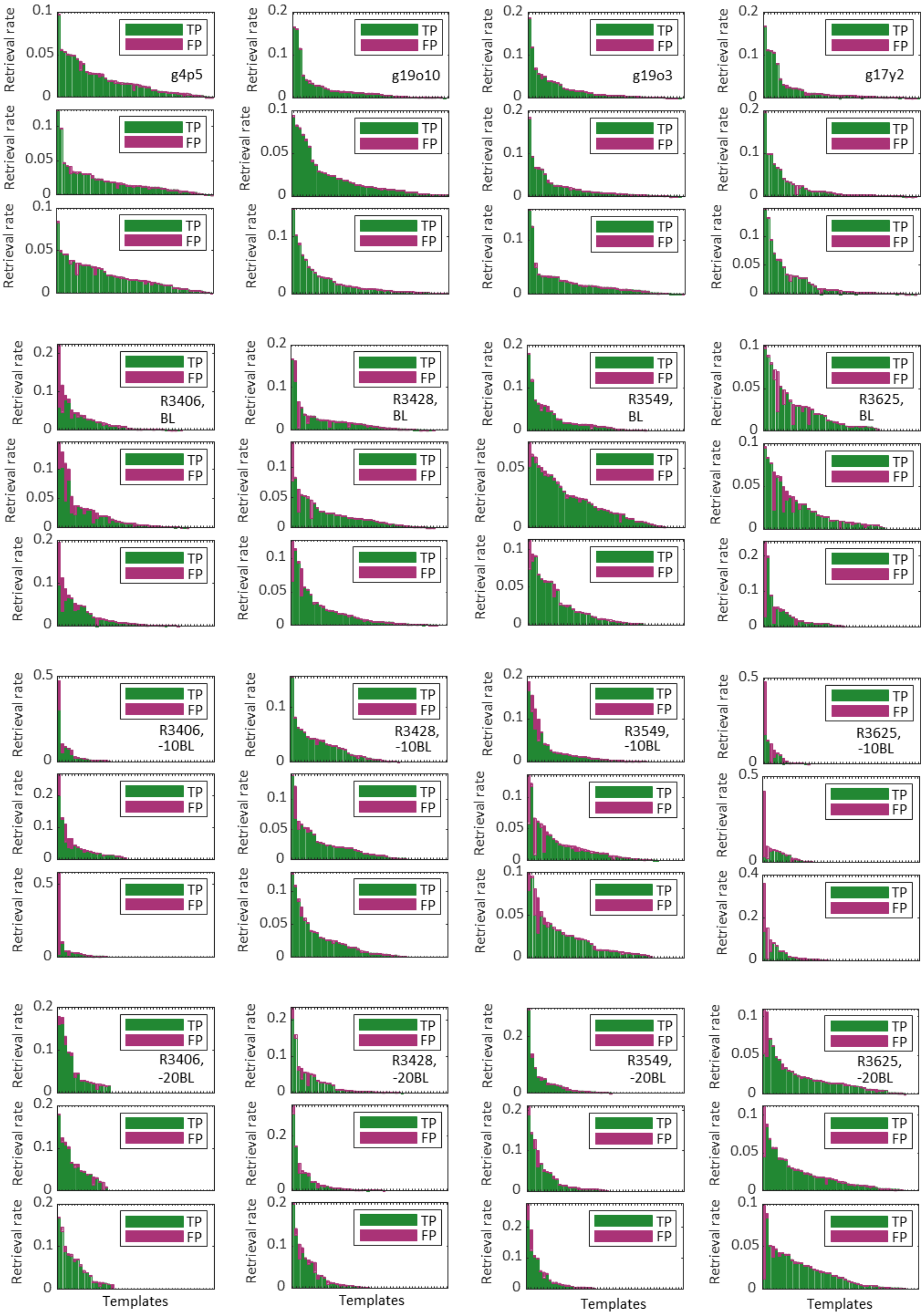

Retrieval rate
TP
FP
g4p5
g19o10
g19o3
g17y2
R3406, BL
R3428, BL
R3549, BL
R3625, BL
R3406, -10BL
R3428, -10BL
R3549, -10BL
R3625, -10BL
R3406, -20BL
R3428, -20BL
R3549, -20BL
R3625, -20BL
0
0.05
0.1
0.2
0.4
0.5
Templates

Figure AIII.3[32]: Extended set of histograms of retrieval rates across templates, using the WHOLE approach with the Spearman distance and 50 templates (3 retrieval replicates for each bird). The top row (consisting of 3 panels for each retrieval replicate) shows adult birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure 1.4a-c for a detailed description.

[32] Figure is displayed on the previous page.

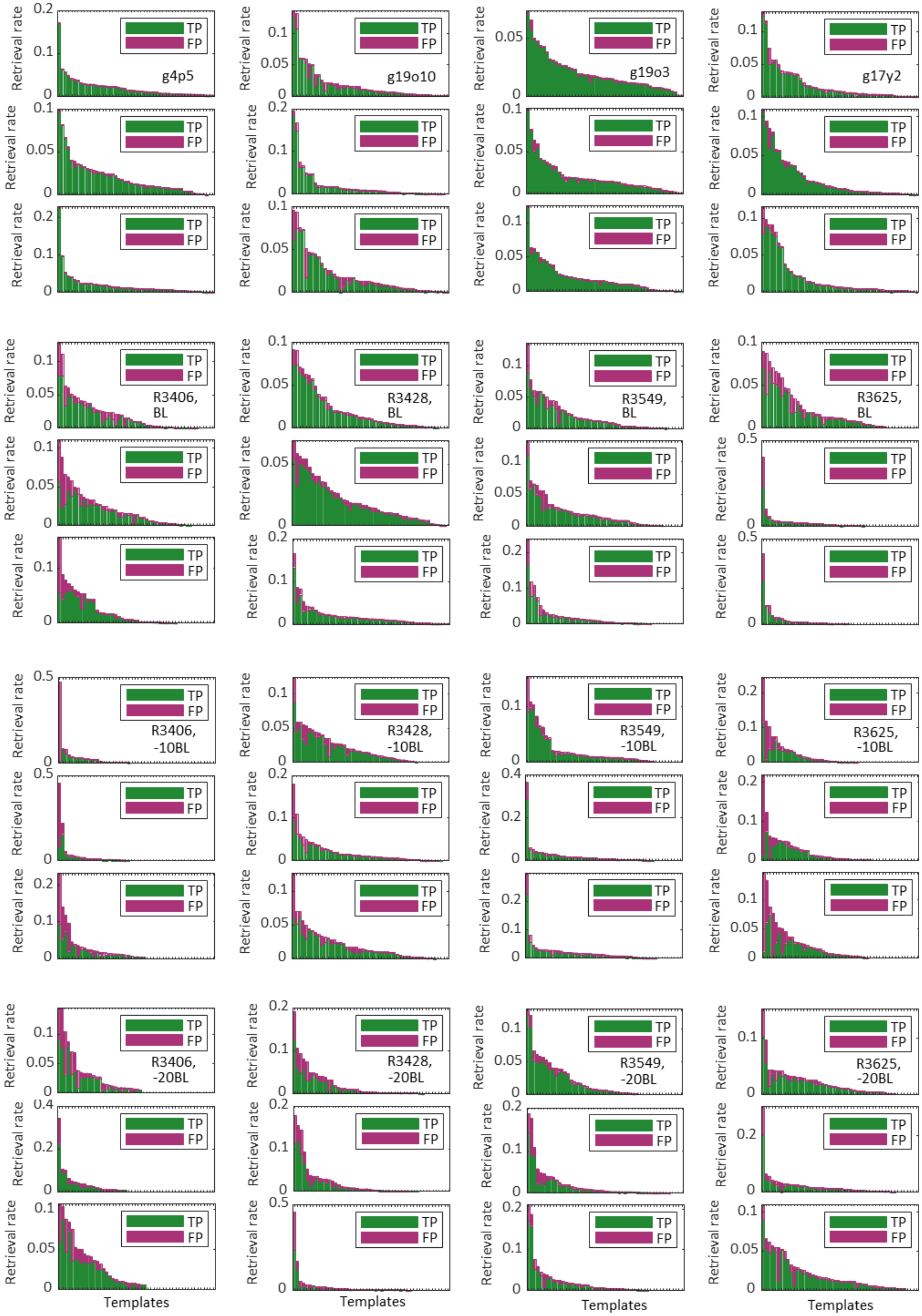

Retrieval rate
TP
FP
g4p5
g19o10
g19o3
g17y2
R3406,
BL
R3428,
BL
R3549,
BL
R3625,
BL
R3406,
-10BL
R3428,
-10BL
R3549,
-10BL
R3625,
-10BL
R3406,
-20BL
R3428,
-20BL
R3549,
-20BL
R3625,
-20BL
Templates

Figure AIII.4[33]: Extended set of histograms of retrieval rates across templates, using the PART approach with the Spearman distance and 50 templates (3 retrieval replicates for each bird). The top row (consisting of 3 panels for each retrieval replicate) shows adult birds, while the subsequent rows show juveniles at different ages relative to baseline. See Figure 1.4a-c for a detailed description.

[33] Figure is displayed on the previous page.

Appendix IV – Protocol for manually annotating vocalizations in accelerometer data (“Chapter 2”)

Our protocol of manually annotating vocalizations by visual inspection of spectrograms comprises two steps:

1. Segment vocalization onsets and offsets.
2. Cluster the segmented vocalizations into syllable and call types.

Unless otherwise specified, we used solely the accelerometer recordings of a given bird to annotate its vocalizations, disregarding the other data channels (to save time). In the following, we describe the choices we made.

Segmentation

Following the guidelines for segmenting vocalizations in single-bird data that we defined in Appendix I, we reduce the segmentation task to two decisions that we take:

a) Whether a sound is vocal or non-vocal (Figure AIV.1a).
b) Whether two consecutive vocalizations are separated by a gap or not (Figure AIV.1b).

These decisions we take by inspecting spectrograms. As for the first decision, we judge whether a sound is vocal (Figure AIV.1a, first column) or non-vocal (Figure AIV.1a, last two columns) by inspecting spectrograms of accelerometer signals and if needed, also by listening to the sounds. Zebra finch vocalizations are easiest to detect when they contain pure tones or harmonic stacks that extend over significant periods of time. More challenging are soft notes (Figure AIV.1a, “2”) that extend beyond the limits of loud vocal activity. We do not know the physical causes of these soft sounds, but we decided to not mark them as vocal, unless they were loud enough and were distinguishable from noises such as wing flaps.

The second decision about dividing song motifs into constituent syllables depends on whether in most motifs, two consecutive syllables are separated by a gap that interrupts the continuous vocal activity (Figure AIV.1b, first column). Exceptions we make to song syllables that contain high-frequency notes (Figure AIV.1b, second column), as these notes often are not picked up by accelerometers. In such cases, we set approximate syllable boundaries by inspecting spectrograms of microphone recordings.

Clustering

We annotated the segmented vocalizations according to the nomenclature of Elie and Theunissen (Elie & Theunissen, 2016, 2018, 2020). Often, we struggled with assigning calls to different types, because of presence of intermediate forms between “tet”, “nest”, and “whine” calls when we projected the vocalizations onto a plane using UMAP (McInnes et al., 2018), Figure AIV.2[34]. A detailed description of this projection is given in the following.

[34] Possible reasons are that our dataset contains many more vocalizations (more than 50’000 vocalizations compared to 3433 vocalizations (or sequences thereof) in (Elie & Theunissen, 2020)).

For each bird, first we extracted spectrogram snippets in an expert-chosen frequency range, usually from around 500 Hz to 2-3 kHz. The snippets were chosen from the onset of a vocalization until time lag d after the onset, where d is given in each bird by the duration of its longest vocalization (all renditions taken into account). To exclude that more than one vocalization overlaps with any snippet, we pad shorter vocalizations as follows. Any gap between a vocalization offset and the end of the snippet is padded with a manually set minimal spectrogram value. This minimal value is manually chosen such that low-intensity fluctuations are not resolved. The resulting rectified and padded spectrogram snippets are then down-projected to 50 principal component coefficients (PCCs) calculated for each bird separately. The coefficients of a given bird we then project onto the embedding plane using UMAP (McInnes et al., 2018).

Using this technique, in particular the "nest" and "whine" calls always populated a shared cluster (n=6/6 birds; Figure AIV.3-AIV.4); for this reason, we decided not to distinguish them. Our final dataset comprises the vocalization categories: "tet" and "nest/whine" calls for each gender, as well as male "song syllables", "introductory notes", "distance calls", and "wsst" calls (Figure AIV.5).

In Figure AIV.1, we depict decision boundaries between different call types, with an example sharp boundary shown in the first column, where two "nest/whine" calls are separate (top 4 panels) and then joined together (lower 4 panels), which might be a strategy to produce a "tet" call. This however, is not the only possible transition from "nest/whine" to "tet" calls. While in females the "tets" are highly stereotyped, with a characteristic step-like upsweep at the call onset, in some birds the morphing to "nest/whine" calls is more gradual (and continuous with the "nest" extreme rather than the "whine" extreme; Figure AIV.3-AIV.4). We had to define a somewhat arbitrary decision for these cases: we labelled vocalizations as "tet" when they consisted of harmonic stacks with a tendency of constant pitch or upsweeps, but no downsweeps. Note, that however, some exceptional "tet" calls with downsweeps were found in well-separated clusters (Figure AIV.5, last three rows of female "tet" calls).

Unfortunately, we noted an inconsistency in the zebra finch literature: our "tet" calls are sometimes referred to as "stacks", while the "tet" label is used for short calls with downsweeping harmonics (Gill et al., 2015; Zann, 1996), residing at a pole within our "nest/whine" clusters.

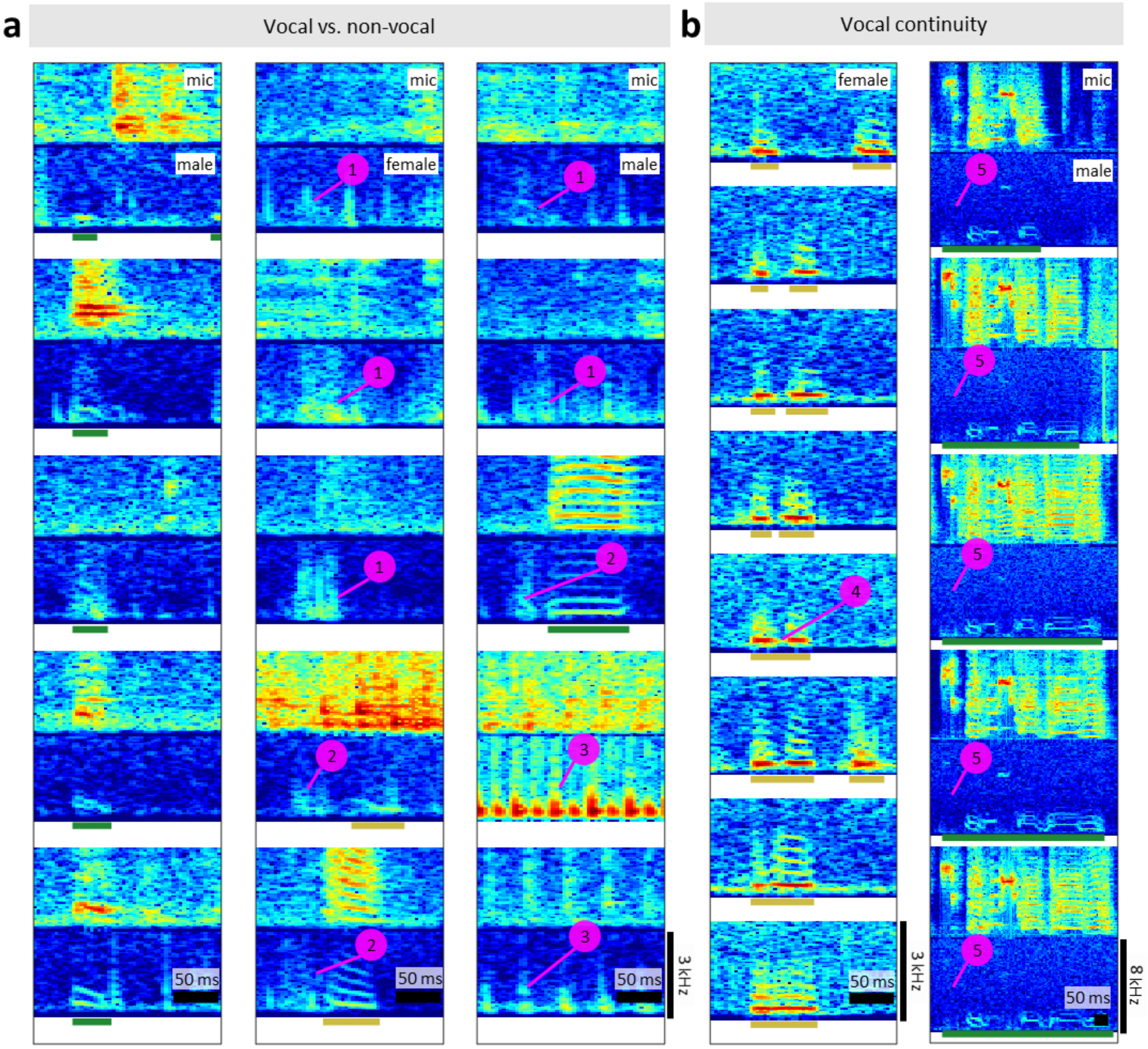


Figure AIV.1: Vocalizations and noises near the decision boundary of the vocal segmentation task. **(a)** Compared to vocalizations (left), non-vocal signals ("1") near the decision boundary often lack significant presence of tonal sounds, both in accelerometer as well as in microphone data. The same is true for soft noises ("2") just before vocalizations. Indicated are also wing flaps ("3"). **(b)** Vocal continuity. Left: Two consecutive calls (top 4 examples) become one vocalization (bottom 4 examples) due to a bridging sound of sufficient signal-to-noise ratio. The example at the decision boundary is indicated by the number "4". Right: Examples, where assessment of the full extent of a vocalization requires inspection of the microphone (mic) recordings, because the initial high-frequency note ("5") is not visible on the male-attached accelerometer signal. **(a,b)** Vocal segments are indicated by horizontal bars (yellow for females, green for males) below the corresponding accelerometer spectrograms. Black bars indicate time and frequency units.

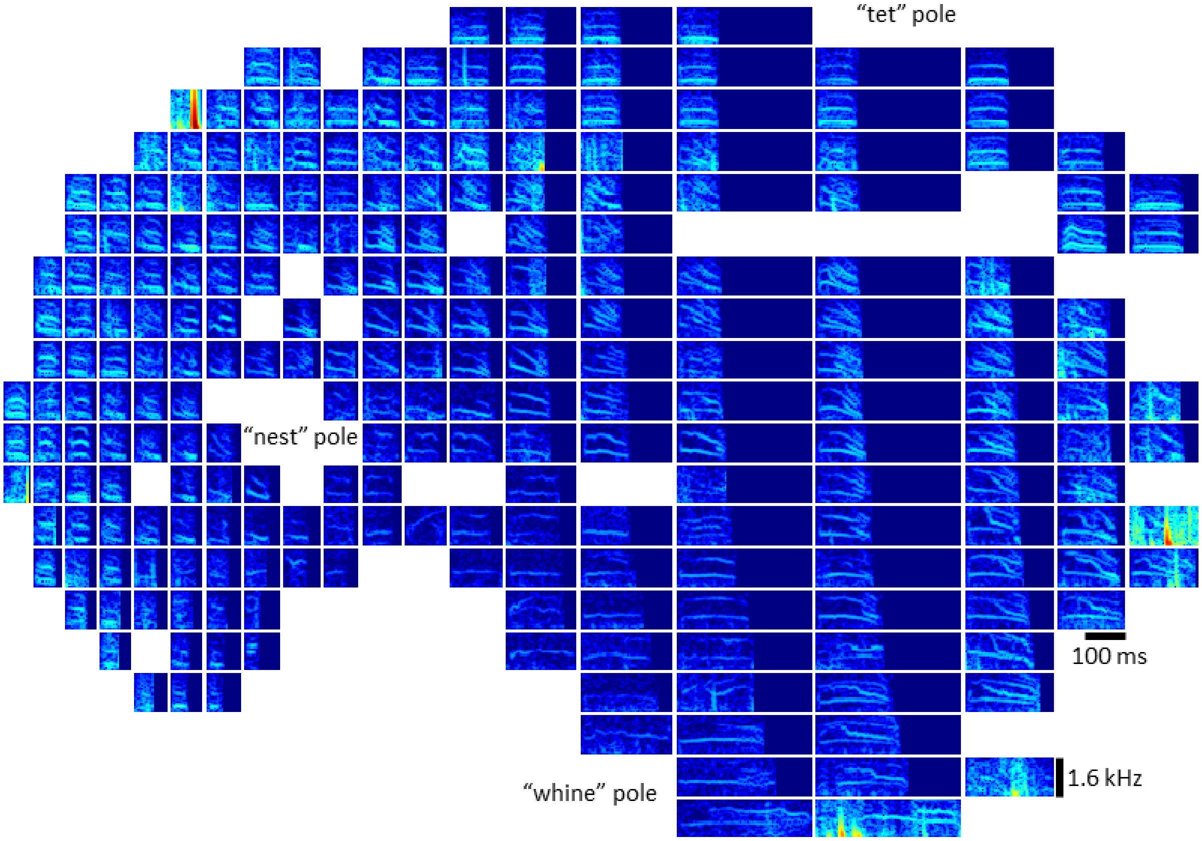


Figure AIV.2: Gradual "tet", "nest", and "whine" calls of an exemplary male populate a single cluster when projected onto a plane using the dimensionality reduction technique UMAP. Initially, all shown spectrograms had the same dimension (zero-padding vocalizations to fit the length of the longest call), however, we removed all-zero columns across the entire figure for nicer visualization. After zero-padding, we computed 50 principal components of the spectrograms, which we then used to compute the UMAP.

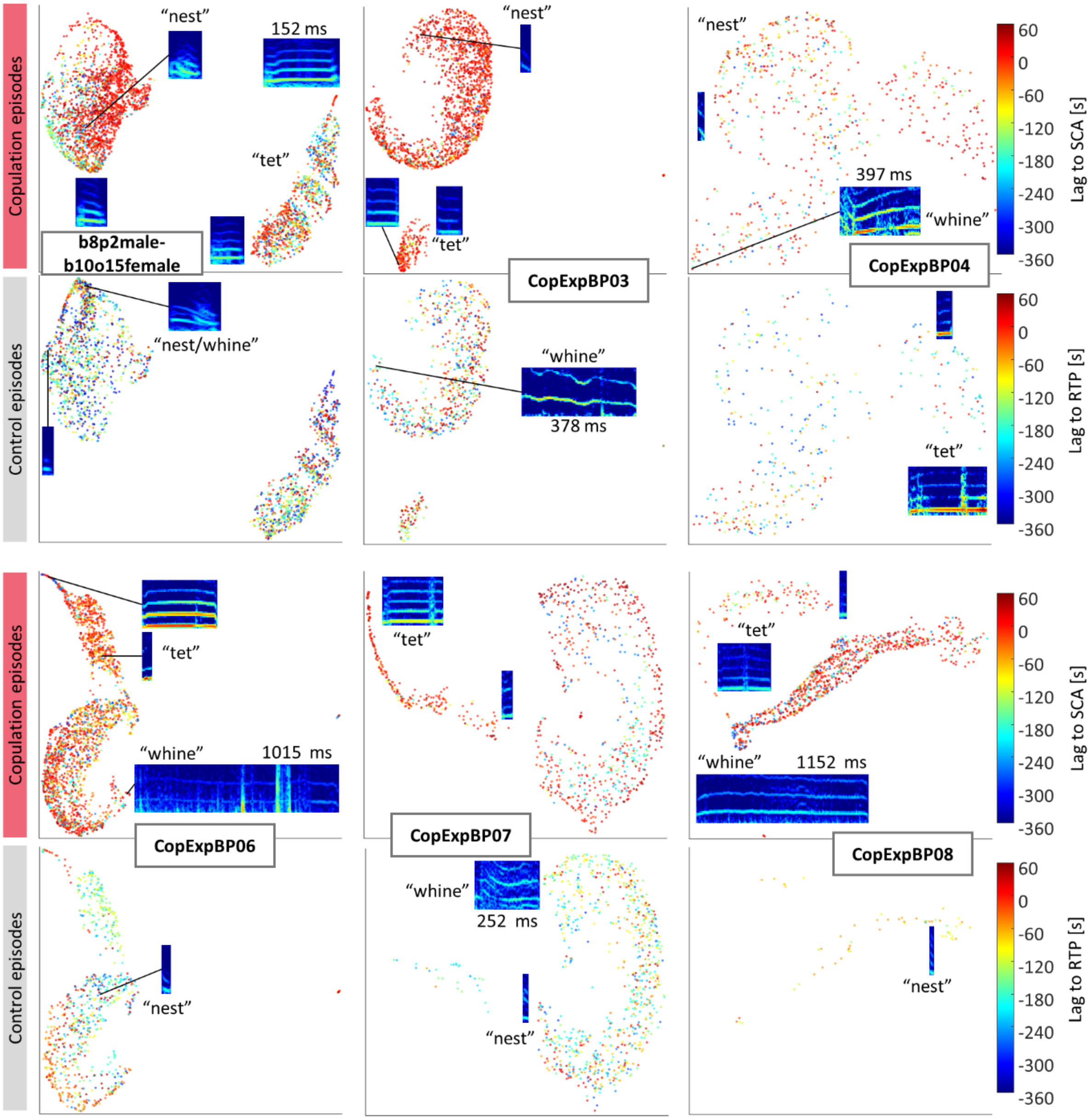


Figure AIV.3: UMAP projections of female vocalizations. For each of the six experiment replicates, we compute an UMAP projection from all annotated female vocalizations and visualize vocalizations occurring within the two episodes groups ("copulation" and "control"), in two separate panels. The proximity to the reference event (solicited copulation attempt, SCA or random time point, RTP) is indicated with a color code (e.g., red color indicating high proximity to the SCA in control episodes). Example spectrograms show some of the variation within single clusters. For better visualization, the spectrogram time units vary across experiment replicates (for each replicate, the duration of one example vocalization is labelled).

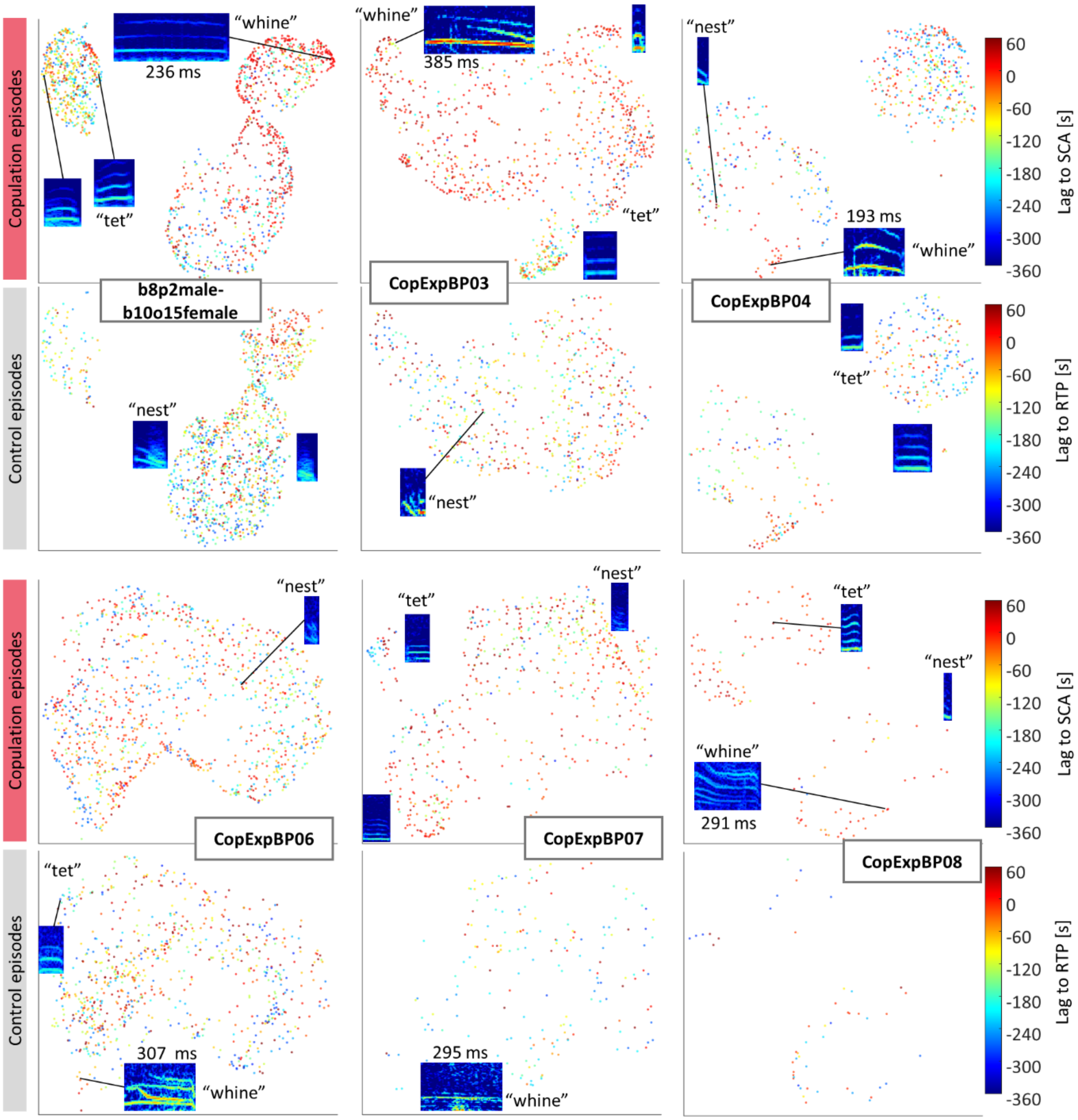


Figure AIV.4: UMAP projections of male "tet", "nest", and "whine" calls. For each of the six experiment replicates, we compute an UMAP projection from all annotated male vocalizations and visualize vocalizations occurring within the two episodes groups ("copulation" and "control"), in two separate panels. The proximity to the reference event (solicited copulation attempt, SCA or random time point, RTP) is indicated with a color code (e.g., red color indicating high proximity to the SCA in control episodes). Example spectrograms show some of the variation within single clusters. For better visualization, the spectrogram time units vary across experiment replicates (for each replicate, the duration of one example vocalization is labelled).

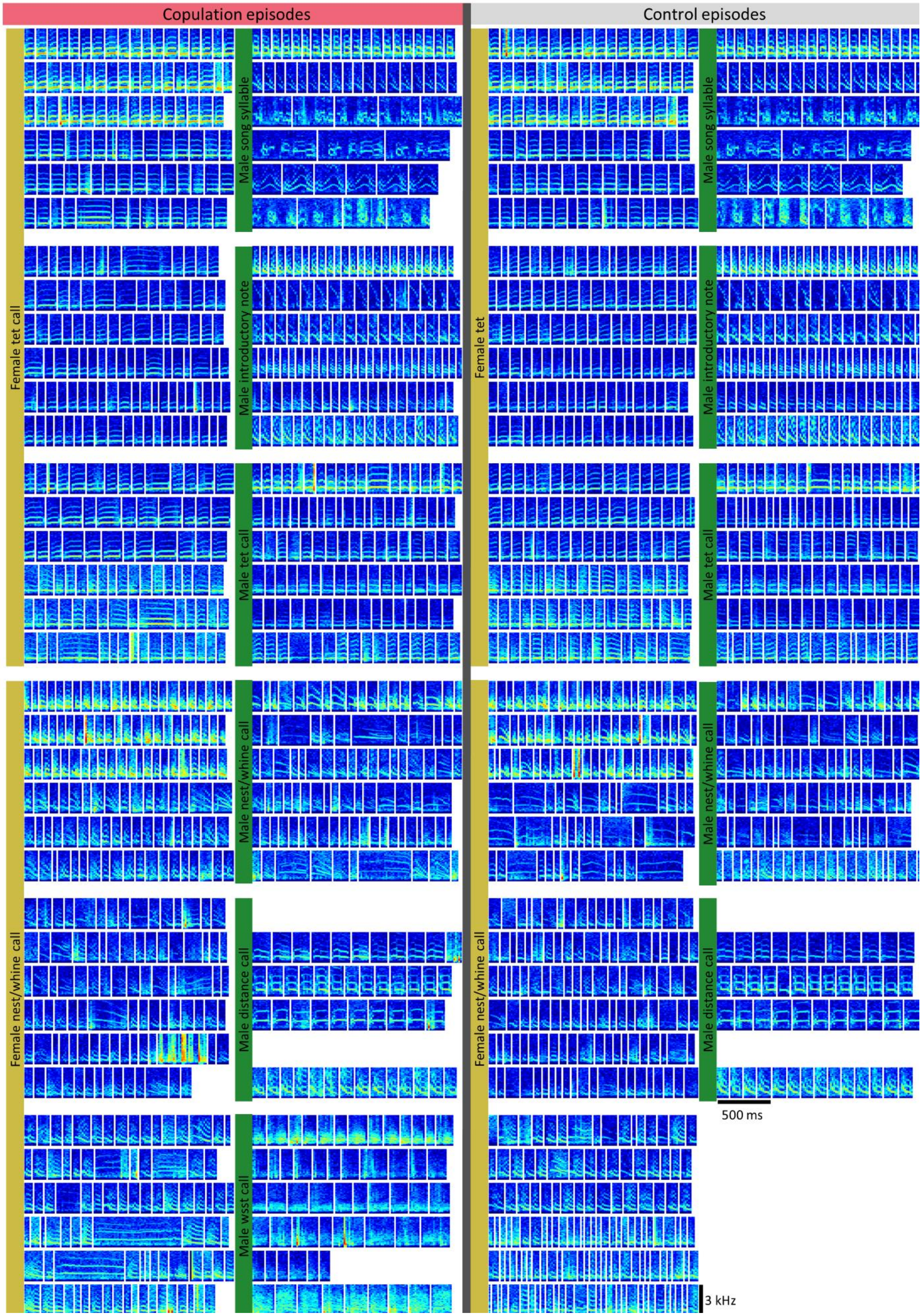

Copulation episodes
Control episodes
Female tet call
Male song syllable
Male introductory note
Male tet call
Female tet
Female nest/whine call
Male nest/whine call
Male distance call
Male wsst call
500 ms
3 kHz

Figure AIV.5[35]: Overview of vocalization categories distinguished in our dataset. Random example spectrograms of each vocalization category, drawn from copulation and control episodes of all birds (n=6 per gender). Black bars indicate time and frequency units.

[35] Figure is displayed on the previous page.

Appendix V – Non-vocal behaviors (“Chapter 2”)[36]

Table AV.1 lists the non-vocal behaviors that have been annotated in the work presented in “Chapter 2”.

Table AV.1: Non-vocal behaviors annotated within COP / CTRL episodes.

Category	Behavior	Description
Courtship behaviors	Beak wipe	Wiping beak on or above the perch.
	Cloacal contact	Male and female cloaca touch; only behavior that gets annotated without a subject (male or female).
	Directed	Male sings in close proximity to female and with body generally facing the female.
	End copulation	Bird unmounts or escapes copulation.
	Head and tail bent	Head and tail twist towards the partner.
	Hop	Hopping on perch or floor. Hop on floor only annotated if not seemingly associated with distractions like looking for food (head high and not looking at floor).
	Mounting	Bird mounts its partner.
	Tail-quiver	Lowering body into a horizontal position, with feathers rather sleeked and legs bent, then vibrates tail extremely rapidly in the vertical plane.
	Turn-around	Turning body axis roughly 180˚.
	Undirected	Male sings far away from female and/or with body facing away from the female (meaning female is directly behind male’s back).
Other behaviors	Allopreening	One bird preens another bird.
	Approaching	Flying/moving towards the partner.
	Chasing	Flight at conspecific forcing it to leave its location.
	Clumping	Birds sit in direct physical contact with each other.
	Fighting	Beak-fencing or full body fight.
	Fleeing	Flying/moving away from conspecific
	Nest building	Collecting nest material to take to the nest.
	Nesting	Bird is inside nest (perch next to nest entry does not count). Behaviors inside the nest are not annotated as visibility inside the nest varies across groups (opaque vs transparent nest sides vs nest cameras).
	Pecking	Biting or striking conspecific with beak.
	Sleeping	bird remains still with eyes mostly closed (birds will often briefly open eyes while sleeping).

[36] Information provided by Dr. Mariana da Rocha.

Appendix VI – Extended set of conditional rate curves ("Chapter 2")

Conditional rate functions (CRF) which are well defined (do not contain all not-a-number values) in at least one bird, and are not shown in Figure 2.1, are supplemented here, Figure AVI.1. Ill-defined CRFs (that have no occurrence in any control episode) have been obtained for the following behaviors: "cloacal contact", "end copulation", "head and tail bent", "mounting", "sleeping", and female "chasing" (see "Chapter 2" and "Appendix V" for more details).

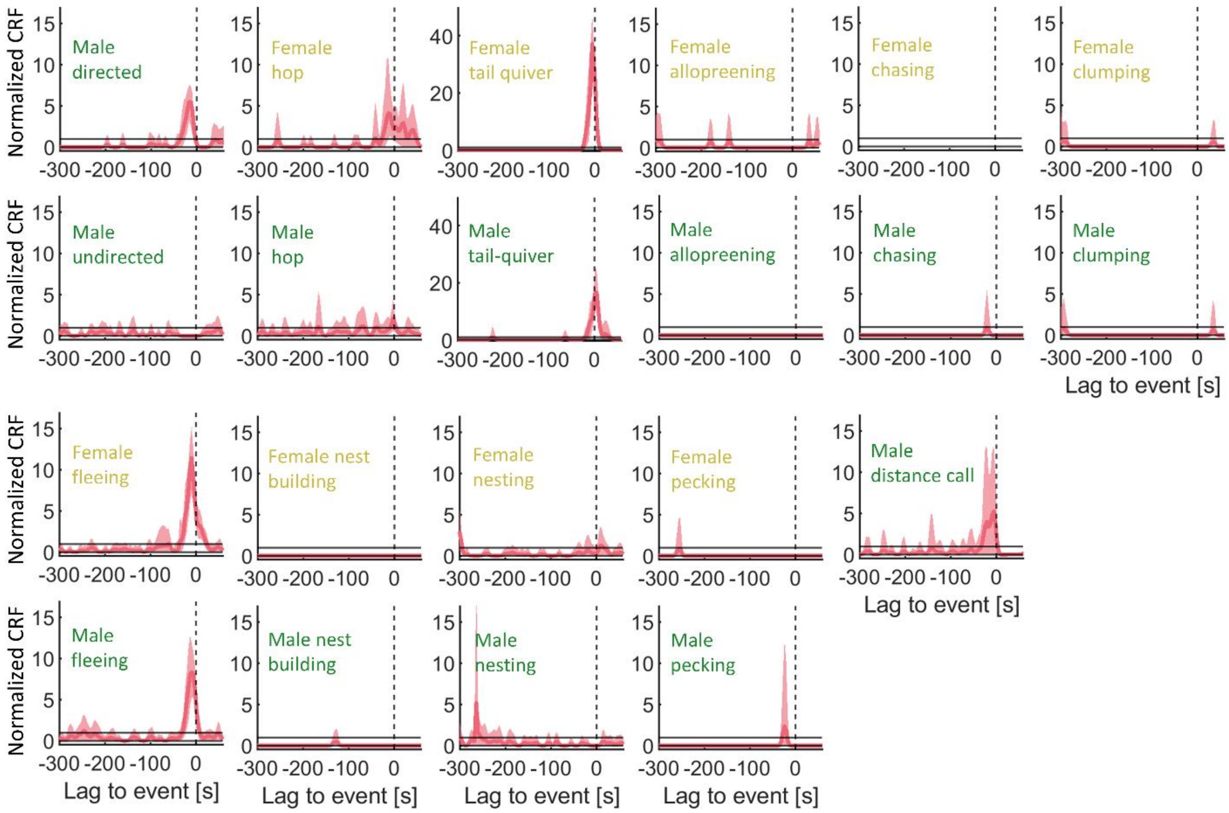


Figure AVI.1: Extended set of conditional rate functions. See Figure 2.1 for a detailed description. Note, that this figure shows 10 non-vocal behaviors for both sexes, and, in addition, the "male distance call". Furthermore, the axis for the "tail-quiver" plots is altered, to show the large CRF peaks around copulations.

Bibliography

Adeniyi, D. A., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108.

Andalman, A. S., & Fee, M. S. (2009). A basal ganglia-forebrain circuit in the songbird biases motor output to avoid vocal errors. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(30), 12518–12523.

Anderson, S. E., Dave, A. S., & Margoliash, D. (1996). Template-based automatic recognition of birdsong syllables from continuous recordings. *The Journal of the Acoustical Society of America*, *100*(2 Pt 1), 1209–1219.

Andoni, A., & Indyk, P. (2008). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, *51*(1), 117–122.

Anisimov, V. N., Herbst, J. A., Abramchuk, A. N., Latanov, A. V., Hahnloser, R. H. R., & Vyssotski, A. L. (2014). Reconstruction of vocal interactions in a group of small songbirds. *Nature Methods*, *11*(11), 1135–1137.

Arnold, A. P. (1975). The effects of castration and androgen replacement on song, courtship, and aggression in zebra finches (Poephila guttata). *The Journal of Experimental Zoology*, *191*(3), 309–326.

Ausra, J., Munger, S. J., Azami, A., Burton, A., Peralta, R., Miller, J. E., & Gutruf, P. (2021). Wireless battery free fully implantable multimodal recording and neuromodulation tools for songbirds. *Nature Communications*, *12*(1), 1968.

Avey, M. T., Phillmore, L. S., & MacDougall-Shackleton, S. A. (2005). Immediate early gene expression following exposure to acoustic and visual components of courtship in zebra finches. *Behavioural Brain Research*, *165*(2), 247–253.

Bairlein, F. (2016). Migratory birds under threat. *Science*, *354*(6312), 547–548.

Bass, A. H., Gilland, E. H., & Baker, R. (2008). Evolutionary origins for social vocalization in a vertebrate hindbrain-spinal compartment. *Science*, *321*(5887), 417–421.

Batthyany, B. (2012). *Ich kenne dich, ich habe dich spielen gehört*. Schweizer Radio und Fernsehen.

Becker, A., Ducas, L., Gama, N., & Laarhoven, T. (2016). New directions in nearest neighbor searching with applications to lattice sieving. In R. Krauthgamer (Ed.), *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 10–24). Presented at the Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Bharati, I. S., & Goodson, J. L. (2006). Fos responses of dopamine neurons to sociosexual stimuli in male zebra finches. *Neuroscience*, *143*(3), 661–670.

Biegler, M. T., Fedrigo, O., Collier, P., Mountcastle, J., Haase, B., Tilgner, H. U., & Jarvis, E. D. (2022). Induction of an immortalized songbird cell line allows for gene characterization and knockout by CRISPR-Cas9. *Scientific Reports*, *12*(1), 4369.

Birkhead, T. R., Clarkson, K., & Zann, R. (1988). Extra-pair courtship, copulation and mate guarding in wild zebra finches taeniopygia guttata. *Animal Behaviour*, *36*(6), 1853–1855.

Birkhead, T. R., Hunter, F. M., & Pellatt, J. E. (1989). Sperm competition in the zebra finch, Taeniopygia guttata. *Animal Behaviour*, *38*(6), 935–950.

Bischof, H.-J., Böhner, J., & Sossinka, R. (1981). Influence of External Stimuli on the Quality of the Song of the Zebra Finch (Taeniopygia guttata castanotis Gould). *Zeitschrift für Tierpsychologie*, *57*(3–4), 261–267.

Böhner, J. (1983). Song learning in the zebra finch (taeniopygia guttata): Selectivity in the choice of a tutor and accuracy of song copies. *Animal Behaviour*, *31*(1), 231–237.

Bolhuis, J. J., Okanoya, K., & Scharff, C. (2010). Twitter evolution: converging mechanisms in birdsong and human speech. *Nature Reviews. Neuroscience*, *11*(11), 747–759.

Brainard, M. S., & Doupe, A. J. (2002). What songbirds teach us about learning. *Nature*, *417*(6886), 351–358.

Brainard, M. S., & Doupe, A. J. (2000). Interruption of a basal ganglia-forebrain circuit prevents plasticity of learned vocalizations. *Nature*, *404*(6779), 762–766.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1–20.

Brock, A., Donahue, J., & Simonyan, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv*.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, perception & psychophysics*, *77*(5), 1465–1487.

Brooker, S. A., Stephens, P. A., Whittingham, M. J., & Willis, S. G. (2020). Automated detection and classification of birdsong: An ensemble approach. *Ecological Indicators*, *117,* 106609.

Bryant, G. A. (2021). The evolution of human vocal emotion. *Emotion review*, *13*(1), 25–33.

Burkett, Z. D., Day, N. F., Peñagarikano, O., Geschwind, D. H., & White, S. A. (2015). VoICE: A semi-automated pipeline for standardizing vocal analysis across models. *Scientific Reports*, *5*, 10237.

Burt, J. M., Campbell, S. E., & Beecher, M. D. (2001). Song type matching as threat: a test using interactive playback. *Animal Behaviour*, *62*(6), 1163–1170.

Burt, J. M., O'Loghlen, A. L., Templeton, C. N., Campbell, S. E., & Beecher, M. D. (2007). Assessing the Importance of Social Factors in Bird Song Learning: A Test Using Computer-Simulated Tutors. *Ethology*, *113*(10), 917–925.

Buxton, R. T., & Jones, I. L. (2012). Measuring nocturnal seabird activity and status using acoustic recording devices: applications for island restoration. *Journal of field ornithology*, *83*(1), 47–60.

Carouso-Peck, S., & Goldstein, M. H. (2019). Female Social Feedback Reveals Non-imitative Mechanisms of Vocal Learning in Zebra Finches. *Current Biology*, *29*(4), 631-636.e3.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *52*(4), 431–443.

Charlesworth, J. D., Tumer, E. C., Warren, T. L., & Brainard, M. S. (2011). Learning the microstructure of successful behavior. *Nature Neuroscience*, *14*(3), 373–380.

Charlesworth, J. D., Warren, T. L., & Brainard, M. S. (2012). Covert skill learning in a cortical-basal ganglia circuit. *Nature*, *486*(7402), 251–255.

Chaverri, G., Ancillotto, L., & Russo, D. (2018). Social communication in bats. *Biological Reviews of the Cambridge Philosophical Society*, *93*(4), 1938–1954.

Ciaburri, I., & Williams, H. (2019). Context-dependent variation of house finch song syntax. *Animal Behaviour*, *147*, 33–42.

Clemens, J. (2021). Zebra finch - train and test data. *GRO.data*, *V1*.

Cohen, Y., Nicholson, D. A., Sanchioni, A., Mallaber, E. K., Skidanova, V., & Gardner, T. J. (2022). Automated annotation of birdsong with a neural network that segments spectrograms. *eLife*, *11*.

Couchoux, C., Aubert, M., Garant, D., & Réale, D. (2015). Spying on small wildlife sounds using affordable collar-mounted miniature microphones: an innovative method to record individual daylong vocalisations in chipmunks. *Scientific Reports*, *5*, 10118.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27.

Cusano, D. A., Noad, M. J., & Dunlop, R. A. (2021). Fuzzy clustering as a tool to differentiate between discrete and graded call types. *JASA express letters*, *1*(6), 061201.

Cuzick, J., & Edwards, R. (1990). Spatial clustering for inhomogeneous populations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *52*(1), 73–96.

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

Darwin, C. (1859). *On the Origin of Species*. London, UK: John Murray.

Darwin, C. (1871). *The Descent of Man, and Selection in Relation to Sex*. London, United Kingdom: John Murray.

Derégnaucourt, S. (2011). Birdsong learning in the laboratory, with especial reference to the song of the Zebra Finch (Taeniopygia guttata). *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems*, *12*(2), 324–350.

Digby, A., Towsey, M., Bell, B. D., & Teal, P. D. (2013). A practical comparison of manual and autonomous methods for acoustic monitoring. *Methods in Ecology and Evolution*, *4*(7), 675–683.

Dittus, W. P. J. (1984). Toque macaque food calls: Semantic communication concerning food distribution in the environment. *Animal Behaviour*, *32*(2), 470–477.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv*.

Doupe, A. J., & Konishi, M. (1991). Song-selective auditory circuits in the vocal control system of the zebra finch. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(24), 11339–11343.

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, *22*, 567–631.

Doya, K., & Sejnowski, T. J. (1995). A Novel Reinforcement Model of Birdsong Vocalization Learning.

Elfenbein, H. A., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: a meta-analysis. *Psychological Bulletin*, *128*(2), 203–235.

Elie, J. E., & Theunissen, F. E. (2020). Vocal repertoires from adult and chick, male and female zebra finches (Taeniopygia guttata). *Figshare*.

Elie, J. E., Mariette, M. M., Soula, H. A., Griffith, S. C., Mathevon, N., & Vignal, C. (2010). Vocal communication at the nest between mates in wild zebra finches: a private vocal duet? *Animal Behaviour*, *80*(4), 597–605.

Elie, J. E., & Theunissen, F. E. (2015). Meaning in the avian auditory cortex: neural representation of communication calls. *The European Journal of Neuroscience*, *41*(5), 546–567.

Elie, J. E., & Theunissen, F. E. (2016). The vocal repertoire of the domesticated zebra finch: a data-driven approach to decipher the information-bearing acoustic features of communication signals. *Animal Cognition*, *19*(2), 285–315.

Elie, J. E., & Theunissen, F. E. (2018). Zebra finches identify individuals using vocal signatures unique to each call type. *Nature Communications*, *9*(1), 4026.

Elie, J. E., & Theunissen, F. E. (2019). Invariant neural responses for sensory categories revealed by the time-varying information for communication calls. *PLoS Computational Biology*, *15*(9), e1006698.

Evans, C. S., & Marler, P. (1991). On the use of video images as social stimuli in birds: audience effects on alarm calling. *Animal Behaviour*, *41*(1), 17–26.

Fee, M. S., & Goldberg, J. H. (2011). A hypothesis for basal ganglia-dependent reinforcement learning in the songbird. *Neuroscience*, *198*, 152–170.

Fischer, J. (1998). Barbary macaques categorize shrill barks into two call types. *Animal behaviour*, *55*(4), 799–807.

Fitch, W. T. (2017). Empirical approaches to the study of language evolution. *Psychonomic Bulletin & Review*, *24*(1), 3–33.

Gadagkar, V., Puzerey, P. A., Chen, R., Baird-Daniel, E., Farhang, A. R., & Goldberg, J. H. (2016). Dopamine neurons encode performance error in singing birds. *Science*, *354*(6317), 1278–1282.

Garcia, V., Debreuve, E., & Barlaud, M. (2008). Fast k nearest neighbor search using GPU. *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 1–6). Presented at the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), IEEE.

George, J. M., Jin, H., Woods, W. S., & Clayton, D. F. (1995). Characterization of a novel protein regulated during the critical period for song learning in the zebra finch. *Neuron*, *15*(2), 361–372.

Gill, L. F., Goymann, W., Ter Maat, A., & Gahr, M. (2015). Patterns of call communication between group-housed zebra finches change during the breeding cycle. *eLife*, *4*, e07770.

Gill, S. A., & Bierema, A. M.-K. (2013). On the meaning of alarm calls: A review of functional reference in avian alarm calling. *Ethology : formerly Zeitschrift fur Tierpsychologie*, *119*(6), 449–461.

Goffinet, J., Brudner, S., Mooney, R., & Pearson, J. (2021). Data from: Low-dimensional learned feature spaces quantify individual and group differences in vocal repertoires. *Duke Research Data Repository*.

Goldstein, M. H., King, A. P., & West, M. J. (2003). Social interaction shapes babbling: testing parallels between birdsong and speech. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(13), 8030–8035.

Goldstein, M. H., & Schwade, J. A. (2008). Social feedback to infants' babbling facilitates rapid phonological learning. *Psychological Science*, *19*(5), 515–523.

Goller, F., & Daley, M. A. (2001). Novel motor gestures for phonation during inspiration enhance the acoustic complexity of birdsong. *Proceedings. Biological Sciences / the Royal Society*, *268*(1483), 2301–2305.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., et al. (2014). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139–144.

Goodson, J. L., Kabelik, D., Kelly, A. M., Rinaldi, J., & Klatt, J. D. (2009). Midbrain dopamine neurons reflect affiliation phenotypes in finches and are tightly coupled to courtship. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(21), 8737–8742.

Gouzoules, S., Gouzoules, H., & Marler, P. (1984). Rhesus monkey (Macaca mulatta) screams: Representational signalling in the recruitment of agonistic aid. *Animal Behaviour*, *32*(1), 182–193.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., & Couzin, I. D. (2019). DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, *8*.

Greif, S., & Yovel, Y. (2019). Using on-board sound recordings to infer behaviour of free-moving wild animals. *Journal of Experimental Biology*, *222*(Pt Suppl 1).

Griffith, S. C., & Buchanan, K. L. (2010). The Zebra Finch: the ultimate Australian supermodel. *Emu*, *110*(3), v–xii.

Guillaumin, M., Mensink, T., Verbeek, J., & Schmid, C. (2009). TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. *2009 IEEE 12th International Conference on Computer Vision* (pp. 309–316). Presented at the 2009 IEEE 12th International Conference on Computer Vision (ICCV), IEEE.

Guo, J., Fan, Y., Pang, L., Yang, L., Ai, Q., Zamani, H., Wu, C., et al. (2019). A Deep Look into neural ranking models for information retrieval. *Information Processing & Management*, 102067.

Gurney, M. E., & Konishi, M. (1980). Hormone-induced sexual differentiation of brain and behavior in zebra finches. *Science*, *208*(4450), 1380–1383.

Haesler, S., Rochefort, C., Georgi, B., Licznerski, P., Osten, P., & Scharff, C. (2007). Incomplete and inaccurate vocal imitation after knockdown of FoxP2 in songbird basal ganglia nucleus Area X. *PLoS Biology*, *5*(12), e321.

Hahnloser, R. H. R., Kozhevnikov, A. A., & Fee, M. S. (2002). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature*, *419*(6902), 65–70.

Harding, C. F., Sheridan, K., & Walters, M. J. (1983). Hormonal specificity and activation of sexual behavior in male zebra finches. *Hormones and Behavior*, *17*(1), 111–133.

Hauber, M. E., Louder, M. I., & Griffith, S. C. (2021). Neurogenomic insights into the behavioral and vocal development of the zebra finch. *eLife*, *10*.

Hauser, M. D. (1998). Functional referents and acoustic similarity: field playback experiments with rhesus monkeys. *Animal behaviour*, *55*(6), 1647–1658.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science*, *298*(5598), 1569–1579.

Haykin, S., & Chen, Z. (2005). The cocktail party problem. *Neural Computation*, *17*(9), 1875–1902.

Heckman, J. J., Proville, R., Heckman, G. J., Azarfar, A., Celikel, T., & Englitz, B. (2017). High-precision spatial localization of mouse vocalizations during social interaction. *Scientific Reports*, *7*(1), 3017.

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181–184.

Hoffmann, S., Trost, L., Voigt, C., Leitner, S., Lemazina, A., Sagunsky, H., Abels, M., et al. (2019). Duets recorded in the wild reveal that interindividually coordinated motor control enables cooperative behavior. *Nature Communications*, *10*(1), 2577.

Ikebuchi, M., & Okanoya, K. (1999). Male Zebra Finches and Bengalese Finches Emit Directed Songs to the Video Images of Conspecific Females Projected onto a TFT Display. *Zoological Science*, *16*(1), 63–70.

Immelmann, K. (1969). Song development in the zebra finch and other estrildid finches. In R. A. Hinde (Ed.), *Bird Vocalisations* (pp. 64–74).

Indyk, P., & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC '98* (pp. 604–613). Presented at the thirtieth annual ACM symposium, New York, New York, USA: ACM Press.

James, L. S., Fan, R., & Sakata, J. T. (2019). Behavioural responses to video and live presentations of females reveal a dissociation between performance and motivational aspects of birdsong. *Journal of Experimental Biology*, *222*(Pt 16).

Jarvis, E. D. (2004). Learned birdsong and the neurobiology of human language. *Annals of the New York Academy of Sciences*, *1016*, 749–777.

Jarvis, E. D. (2006). Selection for and against vocal learning in birds and mammals. *Ornithological science*, *5*(1), 5–14.

Jarvis, E. D. (2019). Evolution of vocal learning and spoken language. *Science*, *366*(6461), 50–54.

Jarvis, E D, Scharff, C., Grossman, M. R., Ramos, J. A., & Nottebohm, F. (1998). For whom the bird sings: context-dependent gene expression. *Neuron, 21*(4), 775–788.

Johnson, M., Aguilar de Soto, N., & Madsen, P. T. (2009). Studying the behaviour and sensory ecology of marine mammals using acoustic recording tags: a review. *Marine Ecology Progress Series*, *395*, 55–73.

Karita, S., Chen, N., Hayashi, T., Hori, T., Inaguma, H., Jiang, Z., Someki, M., et al. (2019). A comparative study on transformer vs RNN in speech applications. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 449–456). Presented at the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE.

Katz, J., Hafner, S. D., & Donovan, T. (2016). Tools for automated acoustic monitoring within the R package monitoR. *Bioacoustics*, *25*(2), 197–210.

Kaufman, L., & Rousseeuw, P. J. (Eds.). (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Kershenbaum, A., Blumstein, D. T., Roch, M. A., Akçay, Ç., Backus, G., Bee, M. A., Bohn, K., et al. (2016). Acoustic sequences in non-human animals: a tutorial review and prospectus. *Biological Reviews of the Cambridge Philosophical Society*, *91*(1), 13–52.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2019). Generalization through Memorization: Nearest Neighbor Language Models. *arXiv*.

King, S. L. (2015). You talkin' to me? Interactive playback is a powerful yet underused tool in animal communication research. *Biology Letters*, *11*(7).

Kollmorgen, S., Hahnloser, R. H. R., & Mante, V. (2020). Nearest neighbours reveal fast and slow components of motor learning. *Nature*, *577*(7791), 526–530.

Konishi, M. (1965). The role of auditory feedback in the control of vocalization in the white-crowned sparrow. *Zeitschrift fur Tierpsychologie*, *22*(7), 770–783.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 9096–9101.

Kuhl, P. K. (2007). Is speech learning "gated" by the social brain? *Developmental Science*, *10*(1), 110–120.

Kutylowski, J. (2017). DeepL Translator. URL: https://www.deepl.com/translator [retrieved on Dec 2022].

Law, G., & Kitchener, A. C. (2017). Environmental enrichment for Killer whales *Orcinus orca* at zoological institutions: untried and untested. *International Zoo Yearbook*, *51*(1), 232–247.

Lewis, R. N., Williams, L. J., & Gilman, R. T. (2021). The uses and implications of avian vocalizations for conservation planning. *Conservation Biology*, *35*(1), 50–63.

Liang, F., Das, V., Kostyuk, N., & Hussain, M. M. (2018). Constructing a Data-Driven Society: China's Social Credit System as a State Surveillance Infrastructure. *Policy & Internet*.

Lipkind, D., Geambasu, A., & Levelt, C. C. (2020). The development of structured vocalizations in songbirds and humans: A comparative analysis. *Topics in cognitive science*, *12*(3), 894–909.

Lipkind, D., Marcus, G. F., Bemis, D. K., Sasahara, K., Jacoby, N., Takahasi, M., Suzuki, K., et al. (2013). Stepwise acquisition of vocal combinatorial capacity in songbirds and human infants. *Nature*, *498*(7452), 104–108.

Lipkind, D., & Tchernichovski, O. (2011). Quantification of developmental birdsong learning from the subsyllabic scale to cultural evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 3*, 15572–15579.

Lipkind, D., Zai, A. T., Hanuschkin, A., Marcus, G. F., Tchernichovski, O., & Hahnloser, R. H. R. (2017). Songbirds work around computational complexity by learning song vocabulary independently of sequence. *Nature Communications*, *8*(1), 1247.

Liu, X., Gao, J., He, X., Deng, L., Duh, K., & Wang, Y.-Y. (2015). Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 912–921). Presented at the Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: Association for Computational Linguistics.

Ljubičić, I., Hyland Bruno, J., & Tchernichovski, O. (2016). Social influences on song learning. *Current Opinion in Behavioral Sciences*, *7*, 101–107.

Lorenz, C., Hao, X., Tomka, T., Ruettimann, L., & Hahnloser, R. (2022). Extracting extended vocal units from two neighborhoods in the embedding plane. *BioRxiv*.

MacDougall-Shackleton, S. A., Hulse, S. H., & Ball, G. F. (1998). Neural bases of song preferences in female zebra finches (Taeniopygia guttata). *Neuroreport*, *9*(13), 3047–3052.

Macpherson, T., Churchland, A., Sejnowski, T., DiCarlo, J., Kamitani, Y., Takahashi, H., & Hikida, T. (2021). Natural and Artificial Intelligence: A brief introduction to the interplay between AI and neuroscience research. *Neural Networks*, *144*, 603–613.

Marler, P., Dufty, A., & Pickert, R. (1986a). Vocal communication in the domestic chicken: I. Does a sender communicate information about the quality of a food referent to a receiver? *Animal Behaviour*, *34*, 188–193.

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

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.

Marques, T. A., Thomas, L., Martin, S. W., Mellinger, D. K., Ward, J. A., Moretti, D. J., Harris, D., et al. (2013). Estimating animal population density using passive acoustics. *Biological Reviews of the Cambridge Philosophical Society*, *88*(2), 287–309.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., & Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, *21*(9), 1281–1289.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*.

Mitra, B., & Craswell, N. (2017). Neural Models for Information Retrieval. *arXiv*.

Mizuhara, T., & Okanoya, K. (2020). Do songbirds hear songs syllable by syllable? *Behavioural Processes*, *174*, 104089.

Mooney, R. (2009). Neural mechanisms for learned birdsong. *Learning & Memory*, *16*(11), 655–669.

Morita, T., Koda, H., Okanoya, K., & Tachibana, R. O. (2021). Measuring context dependency in birdsong using artificial neural networks. *PLoS Computational Biology*, *17*(12), e1009707.

Morris, D. (1954). The Reproductive Behaviour of the Zebra Finch (Poephila guttata), with Special Reference to Pseudofemale Behaviour and Displacement Activities on JSTOR. *Behaviour*, *6*(4), 271–322.

Mouterde, S. C., Theunissen, F. E., Elie, J. E., Vignal, C., & Mathevon, N. (2014). Acoustic communication and sound degradation: how do the individual signatures of male and female zebra finch calls transmit over distance? *Plos One*, *9*(7), e102842.

Muja, M., & Lowe, D. (2009). FAST APPROXIMATE NEAREST NEIGHBORS WITH AUTOMATIC ALGORITHM CONFIGURATION. *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications* (pp. 331–340). Presented at the International Conference on Computer Vision Theory and Applications, SciTePress - Science and and Technology Publications.

Nassif, A. B., Shahin, I., Attili, I., Azzeh, M., & Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. *IEEE access : practical innovations, open solutions*, *7*, 19143–19165.

Nordeen, K. W., & Nordeen, E. J. (1988). Projection neurons within a vocal motor pathway are born during song learning in zebra finches. *Nature*, *334*(6178), 149–151.

Nottebohm, F, & Arnold, A. P. (1976). Sexual dimorphism in vocal control areas of the songbird brain. *Science*, *194*(4261), 211–213.

Nottebohm, F. (1972). The origins of vocal learning. *The American Naturalist*, *106*(947), 116–140.

Okanoya, K. (2002). Sexual display as a syntactical vehicle: the evolution of syntax in birdsong and human language through sexual selection. In A. Wray (Ed.), *Transition to Language* (pp. 46–63).

Okobi, D. E., Banerjee, A., Matheson, A. M. M., Phelps, S. M., & Long, M. A. (2019). Motor cortical control of vocal interaction in neotropical singing mice. *Science*, *363*(6430), 983–988.

Okubo, T. S., Mackevicius, E. L., Payne, H. L., Lynch, G. F., & Fee, M. S. (2015). Growth and splitting of neural sequences in songbird vocal development. *Nature*, *528*(7582), 352–357.

Olveczky, B. P., Andalman, A. S., & Fee, M. S. (2005). Vocal experimentation in the juvenile songbird requires a basal ganglia circuit. *PLoS Biology*, *3*(5), e153.

Pearre, B., Perkins, L. N., Markowitz, J. E., & Gardner, T. J. (2017). A fast and accurate zebra finch syllable detector. *Plos One*, *12*(7), e0181992.

Pearre, B. (2017). Zebra Finch Syllable Detector. *OSF*.

Perez, E. C., Elie, J. E., Boucaud, I. C. A., Crouchet, T., Soulage, C. O., Soula, H. A., Theunissen, F. E., et al. (2015). Physiological resonance between mates through calls as possible evidence of empathic processes in songbirds. *Hormones and Behavior*, *75*, 130–141.

Perez, E. C., Elie, J. E., Soulage, C. O., Soula, H. A., Mathevon, N., & Vignal, C. (2012). The acoustic expression of stress in a songbird: does corticosterone drive isolation-induced modifications of zebra finch calls? *Hormones and Behavior*, *61*(4), 573–581.

Perez, E. C., Fernandez, M. S. A., Griffith, S. C., Vignal, C., & Soula, H. A. (2015). Impact of visual contact on vocal interaction dynamics of pair-bonded birds. *Animal Behaviour*, *107*, 125–137.

Petkov, C. I., & Jarvis, E. D. (2012). Birds, primates, and spoken language origins: behavioral phenotypes and neurobiological substrates. *Frontiers in evolutionary neuroscience*, *4*, 12.

Pika, S., Wilkinson, R., Kendrick, K. H., & Vernes, S. C. (2018). Taking turns: bridging the gap between human and animal communication. *Proceedings. Biological Sciences / the Royal Society*, *285*(1880).

Plato. (1925). Symposium. In H. N. Fowler (Tran.), *Plato in Twelve Volumes* (Vol. 9). Cambridge / London: Harvard University Press / W. Heinemann Ltd.

Ravbar, P., Lipkind, D., Parra, L. C., & Tchernichovski, O. (2012). Vocal exploration is locally regulated during song learning. *The Journal of Neuroscience*, *32*(10), 3422–3432.

Richards, N. M. (2013). The Dangers of Surveillance. *Harvard Law Review*.

Riede, T., Schilling, N., & Goller, F. (2013). The acoustic effect of vocal tract adjustments in zebra finches. *Journal of Comparative Physiology. A, Neuroethology, Sensory, Neural, and Behavioral Physiology*, *199*(1), 57–69.

Roberts, T. F., Gobes, S. M. H., Murugan, M., Ölveczky, B. P., & Mooney, R. (2012). Motor circuits are required to encode a sensory model for imitative learning. *Nature Neuroscience*, *15*(10), 1454–1459.

Robinson, J. G. (1979). An analysis of the organization of vocal communication in the titi monkey Callicebus moloch. *Zeitschrift fur Tierpsychologie*, *49*(4), 381–405.

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., et al. (2019). Decline of the North American avifauna. *Science*, *366*(6461), 120–124.

Rüttimann, L., Rychen, J., Tomka, T., Hörster, H., Rocha, M. D., & Hahnloser, R. H. (2022). Multimodal system for recording individual-level behaviors in songbird groups. *BioRxiv*.

Rychen, J., Rodrigues, D. I., Tomka, T., Rüttimann, L., Yamahachi, H., & Hahnloser, R. H. R. (2021). A system for controlling vocal communication networks. *Scientific Reports*, *11*(1), 11099.

Sainburg, T., & Gentner, T. Q. (2021). Toward a computational neuroethology of vocal communication: from bioacoustics to neurophysiology, emerging tools and future directions. *Frontiers in Behavioral Neuroscience*, *15*, 811737.

Sainburg, T., Thielk, M., & Gentner, T. Q. (2020). Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, *16*(10), e1008228.

Saravanan, V., Berman, G. J., & Sober, S. J. (2020). Application of the hierarchical bootstrap to multi-level data in neuroscience. *Neurons, behavior, data analysis and theory*, *3*(5).

Scalera, A., & Tomaszycki, M. L. (2018). Acute exogenous corticosterone treatments have few effects on courtship and pair bonding in zebra finches. *General and Comparative Endocrinology*, *268*, 121–127.

Scherer, K. R., Johnstone, T., & Klasmeyer, G. (2003). Vocal expression of emotion. *Handbook of affective sciences* (pp. 433–456). Oxford University Press.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of Voice*, *9*(3), 235–248.

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

Shannon, C. E., Weaver, W., Blahut, R. E., & Hajek, B. (1949). *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Sharma, A., & Suryawanshi, A. (2016). A Novel Method for Detecting Spam Email using KNN Classification with Spearman Correlation as Distance Measure. *International Journal of Control and Automation*, *136*(6), 28–35.

Shen, J.-X., & Xu, Z.-M. (2016). The Lombard effect in male ultrasonic frogs: Regulating antiphonal signal frequency and amplitude in noise. *Scientific Reports*, *6*, 27103.

Slobodchikoff, C. N., Kiriazis, J., Fischer, C., & Creef, E. (1991). Semantic information distinguishing individual predators in the alarm calls of Gunnison's prairie dogs. *Animal Behaviour*, *42*(5), 713–719.

Smith, T. G. L. H. A. P. (1965). Communication between Dolphins in Separate Tanks by Way of an Electronic Acoustic Link. *Science*, *150*(August), 1839.

Sober, S. J., & Brainard, M. S. (2009). Adult birdsong is actively maintained by error correction. *Nature Neuroscience*, *12*(7), 927–931.

Sossinka, R., & Böhner, J. (1980). Song Types in the Zebra Finch *Poephila guttata castanotis*. *Zeitschrift für Tierpsychologie*, *53*(2), 123–132.

Spearman, C. (1906). 'FOOTRULE' FOR MEASURING CORRELATION. *British Journal of Psychology, 1904-1920*, *2*(1), 89–108.

Steinfath, E., Palacios-Muñoz, A., Rottschäfer, J. R., Yuezak, D., & Clemens, J. (2021). Fast and accurate annotation of acoustic signals with deep neural networks. *eLife*, *10*.

Stidsholt, L., Johnson, M., Beedholm, K., Jakobsen, L., Kugler, K., Brinkløv, S., Salles, A., et al. (2019). A 2.6-g sound and movement tag for studying the acoustic scene and kinematics of echolocating bats. *Methods in Ecology and Evolution*, *10*(1), 48–58.

Stowell, D., Gill, L., & Clayton, D. (2016). Detailed temporal structure of communication networks in groups of songbirds. *Journal of the Royal Society, Interface*, *13*(119).

Stowell, D., Wood, M., Stylianou, Y., & Glotin, H. (2016). Bird detection in audio: A survey and a challenge. *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1–6). Presented at the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), IEEE.

Suzuki, R., Buck, J. R., & Tyack, P. L. (2006). Information entropy of humpback whale songs. *The Journal of the Acoustical Society of America*, *119*(3), 1849–1866.

Suzuki, T. N. (2016). Semantic communication in birds: evidence from field research over the past two decades. *Ecological research*, *31*(3), 307–319.

Swaddle, J. P., McBride, L., & Malhotra, S. (2006). Female zebra finches prefer unfamiliar males but not when watching noninteractive video. *Animal Behaviour*, *72*(1), 161–167.

Takahashi, D. Y., Liao, D. A., & Ghazanfar, A. A. (2017). Vocal learning via social reinforcement by infant marmoset monkeys. *Current Biology*, *27*(12), 1844-1852.e6.

Takahashi, D. Y., Narayanan, D. Z., & Ghazanfar, A. A. (2013). Coupled oscillator dynamics of vocal turn-taking in monkeys. *Current Biology*, *23*(21), 2162–2168.

Tchernichovski, O., Lints, T., Mitra, P. P., & Nottebohm, F. (1999). Vocal imitation in zebra finches is inversely related to model abundance. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(22), 12901–12904.

Tchernichovski, O., Mitra, P. P., Lints, T., & Nottebohm, F. (2001). Dynamics of the vocal imitation process: how a zebra finch learns its song. *Science*, *291*(5513), 2564–2569.

Tchernichovski, O., Nottebohm, F., Ho, C. E., Pesaran, B., & Mitra, P. P. (2000). A procedure for an automated measurement of song similarity. *Animal behaviour*, *59*(6), 1167–1176.

Tchernichovski, O., & Nottebohm, F. (1998). Social inhibition of song imitation among sibling male zebra finches. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8951–8956.

Templeton, C. N., Greene, E., & Davis, K. (2005). Allometry of alarm calls: black-capped chickadees encode information about predator size. *Science*, *308*(5730), 1934–1937.

Ter Maat, A., Trost, L., Sagunsky, H., Seltmann, S., & Gahr, M. (2014). Zebra finch mates use their forebrain song system in unlearned call communication. *Plos One*, *9*(10), e109334.

Theunissen, F. E., Mouterde, S., & Mathevon, N. (2013). A neuroethological analysis of the information in propagated communication calls. Proceedings of meetings on acoustics (pp. 010032–010032). Presented at the ICA 2013 Montreal, ASA.

Tian, L. Y., & Brainard, M. S. (2017). Discrete Circuits Support Generalized versus Context-Specific Vocal Learning in the Songbird. *Neuron*, *96*(5), 1168-1177.e5.

Toutounji, H., Zai, A. T., Tchernichovski, O., Hahnloser, R. H. R., & Lipkind, D. (2022). Learning the action inventory of a complex skill via an intrinsic reward. *BioRxiv*.

Tumer, E. C., & Brainard, M. S. (2007). Performance variability enables adaptive plasticity of “crystallized” adult birdsong. *Nature*, *450*(7173), 1240–1244.

Tyack, P. L. (2020). A taxonomy for vocal learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1789), 20180406.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., et al. (2017). Attention Is All You Need. *arXiv*.

Vignal, C., Mathevon, N., & Mottin, S. (2004). Audience drives male songbird response to partner's voice. *Nature*, *430*(6998), 448–451.

Volman, S. F., & Khanna, H. (1995). Convergence of untutored song in group-reared zebra finches (Taeniopygia guttata). *Journal of Comparative Psychology*, *109*(3), 211–221.

Wang, D., & Zheng, T. F. (2015). Transfer learning for speech and language processing. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1225–1237). Presented at the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), IEEE.

Warren, T. L., Tumer, E. C., Charlesworth, J. D., & Brainard, M. S. (2011). Mechanisms and time course of vocal learning and consolidation in the adult songbird. *Journal of Neurophysiology*, *106*(4), 1806–1821.

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., et al. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762.

Xie, Y., Wang, Y., Nallanathan, A., & Wang, L. (2016). An Improved K-Nearest-Neighbor Indoor Localization Method Based on Spearman Distance. *IEEE signal processing letters*, *23*(3), 351–355.

Yu, J., Li, J., Yu, Z., & Huang, Q. (2020). Multimodal Transformer With Multi-View Visual Representation for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(12), 4467–4480.

Zann, R. A. (1996). *The Zebra Finch: A Synthesis of Field and Laboratory Studies*. Oxford University Press.

Zhao, W., Garcia-Oscos, F., Dinh, D., & Roberts, T. F. (2019). Inception of memories that guide vocal learning in the songbird. *Science*, *366*(6461), 83–89.

Zuberbühler, K., Cheney, D. L., & Seyfarth, R. M. (1999). Conceptual semantics in a nonhuman primate. *Journal of Comparative Psychology*, *113*(1), 33–42.

Zuboff, S. (2015). Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, *30*(1), 75–89.