Aspects of Room Acoustics, Vision and Motion in the Human Auditory Perception of Space

Michael Schutte

Graduate School of
Systemic Neurosciences
LMU Munich

Dissertation at the
Graduate School of Systemic Neurosciences
Ludwig-Maximilians-Universität München

München, 23 April 2021

SUPERVISOR
Prof. Dr. Benedikt Grothe
Division of Neurobiology
Department Biology II
Ludwig-Maximilians-Universität München

FIRST REVIEWER Prof. Dr. Benedikt Grothe
SECOND REVIEWER Prof. Dr. Christian Leibold

DATE OF SUBMISSION 23 April 2021
DATE OF DEFENSE 12 July 2021

Irgendwas mit Hören

Michael Schutte

This dissertation is dedicated

to Tom and to Lutz

who would have liked each other

Abstract

The human sense of hearing contributes to the awareness of where sound-generating objects are located in space and of the environment in which the hearing individual is located. This auditory perception of space interacts in complex ways with our other senses, can be both disrupted and enhanced by sound reflections, and includes safety mechanisms which have evolved to protect our lives, but can also mislead us. This dissertation explores some selected topics from this wide subject area, mostly by testing the abilities and subjective judgments of human listeners in virtual environments.

Reverberation is the gradually decaying persistence of sounds in an enclosed space which results from repeated sound reflections at surfaces. The first experiment *(Chapter 2)* compared how strongly people perceived reverberation in different visual situations: when they could see the room and the source which generated the sound; when they could see some room and some sound source, but the image did not match what they heard; and when they could not see anything at all. There were no indications that the visual image had any influence on this aspect of room-acoustical perception.

The potential benefits of motion for judging the distance of sound sources were the focus of the second study *(Chapter 3)*, which consists of two parts. In the first part, loudspeakers were placed at different depths in front of sitting listeners who, on command, had to either remain still or move their upper bodies sideways. This experiment demonstrated that humans can exploit motion parallax (the effect that closer objects appear faster to a moving observer than farther objects) with their ears and not just with their eyes. The second part combined a virtualisation of such sound sources with a motion platform to show that the listeners' interpretation of this auditory motion parallax was better when they performed this lateral movement by themselves, rather than when they were moved by the apparatus or were not actually in motion at all.

Two more experiments were concerned with the perception of sounds which are perceived as becoming louder over time. These have been called "looming", as the source of such a sound might be on a collision course. One of the studies *(Chapter 4)* showed that western diamondback rattlesnakes *(Crotalus atrox)* increase the vibration speed of their rattle in response to the approach of a threatening object. It also demonstrated that human listeners perceive (virtual) snakes which engage in this behaviour as especially close, causing them to keep a greater margin of safety than they would otherwise. The other study *(section 5.6)* was concerned with the well-known looming bias of the sound localisation system, a phenomenon which leads to a sometimes exaggerated, sometimes more accurate perception of approaching compared to receding sounds. It attempted to find out whether this bias is affected by whether listeners hear such sounds in a virtual enclosed space or in an environment with no sound reflections. While the results were inconclusive, this experiment is noteworthy as a proof of concept: It was the first study to make use of a new real-time room-acoustical simulation system, liveRAZR, which was developed as part of this dissertation *(Chapter 5)*.

Finally, while humans have been more often studied for their unique abilities to communicate with each other and bats for their extraordinary capacity to locate objects by sound, this dissertation turns this setting of priorities on its head with the last paper *(Chapter 6)*: Based on recordings of six pale spear-nosed bats *(Phyllostomus discolor)*, it is a survey of the identifiably distinct vocalisations observed in their social interactions, along with a description of the different situations in which they typically occur.

Zusammenfassung

Das menschliche Gehör trägt zum Bewusstsein dafür bei, wo sich schallerzeugende Objekte im Raum befinden und wie die Umgebung beschaffen ist, in der sich eine Person aufhält. Diese auditorische Raumwahrnehmung interagiert auf komplexe Art und Weise mit unseren anderen Sinnen, kann von Schallreflektionen sowohl profitieren als auch durch sie behindert werden, und besitzt Mechanismen welche evolutionär entstanden sind, um unser Leben zu schützen, uns aber auch irreführen können. Diese Dissertation befasst sich mit einigen ausgewählten Themen aus diesem weiten Feld und stützt sich dabei meist auf die Testung von Wahrnehmungsfähigkeiten und subjektiver Einschätzungen menschlicher Hörer/-innen in virtueller Realität.

Beim ersten Experiment *(Kapitel 2)* handelte es sich um einen Vergleich zwischen der Wahrnehmung von Nachhall, dem durch wiederholte Reflexionen an Oberflächen hervorgerufenen, sukzessiv abschwellenden Verbleib von Schall in einem umschlossenen Raum, unter verschiedenen visuellen Umständen: wenn die Versuchsperson den Raum und die Schallquelle sehen konnte; wenn sie irgendeinen Raum und irgendeine Schallquelle sehen konnte, dieses Bild aber vom Schalleindruck abwich; und wenn sie gar kein Bild sehen konnte. Dieser Versuch konnte keinen Einfluss eines Seheindrucks auf diesen Aspekt der raumakustischen Wahrnehmung zu Tage fördern.

Mögliche Vorteile von Bewegung für die Einschätzung der Entfernung von Schallquellen waren der Schwerpunkt der zweiten Studie *(Kapitel 3)*. Diese bestand aus zwei Teilen, wovon der erste zeigte, dass Hörer/-innen, die ihren Oberkörper relativ zu zwei in unterschiedlichen Abständen vor ihnen aufgestellten Lautsprechern auf Kommando entweder stillhalten oder seitlich bewegen mussten, im letzteren Falle von der Bewegungsparallaxe (dem Effekt, dass sich der nähere Lautsprecher relativ zum sich bewegenden Körper schneller bewegte als der weiter entfernte) profitieren konnten. Der zweite Teil kombinierte eine Simulation solcher Schallquellen mit einer Bewegungsplattform, wodurch gezeigt werden konnte, dass die bewusste Eigenbewegung für die Versuchspersonen hilfreicher war, als durch die Plattform bewegt zu werden oder gar nicht wirklich in Bewegung zu sein.

Zwei weitere Versuche gingen auf die Wahrnehmung von Schallen ein, deren Ursprungsort sich nach und nach näher an den/die Hörer/-in heranbewegte. Derartige Schalle werden auch als „looming" („anbahnend") bezeichnet, da eine solche Annäherung bei bedrohlichen Signalen nichts Gutes ahnen lässt. Einer dieser Versuche *(Kapitel 4)* zeigte zunächst, dass Texas-Klapperschlangen *(Crotalus atrox)* die Vibrationsgeschwindigkeit der Schwanzrassel steigern, wenn sich ein bedrohliches Objekt ihnen nähert. Menschliche Hörer/-innen nahmen (virtuelle) Schlangen, die dieses Verhalten aufweisen, als besonders nahe wahr und hielten einen größeren Sicherheitsabstand ein, als sie es sonst tun würden. Der andere Versuch *(Abschnitt 5.6)* versuchte festzustellen, ob die wohlbekannte Neigung unserer Schallwahrnehmung, näherkommende Schalle manchmal übertrieben und manchmal genauer einzuschätzen als sich entfernende, durch Schallreflektionen beeinflusst werden kann. Diese Ergebnisse waren unschlüssig, jedoch bestand die Besonderheit dieses Versuchs darin, dass er erstmals ein neues Echtzeitsystem zur Raumakustiksimulation (liveRAZR) nutzte, welches als Teil dieser Dissertation entwickelt wurde *(Kapitel 5)*.

Abschließend *(Kapitel 6)* wird die Schwerpunktsetzung auf den Kopf gestellt, nach der Menschen öfter auf ihre einmaligen Fähigkeiten zur Kommunikation miteinander untersucht werden und Fledermäuse öfter auf ihre außergewöhnliches Geschick, Objekte durch Schall zu orten: Anhand von Aufnahmen von sechs Kleinen Lanzennasen *(Phyllostomus discolor)* fasst das Kapitel die klar voneinander unterscheidbaren Laute zusammen, die diese Tiere im sozialen Umgang miteinander produzieren, und beschreibt, in welchen Situationen diese Lauttypen typischerweise auftreten.

Contents

Abstract i

Zusammenfassung iii

Contents v

List of figures vii

List of tables ix

List of abbreviations xi

Chapter 1
Introduction 1

1.1 Acoustics, room acoustics and human hearing . . . 1
1.2 Sound localisation . . . 5
1.3 Digital processing of sound . . . 11
1.4 Perception of reflections and room acoustics . . . 14
1.5 Multimodal aspects of the auditory perception of space . . . 19
1.6 Acoustic communication within and between animal species . . . 22
1.7 Overview . . . 24

Chapter 2
Stationary listeners & stationary sources 27

2.0 Author contributions . . . 27
2.1 Introduction . . . 28
2.2 Methods . . . 29
2.3 Results . . . 31
2.4 Discussion . . . 32
2.5 Acknowledgments . . . 33
2.6 References and links . . . 33

Chapter 3
Moving listeners & stationary sources 35

3.0 Author contributions . . . 35
3.1 Introduction . . . 36
3.2 Results . . . 37
3.3 Discussion . . . 38
3.4 Materials and Methods . . . 40
3.5 Acknowledgments . . . 41
3.6 References . . . 41
3.7 Supporting Information . . . 42

Chapter 4
Stationary listeners & moving sources **45**

4.0 Author contributions . . . 45
4.1 Main text . . . 46
4.2 Supplementary materials and methods . . . 53

Chapter 5
Moving listeners & moving sources **61**

5.1 Introduction . . . 61
5.2 The RAZR model . . . 69
5.3 liveRAZR: A real-time implementation of RAZR . . . 74
5.4 Verification of liveRAZR . . . 81
5.5 Runtime analysis of liveRAZR . . . 83
5.6 An application of liveRAZR: Looming in rooms . . . 85
5.7 Discussion . . . 89

Chapter 6
Listeners & sources with moving ears **91**

6.0 Author contributions . . . 91
6.1 Introduction . . . 92
6.2 Materials and methods . . . 93
6.3 Results . . . 96
6.4 Discussion . . . 100
6.5 Conclusions . . . 102
6.6 Funding . . . 103
6.7 Acknowledgments . . . 103
6.8 References . . . 103
6.9 Supplementary material . . . 106

Chapter 7
Discussion **111**

7.1 Vision and the perception of room acoustics . . . 111
7.2 Translational motion and the auditory perception of distance . . . 113
7.3 Stimulus characteristics and the auditory perception of looming sounds . . . 116
7.4 Social communication in bats and the study of vocal learning . . . 118
7.5 Room-acoustical simulation in real time and a brief outlook . . . 118

Bibliography **121**

Acknowledgments **149**

List of publications **151**

Copyright and terms of use **153**

Affidavit **155**

Author contributions **157**

List of figures

Chapter 1
Introduction

1.1 Bregman's (1994) analogy for auditory scene analysis 1
1.2 Anatomy of the human ear 4
1.3 Simplified circuit diagram of the parts of the ascending auditory pathway with relevance to sound localisation 6
1.4 Cues for sound localisation in azimuth and elevation 7
1.5 Level and direct-to-reverberant ratio cues for sound localisation in depth 9
1.6 Two examples for the perceptual suppression of reflected sounds 17

Chapter 2
Stationary listeners & stationary sources

2.1 Stimuli and experimental setup 30
2.2 Individual subjects' reverberation ratings for the auditory room identities bedroom, office and factory 31
2.3 Paired comparisons between each subject's average of five ratings in audiovisually congruent conditions vs otherwise identical conditions which differ visually in one aspect 33

Chapter 3
Moving listeners & stationary sources

3.1 Illustration of the experimental setup, the stimuli, and psychophysical results to demonstrate auditory motion parallax in Exp. I 37
3.2 Illustration of the setup and paradigm of Exp. II and the hypothesis. 38
3.3 Exemplary performance and fitted psychometric functions for depth discrimination of two alternating sound sources as a function of their distance difference in Exp. II 38
3.4 Psychophysical performance thresholds for sound-source distance in Exp. II . . . 39
3.S1 Tracks of horizontal and vertical head motion of the subject relative to the stationary sound sources 43

Chapter 4
Stationary listeners & moving sources

4.1 Acoustic properties of rattlesnake rattling 47
4.2 Effect of constant approach velocity on low and high modulation frequency modes of rattling responses 48
4.3 Psychophysical experiments in a virtual reality environment reveal that adaptive rattling generates an underestimation of distance in human subjects 50
4.S1 Spectrogram of a rattlesnake acoustic display evoked by an artificial human torso which was moved towards the snake 54

4.S2 Experimental design 55
4.S3 Response variability to looming stimuli across and within snakes 56
4.S4 Effect of decreasing approach velocity on the low and high frequency phases of the rattling responses 59
4.S5 Individual human psychoacoustical data for all even subjects 60

Chapter 5
Moving listeners & moving sources

5.1 Illustration of ray tracing 62
5.2 Illustration of beam tracing 63
5.3 Illustration of the image-source model 64
5.4 Block diagram of the signal processing related to the image-source model (ISM) implemented in RAZR 69
5.5 Block diagram of the feedback-delay network (FDN) implemented in RAZR. . . . 71
5.6 Positions assigned to the twelve default FDN channels in RAZR 72
5.7 Block diagram of the mapping from ISM to FDN 73
5.8 Synoptic block diagram of liveRAZR 75
5.9 Chart of the fast convolution process as implemented in the array and merging spatialisers in liveRAZR 80
5.10 Comparison between the impulse responses generated by RAZR and liveRAZR for two identical configurations 82
5.11 Runtimes of liveRAZR relative to 1 s of input signal as a function of the number of virtual sound sources, for various spatialiser configurations. 84
5.12 Results of a linear model fitted to the runtime data presented in Figure 5.11 84
5.13 Stimulation in the "looming in rooms" 2-AFC experiment 86
5.14 Results from the "looming in room" experiment 88

Chapter 6
Listeners & sources with moving ears

6.1 Schematic of the setup 94
6.2 Example syllables from the eight commonly occurring classes 97
6.3 Boxplots of four selected spectral and temporal parameters 98
6.4 Confusion matrix depicting the distinguishability of the suppressed fundamental frequency class and the eight commonly occurring syllable classes 98
6.5 Behavioral context of the common syllable classes 99
6.6 Examples of syllables from rarely occurring classes 100
6.7 Spectral centroid frequencies of all analyzed syllables 100
6.8 Spectrograms of syllables with a suppressed fundamental frequency resembling syllables from other classes 101
6.9 Example spectrograms from three syllable trains 101
6.S1 Syllable diversity in the commonly occurring syllable classes 106

List of tables

Chapter 4
Stationary listeners & moving sources

4.S1 Summary of average slopes and durations of low-frequency and high-frequency components of the rattling behavior elicited by the visual looming stimulus with a constant velocity profile at four different approach velocities 57
4.S2 Summary of average slopes and durations of low-frequency and high-frequency components of the rattling behavior elicited by the visual looming stimulus with a decreasing velocity profile at three different approach velocities. 57

Chapter 5
Moving listeners & moving sources

5.1 Specifications of the rooms which were used for the verification of liveRAZR based on a comparison to RAZR 82

Chapter 6
Listeners & sources with moving ears

6.S1 Measured and calculated acoustic parameters of the common syllable classes . . . 107
6.S2 Measured and calculated acoustic parameters of the rare syllable classes and the suppressed fundamental frequency class 108
6.S3 Behavioral contexts scored for 20 syllables per class 109

List of abbreviations

2-AFC	2-alternative forced choice
AM	amplitude modulation; in Chapter 3: active motion
ANOVA	analysis of variance
ASIO	Audio Stream Input/Output (Steinberg Media Technologies GmbH, Hamburg, Germany)
CN	cochlear nucleus
CS	composite syllable
DCN	dorsal cochlear nucleus
DFT	discrete Fourier transform
DNLL	dorsal nucleus of the lateral lemniscus
DRR	direct-to-reverberant ratio
EEG	electroencephalogram
EMM	estimated marginal mean
FDN	feedback-delay network
FFT	fast Fourier transform
FIR	finite impulse response
HE	high entropy
HF	high frequency
hFM	hooked frequency modulated
HOB	head-orientation benefit
HRTF	head-related transfer function
IACC	interaural cross-correlation
IC	inferior colliculus
IIR	infinite impulse response
ILD	interaural level difference
IR	impulse response
ISM	image-source model
ITD	interaural time difference
JND	just-noticeable difference
LDA	linear discriminant analysis
lDFM	linearly downward frequency modulated
LF	low frequency
LNTB	lateral nucleus of the trapezoid body
LR	linear regression
LSO	lateral superior olive
LTI	linear and time-invariant
MMAA	minimum moving audible angle
MNTB	medial nucleus of the trapezoid body
MRI	magnetic resonance imaging
MSO	medial superior olive
nlDFM	non-linearly downward frequency modulated
NM	no motion
OSC	Open Sound Control
PF	peak frequency
PM	passive motion
PSD	power spectral density
QR	quadratic regression
qCF	quasi-constant frequency (_so: with a steep onset; _n: noisy; _nso: noisy and with a steep onset)
RMSE	root-mean-square error
RT_{60}	reverberation time
SC	superior colliculus
SCF	spectral centroid frequency
SFM	sinusoidally frequency modulated
SF	suppressed fundamental
SPL	sound pressure level
SSM	sound-source motion
TV	television
VAE	virtual acoustic environment
VAS	virtual acoustic space
VBAP	vector-base amplitude panning
VCN	ventral cochlear nucleus
VR	virtual reality

CHAPTER I

Introduction

This dissertation in auditory neuroscience explores some selected facets of how the human sense of hearing manages to turn the simple one-dimensional vibrations of the two eardrums into a three-dimensional perception (as suggested by Figure 1.1). It inevitably touches on a range of subjects beyond this field, particularly on aspects of physics and computing. Rather than just summarising the scientific state of the art, this first chapter is therefore also an attempt to introduce readers who might mostly be familiar with one field to fundamental concepts of the others.[1]

1.1 Acoustics, room acoustics and human hearing

Sound is an oscillation in pressure which propagates through a gas, liquid or solid. Depending on the *frequency* (number of oscillation cycles per unit of time) and *amplitude* (amount of the pressure change) of such an oscillation, sound can be heard by animals. In the context of human hearing, the propagation medium is usually air. The frequency range of human hearing is often stated

[1]Note that this introductory chapter does not report on any original research by the author. Wherever I give no explicit citation, any definition or statement concerning the fundamentals of acoustics can be found in a suitable textbook such as Kuttruff (2007). Similarly, for well-known facts regarding the analysis, synthesis and manipulation of sound with a computer, the reader is referred to textbooks on digital audio signal processing like Zölzer (2008).

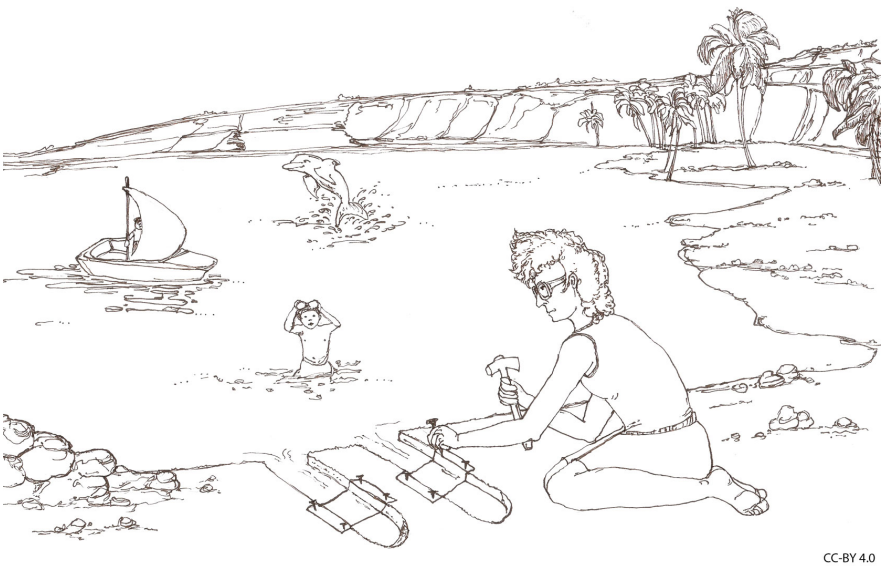

Figure 1.1 Bregman (1994) has compared the human capabilities of *auditory scene analysis* with telling the number, locations and properties of the objects on a lake, just by observing the motion of two handkerchiefs (in analogy to the left and right eardrums) stretched across two small channels at the lakeside. Drawing by an anonymous artist, provided by Fabian Brinkmann (2019), reproduced here in accordance with the Creative Commons Attribution 4.0 License.

as 20–20 000 Hz, although with sufficiently high amplitudes, a human listener may still be able to hear sounds that lie substantially outside this range (Ashihara, 2007; Whittle *et al.*, 1972). The exact bounds also vary between individuals, as well as over the lifespan of an individual due to age and disease (Gates and J. H. Mills, 2005).

The pressure changes required for hearing are minute: One metre away from a person talking relatively loudly, for example, the air is rapidly compressed and decompressed by approximately 0.2 parts per million, a sound pressure of 20 mPa added to the static air pressure of 1013.25 hPa. For convenience, sound pressure is usually expressed as a scaled logarithm of the ratio of the pressure relative to $p_0 = 20$ µPa, resulting in the so-called *sound pressure level* in decibels (dB):

$$L_p = 20 \log_{10} \frac{p}{p_0}$$

The exemplary 20 mPa are 1000 times higher than this reference, a level of $20 \log_{10} 10^3 = 60$ dB. Because L_p may be calculated relative to any reference pressure, the suffix notation SPL is frequently used to indicate that $p_0 = 20$ µPa, as in $L_p = 60$ dB SPL.

1.1.1 Propagation delay, geometric attenuation and atmospheric absorption

In air, sound waves travel with a speed of approximately 343 m/s, or about 1 m per 2.9 ms (at 20 °C). In other words, each metre of distance between a sound source and a receiver (such as a human listener) delays the sound by 2.9 ms. Moreover, due to the spherical spread of sound waves in air, the energy emitted by the source is geometrically "diluted" more and more with increasing distance from the source. In terms of sound pressure, this effect is described by the inverse distance law

$$p \propto \frac{1}{r}$$

where r is the distance between the source and the receiver. It means that for every doubling of distance, the sound pressure arriving at the receiver is halved—or, equivalently, the sound pressure level is changed by $20 \log_{10} \frac{1}{2} \approx -6$ dB.

The propagation of sound through a medium is also subject to losses. Air absorbs sound due to friction and relaxation, processes which depend on atmospheric conditions and are summarised as atmospheric absorption. Per ISO standard 9613-1 (International Organization for Standardization, 1993), at standard room conditions, every 100 m of distance lead to an attenuation of 10 dB for sound frequencies above 8 kHz. A physical model of atmospheric absorption as a function of static air pressure, humidity and sound frequency was developed by Bass *et al.* (1995).

1.1.2 Reflections and reverberation

In many everyday environments, sound waves emitted by a source arrive at the location of a receiver not only via the direct path between them, but also after bouncing off of reflective surfaces. These can be walls, the floor and ceiling of a room, objects such as furniture, parts of a listener's own or other individuals' bodies, *etc*. Since any propagation path which includes reflective surfaces is inevitably longer than the direct path, these reflected waves arrive at the receiver later and with greater attenuation than the unreflected sound. Moreover, real-world objects are not ideal reflectors: A portion of the sound energy is absorbed instead. This process depends on sound frequency. The walls of a room, for example, usually absorb more energy at higher than at lower frequencies. This effect can be described by frequency-dependent *absorption* or, inversely, *reflection coefficients*.

Sound waves in a room that have only been reflected a few, perhaps two or three times are called *early reflections*. These are special in that they are relatively distinct: When a brief click sound is emitted in an enclosed space and the response of the room is recorded with a microphone, they can

be seen in the waveform as clear peaks which appear like delayed and downscaled copies of the direct sound at the beginning of the recording. As these early reflections are repeatedly reflected further, the sound captured by the microphone becomes more and more stochastic at later points in time. The individual contributions of the many propagation paths of the sound signal through space can no longer be separated, resulting in a continuous *reverberation* that decays over time due to absorption.

The time it takes for the response of a room to decrease by 60 dB in level after the sound source is turned off is an important room-acoustic characteristic, known as *reverberation time* or RT_{60}. Wallace C. Sabine (in the late 19th century) and Eyring (1930) have both presented formulas which allow the estimation of reverberation time given the absorption coefficients of the surfaces in a room, as have other authors. In an influential paper, Schroeder (1965) has described a method for measuring it. Another key quantity is the *direct-to-reverberant ratio* or DRR, which indicates (usually in dB) the amplitude of the direct sound in relation to the room response. Because the inverse distance law holds for direct sound, but not for reverberation, the DRR depends not only on the room, but also on the distance between source and receiver. This means that the reverberant sound becomes relatively more prominent as a sound source is moved further away from a receiver, such that the DRR decreases.

1.1.3 Frequency spectrum

A *pure tone* is a sound in which the relationship between time and amplitude can be described by a sine wave. Human listeners perceive a pure tone as having a very clear *pitch* related to its frequency. It is possible to represent any sound as a superposition (sum) of such sinusoids with different frequencies. Each constituent sinusoid can then be fully characterised by two properties, its *magnitude* (related to the amplitude) and its *phase offset* (describing how much it is shifted in time). When these properties are analysed for an arbitrary sound as functions of frequency, by effectively decomposing the sound into all the sinusoids it contains, these functions are called the *magnitude spectrum* and *phase spectrum*, respectively.

Spectrograms are often used to visualise the variation of a sound over time. These two-dimensional diagrams have one axis representing time and the other axis representing frequency, while the magnitude[2] at a given time and frequency is indicated by a colour or greyscale. They can be understood as many short-time spectrums, stacked (usually) horizontally and displayed in a visually compressed manner.

The lowest frequency for which a sinusoid is present is the frequency at which the entire wave repeats. This is called the *fundamental frequency*. For many natural sound sources, sinusoids whose frequencies are integer multiples of the fundamental frequency are present in a signal too. The frequencies of these are called *harmonics*, with the fundamental frequency f_0 called the *first harmonic*, and the multiple nf_0 called the *n-th harmonic*. A *(harmonic) complex* is a sound which consists only of harmonics. Notably, the sine wave corresponding to the fundamental frequency may be absent, as was the case for the sounds used in Chapter 3. In humans and other animals, under the right circumstances, such a sound with a *missing fundamental* still elicits the same pitch percept as the fundamental frequency would by itself (*periodicity pitch*; Seebeck, 1841).

1.1.4 Periphery of the human auditory system

A sound arriving at a human listener's outer ear first undergoes reflections at the *pinna* or *auricle*, the visible cartilaginous part of the organ (see Figure 1.2, left). These reflections not only "funnel" the sound into the *ear canal* (Ekdale, 2016), but also facilitate sound localisation (see section 1.2.2). The resulting pressure wave then travels to, and deflects, the *tympanic membrane* or *eardrum*. This membrane constitutes the interface between the outer and the *middle ear*, whose three small bones

[2] Other quantities, such as the phase offset, could conceivably be visualised following the same principle, but this is rarely done in practice.

(*ossicles*), acting like a piston, efficiently transmit the incoming air pressure wave into the fluid-filled *inner ear* (Mason, 2016). This cavity within the temporal bone includes the spiral-shaped *cochlea*, which contains the *organ of Corti*, the tissue which carries the mechano-sensory *inner hair cells* that convert mechanical deflections (due to sound) of their bundles of their *hair bundles* or *stereocilia* into bioelectrical activity. In the form of *action potentials* (rapid changes of the difference in electrical potential between the inside and the outside of a cell), this activity reaches the brain via the *auditory nerve* at the *cochlear nucleus* in the brainstem (Rhode and Greenberg, 1992).

The organ of Corti is located on the *basilar membrane*, a flexible structure which spans the coil of the cochlea, decreasing in stiffness and increasing in width from its base (closest to the middle ear) to its apex (the center of the spiral). This leads to different resonant frequencies for different regions of the basilar membrane, highest at the base and lowest at the apex, and consequently to frequency-specific deflections arriving at the inner hair cells depending on their place along the membrane (Ruggero, 1992). Together with variability in the properties of hair cells and the fibres of the auditory nerve (Mann and Kelley, 2011), the result is *tonopy*: Beginning in the cochlea, information for different frequencies is processed in different spatial locations, whereby neighbouring regions deal with similar frequencies (see Figure 1.2, right). This arrangement is carried through to higher levels of auditory processing in the central nervous system, all the way from the brainstem to the cortex (Romani *et al.*, 1982).

In this introduction, I have confined the description of the role of the brain in auditory perception to cover spatial aspects only, starting in section 1.2 for sound localisation. See Figure 1.3 for a diagram of the ascending pathway (*i.e.*, the circuit directed from the cochlea toward the cortex).

1.1.5 Loudness

Loudness is a perceptual quality of sound that is correlated with, but not completely determined by sound level (see Epstein and Marozeau, 2010). It appears to be related to overall neuronal activity, such that, due to the frequency analysis in the periphery of the auditory system, sounds with the same overall intensity are perceived as louder when their magnitude spectrum is more spread out

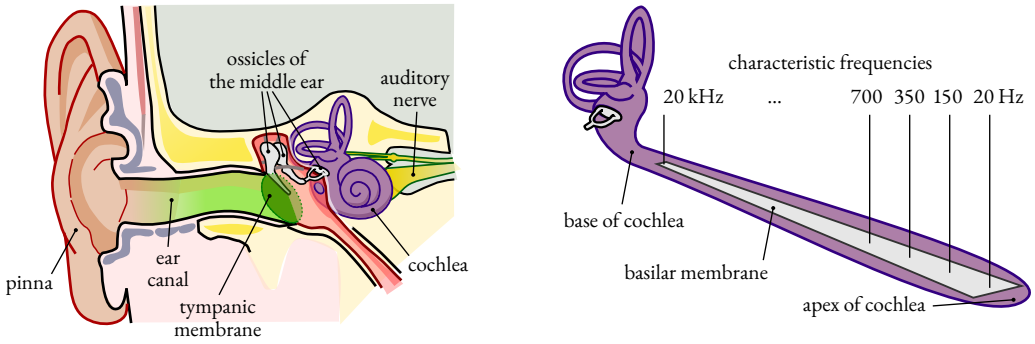

Figure 1.2 Anatomy of the human ear. **Left:** The outer (skin colour and green), middle (pink), and inner ear (purple). **Right:** Schematic of an unrolled cochlea with the basilar membrane (not to scale), annotated with some characteristic frequencies which lead to maximal excitation of the auditory nerve fibres emanating from each point along the organ of Corti (sitting on the basilar membrane; not shown). This tonotopic arrangement, *i.e.*, adjacent nerve cells representing adjacent frequencies, is carried through to the brain (rather than, for example, a representation of spatial location). Figure based on Chittka and Brockmann (2005), vectorised by Inductiveload on Wikimedia Commons (2009), modified and reproduced here in accordance with the Creative Commons Attribution 2.5 License.

(Zwicker *et al.*, 1957). Duration also plays a role (see Buus *et al.*, 1997), as do so-called context effects such as adaptation (see Canévet *et al.*, 1983) and fatigue (Hirsh and Ward, 1952). Moreover, the second ear has long been thought to contribute equally to the first, such that occluding one ear would halve loudness, but this ratio is now thought to be lower than 2:1 ("imperfect summation"; see Marozeau *et al.*, 2006).

Historical attempts to predict subjective loudness from sound pressure alone include a logarithmic function (Fechner, 1860) and a sum compressive power function over frequency bands (S. S. Stevens, 1961). While suitably accurate for simple sounds such as one or a combination of a few pure tones, modern models aim to also yield good predictions for more complex stimuli, and to account for the aspects mentioned above. The "Cambridge" series of loudness models, summarised by B. C. Moore (2014), is notable here. Its version due to B. C. Moore and Glasberg (2007) serves as the basis for the ISO standard 532-2 (International Organization for Standardization, 2016), whereas its most recent revision due to B. C. Moore, Glasberg, *et al.* (2016) underlies the draft of the projected standard 532-3. The latter specifically addresses imperfect summation and variation of sound level over time.

A unit of loudness is the *phon*. A value in phon specifies the level (in dB SPL) of a pure tone with a frequency of 1 kHz which is perceived as just as loud as the sound being described.

1.2 Sound localisation

Sound localisation is the ability of an animal to identify the location of a sound source. It is of crucial importance for the evasion of threats such as predators (Pollack, 2014), for the localisation and tracking of prey by predators (Payne, 1971), and for acoustic communication (Bronkhorst, 2000). The tonotopical arrangement of auditory processing in the brain, as introduced in section 1.1.4, prevents this from being a straightforward task: Different areas in the cochlea and brain correspond to different sound frequencies, not to the location of the sound source in space. This is in contrast to the visual system, where the structural arrangement of nerve cells does correspond to the place where light entered the eye (Daniel and Whitteridge, 1961; Holmes, 1918).

Large parts of this dissertation are concerned with sound localisation in humans. The subject has been studied extensively by means of *psychophysics*, *"the analysis of perceptual processes by studying the effect on a subject's experience or behaviour of systematically varying the properties of a stimulus along one or more physical dimensions"* (Bruce *et al.*, 2003). Middlebrooks and Green (1991) provided a well-known overall summary of the topic from this perspective. Grothe *et al.* (2010) reviewed the underlying physiological mechanisms in mammals; see Figure 1.3 for a simplified replication of their summary circuit diagram, presented here with schematic brain slices for anatomical reference.

1.2.1 Sound localisation in the horizontal plane

The localisation of sound sources at different places in the left–right (*azimuthal*) dimension plays a role in several ways throughout this dissertation. In Chapter 2, for example, differences between the azimuthal angles of a sound source and its suggested visual location were created to elicit an incongruence between visual and auditory perception. In some of the experimental conditions described in Chapter 3, by moving their upper bodies, listeners generated changes in the relative horizontal location of two stationary sound sources themselves. Finally, the room-acoustical simulation software presented in Chapter 5 must be capable of synthesising sound signals which, when presented to a listener via headphones, will be perceived as coming from any desired position in the virtual room. To see how this can be achieved, some understanding of the underlying mechanisms is required.

Compared to the high–low and distance dimensions, localisation in azimuth has been investigated in the greatest detail. The most effective mechanisms underlying this aspect rely on *binaural* cues, *i.e.*, differences in the sounds arriving at the left *vs.* the right ear. Azimuthal localisation is highly

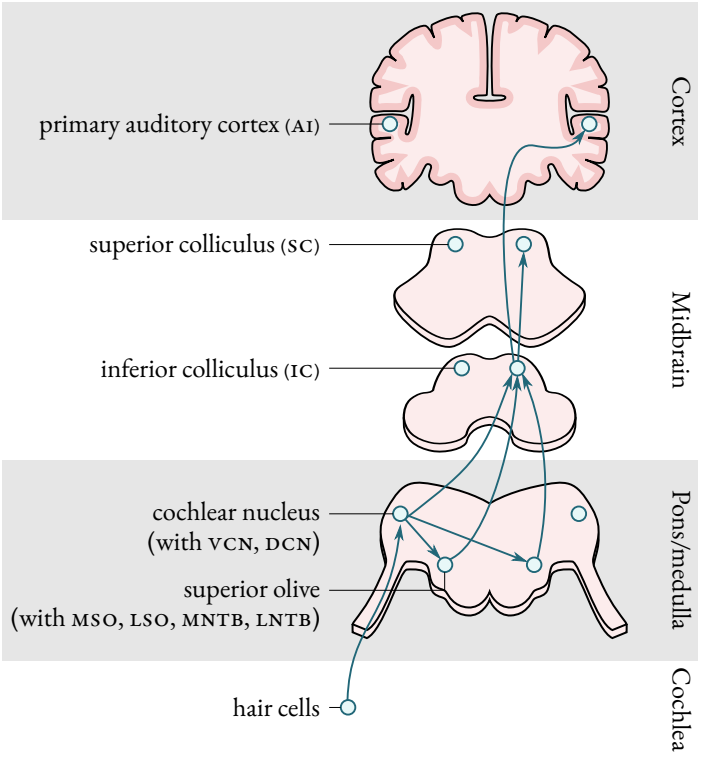

Figure 1.3

Simplified circuit diagram (following Grothe *et al.*, 2010) of the parts of the ascending auditory pathway with relevance to sound localisation. The superior colliculus is included due to its significance in multimodal integration (see section 1.5). Brain slices (horizontal in brainstem and midbrain, coronal in cortex) and location of the nuclei are schematic and not to scale. VCN/DCN: Ventral/dorsal cochlear nucleus. MSO/LSO: Medial/lateral superior olive. MNTB/LNTB: Medial/lateral nucleus of the trapezoid body.

accurate, with typical errors of just a few degrees of angle (see Blauert, 1997), and listeners can detect angular differences as small as a single degree in azimuth (A. W. Mills, 1958). Lord Rayleigh (Strutt, 1907) famously formulated the *duplex theory*, which postulates that this system is based on two types of interaural (between-ear) differences, called ITDs (interaural time differences) and ILDs (interaural level differences), and that their relative effectiveness depends on sound frequency.

ITDs (see Figure 1.4, left) arise because sounds which originate somewhere else than directly in front or behind the listener arrive earlier at one ear compared to the other. For this cue to be unambiguous, the sound needs to contain components whose frequencies are sufficiently low. If all components have a frequency of about 2 kHz or more, it becomes unreliable because some naturally occurring ITDs (up to approximately 700 µs for typical distances between a pair of ears) are long enough to potentially fit entire additional cycles of the oscillation. Headphone studies have demonstrated that human listeners can detect ITDs of just 10 µs at 1 kHz (Klumpp and Eady, 1956; Zwislocki and Feldman, 1956); this is remarkable considering that the transmission of neuronal information typically involves time constants on the order of milliseconds. A range of neuronal mechanisms is implicated in facilitating this temporal precision. One of them is *phase locking*, whereby the timing of the action potentials of auditory nerve fibres is synchronised to a fibre-specific phase of an oscillation (Galambos and H. Davis, 1943; Palmer and Russell, 1986). A nerve fibre which preferentially responds to a sound frequency of 500 Hz might fire twice per second, always precisely at the same time in the oscillation cycle, for example whenever the stereocilia of the corresponding inner hair cell are maximally deflected. This phase locking is carried through to the brainstem, specifically to the *cochlear nucleus* (CN) with its bushy cells (Joris *et al.*, 1994) and onward to the *medial nucleus of the trapezoid body* (MNTB) via a large and very fast synapse known as the calyx of Held (Englitz *et al.*, 2009; Kopp-Scheinpflug *et al.*, 2003; P. H. Smith *et al.*, 1998). Neurons originating in these three structures project, from both sides of the brain and thus from both ears, to each (left and right) *medial superior olive* (MSO; Cant and Hyson, 1992; Kuwabara and Zook, 1992). The results of processing in the MSO can be seen in neurons which are tuned to have an activity peak at a specific ITD each (Goldberg and P. B. Brown, 1969).

For higher frequencies, ILDs (see Figure 1.4, middle) constitute a more salient cue than ITDs (Feddersen *et al.*, 1957). These depend on a "shadowing effect" of the head: For wavelengths not longer than the head diameter, corresponding to these approximate 2 kHz, some of the sound from a laterally located source is reflected and does not reach the more distant ear. This leads to a higher sound pressure at the closer ear, human listeners being able to detect differences of around 1 dB at 1 kHz (A. W. Mills, 1960). Tuning of neurons to ILDs arises in the brainstem too, chiefly in the *lateral superior olive* (LSO). It is understood to be caused by an interaction between excitatory inputs from one CN (Cant and Casseday, 1986) and, again via the MNTB, inhibitory inputs from the other (Wenthold *et al.*, 1987). These counteract each other and create a difference signal between the activities associated with each ear (Boudreau and Tsuchitani, 1968; M. J. Moore and Caspary, 1983).

A second, entirely disparate mechanism is available for sound localisation in the horizontal plane. It is less precise, based on a *monaural* cue (one that does not depend on a between-ear comparison), and essentially functions in the same way as sound localisation on the high–low axis. This is a spectral cue, and its origin can be explained by the concept of head-related transfer functions.

1.2.2 Head-related transfer functions and sound localisation in elevation

As briefly mentioned before, sound waves are reflected at the pinna of the external ear before they enter the ear canal. The exact pattern of reflections depends on the location of the sound source in space. While the pinna is too small in size to generate reflections which could be heard separately in time, the superposition of the direct sound and the reflections leads to a modification of the frequency spectrum, thus generating sound source location-dependent spectral cues (see Middlebrooks and Green, 1991). Expressed as a frequency-dependent ratio between emitted sound level and sound level at the eardrum, these spectral changes are summarised as one *head-related transfer function* (HRTF, see Figure 1.4, right) for each sound source position relative to the head (typically specified as a pair of angles: azimuth and elevation) and for each ear. Due to anatomical variability, HRTFs differ substantially between listeners, though there are overall patterns which generalise well across humans (Middlebrooks, Makous, *et al.*, 1989). Neuronally, basic spectral processing in the CN (E. D. Young *et al.*, 1992) and higher-level integration in the *inferior colliculus* (IC, K. A. Davis *et al.*, 2003) are implicated in HRTF-based localisation.

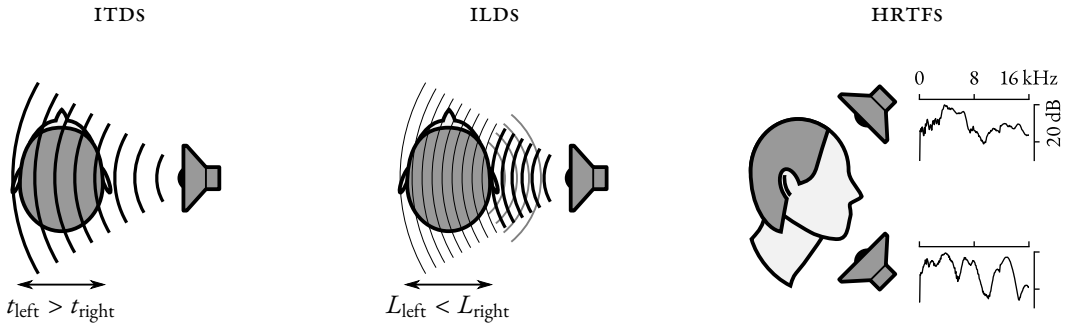

Figure 1.4 Cues for sound localisation in azimuth and elevation. **ITDs:** Sound from an off-center source reaches the nearer ear before it reaches the farther ear. Due to ambiguities which arise at higher sound frequencies, this cue is most effective below about 2 kHz. **ILDs:** Higher-frequency sounds (above about 2 kHz) are subject to a shadowing effect of the head, such that the level of the sound arriving at the nearer ear is higher than at the farther ear. **HRTFs**: Reflections at the head and pinna modify the magnitude spectra of the arriving sound, in a manner that is dependent on the elevation (and azimuth) angle of the source. Figure modelled after Grothe *et al.* (2010).

The information from spectral cues is particularly useful to resolve ambiguities between source locations in front of, behind, below, or above the listener, all of which generate the same binaural cues (cone of confusion, see *e.g.* Blauert, 1997). As such, the spectrum is the primary characteristic used for localisation in elevation, though they can also facilitate localisation in azimuth when binaural comparisons are not possible or unreliable (such as when one ear is occluded; Fisher and Freedman, 1968). As evidence for this mechanism, Jongkees and Groen (1946) and Roffler and Butler (1968) found that the localisation of sound in elevation is impaired when the pinnae of listeners were tied flat against the head. Batteau (1967) proposed "characteristic reverberation" created by the pinna as a possible cue. Based on pairwise individual microphone recordings taken at the entrance of the ear canals of ten human listeners, one recorded with noise from a loudspeaker in front and the other with a loudspeaker behind, Blauert (1969) was able to alter the spectra of the noises emitted by two loudspeakers such that all ten listeners were convinced that the sound came from the front or from the back, depending on which of the two recordings was employed for the spectral modification. In another experiment reported in the same study, he demonstrated that spectral cues can also create the illusion of a source being located above the listener's head. These results not only identified spectral cues as crucial for high–low/front–back localisation, but also that the location-dependent spectral modifications expected by the auditory system can easily be confounded by the actual spectrum of the emitted sound.

HRTFs can be modelled as digital filters (specifically as FIR filters; see introduction in section 1.3.3) and thus applied to sound signals with a computer. The liveRAZR simulation software which I developed as part of this dissertation (see Chapter 5) includes this functionality, so that sound sources can be virtually placed on the high–low/front–back axes for listeners wearing headphones.

1.2.3 Sound localisation in depth

A listener's estimation of their distance to a sound source (that is, the depth of the sound source from the listener's point of view) is a topic of Chapters 3–5. Moreover, while it is less of a priority there, Chapter 2 also features sound sources virtually positioned at two different distances. This aspect of sound localisation is therefore a central one for this dissertation as a whole. Recent comprehensive reviews of auditory distance perception have been provided by Zahorik, Brungart, *et al.* (2005) and by Kolarik, B. C. Moore, *et al.* (2016).

Cues for depth are plentiful, but some of them are relative: They are only useful in comparison to some reference. An obvious example is given by sound level (see Figure 1.5, left). Considering the geometrical dilution described by the inverse distance law, a listener could (in an *anechoic* environment, *i.e.*, one in which sound does not undergo reflections) theoretically calculate their distance from a sound source exactly, but this is only possible with prior knowledge about the level of the sound at a well-known distance. Von Békésy (1949) and Gamble (1909) have demonstrated the expected association between increase in sound level and decrease in perceived distance, but Mershon and King (1975) found that when one group of listeners was first presented with a sound at a given level, and another group was first presented with the same sound 10 dB higher in amplitude, listeners from both groups did not differ in their distance estimations. In a second trial, however, where the sound level was increased for the first group and decreased for the second, their responses deviated; clearly, the listeners each compared the second stimulus to the first. Long-term experience with the same category of sound also influences distance judgments. For example, playbacks of recordings of shouted speech caused listeners to perceive the source as further away than whispered speech at a comparable level, consistent with the higher output power associated with shouting (Philbeck and Mershon, 2002). Similar considerations apply for the effects of atmospheric attenuation (see section 1.1.1). It can be an effective cue for depth, at least at distances where this kind of attenuation is sufficiently pronounced (von Békésy, 1938; Butler *et al.*, 1980; Coleman, 1968), but listeners required experience with a sound

to be able to assess whether its spectral properties are inherent to the emitted signal or due to distance (Little *et al.*, 1992).

Conversely, reverberation provides an absolute cue (see Figure 1.5, right). In a study using loudspeakers positioned at different depths in an *echoic* room, Mershon and Bowers (1979) showed that, despite lack of familiarity with the stimulus and with the room environment, listeners who heard a sound from further away in the first trial of their experimental session responded with greater perceived distances than listeners who started their sessions with a closer source. The predictable variability of the direct-to-reverberant ratio with distance, see section 1.1.2, is generally accepted as the explanation (Bronkhorst and Houtgast, 1999; Kopčo and Shinn-Cunningham, 2011; Mershon, Ballenger, *et al.*, 1989). Zahorik (2002a) varied level and DRR cues systematically in a virtual acoustic environment and found that when asked to provide distance estimates, listeners weighted the cues differently depending on stimulus type. Specifically, they relied more strongly on absolute sound level for speech, with which they can be assumed to be highly familiar, whereas they were more reliant on DRR when presented with artificial noise bursts. The acoustical properties of the environment also play a role, such that DRR cues were found to be more effective for the discrimination of the auditory distances of pairs of speech stimuli in a more reverberant room compared to a less reverberant one (RT_{60} = 0.7 s *vs.* 0.4 s; Kolarik, Cirstea, *et al.*, 2013).

It is notable that distance perception can be improved by reverberation at all. This stands in contrast to other aspects of auditory perception, such as localisation in azimuth (Hartmann, 1983) or the identification of speech sounds (Gelfand and Silman, 1979; Nábělek and Dagenais, 1986), which tend to be negatively affected by room acoustics. The effect of reverberation on *amplitude modulation* (AM; low-frequency periodic changes of stimulus amplitude over time) can be considered in this context: As DRR decreases with increasing distance, reverberant sound increasingly "fills in" the dips in signal amplitude (decreases the *modulation depth*). D. O. Kim, Zahorik, *et al.* (2015) have shown that such dips in the source signal are in fact necessary for a distance-dependent activity which they have discovered in some rabbit IC neurons. They also found the pattern of auditory distance judgments by humans to be consistent with this neuronal activity. Most research concerning the neuronal basis of auditory distance perception, however, has focused on the role of cortical areas (*e.g.* Kopčo, Huang,

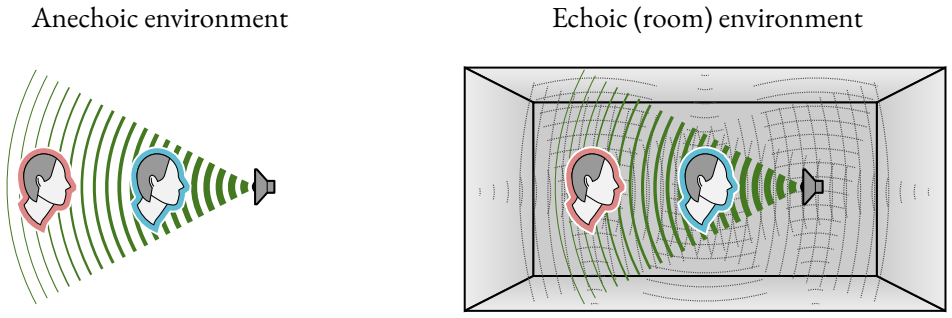

Figure 1.5 Level and direct-to-reverberant ratio cues for sound localisation in depth. **Anechoic environment:** Geometric attenuation (the inverse distance law) leads to a halving of sound levels for every doubling of distance, such that a higher sound level arrives at the right (blue) listener compared to the left (red) listener. If the listeners are familiar with the output of the sound source, the right listener will consequently perceive the source as nearer than the left listener listener. **Echoic (room) environment:** The inverse distance law holds for the direct path from the sound source to the listener, but the level of the repeated reflections from the walls (reverberation) decreases with distance at a far lower rate. The proportion of the direct *vs.* reverberant sound levels can act as an absolute cue, *i.e.*, one which the listeners can exploit without prior familiarity with the sound source.

et al., 2012; Kopčo, Doreswamy, *et al.*, 2020) largely consistent with the so-called *auditory "where" pathway* (Rauschecker and Tian, 2000). In fact, as opposed to the computation of azimuth and (to a lesser extent) elevation information, there is no evidence in the literature of distance coding in any pre-midbrain nuclei.

In a number of ways, the ability of humans to judge auditory distance is arguably less remarkable than that of localisation in azimuth and elevation. For example, Zahorik (2002b) and Larsen *et al.* (2008) have both determined just-noticeable differences (JNDs) in DRR and found relatively high values in the order of 6 dB to be required for DRR-based auditory distance discrimination in many circumstances—a rather high threshold, especially considering that differences of under 0.5 dB can be detected in the case of (anechoic) noise (Miller, 1947). More generally, depth estimates are biased, towards overestimation at distances near the listener (up to approximately 2 m) and towards underestimation at larger distances, a perceptual compression that can be modelled with a power law (Zahorik, 2002a; Zahorik, Brungart, *et al.*, 2005). They are also less precise and more variable compared to visually based judgments of distance (Anderson and Zahorik, 2014). These biases and high variances may well be because of the high level at which distance perception seems to arise within the central nervous system, sufficiently late in the processing stream for cognitive effects and influences from other sensory systems to play a significant role—a complexity that, in my opinion, makes auditory distance perception an interesting research topic in which many important questions are still unanswered.

1.2.4 Perception of moving sounds

The sources of sound that humans encounter in everyday life—and indeed some sources which a few particular humans encountered in the experiments described in Chapters 3, 4 and section 5.6—are not always stationary. Partly due to the difficulty of creating suitable stimuli, less is known about auditory spatial perception in such dynamic settings. Some of the literature on this complex topic was reviewed by Carlile and Leung (2016).

Just as for the localisation of stationary sources, much of the research on this topic has focused on binaural perception and thus on the perception of changes in azimuth. How well listeners are able to detect a rotation of a sound source around them depends on stimulus duration and source velocity. Aggregated data from eight studies (see Carlile and Leung, 2016) showed that movement angle JNDs can reach levels on the same order as the 1° for two stationary stimuli presented in sequence (see section 1.2.1), but listeners never fared better, and indeed much worse for short and/or fast stimuli.

The perception of sound source speed itself is interesting in its own right. Psychophysical studies have found ILD cues to be more effective than ITD cues for listeners to distinguish between different angular velocities, with JNDs of around 2° and 11°, respectively (Altman, Romanov, *et al.*, 1988; Altman and Viskov, 1977). For *translational motion* (*i.e.*, displacement in space), changes in level and frequency shifts (caused by the changed distance between the wavefronts emitted by a moving source, the *Doppler effect*) appear to be more salient, even if the trajectory of the source is variable in azimuth (Lutfi and Wang, 1999).

■ **The looming bias.** Sound sources in translational motion are also subject to another bias in distance perception. Neuhoff (1998) presented twelve listeners with sounds which either increased or decreased in sound level over time and had them indicate the perceived change of level on an unmarked scale. He found that increasing levels were associated with significantly higher perceived changes. Since level acts as a distance cue, this might affect auditory judgments of depth. This is in fact the case: Listeners perceived the motion trajectory to be longer, and the end point of the trajectory to be nearer, when a loudspeaker emitting a sound with a constant amplitude was moved towards them *vs.* when it was moved away from them (Neuhoff, 2001).

Such asymmetries in the perception of level changes have been confirmed for estimations of overall loudness (Stecker and Hafter, 2000; Susini, McAdams, *et al.*, 2007), of motion speed (Neuhoff, 2016), of duration (Schlauch *et al.*, 2001), as well as for emotional response if the sound is perceived as unpleasant (Tajadura-Jiménez *et al.*, 2010). The implication of decreasing distance by increasing level has been suggested by Neuhoff (1998, 2001) as the likely evolutionary cause for this bias: The argument is that sounds rising in level correspond to *looming* danger, and that it is selectively advantageous for an animal to perhaps overestimate the imminence of danger. This is consistent with the findings that rising-level, but not constant-level or falling-level acoustic stimuli promoted activity in the amygdala (Bach, Schachinger, *et al.*, 2008), in a network of cortical areas associated with auditory movement and attention (Hall and D. R. Moore, 2003; Seifritz *et al.*, 2002), and (in marmosets) higher auditory-cortical activity overall (Lu *et al.*, 2001).

It should be noted that the interpretation of such findings as a bias for looming appears to be largely accepted, but has also been met with objections. A simple *"bias for end level"* has been proposed as a alternative explanation (Susini, Meunier, *et al.*, 2010; R. Teghtsoonian *et al.*, 2005). While this idea might be appealing on grounds of parsimony, I believe that it is inadequate: It cannot explain the particular patterns of neuronal activation, or indeed some behavioural results in studies that include suitable control conditions (*e.g.* Olsen *et al.*, 2010).

A converse aspect of auditory motion perception will be covered in section 1.5.3, namely the case of listener rather than sound source motion, which leads to the same type of changes to sound arriving at the ears. Indeed, the disambiguation between the two situations is nontrivial.

1.3 Digital processing of sound

No chapter of this dissertation could have coped without a computer to generate and/or analyse sound signals: Digital processing was necessary to simulate the acoustics of enclosed environments (Chapters 2 and 5); to present virtual sound sources in motion (Chapters 3–5); to precisely control the properties of experimental stimuli (Chapters 2–5); and to make sense of animal-generated sounds captured with microphones (Chapters 4 and 6). On the following pages, I will present a rather technical introduction to some basic computational techniques which will appear repeatedly later on. These fundamentals are especially important in the context of the room-acoustical simulation software described in Chapter 5.

1.3.1 Signal representation and filtering

To store, extract information from, alter or synthesise sound with the aid of a computer, an analog acoustic signal must be represented in a digital form. Digital signals are discrete approximations of continuous sound waves in two ways: Firstly, they are *sampled*, meaning that a finite number of readings of sound pressure (or rather of an amplitude quantity approximately proportional to it, such as the voltage obtained from a microphone or used to drive a loudspeaker) are taken at particular points in time. A typical *sampling rate* for sound is 44.1 kHz, meaning that 44 100 readings (*samples*) are available for each second of signal, or one sample for each 22.7 µs. Secondly, they are *quantised*, *i.e.*, the amplitude is translated into a number with a fixed amount of binary digits (the *bit depth*). 16 bits, for example, allow 65 536 different amplitude values to be distinguished; at 32 bits, there are almost 4.3 million possible distinct values. At sufficiently high sampling rates and bit depths, the translation from an analog to a digital signal, or vice versa, only introduces errors that lie below the human threshold of perception. The required sampling rate can be derived from the *Whittaker–Shannon sampling theorem* (Shannon, 1948; Whittaker, 1915), one formulation of which states that a time-continuous signal can be perfectly reconstructed from a time-discrete signal if the sampling rate $f_S \geq 2f_{\max}$, where $f_{\max}$ is the maximum frequency present in the signal. This is clearly the case for $f_S = 44.1\,\text{kHz}$ and $f_{\max\,(\text{human hearing})} = 20\,\text{kHz}$.

One of the fundamental operations in sound processing is *filtering*. In this context, a filter is a process which accepts a signal as its input, suppresses unwanted or enhances desirable aspects of the signal, and provides the thus modified signal as its output. The most commonly used filters are *linear and time-invariant* (LTI), with the mathematical implication that they can only linearly scale (attenuate/amplify) as well as delay component parts of a sound differently depending on frequency. Among others, there are *high-pass* and *low-pass* filters, blocking all frequencies below or above a certain *cut-off frequency*, respectively; *band-pass* and *band-stop* filters, blocking all frequencies outside or inside some range, respectively; and *high-shelf* and *low-shelf* filters, which act much like high-pass and low-pass ones, but only aim to somewhat reduce rather than to completely reject the affected signal components. The *magnitude response* of a filter, much like a magnitude spectrum, characterises exactly by which factor it amplifies a constituent sinusoid of a given frequency. Finally, *all-pass* filters do not attenuate sound at any frequency. They instead shift constituent sinusoids differently in phase offset (described by their *phase response*). Intuitively, they delay the signal at some frequencies more than at others. Magnitude and phase response together are known as the *frequency response*.

Any combination of such LTI filters is an LTI filter itself, so a filter may be highly complex in how it affects sound at certain frequencies. For instance, the characteristic pattern of reflections due to a human's torso, head and outer ears can be described as a filter too (see section 1.2.2), as can the acoustics of a room.

Filters may be implemented in analogue electronic circuits, but in the context of this dissertation, they are always understood as digital. This means that they can simply be described as a sequence of arithmetic operations, where the quantised samples of the signal act as some of the operands. The LTI constraint strongly limits the arithmetic operations that are permissible, which offers a great advantage: When the input signal to the filter is an *impulse* (zero everywhere except for the first sample, which has a well-defined nonzero value such as 1), its corresponding output (the *impulse response* or IR) is enough to fully characterise its behaviour. If two LTI filters have identical impulse responses, their outputs are identical for any other input signal too.

1.3.2 Infinite impulse response (IIR) filters

If a digital filter uses previous samples of its own output signal to calculate the values of new output samples, it is called *recursive*, and there is no output sample after which all other samples of its impulse response will be all zero.[3] Every IIR filter can be fully described by a sequence of *feedforward coefficients* $b_0, b_1, \ldots, b_n$ and *feedback coefficients* $a_0, a_1, \ldots, a_m$. For an input signal with samples $x_0, x_1, \ldots$, a sample of the output signal $y_0, y_1, \ldots$ is calculated as follows:

$$y_i = \frac{1}{a_0}\big(b_0 x_i + b_1 x_{i-1} + \ldots + b_n x_{i-n} - a_1 y_{i-1} - a_2 y_{i-2} - \ldots - a_m y_{i-m}\big)$$

where $x_j = y_j = 0$ for all $j < 0$. At least one a_k must be nonzero for $k > 0$, otherwise, the impulse response of the filter becomes finite (see below).

Usually, n is equal to m and called the *order* of the filter. For $n = m = 2$, an IIR filter is called *biquadratic*, or *biquad* for short. Any IIR filter of order 4 can be created from two biquads by making the output signal of one the input signal to the other, and an IIR filter of any even order k by *cascading* $\frac{k}{2}$ biquads in this way. The constituent biquads are then sometimes termed *second-order sections*.

The coefficients $a_0, a_1, \ldots$ must be chosen carefully to make sure that the impulse response of an IIR filter approaches 0 over time. If it does not, the filter is *unstable*, and its output will instead grow towards infinity for a subset of input signals. This is a catastrophic failure state from which the

[3] In practice, there will be such an output sample, because computers cannot store numbers that are arbitrarily close to zero. Eventually, small numbers are rounded to zero. However, this does not matter for a theoretical description of the characteristics of IIR filters.

filter cannot recover over time, making it entirely unfit for purpose. When instability results from numerical issues (because coefficients, signal samples and/or results of intermediate calculations are handled with insufficient precision), second-order sections can offer a stable alternative, as numerical errors are less likely to compound at lower filter orders.

1.3.3 Finite impulse response (FIR) filters

The calculation of the output samples of an FIR filter follows that of an IIR filter as shown above, with the constraint that $a_1 = a_2 = \ldots = a_m = 0$, and $a_0 = 1$ by convention, such that they are fully (and uniquely) described by their feedforward coefficients (simply called coefficients here). As FIR filters do not use their own output samples as operands, they are also called *nonrecursive*. This limitation has important consequences: Firstly, the sequence of coefficients of an FIR filter is equal to its impulse response, and the impulse response thus "terminates" after $n+1$ samples. That is, $y_k = 0$ for all $k > n+1$ if the input signal is an impulse, where n is the filter order. Indeed, for an input signal which is l samples long, $y_k = 0$ for $k > n + l$, such that FIR filters are always guaranteed to be stable.

Secondly, the operation performed by an FIR filter can be described mathematically as a (discrete) *convolution* $b * k$ of the sequence of input signal samples x with the *finite* sequence of coefficients b:[4]

$$y_i = (b * x)_i = \sum_{k=0}^{n} b_k x_{i-k}$$

Intuitively, every sample of the input signal is thus multiplied by every coefficient, followed by a summation of all products which correspond to the same sample of the output signal. The view of FIR filtering as convolution is advantageous because of the convolution theorem (see section 1.3.4).

Finally, FIR filters facilitate a lot of control over their frequency response and can be very intuitively designed to achieve very complex results. This flexibility comes at the cost of requiring higher filter orders to achieve some desirable behaviours in many frequency response profiles for which highly efficient IIR designs are available. For simple filtering operations like high-pass, low-pass, *etc.*, IIR filters thus come at a greatly lower computational cost compared to equivalent FIR filters. On the other hand, a single FIR can accurately describe, for example, an HRTF, or the transformation of a sound from a given source location to a given receiver location by an arbitrary room environment by reflection, diffraction, *etc.* (a *room impulse response*).

1.3.4 Fourier transform

Any digital representation of a sound wave can be decomposed into the sinusoids it contains. This is accomplished by the *discrete Fourier transform* (DFT), which derives the *frequency-domain representation* $\hat{x}$ of a signal x, both finite sequences of length N with samples numbered from 0:

$$\hat{x}_k = \sum_{n=0}^{N-1} x_n \left(\cos \frac{2\pi k n}{N} + i \sin \frac{2\pi k n}{N} \right) \qquad \text{for } k = 0, \ldots, N+1$$

$\hat{x}$ is complex-valued and represents the frequency spectrum of the signal. $\hat{x}_k$ refers to the frequency given by $\frac{kf_S}{N}$. Each value of the discrete magnitude and phase spectrum can be obtained from the absolute value $|\hat{x}_k|$ and the argument $\arg \hat{x}_k$, respectively. The DFT is an important building block in the visualisation of sound through spectrograms, many of which will be presented in Chapter 6.

The (circular) *convolution theorem* states that under certain preconditions, for a convolution between two sequences $c = a * b$, the values of the corresponding Fourier-transformed sequences are

[4] The operation of an IIR filter can be described by a convolution of the input signal too, but with the *infinite* impulse response, which makes this characterisation less useful in practice.

simply multiplied:

$$\hat{c}_k = \hat{a}_k \hat{b}_k$$

This is useful because the runtime of convolution with the formula introduced above, called *direct* or *time-domain* convolution, grows quadratically with signal length: For each doubling of signal length, the computational effort is quadrupled (see *e.g.* Wefers, 2014 for a derivation). Multiplication, on the other hand, is fast, and the number of multiplications required is simply a multiple of signal length. The theorem can be applied to great advantage because the DFT can be easily inverted, and because there are efficient divide-and-conquer numerical algorithms to compute it and its inverse (*fast Fourier transforms*, FFT; Cooley and Tukey, 1965 and variants thereof). The method of transforming sequences with an FFT, multiplying their elements, and inversely transforming them with another FFT, is known as *frequency-domain* or *fast convolution*.

1.3.5 Virtual acoustic space and virtual acoustic environments

The computational approaches introduced in this section, as well as any other suitable digital processing method, can be employed to generate sound signals which, when played to a listener using an appropriate arrangement of loudspeakers or with a pair of headphones, are consistent with an acoustic environment that the listener is not actually in. The scientific aim of such a manipulation is to gain maximum control over the auditory cues provided to them, such as to study their behavioural or neuronal responses to well-defined stimuli, while they feel immersed and present in a virtual space (see Begault, 1994). This environment can be a recording of a real space (acquired, for example, with a microphone array or with an artificial head; see *e.g.* Meyer and Elko, 2004; Paul, 2009), a purely synthetic space, or a combination thereof. Regardless of the origin, the technique is known by several names such as *virtual acoustic display*, *virtual acoustic environment*, or *virtual acoustic space*. I will prefer the later term in this dissertation (and abbreviate it as VAS) for the general concept. This includes the relatively simple implementations common in psychophysics, where *e.g.* a single sound source is presented at a virtual location in space. I reserve the term virtual acoustic environment, or VAE, for more complex realisations of this idea, especially when they include room acoustics.

Core manipulations include the generation of auditory spatial cues, which in a headphone-based approach includes ITDs, ILDs, and the application of HRTFs. VAS techniques can be implemented in an interactive manner by tracking the head of the listener and rapidly updating the acoustic signals according to the momentary position and orientation. Such dynamic changes are known to improve the impression of realism (Savioja, Huopaniemi, *et al.*, 1999; Wenzel, 1992).

Some technical details regarding the implementation of VAEs will be discussed in the context of room-acoustical simulation in Chapter 5.

1.4 Perception of reflections and room acoustics

Chapter 2 is concerned with the human auditory perception of the listener's surroundings, rather than with the source of the sound itself. I undertook this study to look for a possible process of reverberation suppression based on visual information, motivated by the existing knowledge about purely auditory mechanisms of this kind. Furthermore, such processes might have been active in the experiment on looming sounds inside *vs.* outside a room, as described in section 5.6. To provide some context for these efforts, this section here summarises the literature on the perception on echoic environments, beginning with a centuries-old observation about individual sound reflections.

1.4.1 Suppression of single reflections

The best-known finding on auditory perception in the presence of sound reflections is probably the *precedence effect*. Wallach *et al.* (1949) noted that *"the repetition of essentially the same stimulus in*

quick sequence leads to a single auditory experience that is qualitatively not very different from the experience resulting from a single stimulus alone" and are credited with coining the term, though the fundamental idea is older; M. B. Gardner (1968a) traces it back to the mid-19th century (Henry, 1851). In a widely cited review, Litovsky, Colburn, *et al.* (1999) provide a framework which encompasses several individual aspects of what may, depending on the author, all be summarised as the precedence effect:

Fusion: When a sound (the leading sound or "lead") and a repetition of that sound from another point in space (such as its reflection at a surface; the lagging sound or "lag") are presented in close temporal succession, listeners report that they can only hear a single sound if the delay between the sounds is short enough.[5] The maximum delay up to which fusion is active is known as the *echo threshold* and depends on the stimulus type, with values as low as 5 ms for clicks (Freyman *et al.*, 1991), and as high as 50 ms for speech (Haas, 1951; Lochner and Burger, 1958). Fusion was originally studied with lead and lag differing in azimuth and has often been viewed in the context of ITD and ILD processing, but it also occurs when both sounds originate in front of the listener at different elevations (Rakerd *et al.*, 2000). Consequently, it is not an exclusively binaural phenomenon.

Fusion "builds up" as listeners gain familiarity with the spatial pattern of a pair of leading and lagging sounds that are emitted repeatedly (Freyman *et al.*, 1991). Clifton (1987) and Clifton and Freyman (1989) showed furthermore that when the (simulated) spatial locations of the lead and lag suddenly change in the course of the repeated presentation, both sounds can be perceived and the buildup process starts over (see Figure 1.6, left). This is linked to a corresponding increase in the echo threshold (Yang and Grantham, 1997).

Localisation dominance: At lead–lag delays greater than about 1 ms but not exceeding the echo threshold, the lead is weighted more strongly than the lag when it comes to sound localisation in azimuth and elevation. A model by Shinn-Cunningham *et al.* (1993), who manipulated ITDs between lead and lag by means of headphone stimulation, found weights for the leading sound of 70–100 % for a 1 ms lead–lag separation, and a strong dependence of weights on the difference between the lead and lag ITDs if they were separated by 10 ms.

Lag-discrimination suppression: At lead–lag delays where fusion occurs, the auditory system suppresses spatial information from the lagging sound: When two sounds are presented at a delay within the echo threshold, listeners struggle to discriminate between different locations of the second (despite being able to perceive both). As opposed to fusion, this effect does not appear to be subject to a buildup (Yang and Grantham, 1997).

It should be noted that the precedence effect is usually described as acting on pairs of sounds that are identical except possibly for level. Real-world reflective surfaces, however, absorb sound better at some frequencies than at others. While an experiment with correlated *vs.* uncorrelated pairs of noise bursts has demonstrated that this identity is important for fusion (Perrott *et al.*, 1987), Blauert and Divenyi (1988) observed that the amount of overlap in the frequency spectrum between lead and lag is positively correlated with the extent of lag-discrimination suppression. For typical reflective surfaces in rooms, this overlap is likely sufficient to not hinder precedence phenomena.

The neuronal origins of the precedence effect are not fully understood. Peripheral processes seem to play a role for very short lead–lag delays (Bianchi *et al.*, 2013), but do not appear to be a sufficient explanation in other cases. Human auditory brainstem responses can be identified in electroencephalograms (EEG) for both clicks in a pair if they are spaced at least 2 ms apart, even when the corresponding

[5] It is still considered fusion even if a listener is able to hear a difference when given the opportunity to compare a single sound with a lead–lag stimulus; all that matters for this definition is that they cannot detect two individual auditory events.

percept is fused (Damaschke *et al.*, 2005). Various electrophysiology studies in mammals (*e.g.* Pecka *et al.*, 2007; Tolnai, Beutelmann, *et al.*, 2017; Yin, 1994) and a lesion study in humans (Litovsky, Fligor, *et al.*, 2002) point to the CN, the dorsal nucleus of the lateral lemniscus (DNLL), and to the IC as likely sites of the underlying circuitry. Finally, temporal analyses of cortex EEG showed that the event-related potential N100, with a latency of about 100 ms after stimulation, differed reliably depending on whether a lead–lag pair (with a delay equal to the individual's echo threshold) was perceived as fused or not (Backer *et al.*, 2010; Sanders *et al.*, 2008). Backer *et al.* (2010) also found this evoked potential to be significantly different between fused lead–lag and single click stimuli, hinting at a cortical involvement in echo suppression. Further evidence for context-dependent mechanisms, which allow full information about the lagging sound to be retrieved if the task at hand requires it, comes from a behavioural experiment: *"[T]he classically described asymmetry in the perception of leading and lagging sounds is strongly diminished in an echolocation task."* (Wallmeier, Geßele, *et al.*, 2013).

1.4.2 Suppression of reverberation

The echo threshold is too short to reasonably expect precedence-related phenomena to operate at the timescales usually observed in the acoustics of typical rooms, where reverberation times on the order of hundreds of milliseconds are by no means unusual. However, the human auditory system does seem to compensate for the effects of longer-lasting reverberation to some extent. For example, a mechanism similar to localisation dominance acting over longer timescales has been termed "onset dominance" (see *e.g.* Devore *et al.*, 2009).

In the context of this dissertation, however, the perception of reverberation itself is a more interesting aspect, especially aspects of it which may be considered similar to fusion. In this vein, Djelani and Blauert (2001) measured echo thresholds for noise bursts in a virtual room with a triangular layout. Ahead of each test burst, for which listeners had to judge whether they heard any echoes, three "conditioning" bursts were played back, either (a) in the same acoustical environment, (b) with the acoustics of a different virtual room, or (c) in an anechoic setting. They found that for listeners to be able to detect the reflections, the virtual room for the test burst had to be scaled up in condition (a), such that the virtual reflections arrived later, compared to the other conditions. This is an indication for a buildup of echo suppression in environments with more than one reflector, such as a room.

Watkins (2005a,b) investigated the categorical perception of speech sounds in reverberant environments. The stimuli were words on a continuum between English "sir" and "stir", with nine synthesised intermediate steps between these two recorded endpoints. These words differed mostly in the duration of silence between the fricative /s/ and the vowel /ɜ:/, as plosives like the /t/ in "stir" block the vocal tract completely until their release. They were presented to listeners via headphones, either anechoically or processed to include the response of a recorded room. With increasing reverberation, listeners were more likely to hear "sir", as the reverberant tail produced by reflections of /s/ filled in the silence due to /t/. Watkins (2005a,b) compared the probabilities with which listeners heard "stir" or "sir" when reverberated test words from the stimuli were presented in the context of a carrier phase. They found that the expected effect due to the reverberant tail could be compensated if the carrier phrase had the same room-acoustical characteristics as the test word, but not when the carrier phrase was almost anechoic (yet the test word remained reverberant). They took this as evidence for a echo-suppressing process which requires exposure to the room in order to build up (see Figure 1.6, right), much like the buildup of the precedence effect.

Nielsen and Dau (2010) have argued against this interpretation, on the basis that the data would also be consistent with an interfering effect of the anechoic carrier phrase, in which the stronger amplitude modulation of the carrier might mask (make less apparent, as described in Wojtczak and Viemeister, 2005) the weaker modulation of the test word. They found evidence for their hypothesis in data they recorded in a control condition, which had silence instead of a carrier phrase but did not yield

results significantly different from those with the reverberant carrier. Watkins and Raimond (2013) subsequently rejected this proposition based on data they acquired and which contradicted this evidence when the test word was presented at different levels of reverberation from trial to trial. They suggested that in the control condition introduced by Nielsen and Dau (2010), two test words with the same level of reverberation always followed each other (albeit separated by silence), such that the test word of the previous trial would have acted as a reverberant carrier for each control trial.

Brandewie and Zahorik (2010) employed a similar paradigm to study speech intelligibility and found results consistent with this interpretation, *i.e.*, an improvement in speech reception when listeners were previously exposed to the acoustics of a room. As opposed to Watkins (2005a,b), to whom the effects of monaural and binaural pre-exposure appeared similar in some conditions, Brandewie and Zahorik (2010) only observed this effect for binaural listening. The latter thus speculated that there may be *"separate and perhaps complementary aspects of room de-reverberation: one that relates*

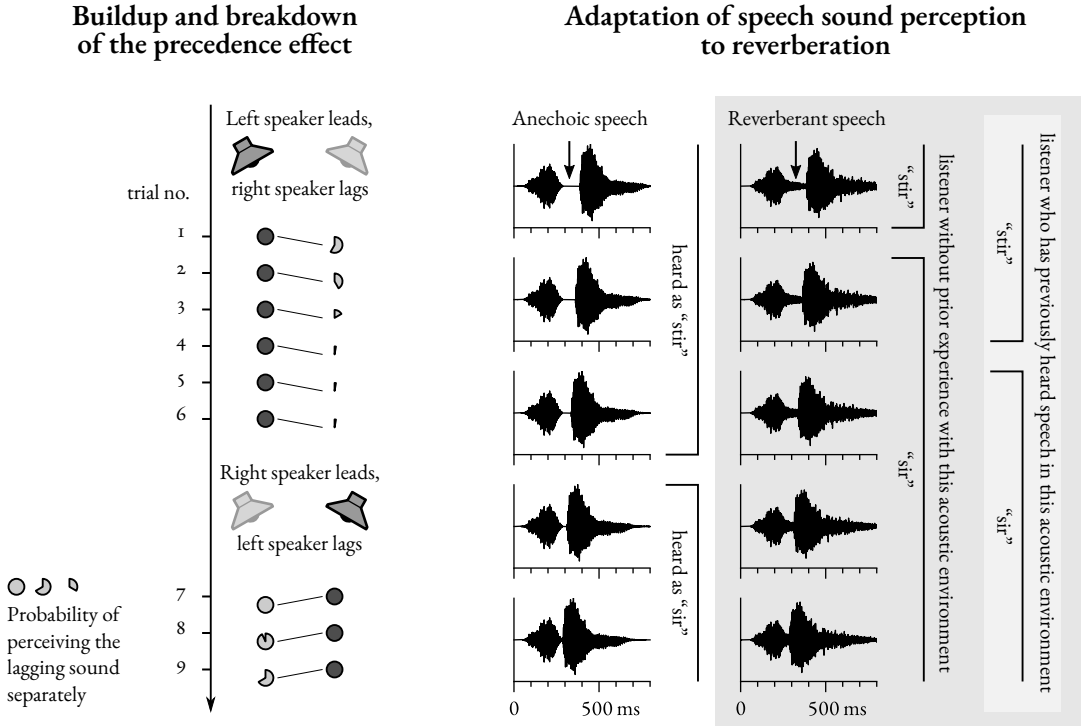

Figure 1.6 Two examples for the perceptual suppression of reflected sounds. These charts are designed for the purpose of illustration and not based directly on psychophysical data. **Buildup and breakdown of the precedence effect:** With repeated presentation of the same lead–lag pattern to one listener, their perception of the two sounds in each trial is increasingly likely to fuse into a single auditory event (Freyman *et al.*, 1991). A sudden reversal of the pattern (the formerly leading sound now lags and vice versa, as in trial 7 here) leads to a "breakdown" of fusion, such that the two sounds are very likely heard separately again (Clifton, 1987). Buildup starts over with a repeated presentation of this new pattern. **Adaptation of speech sound perception to reverberation:** The English words "sir" and "stir" can be distinguished, among other differences, by the duration of silence between the initial consonant /s/ and the vowel /ɜ:/. There is a threshold silence duration above which a listener is more likely to hear "stir". Sound reflections fill in this silent gap with reverberation of the /s/ (see arrows in the upper two waveforms), such that longer durations of silence (measured without reverberation) are required for the recognition of "stir". Listeners with prior auditory experience with the environment can compensate for the effect of reverberation, resulting in a lower threshold (Watkins, 2005a,b, but note the lack of an anechoic control condition in these studies).

to spatial configurations within the room and therefore is facilitated by binaural input, and one that is concerned primarily with removing monaural coloration [that is, changes in spectrum] *caused by room acoustics"*. In follow-up studies, the authors demonstrated that the effect of prior exposure on speech intelligibility depended on the reverberation time of the room, with maximal effect sizes observed at $RT_{60} = 1\,\mathrm{s}$ (Zahorik and Brandewie, 2016), and that sudden changes to late reverberation characteristics, but not to the spatial distribution of early reflections can cause the effect to break down (Brandewie and Zahorik, 2018).

Overall, the suppression of (or adaptation to) reverberation is clearly less of an established fact than the precedence effect. Similar to the complexity of auditory distance perception (see section 1.2.3), I believe this to be due to the intricacy of the required calculations; they likely occur at a high level in the brain and are obviously subject to context effects. It is therefore plausible that input from other sensory systems might play a role, as is the theme of the following chapter.

1.4.3 Perception of room acoustics

A separate but related question to these suppression phenomena, which determine how listeners perceive sounds *in* reverberant environments, is how they perceive such an environment—its acoustical properties—itself. Much of the research in this area has been conducted in the context of architectural acoustics, such as in the planning and evaluation of concert halls and other listening spaces. This topic encompasses many facets, and to cover them in detail would go beyond the scope of this dissertation. Kaplanis *et al.* (2014) have reviewed some of the literature from a rather technical perspective.

Human listeners are able to perceive differences in reverberation time, with a JND of about 3–4 % of RT_{60} (Seraphim, 1958; Tsolias, Davies, *et al.*, 2014). When asked to adjust the reverberation times of one sound to match another, listeners performed better when the sounds were of the same type (*e.g.*, both were speech) than when they were very different (*e.g.*, one was speech and the other a guitar), *i.e.*, they confounded differences in signal with differences in reverberation time (Teret *et al.*, 2017). For the direct-to-reverberant ratio, the JND findings of Larsen *et al.* (2008) and Zahorik (2002b) were already cited in the context of auditory distance perception (see section 1.2.3).

In larger rooms, the times between reflections and absorptions at the walls are longer than in smaller rooms, which—all else being equal—leads to a slower decay of reverberation. Hameed *et al.* (2004) found that RT_{60} is an important cue when listeners are asked to compare the sizes of pairs of simplified simulated rooms, whereas the interpretation of the DRR in this context depended on the individual. Flanagin *et al.* (2017) performed a similar experiment using pairwise comparisons, though rather than using voice recordings as the stimuli, they let the listeners act as the sound sources themselves (they typically produced tongue clicks) and played back the calculated response of scaled versions of one measured real-world room binaurally via headphones. They adaptively varied the scaling factors to find the JND. In contrast to Hameed *et al.* (2004), the overall level of the room impulse responses was randomly varied across the two intervals in each trial, such that loudness was not available as a cue. Flanagin *et al.* (2017) found that listeners could detect changes in room size of approximately 10 %.

Zahorik (2009) presented speech stimuli convolved with pairs of 15 different (measured as well as simulated) binaural room impulse responses to human listeners and asked them to rate the similarity of the two room-acoustical settings in each trial. By means of a multidimensional scaling analysis, he identified two main parameters which affected similarity judgments: One was consistent with RT_{60} and the other with interaural cross-correlation (IACC) at frequencies above 500 Hz. IACC is a parameter which quantifies the similarity of the sounds arriving at both ears and has been linked to the sensation of *spaciousness* or *listener envelopment* (*e.g.* de Vries *et al.*, 2001).

Because the acoustics of a room can only ever be heard when there is excitation from a sound, a prerequisite for the perception of room acoustics is the ability to separate the source signal from the room response. This is not trivial; in fact, the underlying mathematical problem is that of deconvo-

lution, which is ill-posed if both the signal and the impulse response are unknown (blind deconvolution; Stockham *et al.*, 1975). Traer and McDermott (2016) recorded IRs of 271 typical environments, both indoors and outdoors, analysed their common acoustical characteristics, and synthesised new IRs which fit these characteristics and others which did not. In each trial of one psychophysical experiment, they presented three different stimuli to the listeners, two of which were convolved with the same synthetic IR whereas the remaining one differed. They found that listeners were significantly better at identifying the odd one out if the decay characteristics of the artificial IRs matched those of the measured real-world ones *vs.* when they were time-reversed or when they decayed in a physically implausible manner. They suggested that the auditory system *"has internalized the regularities of natural reverberation"* and that *"knowledge of environmental acoustics* [is] *internalized over development or evolution"* (Traer and McDermott, 2016).

1.5 Multimodal aspects of the auditory perception of space

Multisensory or *multimodal perception* is perception which arises from the integration of information from multiple sensory systems (see Bertelson and de Gelder, 2004). According to the theory of Stein and Meredith (1993), this integration is fostered by a coincidence of unisensory stimuli in place and in time, and characterised by relatively weak neuronal responses of the involved sensory systems in isolation which become disproportionately strong when they are combined. Physiological studies in mammals have often been centred around the *superior colliculus* (SC), a midbrain structure in which Meredith and Stein (1983) first described the amplification and attenuation of neuronal activity due to the convergence of inputs from the visual, auditory and somatosensory systems. It is now known that many brain regions, notably including cortical areas which had long been considered specific to one sensory system, exhibit similar behaviours (see Driver and Noesselt, 2008).

Multisensory perception is an expansive—and rapidly expanding—area of research. One reason for this can be found in the great number in which two or more senses could potentially work together in this way: This dissertation alone, for instance, combined auditory stimulation with visual input in some chapters (Chapters 2 and 4) and with motion input in another (Chapter 3). I have limited the introductory overview at hand to interactions between senses in the perception of space, and furthermore to ones in which the auditory system is involved.

One famous example is the *ventriloquism effect* (Howard and Templeton, 1966; Pick *et al.*, 1969): If a sound is emitted at one spatial location, but a compelling visual stimulus suggests that the source is elsewhere, then the overall perception of sound source location is guided by vision. Dominance of vision over audition is a common theme, perhaps because of the great accuracy of visual spatial information (see Witten and Knudsen, 2005) together with *optimal integration*, the theory that multisensory integration weights the information from each sensory modality so as to minimise variance (Ghahramani *et al.*, 1997). This is an explanation for the finding that the ventriloquism effect occurs for greater angular audiovisual discrepancies in elevation than in azimuth (Thurlow and Jack, 1973), considering the higher accuracy of binaural sound localisation. Moreover, Alais and Burr (2004) confirmed that when visual information is made unreliable, such as by blurring, auditory localisation takes over. The idea of optimal integration is also consistent with findings that audition trumps vision in timing-dependent discrimination tasks (Morein-Zamir *et al.*, 2003; Shams *et al.*, 2000), as the auditory system operates with a higher temporal precision.

1.5.1 Audiovisual distance perception

A preferential reliance on visual information in audiovisual distance perception tasks appears to be outside the scope of the ventriloquism term as regularly used in the literature. Such a dominance of vision might be expected because of the high accuracy of visually-based distance estimates (see da Silva, 1985). M. B. Gardner (1968b) indeed described such a phenomenon as the "proximity-image

effect", based on his observation that when five loudspeakers were visibly arranged in depth on a line in front of a listener in an anechoic environment, they perceived all sound stimuli to be coming from the source closest to them, even when a more distant speaker was actually the active one. Mershon, Desaulniers, *et al.* (1980) showed that in a similar setup, when the active sound source was closer to the listener than a visible dummy loudspeaker, but the actual source was hidden from view, overestimations of sound source distance could also occur.

More recently, Zahorik (2001) pointed out that the anechoic listening conditions in those experiments imply an absence of reliable absolute auditory distance cues. He found that in a experiment similar to M. B. Gardner's (1968b), but in a room with $RT_{60} = 0.3$ s where DRR cues were available, the proximity-image effect did not occur. However, Zahorik's (2001) group of listeners who was allowed to see the array of five loudspeakers before and during the auditory stimulation provided more accurate and less variable auditory distance judgments than those who were blindfolded. Calcagno *et al.* (2012) also obtained a boost in accuracy when simple visual distance markers were made available to listeners, even though the sound source was moved freely between these markers and was not visible itself.

In contrast to these experiments which offered no visual cues as to the true location of the sound source, Anderson and Zahorik (2014) provided their subjects with unambiguous visual distance cues (photos of a single loudspeaker in an otherwise empty auditorium) in the V and A+V trials of a study which compared distance perception with purely visual (V), purely auditory (A), and consistent audiovisual stimulation (A+V). They presented the binaural auditory stimuli via headphones in VAS and the visual stimuli via a large screen and found that the typical biases of auditory distance estimation, apparent in the A condition, were eliminated in V and A+V, and that the latter two conditions did not differ in performance.

1.5.2 Audiovisual perception of room acoustics

There is a relative paucity of literature on how the integration of auditory and visual cues shapes the perception of rooms. A relatively straightforward example is due to C. W. Bishop, London, *et al.* (2011) and concerned with the precedence effect. They demonstrated that the suppression of individual echoes can be modulated by visual input: When a diode emitted light just above a loudspeaker playing the leading sound, listeners experienced fusion more often than in an audio-only control condition; when the light indicated the source of the lagging source, fusion is inhibited instead.

McCreery and Calamia (2006) showed human subjects videos of a speaker in three different environments and asked them to adjust room-acoustical parameters according to their expectation of what the scene would sound like in reality. The subjects were able to predict the change in DRR associated with doubling the distance of the speaker from 6 m to 12 m. Similarly, Valente and Braasch (2008) played videos which showed an instrument being played in different rooms and found such an intuition when asking participants to adjust reverberation time (although their adjustments changed with the depicted instrument, for which there is no physical basis). Valente and Braasch (2010) combined different musical performances in VAS with composited videos which suggested that the source of the sound was either a loudspeaker, a performer, or a performer whose sound was amplified by two loudspeakers. They demonstrated that the visual stimulation can strongly alter listeners' judgments of the width of the sound field and of listener envelopment.

When it comes to the perception of room size, Maempel and Jentsch (2013) estimated, on the basis of size judgments obtained from subjects who heard auditory VAS stimuli in mismatch with stereoscopic photographs, that the visual input explained approximately one third of the total variance of the size estimates, whereas the auditory input explained one fifth. Maempel and Horn (2018) compared ratings of a variety of perceptual variables (including aesthetical, geometrical and emotional aspects) in visual, auditory and audiovisual conditions that subjects gave for a string-quartet recording played back in a real room (with or without blindfolds and/or ear muffs) and in a simulation of it

(headphone VAS and/or a stereoscopic photo). Influences of both modalities were apparent on *"loudness"*, *"envelopment"*, *"room brightness"* and *"hue"*. Interestingly, the only significant differences in ratings between the real and the virtual room occurred when the visual stimuli were involved.

1.5.3 Hearing and self-motion

When a listener moves their head or their entire body while a sound source is active, the sounds arriving at their ears change. For every such change, though, there is also a possible movement of the sound source which would produce the same acoustic result. Yet, when one turns their head to the right, they do not usually perceive a sound source to suddenly move circularly around them in a counterclockwise direction, even though the acoustic information might suggest it (see *e.g.* Yost *et al.*, 2015). Clearly, humans are generally able to disambiguate between actual changes in sound source position (in an "objective" and unchanging, *allocentric* spatial reference frame) and apparent changes due to self-motion (in the "subjective", body-centred and ever-changing, *egocentric* spatial reference frame; see *e.g.* Klatzky, 1998 for terminology). To achieve this, the brain must possess a mechanism to integrate auditory spatial information with self-motion information, which in turn arises at least from the vestibular system, the visual system, and neuronal feedback (see DeAngelis and Angelaki, 2012).

That self-motion could be beneficial for spatial hearing was already proposed by Wallach (1939, 1940). He hypothesised that listeners should be able to resolve ambiguities in sound source location, such as those due to the cone of confusion, by slightly rotating the head. This was later confirmed, for example by Bronkhorst (1995), Perrett and Noble (1997a,b), and Wightman and Kistler (1999), though the rotation involved appear to be less slight than Wallach (1940) had assumed (*e.g.*, more than 50° in azimuth in Wightman and Kistler, 1999). Many more authors have studied rotational self-motion in the context of auditory spatial perception in azimuth and elevation; recently, for example, Brimijoin and Akeroyd (2014), Freeman, Culling, *et al.* (2017), Genzel, Firzlaff, *et al.* (2016), and Yost *et al.* (2015). Simpson and Stanton (1973) found that allowing listeners to rotate their heads in all three spatial dimensions (turning and rolling it to the left and right, as well as pitching it to the front and back) did not afford them any benefits in a distance estimation task *vs.* when they had to keep their head stationary.

The underlying mechanisms of this particular aspect have also been studied to some extent. In a human EEG experiment, Altmann *et al.* (2009) found that only a purely egocentric change in the location of a sound source after a listener turned their head triggered a change in the evoked EEG response with a latency of 100–180 ms (termed *mismatch negativity*), whereas a purely allocentric change did not. Both manipulations, however, generated such a change after 220 ms. Based on these observations, they deduced that late, high-level cortical areas are responsible for the invariance of auditory spatial perception to turns of the head (but *cf.* Schechtman *et al.*, 2012, who presented contradictory evidence). In an electrophysiological study of ferrets, Town *et al.* (2017) reported that the majority of spatially sensitive neurons in auditory-cortical encoded an egocentric representation, there was also a small population of neurons which—across individuals and analytical approaches—exhibited allocentric tuning. Wigderson *et al.* (2016) even found head-direction sensitivity long before auditory information reaches the cortex, namely in rat CN. This is a remarkably early point for non-auditory information to enter into auditory processing (but not an isolated case; see Bizley and Dai, 2020 for a review).

The literature on the interaction of translational movements and the auditory perception of space is sparser. It is interesting to note that studies on this topic focus on distance perception instead of localisation in azimuth and elevation. Speigle and Loomis (1993) and Ashmead *et al.* (1995) appear to be the first to look for benefits of active translational movement (by walking) on auditory distance estimation. Both authors acquired distance estimates from the listeners by making them walk to the perceived source location after the sound was turned off, but their results were conflicting: Speigle and Loomis (1993) found no benefit of movement on distance localisation accuracy, but Ashmead

et al. (1995) did. There appears to be no evidence in the literature of any other experiments in this vein which might shed more light on the situation. More recently, Teramoto, Sakamoto, *et al.* (2012) and Teramoto, Cui, *et al.* (2014) found that vestibular acceleration signals, as well as visual flow suggesting translational motion, compress spatial information in the auditory system in the direction of movement.

As a side note, in a reversal of roles, dynamic auditory information can also support the perception of self-motion itself (see Campos *et al.*, 2018), or even create an illusion of self-motion in stationary listeners (*auditory vection*, see Väljamäe, 2009).

1.6 Acoustic communication within and between animal species

Chapters 4 and 6 are concerned with two particular aspects of (non-human) *communication*. This term is in common use in biology, but a widely agreed-upon definition is elusive (see Scott-Phillips, 2008). One popular view stresses the transfer of information between a sending and a receiving organisms (*e.g.* W. J. Smith, 1977; M. Stevens, 2013); another focuses on the sender altering the behaviour of the receiver in a way that provides an advantage to the reproductive success of either individual or both of them (*e.g.* Dawkins and Krebs, 1978; Maynard Smith and Harper, 2003). These approaches may differ in which behaviours can be considered acts of communication. Regardless, there is agreement that communication involves the transfer of mutually understood signals between individuals (see Bradbury and Vehrencamp, 2011), and that possible signals are manifold: They may, for example, be received by chemoreception including olfaction (*e.g.* in the courtship behaviour of the European newt, Malacarne and Giacoma, 1986), vision (*e.g.* the gestures of chimpanzees, Goodall, 1986), or audition. The latter modality is very familiar to us as humans, who possess a highly advanced form of *vocal communication* (in which the power to produce a sound is provided by the exhalation of air from the lungs; see Ploog, 1992) with unique capabilities (Hauser *et al.*, 2002), namely, spoken language. Vocal communication in general, however, is widespread across animals with a vibrating organ integrated into the respiratory tract, such as the larynx in mammals, amphibians and non-avian reptiles and the syrinx in birds (see Fitch, 2000). Moreover, some forms of acoustic communication are non-vocal. Stridulation, as a common example, is the generation of sound by rubbing different parts of the body against each other. It has been subject to extensive research in a wide variety of insects (*e.g.* Alexander, 1962; Markl, 1965; Michael and Rudinsky, 1972), but also occurs in many other animals.

1.6.1 Intraspecies acoustic communication

Intraspecies communication is communication between a sender and receiver which are *conspecific*, *i.e.*, members of the same species. As a direct interaction between such individuals, it is an important aspect of social behaviour and contributes to social organisation, survival and reproduction (see Wilson, 2000). Non-vocal, yet still acoustic intraspecies communication is possible (such as the distress signal of leaf-cutting ants; Markl, 1965), but not relevant to this dissertation. Simple forms of vocal communication, limited to a relatively small number of innate vocalisations, are common among animals with sufficiently advanced vocal production and auditory systems (Bradbury and Vehrencamp, 2011). A well-studied, straightforward example is given by the species-specific croaking of frogs, emitted by males to advertise their fitness and sexual readiness, and used by females to localise conspecific and attractive males (Wells and Schwartz, 2007).

One aim of the study of non-human vocal communication systems has been to gain insights about the evolutionary and developmental origins of human language. More complex animal models are clearly beneficial to achieve this goal. One crucial facet of complexity is described by the concept of *vocal learning*, which is the capability of an animal to acquire new vocalisations by experience, rather than being limited to an innate repertoire (*e.g.* Janik and Slater, 1997). Humans evidently possess this ability (Kuhl and Meltzoff, 1996), as do several birds, on which the vast majority of related literature

has been published (Lattenkamp and Vernes, 2018). The trait has also been found in elephants (Poole *et al.*, 2005) and some marine mammals (*e.g.* Reiss and McCowan, 1993).

Bats are also known to exhibit some degree of vocal learning (Boughman, 1998; Esser, 1994; Esser and U. Schmidt, 1989) and have recently gained attention as particularly interesting model organisms for comparative studies of vocal development in bats, humans, other mammals, and birds (Knörnschild, 2014; Vernes and Wilkinson, 2020). One particular species in this large order, the pale spear-nosed bat or *Phyllostomus discolor*, was the first to be recognised as a vocal learner and was highlighted by Lattenkamp (2020) for their ease of handling and the strong existing foundations of knowledge about their auditory and vocal systems. As an additional contribution to this ground work, Chapter 6 is a survey of their vocal repertoire and the different behaviours associated with their distinguishable vocalisations.

1.6.2 Interspecies acoustic communication

Communication between individuals which belong to different species is called *interspecies communication*. Acoustic or otherwise, there is no shortage of illustrative examples: certain interactions between humans and domesticated animals (*e.g.* Malavasi and Huber, 2016; Pepperberg, 2002; Pilley and Reid, 2011); the mutualism between humans and honey birds, wherein the latter guide the former to bee colonies, aiding both species in accessing a source of food (Spottiswoode *et al.*, 2016); the common understanding of the different alarm calls of one primate species by another (*e.g.* Fichtel, 2004); *etc*. Interspecies communication between prey and predators is also common. It often involves (honest or bluffing) warning signals sent out by a prey animal to fend off a predator.

Poulton (1890) introduced the term *aposematism* for *"an appearance which warns off enemies because it denotes something unpleasant or dangerous"* (quoted in Weldon, 2013), *i.e.*, it refers to an honest advertisment of a capability of self-defence. While originally described in the context of warning colouration, the modern use of the term includes other characteristics such as odours or sounds (*e.g.* Kirchner and Röschard, 1999; J. O. Schmidt, 2004). As one specific example of acoustic aposematism, certain caterpillars, specifically the larvae of some silkmoth, hawkmoth and saturniid species, produce click trains with their mandibles when they are under attack while secreting an unpalatable regurgitant (S. G. Brown *et al.*, 2007). Similarly, some tiger moths react to echolocation calls of approaching bats by producing ultrasonic clicks with an organ on their thorax, a behaviour that can flood the echolocation system with distracting information (*sonar jamming*; Corcoran *et al.*, 2009; Fullard *et al.*, 1979). Hristov and Conner (2005), however, found that big brown bats which are offered tiger moths (without having to hunt them) also learn to be deterred by their clicks after a previous experience of unpalatability, such that the moths' behaviour may be understood as aposematic. Tiger moths are also brightly coloured, a typical form of visual aposematism, whose co-ocurrence with the sound-generating behaviour can be understood as *multimodal signalling* which can increase effectivity in warding off predators that can see and hear (Rowe and Guilford, 1999).

Relatedly, a behaviour is *deimatic* if it serves to startle an attacker in order to cause its retreat (see Umbers *et al.*, 2015). As opposed to aposematism, an animal may engage in deimatic acts misleadingly, *i.e.*, even if it is palatable and does not possess a capability which could cause physical harm to the antagonist. Praying mantises, for example, react to threats with an elaborate deimatic behaviour (Maldonado, 1970) which includes visual components (opening of the mouth, raising of the thorax, *etc.*) as well as sound (created by stridulation with the wings and abdomen).

Deimatic and aposematic are not mutually exclusive attributes, however. Rattlesnakes, in a non-vocal sound-generating behaviour similar to stridulation (though arguably different from it as no rubbing takes place; Bradbury and Vehrencamp, 2011), shake their tails when threatened. This makes the keratinous segments at the tip of the tail collide and create a rattling noise (see Greene, 1969). Because potential foes are startled, this behaviour can be described as deimatic. At the same time,

because rattlesnakes are venomous, the signal can act as a honest warning that fulfills the definition of aposematism too (Fenton and Licht, 1990).

The tail-shaking behaviour involves some of the fastest muscles known in any vertebrate (Martin and Bagby, 1973). Rattling rates of up to 100 cycles per second have been observed, with the exact value depending strongly on body temperature (Chadwick and Rahn, 1954; Martin and Bagby, 1972).[6] Chapter 4 will present evidence that the rattling of western diamondback rattlesnakes also varies in rate as a function of the distance between the snake and an approaching threat, in a way that is suitable to affect auditory depth perception in humans.

1.7 Overview

Beside this introduction and the overall discussion in Chapter 7, this dissertation is made up of five distinct chapters, the first four of which are mainly concerned with the human auditory perception of space. I have light-heartedly organised them (in a scheme that should not be seen as particularly rigorous) into four "quadrants", based on whether or not some sound sources are in motion, and on whether or not the listeners are. The remaining chapter stands well apart by concentrating its attention on vocal communication in bats.

Chapter 2, *"Stationary listeners & stationary sources"*: A reproduction of a peer-reviewed publication of which I am the first author (Schutte *et al.*, 2019), with the original title "The percept of reverberation is not affected by visual room impression in virtual environments".

This chapter describes a study which explored whether simultaneous visual stimulation (based on a head-mounted display, *i.e.*, "VR goggles"), with either matching or misleading stimuli, affects how strongly listeners perceive reverberation in simulated room-acoustic scenes (presented via a ring of loudspeakers in an anechoic room).

Chapter 3, *"Moving listeners & stationary sources"*: A reproduction of a peer-reviewed publication of which I am one of two co-first authors (Genzel, Schutte, *et al.*, 2018), with the original title "Psychophysical evidence for auditory motion parallax".

The two experiments in this chapter investigated the ability of listeners to identify which of two sound sources emitted one sound, and which emitted the other, when the position of the two sources differed only in depth and the listening environment provided few distance cues. The central question was whether the listeners could improve their performance at this task when they moved (orthogonally to the line which connected the two sound sources) instead of remaining stationary.

Chapter 4, *"Stationary listeners & moving sources"*: A preprint of a peer-reviewed publication of which I am one of two co-first authors (Forsthofer, Schutte, *et al.*, 2021), with the original title "Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans".

An experiment on western diamondback rattlesnakes, reported at the beginning of this chapter, established that these animals alter the speed at which they move their rattle in response to the approach of a visual looming object. A subsequent psychophysical experiment (performed in VR, for obvious practical and ethical reasons) was used to test the hypothesis that this newly described rattlesnake behaviour affects the auditory distance perception of human listeners to the benefit of the snakes.

[6] Readers who understand German might find it handy to know that when a rattlesnake starts to engage in its proverbial behaviour, it might still be at it three hours later (Martin and Bagby, 1972).

Chapter 5, *"Moving listeners & moving sources"*: A chapter written specifically for this dissertation, parts of which will be turned into a technical/methodological paper at a later date.

Following a synopsis of common geometry-based approaches to the simulation of room acoustics and of the RAZR (Wendt *et al.*, 2014) simulation model in particular, this chapter describes how I implemented RAZR in a new computer program so that it can be run in a real-time, interactive manner, with freely moveable virtual sound sources and receivers. It concludes with a description of a study of the looming bias in the context of room acoustics, a proof-of-concept experiment for which I first used this new program.

Chapter 6, *"Listeners & sources with moving ears"*: A reproduction of a peer-reviewed publication of which I am the third author (Lattenkamp, Shields, *et al.*, 2019), with the original title "The vocal repertoire of pale spear-nosed bats in a social roosting context".

In the preparation of this chapter, pairs and small groups of adult *Phyllostomus discolor* bats were recorded with microphones while they were going about their daily social lives. The resulting repertoire is the first description of the diverse vocalisations used by this species (and the types of social interactions that each distinguishable vocalisation is associated with) in a communal context. My contribution to this paper was concerned with data analysis and presentation.

CHAPTER 2

Stationary listeners & stationary sources

This chapter is a reproduction of an open-access, peer-reviewed publication which has appeared in *The Journal of the Acoustical Society of America* on 13 March 2019. It can be accessed through the DOI 10.1121/1.5093642, or on the publisher's website under `http://asa.scitation.org/doi/10.1121/1.5093642`. The full citation is

> Michael Schutte, Stephan D. Ewert, and Lutz Wiegrebe (2019). "The percept of reverberation is not affected by visual room impression in virtual environments". In: *The Journal of the Acoustical Society of America* 145.3, EL229–EL235.

2.0 Author contributions

M.S., S.D.E. and L.W. designed the experiment. M.S. set up and performed the experiment, analyzed the data and prepared the figures. M.S., S.D.E. and L.W. interpreted the results. M.S. drafted the manuscript. M.S., S.D.E. and L.W. edited and revised the manuscript and approved its final version.

The percept of reverberation is not affected by visual room impression in virtual environments

Michael Schutte,[1,a)] Stephan D. Ewert,[2] and Lutz Wiegrebe[1]
[1]*Division of Neurobiology, Department Biology II and Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Germany*
[2]*Medical Physics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany*
michael.schutte@uiae.at, stephan.ewert@uni-oldenburg.de, lutzw@lmu.de

Abstract: Humans possess mechanisms to suppress distracting early sound reflections, summarized as the precedence effect. Recent work shows that precedence is affected by visual stimulation. This paper investigates possible effects of visual stimulation on the perception of later reflections, i.e., reverberation. In a highly immersive audio-visual virtual reality environment, subjects were asked to quantify reverberation in conditions where simultaneously presented auditory and visual stimuli either match in room identity, sound source azimuth, and sound source distance, or diverge in one of these aspects. While subjects reliably judged reverberation across acoustic environments, the visual room impression did not affect reverberation estimates.

[CCC]
Date Received: December 3, 2018 **Date Accepted:** February 21, 2019

1. Introduction

Our awareness of space, and subsequent orientation and navigation in space is dominated by the visual system, relying on the high-resolution topographic representation of space on the retinae of our eyes. Nevertheless, our auditory system can contribute: While the visual field is limited to the viewing direction, spatial hearing is omnidirectional, often guiding head and eye orientation.

Surrounding space affects sounds generated therein. It is very rare that we are in completely anechoic spaces. Instead, most man-made enclosed spaces, as well as natural enclosed spaces like caves and natural surroundings like forests (Traer and McDermott, 2016), produce echoes and reverberation that linearly distort the sound on its way from source to receiver. Humans and many other vertebrates possess dedicated perceptual strategies to compensate for sound reflections, summarized as the precedence effect [reviewed in Blauert (1997), Brown *et al.* (2015), and Litovsky *et al.* (1999)]. While some aspects of the precedence effect can be explained as by-products of peripheral auditory processing (Hartung and Trahiotis, 2001), several studies have demonstrated a high-level, cognitive contribution to precedence (Bishop *et al.*, 2014; Clifton, 1987; Clifton and Freyman, 1989; Clifton *et al.*, 2002; Tolnai *et al.*, 2014).

While the precedence effect describes a short-lasting perceptual phenomenon in listening conditions with a single echo that arrives within a few tens of milliseconds, there is also evidence of compensation mechanisms acting in more natural situations, where the effects of reverberation may last for hundreds of milliseconds. Studies have found that familiarity with a room environment can change thresholds in the categorical perception of speech sounds (Watkins, 2005) and improve speech intelligilibity (Brandewie and Zahorik, 2010). These results are consistent with the idea of a "dereverberation" process in auditory perception.

Current theoretical and empirical findings indicate that perceptual information from one sense, such as vision, influences evaluation and perception of information in other senses, such as hearing (Stein and Meredith, 1993). Examples of such "cognitive associations" between modalities can easily be found in everyday situations such as when information from visual senses ("this room looks like a typical concert hall") and auditory senses ("this room sounds like a typical concert hall") combine to form a total impression of the situation. A recent study has shown that auditory distance judgements are affected by vision in audiovisual virtual environments (Postma and Katz, 2017). Moreover, there is evidence that humans can estimate room acoustical

[a)]Author to whom correspondence should be addressed.

parameters based on photos (McCreery and Calamia, 2006), suggesting that we possess an intuitive awareness of the acoustics of rooms which we can see, but not hear.

Supporting the notion of multi-modal integration, it was demonstrated that, well beyond the classical ventriloquism effect, the visual system may also affect the way we deal with reflections of a sound source in enclosed spaces. It was shown that the strength of the precedence effect can be enhanced when the layout of a visual environment is consistent with the acoustically presented sounds and their reflections. Likewise, the precedence effect is diminished when visual and auditory environments are inconsistent (Bishop *et al.*, 2011, 2012). The precedence effect is typically relevant for suppressing spatial information of early reflections. However, it is to date unclear whether the visual impression of a room may affect the perception of later reflections which overlap in the reverberant tail of a room response. Depending on the temporal decay of the reverberation, technically characterized by the reverberation time (RT60) and the energetic ratio in relation to the direct sound [direct-to-reverberant ratio (DRR)] and early reflections, human can judge the perceived reverberation time and level [e.g., Lindau *et al.* (2014)]. Extrapolating from the documented effects of the visual system on classical precedence, we assess here whether perceived reverberation is affected by the congruence of the visual and auditory environments.

To this end, we quantified the extent to which subjects perceive the same auditory environment as less or more reverberant when the visual and auditory environments are congruent in comparison to a condition where the visual environment is not shown, or where it is incongruent with the auditory environment. We recruited latest audio-visual stimulation techniques to create highly realistic and immersive environments, and thus ensure that possible effects may be relevant in every-day listening situations.

2. Methods

2.1 Subjects and reproduction setup

Ten listeners (21–27 years of age, mean 24.3, 5 female) participated in the experiment. They were paid for their participation. Procedures were approved by the ethics committee of the Faculty of Medicine, LMU Munich (project No. 18–327).

The listeners were asked to quantify the perceived degree of reverberation in audiovisual and baseline audio-only conditions on a 1–10 integer rating scale. Five repetitions were measured for each trial. The stimulus in each trial paired one of 12 auditory environments with one of 12 visual environments (or the lack of a visual environment), although not all possible combinations were used (see below).

Listeners were seated in an anechoic chamber (2 m × 2 m base, 2.2 m high) and the auditory stimuli were presented via 36 loudspeakers (Plus XS.2, CANTON Elektronik, Weilrod, DE) mounted at head height near the chamber wall in a horizontal circular arrangement at 10° intervals in azimuth. The speakers were driven by four 12-channel power amplifiers (CI9120, NAD Electronics International, Pickering ON, CA) which received input from a PC via two 24-channel audio interfaces (24I/O, MOTU, Cambridge MA, US) running at a sampling rate of 48 kHz. The loudspeakers were fully equalized in spectral magnitude and phase by application of per-speaker compensation impulse responses. The root-mean-square sound pressure in the loudest conditions was 64 dB sound pressure level.

A head-mounted stereo display (Rift DK2, Oculus VR, Menlo Park CA, US) provided the visual stimuli to the subjects. The reference frames of the virtual visual and auditory environments were aligned with each other through careful placement of the infrared tracking camera. The subject's position in the virtual environment was kept fixed (i.e., the translational component of the head tracking readings was ignored for the real-time updates), and subjects were instructed to rotate their head in the horizontal plane, but not move it translationally, or otherwise rotate it. This was verified by a supervisor from outside the chamber via an infrared camera. Rotational head-tracking data also confirmed that the subjects complied with these instructions.

2.2 Stimuli

Auditory environments were defined by three variables: room identity (bedroom, office room, or factory hall); sound source azimuth (60° left, 0°, or 30° right); and sound source–listener distance (1 or 3 m). We used an improved version of RAZR (Wendt *et al.*, 2016) to simulate the room acoustics for these 18 environments, employing an image-source model for early reflections (up to third order), a scattering module and a feedback delay network for late reverberation. The direct sound and the early reflections were mapped onto 36 channels (corresponding to the locations of the

loudspeakers in the experimental chamber) by means of vector-based amplitude panning (Pulkki, 1997). The late reverberation was mapped onto twelve channels (three per chamber wall).

Visual environments were defined by the three variables room identity, visual source azimuth, and visual source–listener distance (with the same respective sets of possible values as listed above). The visual environments were rendered from 3-D geometric models to stereo equirectangular panoramic images with the CYCLES engine for BLENDER (Blender Foundation and community, 2018). They depict rooms with the same dimensions and wall materials as the corresponding auditory environments, with a TV set placed at the visual source position.

The visual and auditory stimuli and the experimental chamber are illustrated in Fig. 1.

2.3 Procedure

Two types of trials were presented in the experiment: audiovisual trials and audio-only trials. Each audiovisual trial combined an auditory environment with a visual environment such that

(a) the auditory and the visual room identities, source positions and sound source–listener distances are pairwise identical (congruent condition), or that
(b) the visual room identity differs from the auditory room identity while the other variables match (room identity incongruence), or that
(c) the visual source azimuth differs from the sound source azimuth while the other variables match (azimuth incongruence), or that
(d) the visual source–listener distance differs from the sound source–listener distance while the other variables match (distance incongruence).

Note that audiovisual trials can be incongruent in azimuth by up to 90° (by combining a 60° left auditory environment with a 30° right visual environment, or vice versa).

All auditory environments were also presented in audio-only trials, where the visual stimulus was a uniformly black image.

Trials of both types were presented in a random order. In each trial, a German-language speech signal extracted from a TV news show (no background music or other sounds; loudness-normalized according to EBU R 128) drawn randomly from a pool of 8 signals was convolved with the 36 impulse responses (all of which contained early reflections, and 6 of which also included late reverberation) for the corresponding auditory environment and played back. Simultaneously, a corresponding video clip was shown through the head-mounted display on the virtual TV screen at the visual source location. The video did not show the human speaker. The subjects

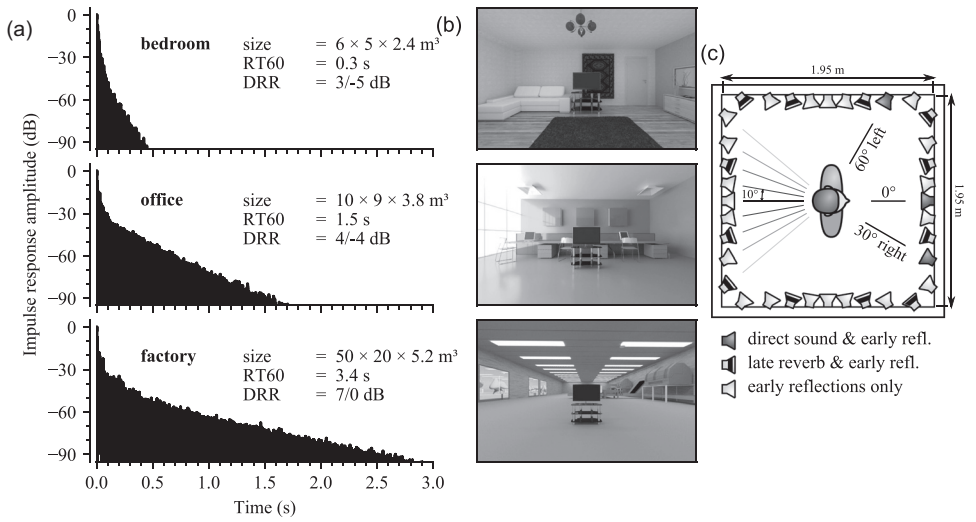

Fig. 1. Stimuli and experimental setup. (a) Estimated impulse responses for each of the three rooms and corresponding room dimensions, broadband reverberation times, and direct-to-reverberant energy ratios (for both the 1 and 3 m sound source distances). (b) Example image of one condition (visual source azimuth = 0°, visual source distance = 3 m) for the same condition. (c) Layout of the speakers inside the experimental chamber. In each trial, exactly one of the dark-shaded speakers played the direct sound from the virtual sound source, and the twelve speakers with a black band emitted late reverberation. All 36 speakers (mounted at 0° elevation) potentially rendered early reflections.

were instructed to look toward the TV screen during stimulus presentation, and to reproduce the up-down orientation of an arrow displayed on the virtual TV screen with the joystick (while the arrow was visible) to ensure that their eyes were open. Subsequently, they were asked to judge the perceived degree of reverberation ("wahrgenommene Verhalltheit") on a purely numeric 1–10 rating scale.

Prior to the experiment, the subjects were given the opportunity to listen to example stimuli representing each value on the rating scale. No visual stimulation took place during this familiarization phase. The familiarization stimuli were the same speech sounds, played through six speakers at 60° offsets in azimuth. The dry speech signal was convolved with random noise impulse responses with an exponentially decaying envelope for each speaker. The impulse responses ranged in broadband reverberation times similarly to those of the synthetic rooms ($RT_{60} = 100$ to 4000 ms). The carrier noise was shaped to be spectrally identical to the average magnitude spectrum of the synthetic room impulse responses.

3. Results

To analyze the results from the audio-only control and audiovisually congruent conditions, we fit a linear model with fixed and random effects (mixed-effects model) to the data, with the rating on the 1–10 scale as the dependent variable. The four independent variables auditory room identity, sound source distance, sound source azimuth, and presence of congruent visual stimulation ("visuals on/off") were included as fixed effects. First-order interactions of these fixed effects were also modelled. The data were grouped by the random factors subject (allowing the slopes for all fixed effects as well as the intercept to vary) and speech signal (allowing only the intercept to vary). All independent variables were treated as categorical.

An adapted F-test, using the approximation of Satterthwaite (1946) for degrees of freedom, revealed a significant ($\alpha = 0.05$) main effect of room identity [$F(2, 9) = 246.93$, $p < 10^{-7}$], and no other significant main effects (azimuth: $p = 0.93$, distance: $p = 0.43$, visuals on/off: $p = 0.45$). Per-subject mean ratings and standard deviations are shown in Fig. 2, panels #1–10 (audio-only control conditions in light grey, congruent conditions in dark grey). Estimated marginal mean differences were 3.20 ± 0.22 (s.e.m.) for bedroom-office, and 2.38 ± 0.26 for office-factory. Two-sided t-tests, again using the Satterthwaite approximation, established all pairwise room identity differences as significant (bedroom-office: $t = 21.08$, $p < 10^{-6}$; office-factory: $t = 9.172$, $p < 10^{-4}$; p-values Tukey-adjusted for multiple comparisons).

Further analysis also showed a significant interaction between sound source distance and room identity [$F(2, 3217) = 18.00$, $p < 10^{-7}$], and no other significant interactions (room identity and azimuth: $p = 0.87$; room identity and visuals on/off: $p = 0.07$; azimuth and sound source distance: $p = 0.14$; azimuth and visuals on/off:

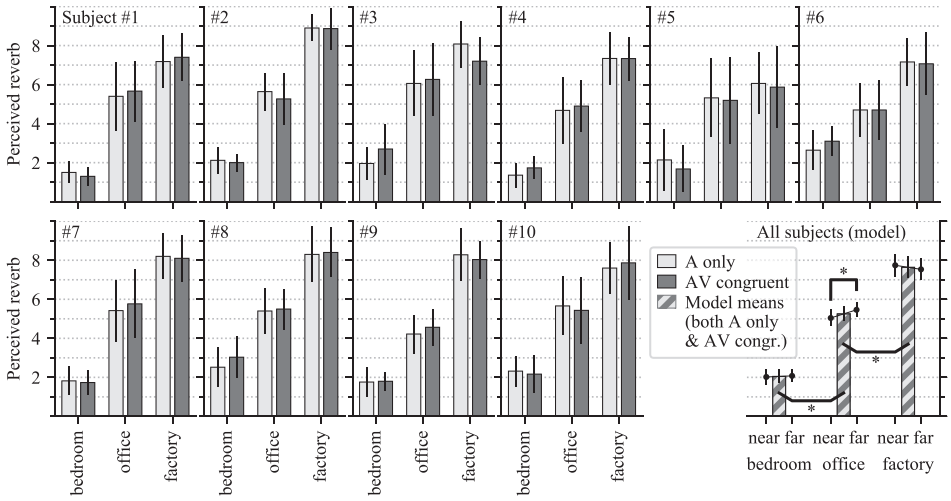

Fig. 2. (1–10) Individual subjects' reverberation ratings for the auditory room identities bedroom, office and factory. Light grey bars: mean scores per room and subject in the audio-only control conditions. Dark grey bars: mean scores per room and subject in the audiovisually congruent conditions. Black lines: Standard deviations. (Bottom right) Striped bars: Estimated marginal mean ratings (based on a mixed-effects model with subject and speech signal as random effects) for the three rooms, across all subjects and both audio-only control and audiovisually congruent conditions. Dots: Means conditioned on sound source distance (1 m vs 3 m). Black lines: 95% confidence intervals. Asterisks indicate significantly different ratings at $\alpha = 0.05$.

$p = 0.37$; distance and visuals on/off: $p = 0.59$). *Post hoc* t-tests conditioned on room identity showed a significant near-far difference in ratings (0.43 ± 0.12) only for the office room ($t = 3.666$, $p = 0.002$; Tukey-adjusted), and no significant differences based on presence or absence of visual stimulation (control conditions perceived as insignificantly more reverberant, 0.12 ± 0.09 for the factory, and insignificantly less reverberant, -0.11 ± 0.09 for the bedroom, -0.14 ± 0.09 for the office).

Thus, statistics confirmed that ratings for the bedroom conditions were overall lower than for office room conditions, and those were in turn lower than for factory room conditions. Perceived reverberation for near distances was only lower than for far distances in the intermediate office room ($RT_{60} = 1.5$ s). Notably, the presence or absence of simultaneous congruent visual stimulation did not affect the perception of reverberation. Modelled overall ratings are shown for the three rooms and six room-distance pairs in the lower right panel of Fig. 2 (control and congruent conditions averaged over).

In order to uncover possible effects of incongruence between the auditory and visual stimuli, we performed two-sided paired t-tests. First, we averaged the ratings from the ten repetitions obtained for each subject and audiovisually congruent condition where the room identity was either "office" or "factory" ("eligible congruent conditions"). We paired these averages with the average ratings over the ten repetitions obtained for the same subject and same auditory condition, but where the visual stimulus suggested a smaller room. In this way, each pairing contrasts two audiovisual conditions from the same listener, with only one difference between them: a fully congruent visual stimulation on one side vs one that differs from the auditory stimulation only in room identity (namely, a smaller visual room identity, "V smaller") on the other. Note that congruent conditions with a room identity of "bedroom" are not eligible in this specific comparison, as there was no smaller visual room identity.

We repeated this method of analysis for the other six types of audiovisually incongruent condition, at a time pairing the respective eligible congruent conditions with conditions in which: the visual stimulus suggests a larger room ("V larger"); the visual source distance is smaller, or larger, than the sound source distance ("V nearer" and "V farther," respectively); or the visual source azimuth differs from the sound source azimuth by 30°/60°/90° ("V 30° off," "V 60° off," and "V 90° off," respectively).

Figure 3 reproduces the raw data used in these seven t-tests in seven scatterplots. A systematic effect of a specific type of audiovisual incongruence would be evident in these plots by a shift of the data points away from the identity line, with a shift towards the horizontal/vertical axis suggesting a higher/lower perceived degree of reverberation in congruent conditions, respectively. We found no type of audiovisual incongruence that was rated significantly differently to congruent conditions by the subjects.

4. Discussion

The current data show that the subjects could reliably judge the degree of reverberation of a presented virtual environment. They show equally clearly that the subjects' reverberation judgements did not depend on whether a visual representation of an auditory environment was provided to them during listening, and that they were also not systematically affected by the congruence or incongruence of the presented auditory and visual environments.

Considering the dominance of the visual system in spatial awareness and its established influence on the perception of single echoes, it might on the one hand appear surprising that judgements of a room-related attribute of sound are entirely unaffected by congruent or incongruent visual input. On the other hand, this result is consistent with recent data which suggest that other spatial parameters (azimuth and compactness of the auditory image) are also assessed by listeners independently of the simultaneous visual impression (Gil-Carvajal *et al.*, 2016). These authors only detected an effect of visual stimulation on sound source distance, where multi-modal integration probably assigns a larger weight to visual information due to the relatively lower reliability of auditory cues for distance perception (Loomis *et al.*, 1998).

In this context, it might be important to note that it was easily possible for our subjects to determine the azimuthal position of the sound source not only as mediated by the visual stimulation (the position of the TV set in the virtual room), but also as mediated by the auditory stimulation. While the audio-visual stimulation technique employed in this experiment is theoretically suitable to elicit a visual capture effect in conditions where the visual source position deviates from the sound source position, this is unlikely to have taken place considering the magnitude of the azimuthal

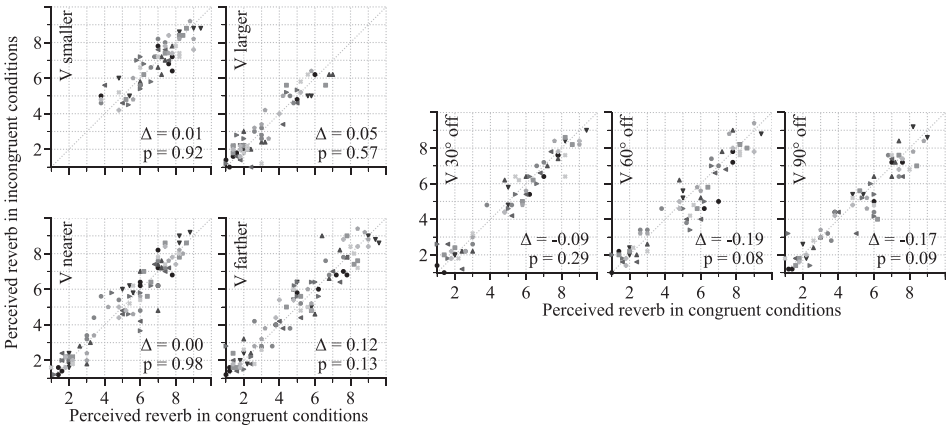

Fig. 3. Paired comparisons between each subject's average of five ratings in audiovisually congruent conditions (x axis) vs otherwise identical conditions which differ visually in one aspect (y axis). The type of divergence is shown in the top left corner of each sub-figure. Δ: mean difference of the points' y and x coordinates (smaller: higher reverberation rating in congruent conditions). p: p-value of a two-sided t-test between all x/y coordinate pairs shown. Marker shapes and greyscales denote individual subjects.

incongruence (30° or more), which exceeds typical thresholds of perceptual fusion (Hendrickx *et al.*, 2015).

It should not be overlooked that despite a lack of statistical significance at $\alpha = 0.05$, test statistics and potential effect sizes are largest when comparing conditions that modulate sound or visual source azimuth. Relatively high score differences and comparatively low p-values are observed between stimulus conditions that have a 0° vs either a 60° or a 90° discrepancy between sound and visual source azimuth, with lower perceived degrees of reverberation in the azimuthally incongruent conditions. We do not believe that this is due to audiovisual interactions, given that a better-ear effect in listening in rooms offers a simpler explanation: Because subjects always looked towards the visual source position in audiovisual conditions, one of their ears was turned towards the sound source when its location diverged in azimuth. This behaviour was found to improve spatial release from masking in a single-speaker, single-masker condition (Grange and Culling, 2016) and may well affect the perception of reverberation.

It should be noted that the initial audio only training for the judgement of reverberation might have primed a listener's focus on the auditory cues while tending to more disregard visual cues. It is unclear whether results might be different for more naive listeners which do not receive audio only training ahead of the task and which judge perceived reverberation of the overall scenario or its effects on perception in a more indirect task.

Acknowledgments

This research was supported by the Munich Center for Neurosciences, the Bernstein Center for Computational Neuroscience Munich, the Graduate School of Systemic Neurosciences, the Deutsche Forschungsgemeinschaft (DFG) FOR 1732 (TPE) to author S.E., and the DFG Grant No. wi1518/17 to author L.W. We thank Baccara Hizli for her assistance in data acquisition. The visual stimuli depicted in Fig. 1 are based on 3D models provided by Rui Teixeira (bedroom; public domain), DragonautX on blendswap.com (office; CC-BY 3.0), and Fatih Eke (factory; CC-BY 3.0).

References and links

Bishop, C. W., London, S., and Miller, L. M. (**2011**). "Visual influences on echo suppression," Curr. Biol. **21**(3), 221–225.

Bishop, C. W., London, S., and Miller, L. M. (**2012**). "Neural time course of visually enhanced echo suppression," J. Neurophys. **108**(7), 1869–1883.

Bishop, C. W., Yadav, D., London, S., and Miller, L. M. (**2014**). "The effects of preceding lead-alone and lag-alone click trains on the buildup of echo suppression," J. Acoust. Soc. Am. **136**(2), 803–817.

Blauert, J. (**1997**). *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).

Blender Foundation and community (**2018**). Blender 2.79b—free and open 3D creation suite, https://www.blender.org/.

Brandewie, E., and Zahorik, P. (**2010**). "Prior listening in rooms improves speech intelligibility," J. Acoust. Soc. Am. **128**(1), 291–299.

Brown, A. D., Stecker, G. C., and Tollin, D. J. (**2015**). "The precedence effect in sound localization," J. Assoc. Res. Otolaryng. **16**(1), 1–28.
Clifton, R. K. (**1987**). "Breakdown of echo suppression in the precedence effect," J. Acoust. Soc. Am. **82**(5), 1834–1835.
Clifton, R. K., and Freyman, R. L. (**1989**). "Effect of click rate and delay on breakdown of the precedence effect," Percept. Psychophys. **46**(2), 139–145.
Clifton, R. K., Freyman, R. L., and Meo, J. (**2002**). "What the precedence effect tells us about room acoustics," Percept. Psychophys. **64**(2), 180–188.
Gil-Carvajal, J. C., Cubick, J., Santurette, S., and Dau, T. (**2016**). "Spatial hearing with incongruent visual or auditory room cues," Sci. Rep. **6**, 37342.
Grange, J. A., and Culling, J. F. (**2016**). "The benefit of head orientation to speech intelligibility in noise," J. Acoust. Soc. Am. **139**(2), 703–712.
Hartung, K., and Trahiotis, C. (**2001**). "Peripheral auditory processing and investigations of the 'precedence effect' which utilize successive transient stimuli," J. Acoust. Soc. Am. **110**(3), 1505–1513.
Hendrickx, E., Paquier, M., Koehl, V., and Palacino, J. (**2015**). "Ventriloquism effect with sound stimuli varying in both azimuth and elevation," J. Acoust. Soc. Am. **138**(6), 3686–3697.
Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (**2014**). "A spatial audio quality inventory (saqi)," Acta Acust. Acust. **100**(5), 984–994.
Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**(4), 1633–1654.
Loomis, J. M., Klatzky, R. L., Philbeck, J. W., and Golledge, R. G. (**1998**). "Assessing auditory distance perception using perceptually directed action," Percept. Psychophys. **60**(6), 966–980.
McCreery, A., and Calamia, P. (**2006**). "Cross-modal perception of room acoustics," J. Acoust. Soc. Am. **120**(5), 3150–3150.
Postma, B. N., and Katz, B. F. (**2017**). "The influence of visual distance on the room-acoustic experience of auralizations," J. Acoust. Soc. Am. **142**(5), 3035–3046.
Pulkki, V. (**1997**). "Virtual sound source positioning using vector base amplitude panning," J. Audio Eng. Soc. **45**(6), 456–466.
Satterthwaite, F. E. (**1946**). "An approximate distribution of estimates of variance components," Biometr. Bull. **2**(6), 110–114.
Stein, B. E., and Meredith, M. A. (**1993**). *The Merging of the Senses* (MIT Press, Cambridge, MA).
Tolnai, S., Litovsky, R. Y., and King, A. J. (**2014**). "The precedence effect and its buildup and breakdown in ferrets and humans," J. Acoust. Soc. Am. **135**(3), 1406–1418.
Traer, J., and McDermott, J. H. (**2016**). "Statistics of natural reverberation enable perceptual separation of sound and space," Proc. Natl. Acad. Sci. **113**(48), E7856–E7865.
Watkins, A. J. (**2005**). "Perceptual compensation for effects of reverberation in speech identification," J. Acoust. Soc. Am. **118**(1), 249–262.
Wendt, T., van de Par, S., and Ewert, S. D. (**2016**). "Perceptually plausible acoustics simulation of single and coupled rooms," J. Acoust. Soc. Am. **140**(4), 3178–3178.

CHAPTER 3

Moving listeners & stationary sources

This chapter is a reproduction of a peer-reviewed publication which has appeared in the *Proceedings of the National Academy of Sciences of the United States of America* (PNAS) on 7 April 2018. I share its first authorship with Daria Genzel. Its full text can be accessed through the DOI 10.1073/pnas.1712058115, or on the publisher's website under `https://www.pnas.org/content/115/16/4264`. The full citation is

> Daria Genzel, Michael Schutte, W. Owen Brimijoin, Paul R. MacNeilage, and Lutz Wiegrebe (2018). "Psychophysical evidence for auditory motion parallax". In: *Proceedings of the National Academy of Sciences* 115.16, pp. 4264–4269.

3.0 Author contributions

M.S. and L.W. designed and set up Experiment I; D.G., W.O.B., P.R.M. and L.W. designed and set up Experiment II. M.S. performed Experiment I; D.G. and P.R.M. performed Experiment II. D.G., M.S., P.R.M. and L.W. analyzed the data and prepared the figures. All authors wrote and revised the manuscript and approved its final version.

Psychophysical evidence for auditory motion parallax

Daria Genzel[a,b,1], Michael Schutte[a,1], W. Owen Brimijoin[c], Paul R. MacNeilage[b,d,2], and Lutz Wiegrebe[a,b,3]

[a]Department Biology II, Ludwig Maximilians University Munich, 82152 Planegg-Martinsried, Germany; [b]Bernstein Center for Computational Neuroscience Munich, 82152 Planegg-Martinsried, Germany; [c]Glasgow Royal Infirmary, Medical Research Council/Chief Scientist Office Institute of Hearing Research (Scottish Section), G31 2ER Glasgow, United Kingdom; and [d]Deutsches Schwindel- und Gleichgewichtszentrum, University Hospital of Munich, 81377 Munich, Germany

Edited by Wilson S. Geisler, University of Texas at Austin, Austin, TX, and approved February 16, 2018 (received for review July 6, 2017)

Distance is important: From an ecological perspective, knowledge about the distance to either prey or predator is vital. However, the distance of an unknown sound source is particularly difficult to assess, especially in anechoic environments. In vision, changes in perspective resulting from observer motion produce a reliable, consistent, and unambiguous impression of depth known as motion parallax. Here we demonstrate with formal psychophysics that humans can exploit auditory motion parallax, i.e., the change in the dynamic binaural cues elicited by self-motion, to assess the relative depths of two sound sources. Our data show that sensitivity to relative depth is best when subjects move actively; performance deteriorates when subjects are moved by a motion platform or when the sound sources themselves move. This is true even though the dynamic binaural cues elicited by these three types of motion are identical. Our data demonstrate a perceptual strategy to segregate intermittent sound sources in depth and highlight the tight interaction between self-motion and binaural processing that allows assessment of the spatial layout of complex acoustic scenes.

depth perception | distance discrimination | spatial hearing | self-motion | auditory updating

Humans' dominant sense for space is vision. The exceptional spatial resolution and acuity of foveal–retinal vision allows for accurate and simultaneous localization of multiple objects in azimuth and elevation (1). The observer's distance to an object, however, is more difficult to assess. In vision, distance of near objects is mainly encoded by binocular disparity which relies on image differences resulting from the spatially separate views of the two eyes onto the object (2–4); these differences become minimal for far objects. Higher visual centers integrate disparity with information arising from monocular cues, many of which provide information about relative depth separation, rather than absolute distance. These include occlusion of one object by another one, relative size, perspective, shading, texture gradients, and blur (5, 6). Important information about relative depth is also added when there is motion of the observer relative to the environment or object: the resulting difference in image motion between features at different depths is termed motion parallax (3, 6). In the case of observer motion, relative depth from motion parallax can be scaled to obtain absolute estimates of object distance if information about speed of observer motion is available, for example, based on vestibular signals (7). Such scaling cues are generally not available when the object moves relative to the stationary observer.

Apart from the visual system, only audition allows locating objects (i.e., sound sources) in the far field beyond the range of touch. As in vision, azimuth and elevation of the sound sources are readily encoded through auditory computation, both binaural (interaural time and level differences, ref. 8) and monaural (elevation-dependent analysis of pinna-induced spectral interference patterns, ref. 9). But again, the distance to a sound source is most difficult to assess: in the absence of reverberation, and without a priori knowledge about the level and spectral composition of the emitted sounds, distance estimation for humans is indeed impossible (10). This is not surprising, considering that an important visual distance cue (binocular disparity) is not available in audition, not least because humans cannot point each of their ears toward a sound source. Some visual depth cues have auditory counterparts, (e.g., blur is related to frequency-dependent atmospheric attenuation, and relative size to loudness), but many others are unavailable (e.g., occlusion, texture gradients, shading).

In reverberant rooms, the ratio of the sound energy in the first wave front relative to the energy reflected from the surfaces is a function of distance and allows the estimation of sound-source distance without motion (11–14). Recent theoretical work has pointed out that motion of the interaural axis (and specifically translational head motion) also allows fixing sound-source distance, through the analysis of auditory motion parallax (15). To date, however, it is unexplored to what extent auditory motion parallax may be exploited by human subjects to perceptually segregate sound sources in distance and how the time-variant binaural cues that are generated by translational head motion are integrated with vestibular and/or proprioceptive cues for auditory distance perception. After an early report that "head movement does not facilitate perception of the distance of a source of sound" (16) work by Loomis and coworkers (17, 18)

Significance

When we cannot see a sound source, it is very difficult to estimate how far away it is. In vision, motion parallax facilitates depth perception in that when we move, nearer objects move more in our visual field than farther objects. Our experiments show that humans can also exploit motion parallax for sound sources. Moreover, we show that, as in the visual system, facilitation of auditory depth perception is stronger when the subjects move actively than when subjects are being moved or when only the sound sources move. We conclude that dedicated mechanisms exist that integrate self-motion with binaural auditory motion cues, allowing humans to exploit auditory motion parallax.

Author contributions: D.G., M.S., W.O.B., P.R.M., and L.W. designed research; D.G., M.S., and P.R.M. performed research; D.G., M.S., P.R.M., and L.W. analyzed data; and D.G., M.S., W.O.B., P.R.M., and L.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: Data have been deposited in the G-Node Network of the Bernstein Center for Computational Neuroscience (available at https://doid.gin.g-node.org/5103a1db3d918f82a1724c8d90f6ca0b/).

See Commentary on page 3998.

[1]D.G. and M.S. contributed equally to this work.

[2]Present address: Department of Psychology, Cognitive and Brain Sciences, University of Nevada, Reno, NV 89523.

[3]To whom correspondence should be addressed. Email: lutzw@lmu.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1712058115/-/DCSupplemental.

Published online March 12, 2018.

has shown that dynamic binaural cues elicited by translational self-motion relative to a stationary sound source may provide some (rather erroneous) information about the absolute distance of a sound source for tested source distances between two and six meters. More recent work has highlighted the interaction of self-motion (real or visually induced) on the perception of auditory space: Teramoto et al. (19, 20) have shown that self-motion distorts auditory space in that space is contracted into the direction of self-motion, regardless of whether the self-motion was real (which provided a vestibular signal) or visually induced (which provides no vestibular input but only visually mediated self-motion information). However, it remains unclear whether self-motion can support the segregation of sound sources in distance through an auditory motion parallax and how proprioceptive and vestibular inputs may contribute to this segregation.

Here we present formal psychophysical data showing that humans can segregate a high-pitched sound source from a low-pitched sound source in distance based on the time-variant binaural perceptual cues associated with motion. The initial demonstration of auditory motion parallax is implemented as a forced-choice experiment with real sound sources positioned at different depths in anechoic space that have been carefully calibrated to eliminate nonmotion-based cues to distance. In a second experiment, we instead elicit differences in perceived depth of sound sources positioned at the same depth by rendering sounds contingent on head tracking, and we show that this exploitation of auditory motion parallax is facilitated by both vestibular and proprioceptive information arising from active self-motion.

Results

Seven subjects were asked to respond whether a high-pitched sound source was closer or farther away than a low-pitched source. The two sound sources were temporally interleaved, i.e., the sum of the sound sources was perceived as alternating in pitch over time at a rate of 10 Hz for each source or 20 Hz for the summed sources (*Materials and Methods*).

Careful steps were taken to eliminate nonmotion-based auditory cues to distance. Consequently, at each position of the sound sources, sound level and spectral content of each sound source was identical when measured either with an omnidirectional microphone at the center of the subjects' interaural axis or when measured binaurally with a Bruel & Kjaer (B&K) 4128C Head-and-Torso Simulator (*Materials and Methods*). An illustration of the experimental setup, the stimulus, and the psychophysical results is shown in Fig. 1.

When the sound sources were separated in distance by only 16 cm (leftmost data in Fig. 1*C*), subjects could not solve the task when they were not allowed to move; performance was around chance level, 50% (black symbols and line). However, when the subjects were allowed to move their heads laterally by ±23 cm (green symbols and line), performance was much better and subjects scored on average 75% correctly even at the smallest presented distance difference of 16 cm. With increasing distance difference between the sound sources, performance quickly improved when active self-motion was allowed while performance stayed rather poor without active motion. Nevertheless, some subjects could discriminate the sound sources without self-motion for larger distance differences. Possible residual distance cues are discussed below. Of the 42 pairs of performances (seven subjects times six distance differences) performance in the active-motion condition (AM) was significantly better than in the no-motion condition (NM) in 29 cases. In no case was performance better without motion than with motion (Fisher's exact test, $P < 0.05$, P values corrected for multiple testing with the Benjamini–Hochberg procedure). These data clearly show that human subjects can easily exploit auditory motion parallax to segregate sound sources in depth. To this end, subjects likely exploit time-variant binaural cues arising from the lateral self-motion: with a given lateral motion, the closer object creates larger binaural cues because it covers a larger range of azimuthal angles relative to the subject's moving head. The role of self-motion and its interaction with dynamic binaural processing is further explored in the following experiment.

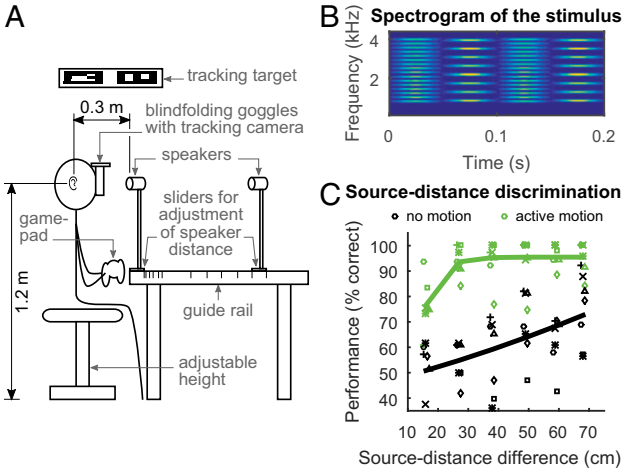

Fig. 1. Illustration of the experimental setup (*A*), the stimuli (*B*), and psychophysical results (*C*) to demonstrate auditory motion parallax in Exp. I. (*A*) Subjects were seated with their interaural axis exactly perpendicular to the axis of two miniature broadband loudspeakers. One randomly chosen speaker emitted the high-pitched sound, the other speaker emitted the low-pitched sound. Head motion in each trial was continuously recorded with a head-tracking system consisting of a tracking camera on the subjects' heads and a tracking target at the ceiling. (*B*) Spectrogram of a 0.2-s section of the intermittent low- and high-pitched stimulation in each trial. The two different pitches are presented by the two speakers at different depths. (*C*) Individual performances (marked by different symbols) and sigmoidal fit to average performance (solid lines) with motion (green) and without motion (black). The data show that with motion, subjects discriminate sound-source distances overall quite well, whereas performance hardly deviated from chance level without motion.

Here sound sources were presented in virtual space via a linear high-resolution loudspeaker array which precluded the use of distance-dependent loudness and reverberation cues, so that it was not necessary to change calibration dependent on virtual-source distance (*Materials and Methods*). The motion conditions were as follows: NM, subjects remained positioned with their head in line with the two sound sources at different distances; AM (Fig. 2 *A* and *B*), subjects actively moved their upper body by about 23 cm left and right following a previously trained motion profile (these two conditions were the same as in the first experiment) (*Materials and Methods*); passive motion (PM) (Fig. 2*C*): subjects did not move, but the subjects were moved by a motion platform such that the subject's head moved in the same way as in the AM condition; and sound-source motion (SSM) (Fig. 2*D*), subjects remained still but the sound sources presented via the array moved such that the relative motion between the sound sources and the subject's head in azimuth was the same as in the AM and PM conditions. Twelve subjects took part in this second experiment.

Without any motion of either the sound sources or the subjects, none of the subjects could reliably determine whether the high-pitched sound source was nearer or farther than the low-pitched sound source. Performance of an example subject in the experimental condition without motion (NM) is represented by the black asterisk in Fig. 3. The failure to discriminate distances is not surprising because loudness cues related to absolute distance were quantitatively removed and the use of the speaker array for virtualization (*Materials and Methods*) precluded the

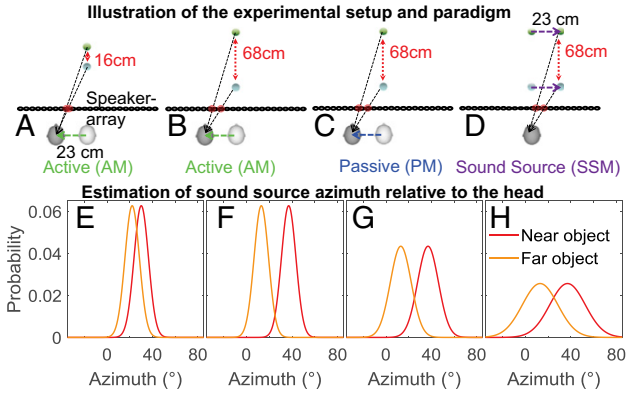

Fig. 2. Illustration of the setup and paradigm of Exp. II and the hypothesis. (*A–D*) Subjects were trained to move parallel to the speaker array with the same motion profile as in Exp. I. Subjects performed the motion either actively (AM) (*A* and *B*), or they were moved by a motion platform (PM) (*C*). In these conditions, tracking of the head motion relative to the array and the virtual sound sources allowed us to update the speaker activation in real time. In the SSM condition (*D*) the sound sources moved along the array but the subjects were stationary. Speaker activation is illustrated by the red area around the speakers. (*E–H*) With increasing depth separation, dynamic binaural cues get stronger (*E* and *F*). AM provides additional information (proprioceptive and efference copy signal) and leads to better discrimination (*F*). During PM, only vestibular signals provide additional information (*G*). Discrimination is therefore worse than for AM, but better than for SSM, where only dynamic binaural cues are present (*H*).

use of differential reverberation cues for distance estimation. When we trained our subjects to move laterally during the presentation of the alternating high- and low-pitched sources, the subjects improved their ability to identify which sound source was nearer. In principle, this question can be answered by identifying the nearer source as the one whose perceived azimuthal angle changes more during the lateral self-motion. An example of depth discrimination performance as a function of source distance difference is shown in Fig. 3. Performance in the AM condition is shown in green. This subject could reliably judge whether the high-pitched source was closer or farther away than the low-pitched source when the closer source was 40 cm and the farther source was 56 cm away from the subject, i.e., the distance difference was only 16 cm. However, performance deteriorated when the subject was passively moved by a motion platform (PM, blue curve), or when the sound sources moved (SSM, purple curve).

The validity of the direct comparison between the motion conditions depends on the precision of the actively executed motion and how well this motion is reproduced by the motion platform. A comparison of the active and passive motion profiles is found in *Supporting Information*.

Distance-difference thresholds are shown in Fig. 4*A*, individual data represented by the colored bars and Fig. 4*B*, medians and interquartiles represented by the box plots. The data clearly show that subjects performed best [just-noticeable distance differences (JNDs) were smallest] when they actively moved in front of the virtual sound sources (AM). Performance was significantly worse when subjects were moved by the motion platform (PM). When the subjects were stationary but the sound sources moved (SSM), thresholds were worst. In this condition, some of the subjects could not solve the task even for the largest source-distance difference, 68 cm. In Fig. 4, data from these subjects are artificially set to a threshold of 80 cm; note, however, that real perceptual thresholds may be larger. In summary, these data confirm that also with virtual sound sources, subjects can resolve distance differences between sound sources quite well when they move in a manner that exploits auditory motion parallax. The fact that they performed worse with passive motion indicates that both proprioceptive and vestibular signals are integrated with dynamic binaural cues to solve the task. Visual cues were unavailable because the subjects were blindfolded. Without proprioceptive and vestibular signals, i.e., without motion of the subject, performance was significantly worse, which shows that the dynamic binaural cues alone (which were the same in all three conditions of Exp. II) do not suffice to provide the best performance. Results in the AM condition compare well across Exps. I and II: The average threshold for 75% correct performance was about 16 cm sound-source difference in Exp. I and 20 cm source difference in Exp. II. This is true although the setups differed substantially (real sound sources in Exp. I vs. simulated sound sources in Exp. II).

Discussion

The current psychophysical experiments support the hypothesis that human subjects can exploit auditory motion parallax to discriminate distances of sound sources. Thus, the capacity to exploit motion parallax to disambiguate sensory scenes is shared between the senses of vision and audition. Importantly, subjects received no trial-to-trial feedback about their performance in Exp. I. When asked to respond to whether the high-pitched sound source was closer or farther than the low-pitched source, subjects appeared to readily exploit motion parallax when they were instructed to move, without extensive training. The perceptual basis for auditory motion parallax is that, through lateral motion of either the objects or the subject, the distance difference between the objects is transferred into time-variant horizontal localization cues. For a given lateral motion, the closer object produces the stronger variation in horizontal localization cues, i.e., interaural time differences (for the lower part of the

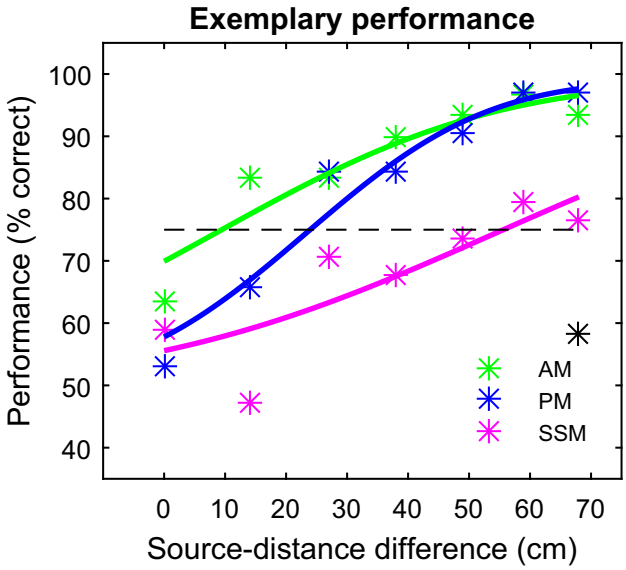

Fig. 3. Exemplary performance (symbols) and fitted psychometric functions (lines) for depth discrimination of two alternating sound sources as a function of their distance difference in Exp. II. Performance is best in the AM condition (green) where the subject performed an active head motion and worse in the SSM condition (purple) where the subject was stationary, but the sound sources moved past him or her. When the subject was moved by the motion platform past the virtual sound sources (PM, blue), performance was intermediate. Without both subject- and sound-source motion (NM), the subject could not solve the task even at the largest distance difference of 68 cm (single black star). Therefore, full psychometric functions were not obtained in the NM condition.

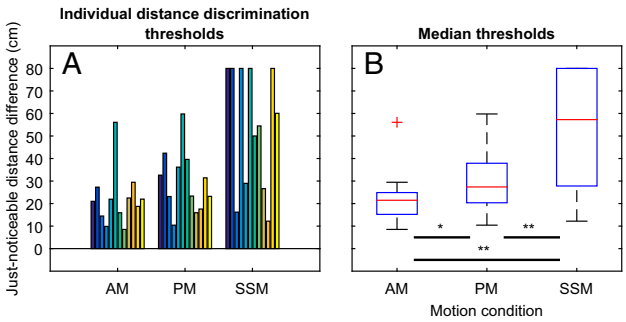

Fig. 4. Psychophysical performance thresholds (just-noticeable differences) for sound-source distance in Exp. II. Individual data are shown by the colored bars in *A*; boxplots of medians (red) and interquartiles (blue boxes) are provided in *B*. Whiskers represent the data range expressed as the 75th percentile plus 1.5 times the difference between the 75th and the 25th percentile (maximum range) and the 25th percentile minus 1.5 times the difference between the 75th and the 25th percentile (minimum range). The red cross in *B* represents the only value outside the whisker range (outlier). Nonparametric paired comparisons (Wilcoxon signed rank tests) show that performance in the AM condition is significantly better than in both the PM (*$P < 0.05$, signed rank = 10) and the SSM condition (**$P < 0.01$, signed rank = 4), and that performance in the PM condition is significantly better than in the SSM condition (**$P < 0.01$, signed rank = 6).

stimulus spectrum below about 1 kHz) and level differences (for the higher-frequency parts) change faster for the closer sound source than for the farther source. Thus, while self-motion may have limited value in estimating absolute distance to a single sound source (17, 18), the current experiments demonstrate that self-motion readily supports segregation of sound sources in depth.

The dynamic binaural cues in the current experimental conditions with motion (AM, PM, and SSM) are equally salient, no matter whether the subject moves actively, the passive subject is moved, or the objects move. Nevertheless, the current data show that subjects are more sensitive to distance differences when they move actively than when they are moved or when the objects move.

For the visual system, it has long been known that viewers can use motion parallax to estimate distances of objects (2, 3, 21) not only by humans but also by, e.g., Mongolian gerbils (22). Interestingly, also in vision, distance estimation is better when the viewer moves than when the objects move (23). Thus, the current data corroborate previous conclusions, drawn for the visual system, that self-motion information facilitates the depth segmentation of sensory scenes.

While with virtual sound sources (Exp. II) subjects failed completely to discriminate sound-source distances without motion (cf. Fig. 3), some subjects could discriminate large distance differences between real sound sources (Exp. I, data in black in Fig. 1*C*). Close inspection of binaural room impulse responses recorded from the two sound sources with a head-and-torso simulator indicate that this may be related to residual low-frequency reflections in the experimental booth. The booth was fully lined with acoustic foam of 10-cm thickness, resulting in a lower cutoff frequency of the damping to around 1 kHz. Given that the lower cutoff of our stimulation was at 800 Hz, it is possible that some subjects exploited residual reverberation cues to solve the distance discrimination task even without motion. Nevertheless, the data clearly show that motion-induced perceptual cues are dominant in solving the task.

In purely geometric terms, there is a limit to the extent to which motion parallax may be used to resolve a difference in distance between two sources. Assuming perfect detection, quantification, and temporal integration of an observer's own physical motion, successful source-distance discrimination could only occur if the subject's motion were to result in a difference in subtended angle between the two sources that is equal to or larger than the minimum detectable change in source angle over time. In the auditory system, this limit is imposed by the minimum audible movement angle; in the visual system, it is imposed by the spatial displacement threshold. The fundamental constraint applied by these angular acuity thresholds may be formalized in Eq. **1**:

$$d' = \tan\left(a\tan\left(\frac{d}{x}\right) - \Theta\right) \times x, \qquad [1]$$

where d is the distance of the farther target, x is the amount of lateral motion, Θ is the angular acuity threshold, and d' is the distance to a closer target that is just discriminable.

At ideal source velocities, signal characteristics, and contrasts, the lowest auditory motion detection threshold is roughly 2° (24, 25), whereas the threshold in the visual system is at least 100 times smaller at roughly 1 arcmin, or 0.017° (26). In the framework of Eq. **1** it is clear that the visual system should be more capable of using motion parallax to discriminate distance than the auditory system. By using each modality's values for Θ in Eq. **1**, we can estimate that for a maximum lateral displacement of 23 cm from the loudspeaker axis and a distance of the farther target of 52 cm, the auditory system should begin to detect a difference when the closer target was at about 47 cm. In Exp. II, only our best subject could reliably discriminate 45 cm from 52 cm, i.e., a distance difference of 7 cm in the AM condition. Thus, even with optimal cue combination, our subjects performed worse than predicted from auditory motion detection of a single sound source.

In the visual system, on the other hand, in an equivalent task with the same lateral motion, a difference should become perceivable with the closer object being only 4 mm closer than the farther object. In practice, parallax distance acuity in the visual system may be yet more accurate even than this, due to the ability to compare signals at the two eyes (27) and the use of eye motion itself (28), a mechanism unavailable to the human auditory system. Critically, parallax-based distance discrimination becomes poorer as a function of distance for both visual and auditory objects. Given the lower spatial acuity, this is especially impactful for auditory signals: for a sound source at 4 m and an orthogonal listener motion of 20 cm, a second sound source would have to be about 1.6 m closer to be discriminable in depth.

These computations assume a perfect assessment and use of observer motion. Combination of motion signals with other sensory input is known to be imperfect and this has been established in the visual system (29, 30), auditory system (31), and even the somatosensory system (32). Given this additional source of error, it is likely that the true depth discrimination thresholds are higher than estimated by Eq. **1**. Larger physical motion would necessarily increase distance acuity, however, and the motion limits used here may not accurately reflect natural behavior, particularly for a walking individual.

The current results are in line with previous work showing that dynamic binaural processing works best under the assumption that sound sources are fixed in world coordinates and dynamic binaural changes are assumed to be generated by self-motion: Brimijoin and Akeroyd (33) measured minimum moving audible angles (MMAAs), i.e., the minimum perceivable angle between two (speech) sound sources when both sounds rotated relative to the subject's head. The authors showed that the MMAA was significantly smaller when the subject's head rotated but the sound sources were kept fixed in world coordinates than when the head was kept fixed and the sound sources were rotated around the subject. As in the current study, the authors took care

that dynamic binaural cues were the same in the two experimental conditions.

Given the accumulating evidence suggesting that binaural processing, and even auditory distance computation, is facilitated by self-motion, it is important to consider how this facilitation takes place: we assume that vestibular and proprioceptive cues (and of course visual cues, if available) allow the generation of a prediction about the velocity and position of auditory targets. This prediction acts as additional information, which according to standard cue-integration models (34), leads to a reduced variance in the combined estimate. This argumentation is illustrated in the *Lower* panels of Fig. 2, referenced to the experimental conditions illustrated in the respective *Upper* panels: In the SSM condition (Fig. 2 *D* and *H*), the lack of nonauditory cues results in an imprecise representation of the azimuth of the two sound sources. The distributions overlap significantly, and depth discrimination based on these representations will be poor. In the PM condition (Fig. 2 *C* and *G*), vestibular information is integrated with the auditory information, leading to a decrease in the variance of the representations. In the AM condition (Fig. 2 *B* and *F*), proprioceptive information is also integrated, resulting in a further decrease of the variance. This decrease in variance allows for more reliable discrimination and consequently better thresholds (Fig. 2 *A* and *E*). Overall we argue that auditory motion parallax is a classical illustration of how a combination of cues from different modalities supports object-discrimination performance.

It could be expected then, that passive self-motion leads to less facilitation, and exclusive sound-source motion removes all nonauditory cues. It was suggested that the ratio of motion to visual pursuit encodes depth information from motion parallax better than motion or pursuit alone (35). In the current auditory study, distance discrimination also improved when the subjects were actively moving, and this improvement might be a result of a similar ratio of self-motion to binaural auditory pursuit. The fact that our subjects performed significantly better when they moved actively than when they were moved or when the sound sources moved supports this hypothesis because the motion is less well defined when it lacks the proprioceptive component (passive vs. active motion) and explicit motion information is missing completely when only the sources move (SSM). In the latter condition, subjects are likely to fall back on the use of pursuit information alone and consequently perform still worse. Overall the good correspondence between the current results and those on visual motion parallax support the hypothesis that the current experiments may tap into a dedicated multimodal motion parallax circuit.

Regardless of the exact nature of the underlying circuit, we conclude that distance discrimination in the current study was based solely on parallax cues. While recent studies indicate that the classical binaural cues (interaural time and/or level differences) also depend on distance, at least when the sound source is in the near field, i.e., quite close to the subject (36, 37), these effects cannot account for the present results. Even though the positions of the virtual sound sources were quite close to the subjects (between 30 and 98 cm), near-field effects can be excluded because the loudspeakers used to present the virtual sound sources were very small (membrane diameter of <2.5 cm) and frequencies relatively high (≥800 Hz). With these parameters, the near field extends to no more than 6 cm in front of the array, even when two adjacent speakers are active at a time (38).

Where would such a "sensitivity" to acoustic distance cues be computed in the brain? A possible candidate for neuronal representation of auditory distance might be the auditory "where" pathway. Indeed Kopčo et al. (39) found that the posterior superior temporal gyrus and planum temporal were activated by the above-mentioned auditory distance cues like the direct-to-reverberant ratio (11) and distance-dependent interaural level differences (36). It would be very promising to include active or passive motion into such scanning paradigms and test the extent to which motion enhances neural activity in the spatial–auditory brain areas, however challenging this might be for brain-imaging techniques.

Materials and Methods

The current psychophysical experiments were approved by the Ethics Committee of the Ludwig Maximilians University Munich, project no. 115–10. All subjects signed an informed consent protocol.

Exp. I.

Stimuli. Subjects were required to judge the relative distances of two intermittent sound sources. Each of the sources emitted a train of tone pips with a pip duration of 25 ms and a repetition period of 100 ms. The carrier for the tone pips was a harmonic complex with a fundamental frequency of either 210 Hz (low-pitched source) or 440 Hz (high-pitched source). For reasons detailed in *Supporting Information*, pips were band-pass filtered to cover the same frequency range between 800 and 4,000 Hz, i.e., the fundamental and (at least for f0 = 210 Hz) a few lower harmonics were missing in both sources. The phase of the low-pitched pip train was shifted by 50 ms, relative to the high-pitched train, such that the overall stimulation consisted of a summary pip train with a 50-ms period and periodically alternating pitch. A spectrogram of the summary pip train with alternating pitches is shown in Fig. 1*B*.

In Exp. I, the pips were played back through two miniature speakers (NSW1-205–8A, AuraSound) positioned at different depths in front of the subject in an anechoic chamber. The speakers were mounted on vertical rods that were fitted to mechanical sliders moving in a guide rail (see Fig. 1*A*). This construction allowed the speakers to be precisely positioned in depth while minimizing the mutual acoustic shadowing of the speakers. Relative to the subjects' interaural axis, the source distances were (at increasing level of difficulty) 98/30 cm, 90/31 cm, 82/33 cm, 73/35 cm, 65/38 cm, and 56/40 cm. This resulted in distance differences between the sound sources of 68, 59, 49, 38, 27, and 16 cm. Without the spectral rove (see below), the sound level of the pip trains was 67 dB sound pressure level. The loudspeakers were driven via a stereo amplifier (Pioneer A107) from a PC soundcard. In each trial, the closer loudspeaker pseudorandomly emitted either the low-pitched or the high-pitched pip train, and the more distant loudspeaker emitted the other pip train. Detailed information on our acoustic calibrations is provided in *Supporting Information*.

Procedure. In a one-interval, two-alternative forced choice paradigm with feedback, subjects had to judge whether the high-pitched sound source was closer or farther away from them than the low-pitched source by pressing one of two buttons on a gamepad. The subjects were seated throughout the experiment. Their heads were continuously tracked. Head tracking procedures are detailed in *Supporting Information*. At the beginning of each trial, a 100 ms pure-tone burst at 1 kHz informed the subjects when their head had reached an acceptable position. Then subjects were instructed to either remain in that position during the following 4-s stimulus presentation (for the NM condition) or to make a trained ±23-cm lateral motion (for the AM condition) (*Supporting Information*, *Motion Training and Body Motion Analysis for Exp. I*).

Within each block of 20 trials (10 NM trials and 10 AM trials), loudspeaker positions were fixed. Data were acquired in at least four sessions of six blocks each. Trials were included or excluded based on the respective head tracks (see below), and data acquisition was continued until at least 30 acceptable trials were available for each experimental condition and pair of loudspeaker depths. Reported distance-difference thresholds correspond to the 75% correct value extracted from a cumulative Gaussian distribution fitted to the data.

Subjects. Seven female subjects (ages ranging from 22 to 38 y) participated in the experiment. None of the subjects reported auditory, vestibular, or sensory–motor impairments.

Exp. II. Stimuli, procedure, and data analysis for Exp. II were the same as for Exp. I with the following exceptions:

Stimuli. In contrast to Exp. I, the sound sources were presented with a loudspeaker array which allowed positioning the sound sources in virtual space behind the array. The array consisted of 24 miniature broadband speakers (NSW1-205–8A, AuraSound) spaced at a distance of 4 cm. Each speaker was individually equalized with a 64-point finite impulse response filter to provide a flat magnitude and phase response between 200 Hz and 10 kHz.

The sound presentation was controlled via SoundMexPro (HörTech GmbH) allowing for dynamically adjusting the loudness of each speaker during playback. Sounds were sent out by a multichannel audio interface (MOTU 424 with two HD192 converters, MOTU, Inc.) and amplified with four multichannel amplifiers (AVR 445, Harman Kardon).

Procedure. The subject's head and the motion platform on which the subject was seated (see below) were continuously tracked with a 6-degree-of-freedom tracking system (Optitrack Flex 13, three cameras; NaturalPoint) sampling at 120 frames per second. The readings from the tracking system were used during stimulation to map the virtual sound sources to the speaker array by means of an amplitude panning procedure (for details see *Supporting Information*).

Exp. II was conducted on a 6-degree-of-freedom motion platform (Moog 6DOF2000E). Blindfolded subjects were seated in a padded seat mounted on the platform. All experiments were performed in a darkened room. The PC also controlled the platform. The tracking system sent its acquired data to a second PC. Both computers were connected via Ethernet.

Subjects initiated each trial by positioning their heads facing the middle of the speaker array (between loudspeaker 12 and 13) at a distance to the array of 20 cm. The press of a gamepad button started the presentation of the two sound sources via the speaker array. While the sound sources were on, the subjects or the sound sources moved, depending on the instructed motion condition, which included the two conditions studied in Exp. I (NM and AM) plus two additional conditions. In the PM condition (Fig. 2*C*), subjects did not move their upper body, but the platform moved the subjects such that the subjects' head motion relative to the virtual sound sources was very similar to the trained motion in the AM condition. In the SSM condition (Fig. 2*D*), subjects remained positioned with their heads directed toward the middle speakers but the sound sources presented via the array moved such that the relative motion between the sound sources and the subjects' heads in azimuth was the same as in the AM and PM conditions.

In contrast to Exp. I, the subjects received auditory feedback after every trial, indicating whether their decision was correct or not.

For each of the three motion conditions AM, PM, and SSM, 210 trials were collected per subject, 30 repetitions for each of the seven source-distance differences. A total of 90 additional trials were collected for the NM condition, but only at the largest distance difference of 68 cm. The overall 720 trials were divided into six blocks of 120 trials each. Trials for all conditions and sound-source distances were presented in a predefined randomly interleaved sequence in a given experimental block. Subjects were instructed about what kind of motion was required for the next trial. See *Supporting Information*, *Motion Training and Body Motion Analysis for Exp. II* for details.

Subjects. Twelve subjects, four males and eight females (ages ranging from 21 to 37 y), participated in the experiment. Two of the subjects also took part in Exp. I. None of the subjects reported auditory, vestibular, or sensory–motor impairments.

ACKNOWLEDGMENTS. We thank Benedikt Grothe and the Bernstein Center in Munich for providing excellent research infrastructure. We thank Isabelle Ripp for her assistance in data acquisition. This work was supported by a grant (01GQ1004A) of the Bernstein Center Munich, BMBF (Federal Ministry of Education and Research, Germany), and the Munich Center for Neuroscience.

1. Westheimer G (2005) The resolving power of the eye. *Vision Res* 45:945–947.
2. Rogers B, Graham M (1979) Motion parallax as an independent cue for depth perception. *Perception* 8:125–134.
3. Helmholtz H (1925) *Helmholtz's Treatise on Physiological Optics* (Optical Society of America, New York).
4. Qian N (1997) Binocular disparity and the perception of depth. *Neuron* 18:359–368.
5. Howard IP, Rogers BJ (2002) *Seeing in Depth*, Depth Perception (Porteus, Toronto), Vol 2.
6. Howard IP, Rogers BJ (1995) *Binocular Vision and Stereopsis* (Oxford Univ Press, New York).
7. Dokka K, MacNeilage PR, DeAngelis GC, Angelaki DE (2011) Estimating distance during self-motion: A role for visual-vestibular interactions. *J Vis* 11:2.
8. Rayleigh JWS (1879) XXXI. Investigations in optics, with special reference to the spectroscope. *London Edinburgh Dublin Philos Mag J Sci* 8:261–274.
9. Blauert J (1997) *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA).
10. Coleman PD (1962) Failure to localize the source distance of an unfamiliar sound. *J Acoust Soc Am* 34:345–346.
11. Bronkhorst AW, Houtgast T (1999) Auditory distance perception in rooms. *Nature* 397:517–520.
12. Zahorik P (2002) Direct-to-reverberant energy ratio sensitivity. *J Acoust Soc Am* 112: 2110–2117.
13. Kolarik AJ, Moore BC, Zahorik P, Cirstea S, Pardhan S (2016) Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. *Atten Percept Psychophys* 78:373–395.
14. Bekesy GV (1938) Über die Entstehung der Entfernungsempfindung beim Hören. *Akust Z* 3:21–31.
15. Kneip L, Baumann C (2008) Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis. *J Acoust Soc Am* 124:3108–3119.
16. Simpson WE, Stanton LD (1973) Head movement does not facilitate perception of the distance of a source of sound. *Am J Psychol* 86:151–159.
17. Speigle JM, Loomis JM (1993) Auditory distance perception by translating observers. *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium* (San Jose, CA), pp 92–99.
18. Loomis JM, Klatzky RL, Philbeck JW, Golledge RG (1998) Assessing auditory distance perception using perceptually directed action. *Percept Psychophys* 60:966–980.
19. Teramoto W, Sakamoto S, Furune F, Gyoba J, Suzuki Y (2012) Compression of auditory space during forward self-motion. *PLoS One* 7:e39402.
20. Teramoto W, Cui Z, Sakamoto S, Gyoba J (2014) Distortion of auditory space during visually induced self-motion in depth. *Front Psychol* 5:848.
21. Wexler M, van Boxtel JJ (2005) Depth perception by the active observer. *Trends Cogn Sci* 9:431–438.
22. Ellard CG, Goodale MA, Timney B (1984) Distance estimation in the Mongolian gerbil: The role of dynamic depth cues. *Behav Brain Res* 14:29–39.
23. Panerai F, Cornilleau-Pérès V, Droulez J (2002) Contribution of extraretinal signals to the scaling of object distance during self-motion. *Percept Psychophys* 64:717–731.
24. Saberi K, Perrott DR (1990) Minimum audible movement angles as a function of sound source trajectory. *J Acoust Soc Am* 88:2639–2644.
25. Strybel TZ, Manligas CL, Perrott DR (1992) Minimum audible movement angle as a function of the azimuth and elevation of the source. *Hum Factors* 34:267–275.
26. Lappin JS, Tadin D, Nyquist JB, Corn AL (2009) Spatial and temporal limits of motion perception across variations in speed, eccentricity, and low vision. *J Vis* 9:1–14.
27. McKee SP, Taylor DG (2010) The precision of binocular and monocular depth judgments in natural settings. *J Vis* 10:5.
28. Naji JJ, Freeman TC (2004) Perceiving depth order during pursuit eye movement. *Vision Res* 44:3025–3034.
29. Furman M, Gur M (2012) And yet it moves: Perceptual illusions and neural mechanisms of pursuit compensation during smooth pursuit eye movements. *Neurosci Biobehav Rev* 36:143–151.
30. Freeman TC, Champion RA, Warren PA (2010) A Bayesian model of perceived head-centered velocity during smooth pursuit eye movement. *Curr Biol* 20:757–762.
31. Freeman TC, Culling JF, Akeroyd MA, Brimijoin WO (2017) Auditory compensation for head rotation is incomplete. *J Exp Psychol Hum Percept Perform* 43:371–380.
32. Moscatelli A, Hayward V, Wexler M, Ernst MO (2015) Illusory tactile motion perception: An analog of the visual Filehne illusion. *Sci Rep* 5:14584.
33. Brimijoin WO, Akeroyd MA (2014) The moving minimum audible angle is smaller during self motion than during source motion. *Front Neurosci* 8:273.
34. Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–433.
35. Nawrot M, Stroyan K (2009) The motion/pursuit law for visual depth perception from motion parallax. *Vision Res* 49:1969–1978.
36. Kuwada CA, Bishop B, Kuwada S, Kim DO (2010) Acoustic recordings in human ear canals to sounds at different locations. *Otolaryngol Head Neck Surg* 142:615–617.
37. Kim DO, Bishop B, Kuwada S (2010) Acoustic cues for sound source distance and azimuth in rabbits, a racquetball and a rigid spherical model. *J Assoc Res Otolaryngol* 11:541–557.
38. Weinzierl S (2008) *Handbuch der Audiotechnik* (Springer, Berlin).
39. Kopčo N, et al. (2012) Neuronal representations of distance in human auditory cortex. *Proc Natl Acad Sci USA* 109:11019–11024.

Supporting Information

Genzel et al. 10.1073/pnas.1712058115

A direct comparison of motion profiles for the AM (green) and PM (blue) conditions is shown in Fig. S1. This subject performed the active motion quite precisely and the motion platform could reproduce the motion reliably. This pertains both to the (dominating) azimuthal motion (Fig. S1*A*) and the associated small vertical motion (Fig. S1*B*).

SI Materials and Methods

Reasons for Band-Pass Filtering of the Stimuli in Exp. I. The reason for restricting the frequency range to 800–4,000 Hz is that at the low end (below 800 Hz), we have to avoid near-field effects, some of them being dependent on absolute frequency, not only on the relationship between frequency and sound-source diameter. The reason for the low-pass cutoff at 4 kHz is that for higher frequencies we would have stronger occlusion effects of the farther speaker by the closer speaker (see below). Nevertheless we needed enough bandwidth to accommodate several harmonics to provide a reasonably strong pitch. Due to the dominance region of pitch (1), it is likely that the high-pitched pip train had higher pitch strength than the low-pitched train.

Motion Training and Body Motion Analysis for Exp. I. At the beginning of the first experimental session, each subject was taught to move in a stereotypical way for the AM condition. They had to start moving their upper body such that their head was displaced 23 cm to the left within the first second of stimulus playback, then 46 cm to the right (for a target displacement of 23 cm to the right from the head origin) within the following 2 s, and back to the starting point within the final second, in an overall smooth motion akin to a single period of a sine wave (Fig. S1). This motion profile was trained before every session with feedback from the experimenter, who instructed the subjects to move, started the playback of a 4-s click train similar to the stimulus used in the experiment, and immediately analyzed the head tracking data. This procedure was repeated until the subject was confident that they had memorized the motion profile and the experimenter observed several subsequent trials with acceptable head tracks (correct velocity of motion, displacement to the left and right between 20 and 26 cm, and stable position along the other two spatial axes).

After the session, the head tracks for each trial were analyzed as to whether they met the inclusion criteria: NM trials were excluded when any tracking point acquired during stimulus presentation deviated from the head origin by more than 2 cm along the interaural axis, or when the mean absolute deviation from that point exceeded 1 cm for the trial as a whole. AM trials were excluded when the maximum displacement to the left or to the right along the interaural axis differed by more than 4 cm from the mean displacement for the subject.

Calibrations for Exp. I. To calibrate setup with real sound sources, a measurement microphone (1/2″; BSWA Technology) was positioned at the point corresponding to the middle of the interaural axis of a seated subject. For each pair of loudspeaker depths, acoustic impulse responses of the speakers were measured and compensation impulse responses calculated by pointwise division of the complex discrete Fourier transform (DFT) of an ideal band-pass IR between 200 and 8,000 Hz by the complex DFT of the measured IR. All sounds presented through the two speakers were convolved with the corresponding compensation impulse responses. This procedure equalized loudness, spectrum, and latency differences between the two speakers.

To remove possible residual spectral or loudness cues that may contribute to distance discrimination, we implemented a roving spectral envelope. Specifically, we defined a random spectral envelope by varying loudness across a ±6-dB range in third-octave steps throughout the whole calibrated pass band of the speakers (200–8,000 Hz). Thus, the timbre of the harmonic complexes changed from trial to trial which renders the use of timbre or near-field cues very difficult. The validity of this precise equalization procedure plus the application of the spectral-envelope rove was psychophysically confirmed by the fact that our subjects performed poorly when they were not allowed to move during stimulus presentation.

To check for residual spectral effects that may arise through interaction of the sound sources with the subject's head or torso, we made control measurements replacing the subject with a head and torso simulator (B&K 4128C).

Head Tracking in Exp. I. Tracking was implemented with a camera on the subject's head scanning a target made up of fiducial markers mounted at the ceiling above the subject (2). Stimulus presentation for each trial was started only when the head position did not vary by more than 1 cm along the interaural axis, 1.5 cm along the anterior–posterior axis, or 1.5 cm along the cranial–caudal axis from the required head origin. The head origin was defined at the beginning of each experimental session as that head position where the distance from the interaural axis to the membrane of the front loudspeaker at its closest position was exactly 30 cm, and the head was exactly on axis with the two loudspeakers.

Rendering of Virtual Sound Sources on the Loudspeaker Array in Exp. II (Amplitude Panning Procedure). For each of the two virtual sound sources, the horizontal axis of the speaker array was intersected with the line between the sound source and the most recently acquired head position of the subject. The two speakers closest to this intersection point were activated to simultaneously reproduce the respective sound source. The ratio of their gains was chosen according to the distance of the center points of those loudspeakers to the intersection point: When one activated speaker was at a distance a and the other at a distance b from the intersection point, their gains were set according to the ratio b:a. The combined gain of the two speakers was set to account for geometric attenuation due to the distance between the subject's head position and the virtual sound source. Loudspeaker activations and gain settings were updated at a rate of 100 Hz throughout stimulus playback.

Motion Training and Body Motion Analysis for Exp. II. To ensure comparability of the results between the three conditions that involved motion of the sound sources relative to the subjects' heads (AM, PM, and SSM), both motion training and inclusion criteria for motion trials were more rigorous than in Exp. I. All subjects underwent precise training concerning the active body motion that they had to perform in the AM condition. Small markers on the speaker array indicated the leftmost, rightmost, and middle positions the subjects had to meet in this sequence during their motion. The experimenter informed the subject during training if the motion matched the targeted motion profile.

During the main data acquisition, trials were excluded when they did not meet a nested set of criteria that quantified deviations of the executed motion in that trial from the targeted motion profile. These trials were repeated again at a later time until at least 30 trials per condition were obtained.

When subjects had learned to reliably reproduce body motion with the required displacement and velocity, a further training procedure was initiated. Here subjects were moved by the platform or conducted their learned body motion, but additionally the sound sources were presented with the largest source-distance difference and the subjects had to decide whether the high-pitched source was closer or farther away than the low-pitched source. One training block consisted of 120 trials. This training was necessary because with virtual sound sources and the many different interleaved conditions, it was somewhat harder for the subjects to exploit auditory motion parallax. The main experiment could begin only after a subject's performance in a training block was at least 80% correct.

1. Ritsma RJ (1967) Frequencies dominant in the perception of the pitch of complex sounds. *J Acoust Soc Am* 42:191–198.
2. Garrido-Jurado S, Muñoz-Salinas R, Madrid-Cuevas FJ, Marín-Jiménez MJ (2014) Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognit* 47:2280–2292.

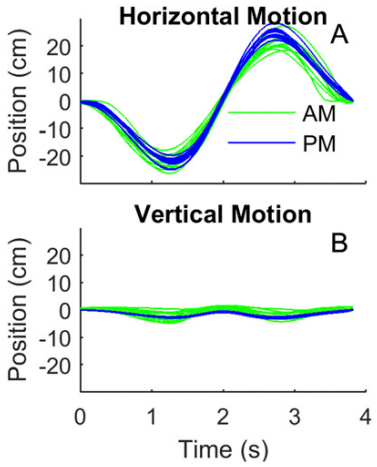

Fig. S1. Tracks of horizontal (*A*) and vertical (*B*) head motion of the subject relative to the stationary sound sources when the subject moved either actively (green) or the subject was moved by the motion platform (blue). Data show that subjects were successfully trained to move quite stereotypically and that the platform captured this stereotypical motion quite well.

CHAPTER 4

Stationary listeners & moving sources

This chapter is a preprint of a peer-reviewed publication which has appeared online (before distribution in a printed issue of the journal) in *Current Biology* on 19 August 2021. I share its first authorship with Michael Forsthofer. The full text of the published version, which has undergone revision since the version printed here, can be accessed through the DOI 10.1016/j.cub.2021.07.018, or on the publisher's website under `https://www.cell.com/current-biology/fulltext/S0960-9822(21)00973-8`. The full citation is

> Michael Forsthofer, Michael Schutte, Harald Luksch, Tobias Kohl, Lutz Wiegrebe, and Boris P. Chagnaud (2021). "Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans". In: *Current Biology*. Published online ahead of print: `https://doi.org/10.1016/j.cub.2021.07.018`.

4.0 Author contributions

B.P.C. planned and designed the study. M.F. and T.K. performed and M.F. analyzed the behavioral rattlesnake experiments. M.S. and L.W. designed, M.S. performed and M.S. and L.W. analyzed the psychophysical experiments. B.P.C., M.F. and M.S. wrote the paper. M.F., M.S., H.L., T.K. and B.P.C. edited and proofread the paper.

Title: Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans

Authors: Michael Forsthofer[(1),#], Michael Schutte[(1),(2),#], Harald Luksch[(3)], Tobias Kohl[(3),§], Lutz Wiegrebe[(1),§] and Boris P. Chagnaud[(1),(4),§]

[#, §] These authors contributed equally.

Affiliations:
(1) Department Biology II, Ludwig-Maximilians-University Munich, Großhaderner Str. 2, 82152 Planegg, Germany
(2) Graduate School of Systemic Neurosciences, Ludwig-Maximilians-University Munich, Großhaderner Str. 2, 82152 Planegg, Germany
(3) Chair of Zoology, School of Life Sciences, Technical University of Munich, Liesel-Beckmann-Str. 4, 85354 Freising, Germany
(4) Institute for Biology, Universitätsplatz 2, Karl-Franzens-University Graz, 8010 Graz, Austria

Correspondence to: Boris P. Chagnaud, Institute for Biology, Universitätsplatz 2, Karl-Franzens-University Graz, 8010 Graz, Austria.
Email: boris.chagnaud@uni-graz.at

Abstract: The estimation of one's distance to a potential threat is essential for any animal's survival. Rattlesnakes inform about their presence by generating acoustic broadband rattling sounds. Combining visual looming stimuli with acoustic measurements, we show that rattlesnakes increase their rattling rate (up to 40 Hz) with decreasing distance of a potential threat, reminiscent of the acoustic signals of sensors while parking a car. Rattlesnakes then abruptly switch to a higher and less variable rate of 60–100 Hz. In a virtual reality experiment, we show that this behavior systematically affects distance judgments by humans: the abrupt switch in rattling rate generates a sudden, strong percept of decreased distance which, together with the low frequency rattling, acts as a remarkable interspecies communication deceptive signal.

One-sentence summary: Adaptive rattling fools distance perception

Competing interests: The authors declare no competing interests.

Data and materials availability: All original files and analysis scripts are available at the University of Graz file deposit system.

4.1 Main text

Our ability to convey information to our personal and work environment enables us to interact in the society. Communication is, however, not restricted to signals within a species but also readily occurs across species. One of the most striking examples of interspecies communication is the acoustic display of rattlesnakes (figures 4.1A, 4.S1A). These snakes generate acoustic signals by clashing a series of keratinous segments onto each other, which are located at the tip of their tails (Fenton and Licht, 1990; Martin and Bagby, 1972). Each tail shake results in a broadband sound pulse that merges into a continuous acoustic signal with fast-repeating tail shakes (figure 4.S1B). This acoustic display is readily recognized by other animals (do Valle and Leão-Vaz, 2005) and serves as an aposematic threat/warning display, likely to avoid being preyed upon or accidentally stepped on (Fenton and Licht, 1990; Reiserer and Schuett, 2016). The probability of a snake to rattle and the acoustic properties of the rattling display depend on various factors such as body temperature, pregnancy, size of the snake and on the

amount of rattle segments (Chadwick and Rahn, 1954; Glaudas *et al.*, 2005; Kissner *et al.*, 1997; Martin and Bagby, 1972; Prior and Weatherhead, 1994; Shine *et al.*, 2002; B. A. Young and I. P. Brown, 1993). It is, however, unknown if snakes actively vary their rattling behavior. Adaptive rattling, for instance, could be used to inform a potential threat about its relative distance to the snake, similar to how distance information from proximity sensors in the rear bumper of a car is encoded in the repetition rate of an acoustic signal.

Here, we test the hypothesis that the western diamondback rattlesnake (Crotalus atrox, Baird and Girard, 1853) can actively vary its rattling behavior in response to distance changes of a potential threat. In a first experiment, we moved a human-like torso towards a stationary snake (figure 4.S1A). Snakes readily initiated their acoustic display, starting with sparse tail shakes that elicited distinct sound pulses (figure 4.S1B). With decreasing torso-snake distance, the frequency of individual sound pulses increased up to a frequency of about 40 Hz, which was followed by a sudden, sharp increase to a higher frequency range (60–100 Hz; figure 4.S1A, red arrow). To avoid acoustic noise generated by the torso motion (motion artifact) and to gain better experimental control, a second experiment was devised: an approaching object was simulated using a visual looming stimulus consisting of a black disk that increased in size with a constant velocity profile (tested at four different velocities) by setting the diameter of the disk proportional to $\frac{1}{x}$ for a decreasing virtual object distance of x (figures 4.1A, 4.S2). Rattlesnakes readily responded to this looming stimulus with the acoustic display described above. While individual snakes showed similar response patterns to multiple stimulus pre-

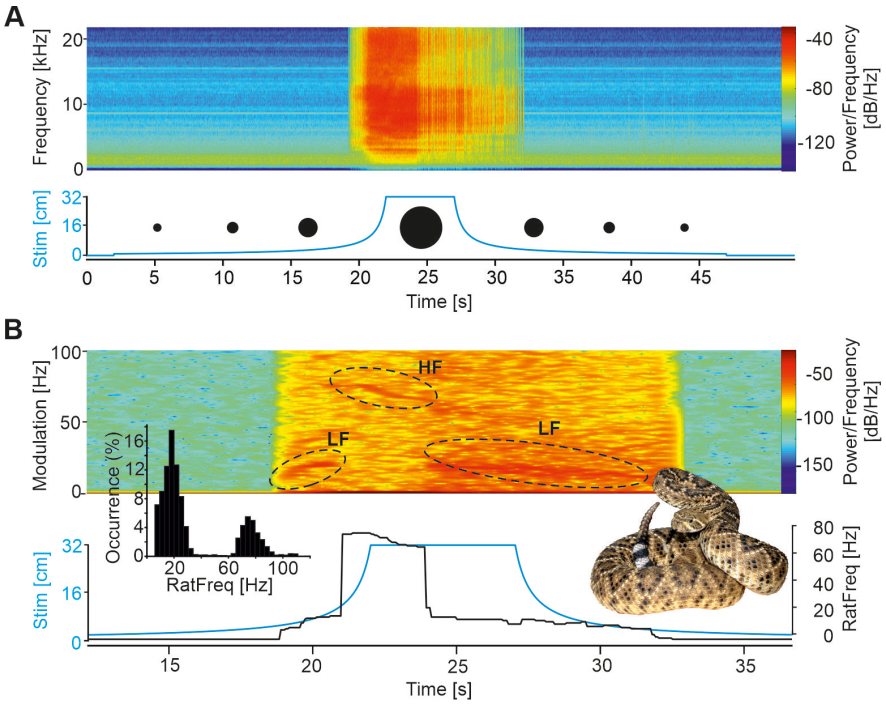

Figure 4.1 Acoustic properties of rattlesnake rattling. **(A)** Spectrogram of a recording of a rattling event (top) triggered by a looming stimulus (blue trace and black circles, bottom) using a constant approach velocity, resulting in a $\frac{1}{\text{distance}}$ increase in stimulus diameter. Black circles are for illustration purposes and not to scale. **(B)** Modulation spectrogram (note different time axis) of rattling depicted in (A) and relationship between rattling modulation frequency (black line, right axis) and looming profile (blue line, left axis). Lower right inset: Snake in striking pose with a raised rattle during rattling. Lower left inset: Histogram of the distribution of rattling modulation frequency (RatFreq) observed in all constant approach velocity experiments (n=197 trials; $N = 25$ snakes). LF: low frequency; HF: high frequency.

sentations, there was considerable variability in responses across snakes in terms of rattling duration and onset (figure 4.S3), a well-known feature in rattlesnakes (Place and Abramson, 2008). Across trials ($n = 197$; $N = 25$ snakes), rattling rate showed a bimodal distribution (figure 4.1B inset), consisting of a low (LF; < 40 Hz) and a high frequency (HF; 60 to 100 Hz) range. Interestingly, rattling rate in the LF mode linearly increased with the increasing visual stimulus, thus carrying information about the relative change of distance between an approaching animal and the snake. The slope of this LF mode change depended on stimulus velocity, with slower stimuli resulting in slower rate increases (Kruskal-Wallis, $p = 0.036, \chi^2 = 8.52$; figure 4.2A; table 4.S1). The duration of rattling in the LF mode also depended on stimulus velocity, with faster stimuli resulting in shorter LF displays before switching to the HF (Kruskal-Wallis, $p = 3.67 \cdot 10^{-5}, \chi^2 = 23.2$; figure 4.2B). In contrast, the HF component of the rattling was independent of the stimulus velocity in terms of both changes in rattling rate (Kruskal-Wallis, $p = 0.98, \chi^2 = 0.2$; figure 4.2C) and duration (Kruskal-Wallis, $p = 0.125$, $\chi^2 = 5.75$; figure 4.2D). While not significant, a trend between the stimulus velocity and the HF duration component is apparent. The HF acoustic display generally continued at a stable rate or slowly decreased over time even when stimulus size was constant (average rate of −1.95 Hz/s at a medium stimulus velocity of 1.1 m/s; table 4.S1). Response latency also decreased on average with increasing

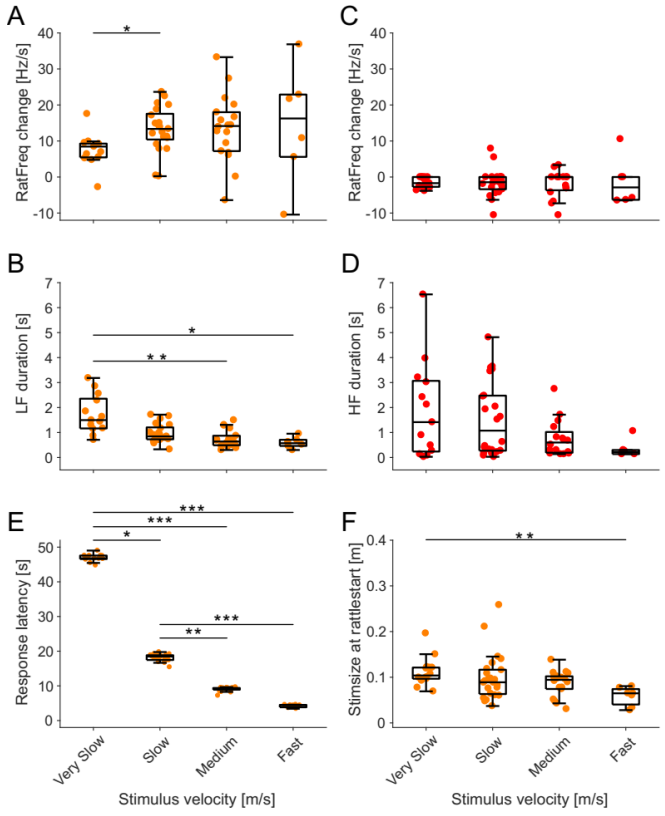

Figure 4.2 Effect of constant approach velocity on low (LF, orange) and high (HF, red) modulation frequency modes of rattling responses. Box and whisker plots of the LF rattling modulation frequency (RatFreq) changes **(A)** and LF mode duration **(B)** for different approach velocities until the shift to the HF mode. Both factors depend on stimulus velocity. Rattling modulation frequency rate changes **(C)** and HF duration **(D)** are less variable and independent of approach velocity for the HF rattling mode. Response latency **(E)** depends on stimulus velocity, while the size of the looming stimulus at rattle onset is independent of stimulus velocity **(F)**. Significance levels (Tukey-Kramer post-hoc test) are indicated by asterisks: $* p \leq 0.05$, $** p \leq 0.01$ and $*** p \leq 0.001$.

approach velocity ($p = 3.67 \cdot 10^{-5}, \chi^2 = 23.2$; figure 4.2E). After the approach phase of the stimulus (*i.e.* when the black disk grew in size), a stationary phase of maximum stimulus size followed, until the black disk decreased in size, mimicking a stopping and a retreating motion of the object, respectively (figures 4.1, 4.S2). In response, snakes generally left the HF rattling mode and changed back to the LF mode until the subsequent end of rattling. In contrast to the rising phase of the stimulus, the rattlesnakes' responses to the stationary and retreating component were highly variable, with some snakes sustaining their HF rattling even until the stimulus had fully disappeared. Besides for the fastest approach velocity (Kruskal-Wallis, $p = 0.036, \chi^2 = 13.52$; figure 4.2F) rattling was initiated independently of the stimulus size, indicating that snakes must have been able to interpret the different approach velocities.

To test whether the rattling rate depended not only on the looming stimulus velocity, but also on the approaching profile, we altered the looming stimulus from a constant velocity to a decreasing velocity profile (figure 4.S2B, C). Snakes ($n = 81$ trials, $N = 13$ snakes) responded to this altered visual stimulus primarily by a decrease in LF slopes which, when compared to the constant approach velocity, was independent of approach duration (Kruskal-Wallis, $p = 0.813, \chi^2 = 0.41$; table 4.S2; figure 4.S4). The HF mode neither correlated in terms of the slope nor the duration to stimulus velocity (Kruskal-Wallis, slope: $p = 0.171, \chi^2 = 3.53$, duration: $p = 0.93, \chi^2 = 0.13$). As the final stimulus size was identical between the constant and the declining velocity profiles, these results demonstrate that rattlesnakes adapt their rattling rate in response to the approach velocity of an object rather than its size.

Why might snakes have evolved to modulate their rattling rate in this way, and why do they switch to the HF mode instead of linearly increasing their rattling rate up until time of contact (which would more honestly advertise the relative distance between the approaching object and themselves)? We hypothesize that the sudden switch to HF mode could serve to create the perception in an approaching animal that contact with the snake is imminent, according to the previously established "rule" of the distance-dependent increase in rattle rate in the LF mode (figure 4.3A, B). To test this hypothesis, we designed an audio-visual virtual environment in which naive human subjects ($N = 11$) were positioned on a chair and were virtually moved through a grass land VR environment while approaching an invisible sound source (the "virtual snake"). This virtual snake emitted broadband sound pulses at either a constant (12 Hz) or at an adaptive rattling rate that depended on the listener–snake distance. These sounds were played back to the listener via a vertical loudspeaker array, with amplitude gains set dynamically to reflect geometric attenuation (closer snakes heard more loudly) and elevation (closer snakes heard from further below). Each trial randomly started at one of six distances (figure 4.3C) and the listeners were asked to stop the automatic approach towards the sound source when they estimated the source to be 1 m away. In the adaptive rattling condition, the virtual snake was programmed to increase its rattling rate from 5 to 20 Hz for a distance decrease from 8 m to 4 m. When the distance undercut 4 m, the virtual snake switched to a HF, distance-independent rattling of 70 Hz (figure 4.3C). Thus, in trials where the starting distance was smaller or equal to 4 m ("short trials"), we compare stopping distances for a time-invariant low rattling frequency (12 Hz) and an equally time-invariant high rattling frequency of 70 Hz. We found a significant difference in the listeners' stopping distances between these two conditions, with LF rattling causing shorter stopping distances than HF rattling (figure 4.3D; repeated-measures ANOVA: $F_{1,10} = 15.47, p = 0.002$). This indicates that the difference between rattling at 12 or 70 Hz by itself leads humans to significantly underestimate their distance to the virtual snake at the higher rattling rates, presumably due to an increase in perceived loudness (S. S. Stevens and Guirao, 1962). In trials with a starting distance of more than 4 m ("long trials"), the stopping distances with adaptive rattling exhibited a clear bimodal distribution when compared to those with constant rattling, with a secondary mode at around 4 m, *i.e.* the time of the sudden rattling rate change (figure 4.3E orange arrow; medians 1.08 m *vs.* 1.21 m;

$F_{1,10} = 8.84$, $p = 0.01$). This suggests that the sudden change from the LF to the HF range indeed acted (whether intended or not) as a deceptive signal about the snake's proximity.

Our data show that the acoustic display of rattlesnakes, which has been interpreted for decades as a simple acoustic warning signal informing about the presence of the snake, is in fact a far more intricate interspecies communication signal. While the LF rattling mode informs the approaching subject in a predictive fashion about its approach towards the snake, the sudden switch to the HF

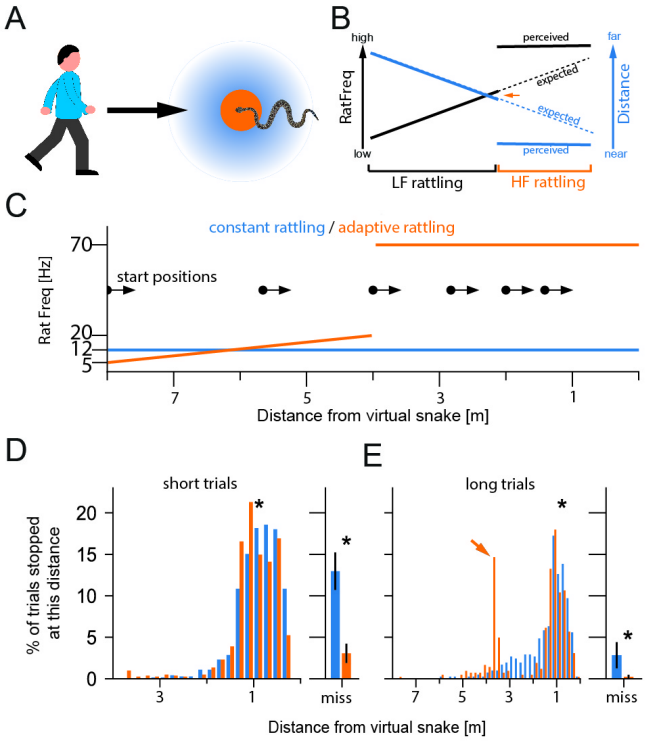

Figure 4.3 Psychophysical experiments in a virtual reality environment reveal that adaptive rattling generates an underestimation of distance in human subjects. **(A)** Schematic drawing indicating the different modes in rattling depending on distance to an approaching object: blue area—distance dependent LF rattling; orange area—HF rattling. **(B)** Schematic drawing of how a listener might predict the time course of the rattling frequency, compared to the rattling frequency they will actually experience, and how they would correspondingly predict their distance to the snake. Black lines indicates rattling modulation frequency and blue lines indicate the perceived distance (expectation in dashed lines respectively). **(C)** Acoustic stimulation paradigm in the virtual environment: Momentary rattling frequency computed from one of two modulation functions (constant *vs.* adaptive) based on the distance between the virtual snake and the position of the listener in the virtual environment. The blue line shows a distance-independent 12 Hz sound in the constant rattling frequency condition; the orange lines represent the adaptive rattling frequency condition with its gradual increase in LF mode up to 4 m, and the jump to HF mode at 4 m. **(D, E)** Histograms of virtual listener–snake distances at which the listeners stopped the trial because they perceived the virtual snake to be exactly 1 m away, in non-miss trials where the starting distance was less than or equal to (D), or greater than 4 m (E). Asterisks indicate significant differences in the distributions for the constant and adaptive rattling frequency conditions ($p < 0.05$ in repeated-measures ANOVA with condition as a within-subject factor). To the right of (D, E), frequency of miss trials (trials which the listener did not stop before the virtual snake was at a distance of 0.2 m) in the respective trials. Asterisks indicate significant differences in miss rate between the two conditions ($p < 0.05$ in one-sided Fisher's exact tests).

mode acts as a smart deceptive signal fooling the listener about its actual distance to the sound source. The misinterpretation of distance by the listener thereby creates a distance "safety margin".

A question remaining is to which aspect of the looming stimulus the snakes responded. As approach velocity and increasing visual stimulation (*i.e.* the diameter of the black disk) both depend on each other, we were not able to separate their contribution. Furthermore, while we have so far described the behavior of the rattlesnakes in terms of a relationship between rattling rate and object distance, it is also conceivable that the rattling rate is an intermediary parameter controlled by the snake to change the perceived loudness of the signal by its recipient. Extracting distance information from sound sources is generally a challenging task (Middlebrooks and Green, 1991). The primary distance cues that could be resolved by the listener in the virtual environment are the distance-dependent elevation cues (due to the snake being heard from a lower angle below the horizon as the listener gets closer) and the geometric attenuation of the rattling by the snake (both factors were included in our virtual acoustic environment). The latter cue is modified by the acoustic display of the snakes: Our acoustic analyses in the looming experiments show that the sound level of single rattle events is rather constant (figures 4.1, 4.S1) but the snake adjusts the number of events, *i.e.*, the rattling rate. While this does not change the physical loudness of the emitted sound, it potentially increases perceived loudness, due to the phenomenon of temporal integration. An analysis using a well-established perceptual-loudness model which accounts for sounds changing over time (B. C. Moore, Glasberg, *et al.*, 2016) showed that the difference between our simulated rattling stimuli at 12 Hz *vs.* 70 Hz would lead to the latter stimulus being perceived as twice as loud by humans.

The human auditory system is biased towards perceiving sounds that increase in loudness as moving faster, and getting closer, than sounds that become quieter (Neuhoff, 1998, 2001). The rattling behavior of the snakes could thus be interpreted as exploiting this bias by exaggerating the loudness increase beyond the purely physical intensification of the sound pressure at a listener's ears due to the approach. While the distance to the snake is not encoded in absolute values in the rattling display (different onset points of rattling and onset times of changes between the LF and the HF mode across snakes, see figure 4.S3), the relationship between approach velocity and rattling rate (*e.g.* figures 4.1A, 4.S1), however, suggests that the relative approach velocity/distance is encoded. This is enough to generate this unique auditory deceptive signal combination which, as shown by our psychophysical experiment, acts as a highly effective interspecies communication system.

4.1.1 References

Baird, Spencer Fullerton and Charles Girard (1853). *Catalogue of North American reptiles in the Museum of the Smithsonian Institution*. Smithsonian Institution.

Chadwick, L.E. and Hermann Rahn (1954). "Temperature dependence of rattling frequency in the rattlesnake, *Crotalus v. viridis*". In: *Science* 119.3092, pp. 442–443.

Fenton, M. Brock and Lawrence E. Licht (1990). "Why rattle snake?" In: *Journal of Herpetology*, pp. 274–279.

Glaudas, Xavier, Terence M. Farrell, and Peter G. May (2005). "Defensive behavior of free-ranging pygmy rattlesnakes *(Sistrurus miliarius)*". In: *Copeia* 2005.1, pp. 196–200.

Kissner, Kelley J., Mark R. Forbes, and Diane M. Secoy (1997). "Rattling behavior of prairie rattlesnakes (*Crotalus viridis viridis*, Viperidae) in relation to sex, reproductive status, body size, and body temperature". In: *Ethology* 103.12, pp. 1042–1050.

Martin, James H. and Roland M. Bagby (1972). "Temperature-frequency relationship of the rattlesnake rattle". In: *Copeia*, pp. 482–485.

Middlebrooks, John C. and David M. Green (1991). "Sound localization by human listeners". In: *Annual Review of Psychology* 42.1, pp. 135–159.

Moore, Brian C.J., Brian R. Glasberg, Ajanth Varathanathan, and Josef Schlittenlacher (2016). "A loudness model for time-varying sounds incorporating binaural inhibition". In: *Trends in hearing* 20, p. 2331216516682698.

Neuhoff, John G. (1998). "Perceptual bias for rising tones". In: *Nature* 395.6698, pp. 123–124.

— (2001). "An adaptive bias in the perception of looming auditory motion". In: *Ecological Psychology* 13.2, pp. 87–110.

Place, Aaron J. and Charles I. Abramson (2008). "Habituation of the rattle response in Western Diamondback rattlesnakes, *Crotalus atrox*". In: *Copeia* 2008.4, pp. 835–843.

Prior, Kent A. and Patrick J. Weatherhead (1994). "Response of free-ranging eastern massasauga rattlesnakes to human disturbance". In: *Journal of Herpetology* 28.2, pp. 255–257.

Reiserer, Randall S. and Gordon W. Schuett (2016). "The Origin and Evolution of the Rattlesnake Rattle: Misdirection, Clarification, Theory, and Progress". In: *Rattlesnakes of Arizona*. Ed. by Gordon W. Schuett, Martin J. Feldner, Charles F. Smith, and Randall S. Reiserer. Vol. 2.

Shine, Richard, Li-Xin Sun, Mark Fitzgerald, and Michael Kearney (2002). "Antipredator responses of free-ranging pit vipers (*Gloydius shedaoensis*, Viperidae)". In: *Copeia* 2002.3, pp. 843–850.

Stevens, Stanley Smith and Miguelina Guirao (1962). "Loudness, reciprocality, and partition scales". In: *The Journal of the Acoustical Society of America* 34.9B, pp. 1466–1471.

do Valle, Anderson Luis and Letícia de Almeida Leão-Vaz (2005). "The defensive reaction of rheas (Rhea americana) to a Rattlesnake Signal". In: *Revista de Etologia* 7.1, pp. 49–50.

Young, Bruce A. and Ilonna P. Brown (1993). "On the acoustic profile of the rattlesnake rattle". In: *Amphibia-Reptilia* 14.4, pp. 373–380.

4.1.2 Acknowledgments

The authors thank Maximilian Bothe, Yvonne Schwarz, Gabriele Schwabedissen, Nora Dallmann, Vivien Lücke and Eva Mardus for technical assistance and Thorin Johnson and Heinrich Römer for comments on a previous version of this manuscript. BPC thanks B. Grothe and H. Straka for generously providing experimental facilities and salaries to LW and BPC. Funding: Funding was provided by the Munich Center for Neurosciences to MS, by the School of Life Sciences Weihenstephan to HL, by the DFG to LW (Wi1518/17) and to BPC (CH 857/2-1).

4.2 Supplementary materials and methods

4.2.1 Animals

Experiments were performed on 30 juvenile (age: 1–2 years old) western diamondback rattlesnakes, Crotalus atrox, Baird and Girard, 1853, of either sex (weight: 53–241 g; snout–vent length: 37.7–69.9 cm). Snakes were kept on a 12:12-hour day:night cycle at a temperature of 25–31 °C with water ad libitum and were fed weekly with dead mice. All experimental animals were bred and kept at the Chair of Zoology of the Technical University of Munich, following the established guidelines for care and maintenance of venomous snakes.

During the course of experiments all snakes were kept solitarily. Snakes were not used for experiments the same day they were fed and were given at least one day to rest after each experimental session. For each session, individual snakes were transferred in a lightproof transport box from the animal facilities into the experimental setup (ambient temperature: 27–32 °C). To motivate snakes to remain at a certain position, they were placed on an elevated platform (30 cm × 30 cm, height: 24.5 cm). A clay pot adjusted to the size of the snakes (diameter: 11 cm/27.5 cm) was provided as shelter. After being placed on the platform, snakes were allowed to acclimatize for 5 minutes before the shelter was removed and the experimental session started. Each session consisted of up to 5 trials with inter-trial intervals of 5 minutes.

4.2.2 Real object stimulation

A Brüel & Kjaer Head and Torso Simulator (HATS, type 4128-C) was mounted on a sled with Teflon runners, which was placed on a guide rail system positioned longitudinally to the experimental platform (figure 4.S1). A wire system, attached to the front and back ends of the sled and running along the guide rail, was used to manually move the HATS towards or away from the experimental platform. A distance sensor positioned at the back wall allowed to monitor the moved distance of the HATS.

After 2 s of pre-stimulus time, the HATS was manually pulled from a starting distance of 1.8 m towards the experimental platform until the snakes initiated HF rattling, down to a possible minimum distance of 0.25 m. Velocities were not constant due to the manually controlled movement and ranged between 0.07 and 0.35 m/s. Luminescent tubes were used as light sources.

4.2.3 Visual stimulation

Visual stimuli were back-projected by a projector (Mitsubishi, XD350U, resolution: 1024–768 px, image refresh rate: 60 Hz, C. atrox electroretinography temporal resolution: 36 Hz; Kohl and B. A. Young, 2011) onto a white screen located 35 cm in front of the center of the platform (figure 4.S2). Custom-written scripts (MATLAB, version 7.11.0, Psychtoolbox) were used to generate visual looming stimuli which consisted of black disks with an increasing diameter over time (approach phase) after which the stimulus remained constant for 5 s (stationary phase). This constant phase was followed by a stimulus size decrease in a mirror-image fashion to the previous increase (retreat phase, figure 4.S2B). Two different visual stimulus paradigms were used: one mimicked an object approaching the snake and departing from it at a constant velocity, the other an object in which the approach velocity decreased over time (figure 4.S2C). For constant approaches, the diameter (d) of the black disk changed with approach duration (t) according to the function

$$d(t) = \frac{\Delta D}{x(t)}$$

where ΔD is the distance of the snake to the screen and $x(t)$ the distance to the virtual approaching object. For constant approach speeds the object started at $d = 3.2$ cm (virtual distance: 11 m, visual angle: 1.47°) and increased to $d = 16.25$ cm (virtual distance: 0.32 m, visual angle: 49.13°, figure 4.S2).

Four different approach durations were tested and velocity directly scaled with approach durations of 50 s (0.2 m/s), 20 s (0.5 m/s), 10 s (1.1 m/s) and 5 s (2.1 m/s). Decreasing approach velocity stimuli started at $d = 0$ cm and increased to a maximum $d = 16.25$ cm in the same time frames of 20 s, 10 s and 5 s. Thus, for both paradigms, 50 s, 20 s, 10 s and 5 s approaches simulated a very slow, slow, medium and fast approach, respectively. Each snake was tested for up to four different velocities per experimental session, snakes that repeatedly left the experimental platform ($N = 1$) or did not or only rarely elicit rattling sound to the visual stimuli ($N = 4$) were not used in further experiments. Consequently, sample sizes differ across different stimulus presentations.

To test whether variation in acoustic responses towards different stimuli was caused by individual differences, we presented several snakes with 5 repetitions of one stimulus and compared the evoked responses to individual stimulations of five different animals. Multiple consecutive repetitions on one animal were done according to normal experimental procedure with 5 minutes inter-trial time (similar to Place and Abramson, 2008) and one resting day between experiments.

4.2.4 Recording of rattling sounds

Rattling sounds were recorded with an electrostatic microphone (frequency range: 20 Hz–31.5 kHz; M215, MicW, Beijing, China) placed 11 cm above the plane of the experimental platform and at a lateral distance of 18 cm (figure 4.S2A), digitized at a rate of 44.1 kHz with an external soundcard (Profire 610, M-Audio, Cumberland, RI, USA) connected to a personal computer. Recordings were saved in the MATLAB MAT file format.

To assure that the snakes' visual field encompassed the screen in which the stimuli were presented, a video camera (Guppy, Allied Vision Technologies, frame rate 10 Hz) was placed above (distance 86 cm) the platform to monitor snake head orientation. Two infrared spotlights (Abus, TV6700, $\lambda = 850$ nm) suspended from the top were used to constantly illuminate the experimental platform.

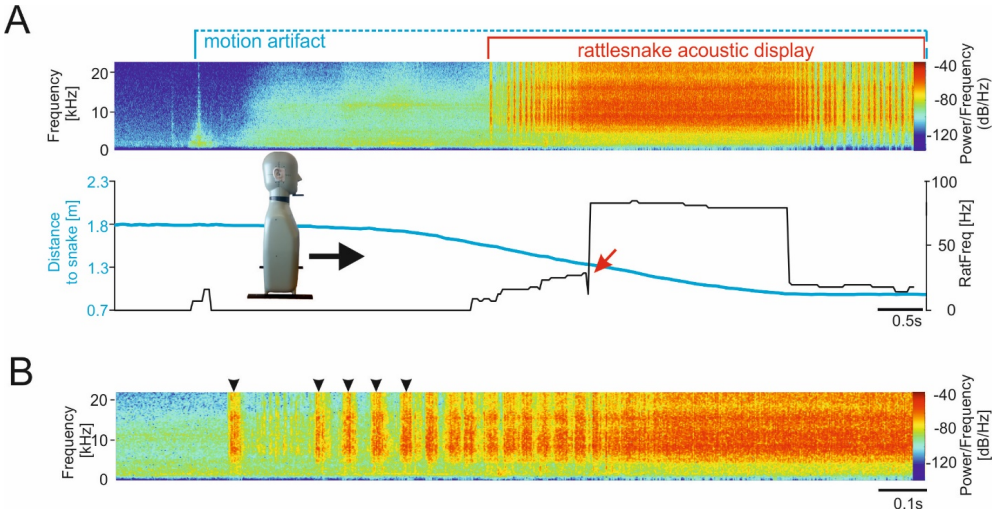

Figure 4.S1 **(A)** Spectrogram (top) of a rattlesnake acoustic display evoked by an artificial human torso which was moved towards the snake. Bottom graph shows the position of the torso relative to the snake (blue line) and the rattling modulation frequency (RatFreq; black line). Red arrow indicates the time point of the jump from the low to the high frequency mode. Note the acoustic artifact (indicated by the dashed blue line) caused by motion of the torso. **(B)** Higher temporal magnification of onset of rattling shown in (A) reveals that individual pulses of tail shakes (black arrowheads indicate the first five) are spectrally similar to high frequency rattling. Note that individual tail shakes merge with increasing frequency into a constant rattling sound.

The angle of the snake head was determined in the video frames just before, directly at, and just after stimulus onset. Only those recordings were analyzed in which the snake head was oriented towards the screen on which the visual stimuli were presented. An orientation of the snake's head directly towards the screen was set as a deviation angle α of 0° (figure 4.S2D). Only recordings with a deviation between −90° and 90° were analyzed. Snake head orientation towards the screen was measured post-hoc. Sound and video acquisition was synchronized via MATLAB.

4.2.5 Sound and video analysis

Custom written software was used to analyze sound recordings (MATLAB). Spectrograms of the upper temporal envelope (calculated as a 200-fold downsampling of the absolute signal with a digital anti-aliasing filter) were generated (window: 128 samples, overlap: 95 %, sample rate: 220.5 Hz, resulting bin size 0.6 s), providing the power spectral density (PSD) of the downsampled signal. From the PSD a modulation spectrogram was generated to allow for extraction of the dominant modulation frequency within each bin. The resulting curve of the dominant modulation frequency represented the modulation frequency of the rattling sound: the snake's tail shake frequency (RatFreq). A high modulation power at 0 Hz due to the rectification of the signal, as well as powerful low frequency artifacts at 1.7 Hz and 3.4 Hz in the modulation spectrogram could, however, mask the rattling frequency as the dominant modulation frequency. This led to the detection of rattling frequencies of 0 Hz, when in reality the snake was rattling. To limit detection of these false zero values during rattling, a lower

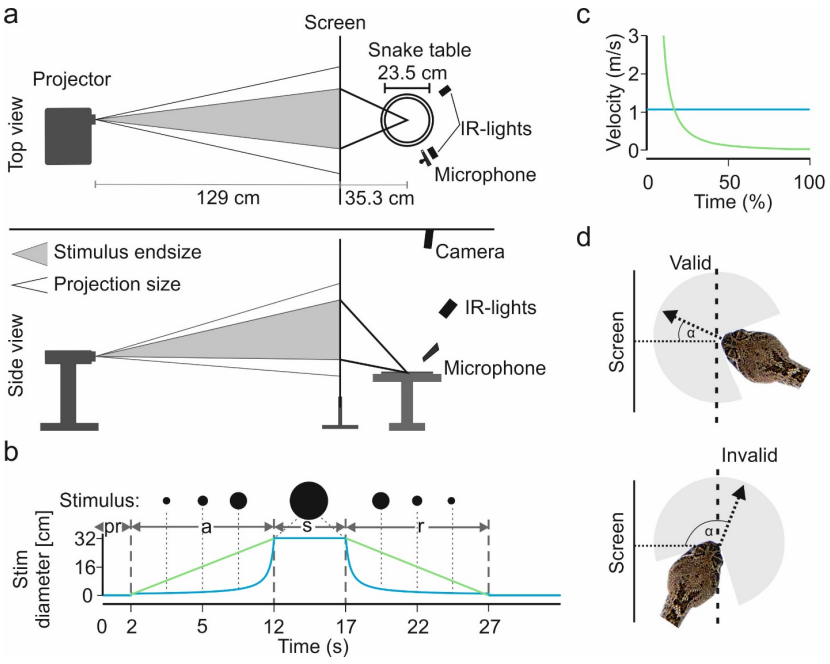

Figure 4.S2 Experimental design. **(A)** Schematic top and side view of the experimental setup showing the position of the projector used to present the visual looming stimuli on a screen and the table position in which the snake rested. **(B)** Schematic drawing showing the relative diameter of the two different types of visual stimuli used in the looming experiments: constant velocity (blue line, black circles) or decreasing velocity (green line) looming stimuli lead to different changes in circle diameter for a stimulus approaching over the course of 10 s. Black circles are not to scale. Stimulation phases are indicated: approach phase (a), stationary (s), retreat (r). **(C)** Velocity profiles of virtual objects approaching at a constant (blue line) and a decreasing (green line) velocity. **(D)** The snake head orientation (dorsal view) was used to determine validity of trials. Estimated visual angle of the snakes (area shaded in gray, from Reinert *et al.*, 1984).

cutoff value was set to remove modulation frequencies extraction below 5.17 Hz. Despite this cutoff, rattling could still be masked by low power modulation frequencies. We therefore excluded rattling sequences from further analysis that contained spontaneous drops of the rattling frequency to the lower analysis threshold during rattling. Since LF and HF sequences from one trial were analyzed independently, sample sizes for HF and LF sequences differ.

The sound level of the signal in decibels (relative to an arbitrary full-scale value) was calculated from the original signal (p) as

$$20 \log_{10} p$$

with identical bin sizes to the modulation frequency extraction (0.6 s) without overlap between bins.

■ **Modulation frequency analysis.** Rattling sounds were characterized by a broadband component to single tail shakes (figure 4.S1B). The absolute frequency range of rattling sound pulses was not analyzed, as the rattling sound itself has been subject to multiple studies already. Our sampling rate was instead adjusted to cover the frequency ranges of the rattling sound containing the most power (Fenton and Licht, 1990). Increasing tail flicking frequency led to a modulation of the spectrum. Several components of rattling sounds were analyzed: duration of sounds, the rate of frequency change and the time of an abrupt change in rattling modulation frequency, as well as the general distribution of rattling frequencies across trials.

LF rattling sequences were identified by searching for modulation frequencies that lay within the LF range (0–40 Hz) and directly preceded modulation frequencies in the HF range (> 60 Hz). Only the first LF rattling sequence per trial was analyzed. A linear regression was done through these sequences, from a frequency of zero preceding rattling initiation (to account for varying starting frequencies) and ending just before the shift to HF. The resulting regression coefficient served as the rate of RatFreq change. HF RatFreq sequences were determined similarly, starting at the beginning of HF rattling following LF rattling and ending before the first shift back below the lower HF limit or the end of the recording.

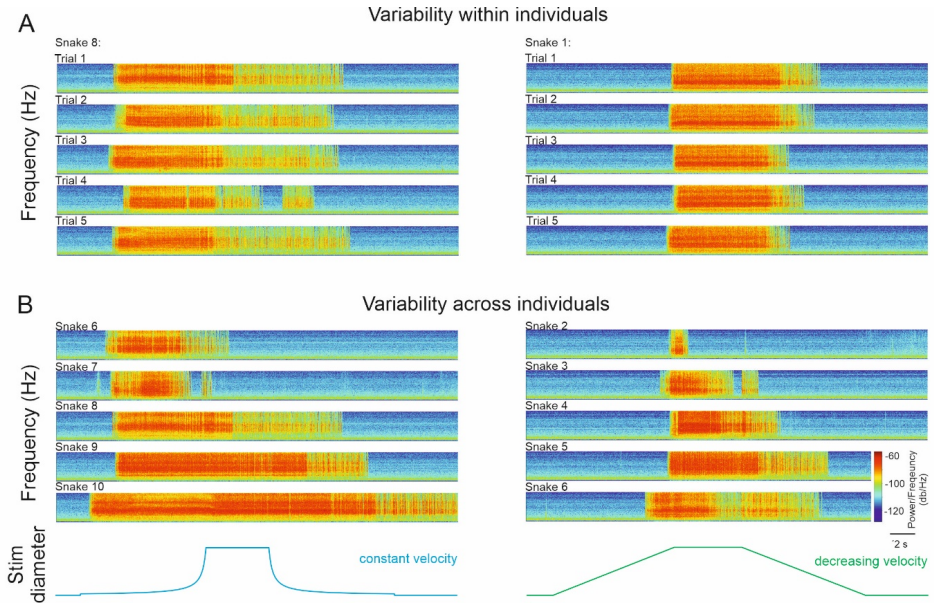

Figure 4.S3 Response variability to looming stimuli across and within snakes. Spectrograms of rattling responses to looming stimuli with constant (left) and decreasing (right) approach velocity across trials within a snake **(A)** and across snakes **(B)**.

		LF averages		HF averages	
approach velocity (constant)		slope (Hz/s)	duration (s)	slope (Hz/s)	duration (s)
very slow	(0.2 m/s)	7.59	1.73	−1.49	1.90
slow	(0.5 m/s)	13.20	0.96	−1.58	1.55
medium	(1.1 m/s)	13.57	0.73	−1.95	0.75
fast	(2.1 m/s)	14.57	0.59	−1.34	0.34

Table 4.S1 Summary of average slopes and durations of low-frequency (LF) and high-frequency (HF) components of the rattling behavior elicited by the visual looming stimulus with a constant velocity profile at four different approach velocities.

	LF averages		HF averages	
approach velocity (decreasing)	slope (Hz/s)	duration (s)	slope (Hz/s)	duration (s)
slow	11.15	1.30	−2.81	2.00
medium	12.60	0.98	−0.78	2.48
fast	12.18	0.85	−1.70	2.63

Table 4.S2 Summary of average slopes and durations of low-frequency (LF) and high-frequency (HF) components of the rattling behavior elicited by the visual looming stimulus with a decreasing velocity profile at three different approach velocities.

To assess the information contained in the rattling behavior elicited by an approaching object, only LF rattling during the approach phase of the stimulus was analyzed. Rattling durations were analyzed using the same start and end criteria used in RatFreq change analysis but were not limited to the end of the approach. In few cases snakes failed to elicit a HF mode (31.8 %) or began rattling in HF mode with no preceding LF mode (4.9 %). These recordings were also omitted from the analysis. Data was then pooled per velocity and significant outliers were identified (generalized extreme studentized deviate test) and removed.

4.2.6 Psychophysics

The human psychoacoustical experiments took place in an anechoic chamber with a 2 m × 2 m base and 2.2 m of height. The human subjects were individually seated on a chair facing a vertical array of five loudspeakers (Plus XS.2, CANTON Elektronik, Weilrod, Germany) at elevations of 0°, 12.5°, 25°, 37.5°, and 50° down, wore a Rift DK2 virtual reality head-mounted display (Oculus VR, Menlo Park, CA, US), and held a joystick in one hand.

The auditory stimuli were synthesized by repeating randomly generated individual rattling sounds at a rattling frequency that depended on the momentary virtual listener–snake distance and the trial condition (constant *vs.* adaptive rattling frequency). The individual rattling sounds were made up of 20 identical linearly decaying sawtooth wave pulses (center frequency 8 kHz, 1 ms duration) which were randomly spaced in time according to an exponential distribution with a rate parameter of 6 ms. The rattling frequency followed one of two functions of virtual listener–snake distance:

1. in the constant rattling frequency condition,

$$f_{\text{const}}(x) = 12\,\text{Hz}$$

2. in the adaptive rattling frequency condition,

$$f_{\text{adaptive}}(x) = \begin{cases} 70\ \text{Hz} & \text{for } x \leq 4\ \text{m} \\ 35\ \text{Hz} - 3.75\ \text{Hz/m} \cdot x & \text{for } 4\ \text{m} < x \leq 8\ \text{m} \end{cases}$$

The stimuli were fed into a 24-channel audio interface (24I/O, MOTU, Cambridge, MA, US) that was connected to the speakers via a 12-channel power amplifier (CI9120, NAD Electronics International, Pickering, ON, CA).

The virtual reality visual stimulus was a binocular rendering of a dim, flat, grassy landscape, presented through the head-mounted display, in which subjects could look around freely by rotating their heads. A bright spot on the ground pointed out a distance of 1 m. There were no visual cues as to the location of the virtual snake. The auditory and visual stimuli were dynamic. A 1 m/s approaching motion of the listener towards the virtual snake was simulated acoustically by a decrease of geometric attenuation and of sound source elevation over time—by vector-base amplitude panning (Pulkki, 1997) between loudspeakers in the vertical array—and visually by the optic flow of a flight through the grassy landscape.

■ **Procedure.** In two half-hour sessions performed at least one day before the main experiment, subjects were familiarized with the virtual audiovisual environment. We provided them with the same visual stimulation as in the main experiment and a similar auditory stimulation (500 ms on/250 ms off train of noise bursts, spectrally identical to the synthetic rattling pulses). In contrast to the main experiment, stimulus presentation automatically stopped when the virtual snake–listener distance reached 1 m, 1.41 m, 2 m, 2.83 m, 4 m, 5.66 m, or 8 m. The listeners "wore" a virtual headlamp and were asked to use it to point at the presumed location of the snake by moving their head. No feedback was given. The third session also lasted for approximately half an hour and constituted the main experiment. At the beginning of each trial, the subjects found themselves in silence and stationary in a new random location of the grassy landscape. After 0.5 s, sound and motion were turned on, until either a virtual snake–listener distance of 0.2 m was hit (miss trial) or the listener pressed a button on the joystick to indicate that they perceived the virtual snake to be 1 m away (non-miss trial). Acoustically, depending on trial condition, the momentary rattling frequency was calculated using either or . The starting distance from the virtual snake was either 1.41 m, 2 m, 2.83 m, 4 m, 5.66 m, or 8 m. The different rattling frequency conditions and starting distances were presented in a randomized order individual to each subject. Each pairing of trial condition and starting distance was measured 20 times. Data for individual subjects is reproduced in figure 4.S5.

Rattlesnake behavioral experiments were approved by the ethics committee of the Chair of Zoology, TUM Freising. Human psychophysics procedures were approved by the ethics committee of the Faculty of Medicine, LMU Munich (project no. 18-327).

4.2.7 References

Baird, Spencer Fullerton and Charles Girard (1853). *Catalogue of North American reptiles in the Museum of the Smithsonian Institution*. Smithsonian Institution.

Fenton, M. Brock and Lawrence E. Licht (1990). "Why rattle snake?" In: *Journal of Herpetology*, pp. 274–279.

Kohl, Tobias and Bruce A. Young (2011). "Electrophysiology of the snake retina". In: *Annual Meeting of the SICB*. Vol. 51. Society for Integrative & Comparative Biology, E71.

Place, Aaron J. and Charles I. Abramson (2008). "Habituation of the rattle response in Western Diamondback rattlesnakes, *Crotalus atrox*". In: *Copeia* 2008.4, pp. 835–843.

Pulkki, Ville (1997). "Virtual sound source positioning using vector base amplitude panning". In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466.

Reinert, Howard K., David Cundall, and Lauretta M. Bushar (1984). "Foraging behavior of the timber rattlesnake, *Crotalus horridus*". In: *Copeia*, pp. 976–981.

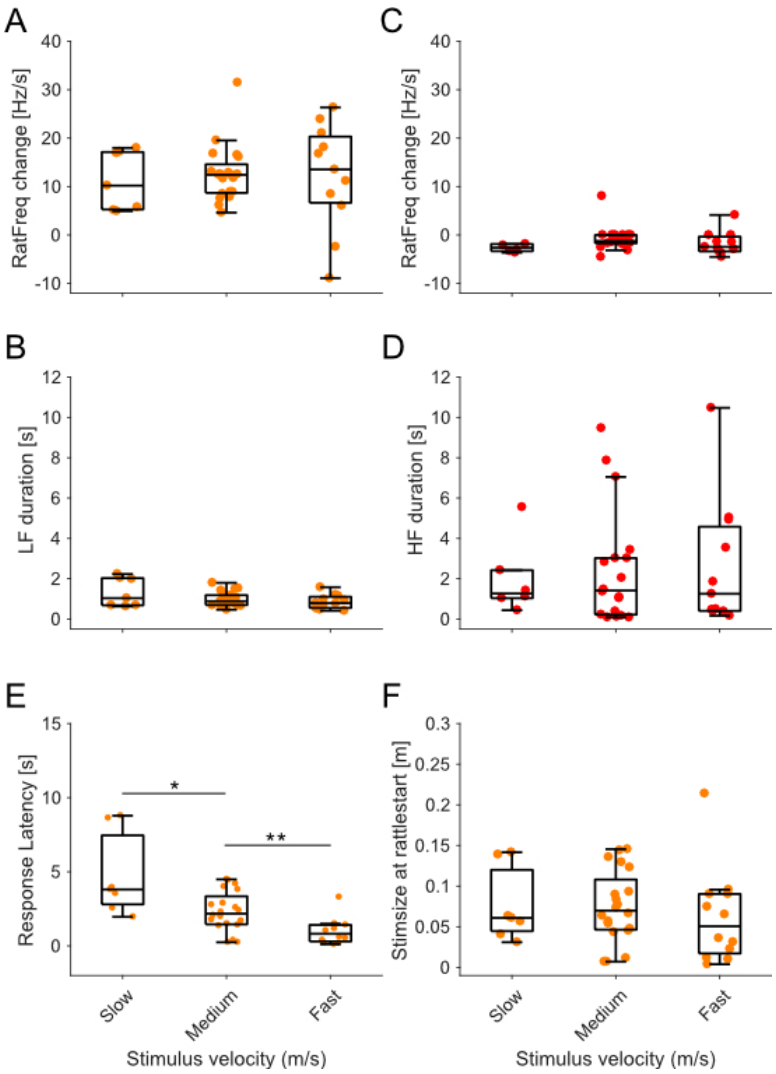

Figure 4.S4 Effect of decreasing approach velocity on the low (LF, orange) and high frequency (HF, red) phases of the rattling responses. Box and whisker plots of the LF rattling frequency changes **(A)** and LF mode duration **(B)** for different approach velocities until the shift to the HF mode. Neither factor depends on stimulus velocity. Modulation frequency changes **(C)** and duration **(D)** are not variable and independent of approach velocity for the HF rattling mode. Response latency **(E)** depends on stimulus velocity (Kruskal-Wallis, $p = 0.0004, \chi^2 = 15.67$), while the size of the looming stimulus at rattling onset is independent of stimulus velocity **(F)**.

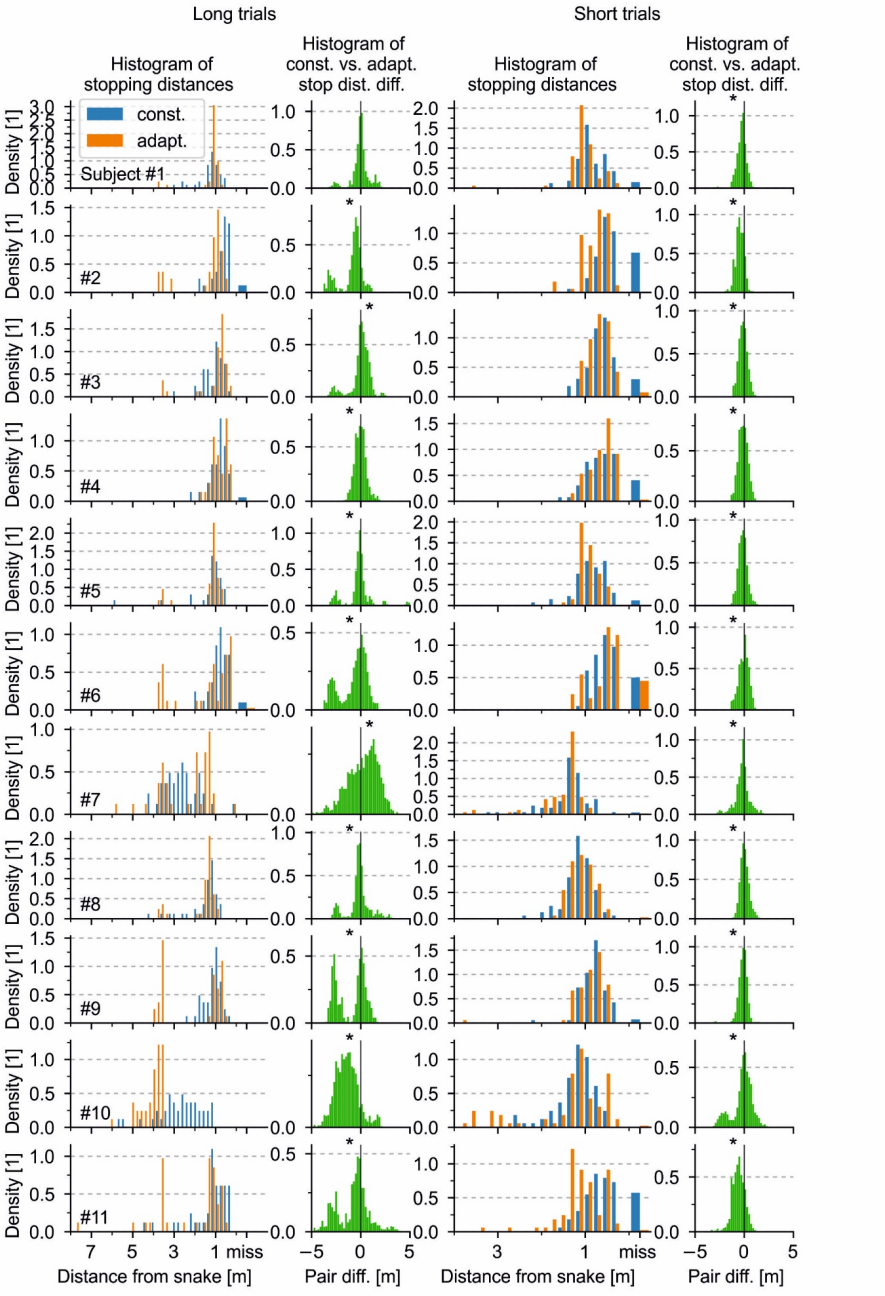

Figure 4.S5 Individual human psychoacoustical data for all eleven subjects. Each row refers to one subject. The first and second columns contain data obtained in trials where the starting distance between the virtual snake and the listener was greater than 4 m, the third and fourth column where it was less than or equal to 4 m. The first and third columns reproduce the histograms of virtual listener–snake distances at which individual listeners stopped the trial because they perceived the virtual snake to be exactly 1 m away, color-coded by condition (in blue for constant, in orange for adaptive rattling frequencies). The two bars at the very right of these plots indicate the prevalence of miss trials. The second and fourth column are histograms of all pairwise differences (between all stopping distances in the adaptive rattling frequency condition and all stopping distance in the constant rattling frequency condition). Asterisks indicate that the distribution of pairwise differences is significantly different from a symmetric distribution around zero (Wilcoxon signed-rank tests, $p < 0.05$). The position of the asterisk indicates the direction of the effect. If it is printed on the left, distances were higher in the adaptive condition, otherwise they were higher in the constant condition.

CHAPTER 5

Moving listeners & moving sources

Chapter 2 described an experiment in which I asked human subjects to quantify one aspect of their perception of the acoustics of a simulated room. The positions of both the sound sources and the listeners in virtual space were kept stationary. In everyday life, of course, both listeners and sound sources are often in motion. Moreover, in contrast to the simple "replay" of predetermined movement trajectories as in the studies presented in Chapters 3 and 4, the exact movement trajectories are not usually known a priori. To facilitate experiments that allow spontaneous sound source and listener motion (within a virtual enclosure whose acoustic response is to be simulated), I created a new implementation of the same room-acoustical model that I already used in Chapter 2, but which can process scene geometry updates in real time and thus allows the simulation of dynamic scenes. This chapter presents this new piece of software alongside a description of the underlying model and a general overview of the idea of room-acoustical simulation.

5.1 Introduction

The concept of simulating the acoustics of an enclosed space can be traced back to the field of architectural acoustics, where it is crucial to predict a variety of parameters (such as the intelligibility of speech in a classroom, or the pleasantness of the music heard by an audience in every seat of a concert hall) before the room is built. Early in the 20th century, it was popular to create scale models out of light-reflecting and light-absorbing materials which could be illuminated to predict sound energy distribution (Rindel, 2002). While methods of this kind could be helpful, for example, to avoid acoustic analogues to "dark spots" appearing in a space meant for listening, the high speed of light makes it impossible to deduce any time-dependent properties of the modelled environment. Specifically, these measurements cannot generate acoustic signals that could be used to produce the auditory percepts required for subjective evaluations. In other words, they cannot be *auralised*, where auralisation is defined as *"the process of rendering audible, by physical or mathematical modelling, the sound field of a source in a space, in such a way as to simulate the binaural listening experience at a given position in the modeled space"* (Kleiner *et al.*, 1993).

Spandöck (1934) was the first to succeed in acoustically exciting the model of a room (at a scale of one fifth of its actual dimensions) and making its response audible, by playing back from and recording onto phonograph cylinders inside the replica. This technique facilitated both quantitative analyses as well as qualitative studies with human listeners (*e.g.* Krauth and Bücklein, 1962). Many of today's computer simulation approaches can be considered similar in concept, in that they typically also "emit sound" at one place in a model and "record" it at another—with, of course, the crucial difference that the physical transmission of sound waves throughout the model is replaced with appropriate binary arithmetics. The nature of these computations varies greatly; the most common examples will be discussed in this section. Note, however, that even some computational methods do not lend themselves to auralisation, and as such are not able to create stimuli for psychophysical experiments. These are not of interest for the type of research projects described in this thesis, and are therefore not considered further. Neither will I discuss the physically most rigorous approaches which provide numerical solutions to the wave equation in time and space (*wave-based* methods),

e.g. via the boundary element method (see Kirkup, 2019) or the finite element method (see Thompson, 2006). This introduction is instead focused on methods which follow geometrical principles, partly because those are the most common in practice, approximately correct for sufficiently short wavelengths of sound, and far less computationally demanding (Savioja and Svensson, 2015; Siltanen, Lokki, and Savioja, 2010)—but mostly because the RAZR simulation model, the focus of this chapter (see section 5.2), is a member of this class.

5.1.1 Computational models based on geometrical principles

Schroeder *et al.* (1962) first introduced digital computers to the field of room-acoustical simulation. Beside a concept of recursive digital networks for artificial reverberation (Schroeder, 1962) which is commonly used to this day, this early work includes a method wherein a *"computer calculates the paths of 300 rays from an omnidirectional source"*, simulating sound-absorbing reflections off of walls while the program *"keeps a running account of the remaining energy"* (quotations from Schroeder, 1969). He envisioned that the method would be useful to *"'preaudit' architectural designs before construction"*, but also already mentioned having used it in a psychophysical study of how the results of a reverberation process are perceived by human listeners. This pioneering work by Schroeder, however, took the shape of a proof of concept with substantial simplifications. For example, it was apparently limited to two-dimensional "rooms". Krokstad *et al.* (1968) are credited with the first practically useful execution of the idea (Kuttruff, 2016; Rindel, 2002; Savioja and Svensson, 2015).

■ **Ray tracing and sound particle simulation.** Methods similar to Schroeder's and Krokstad's are still in use today as one of the main classes of modern room-acoustical simulation techniques. These are commonly called *ray-tracing* algorithms and can—when removed from the context of room acoustics—be considered to be an implementation of an algorithm first put into words by the German Renaissance painter Albrecht Dürer (1525; see Hofmann, 1990 for reproductions). The principle is visualised in Figure 5.1.

Ray tracing is usually done stochastically: A number of rays is sent out in random directions from the location of the sound source, and each of these rays that passes through a receiver volume is captured and included in the target model of the room response. Whenever a ray hits a wall, it is reflected; this may be modelled, for example, by changing the ray's path according to the law of reflection (modelling *specular* reflections only; Figure 5.1, left), or by stochastic scattering, giving rise to an entire range of rays (also modelling *diffuse* reflections; Figure 5.1, right). The sound energy remain-

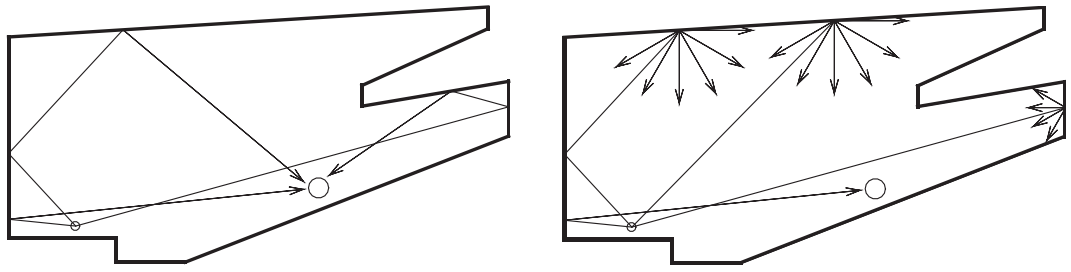

Figure 5.1 Illustration of ray tracing. Simulated rays are sent out from a sound source position (small circle) and reflected at surfaces until they arrive in a receiver volume (larger circle). **Left:** By enforcing the law of reflection, *i.e.*, only considering reflected rays that make the same angle with the surface normal than the incident ray, the simulation can be limited to specular reflections. **Right:** Diffuse reflections can be considered by generating other angles of reflection too. Reproduced from Savioja and Svensson (2015) in accordance with the Creative Commons Attribution 3.0 Unported License.

ing in each ray is typically tracked in frequency bands, appropriately attenuated due to absorption at surfaces (a frequency-dependent process) and distance travelled. A ray can be discarded if its energy falls below a certain threshold across all frequency bands. There is obviously a risk of under-sampling the space when too few rays are emitted in a limited number of directions, such that subsequent applications of a ray-tracing program to the same scene can yield very different results (Kulowski, 1982; Vorländer, 1988). When using a ray tracer, care should therefore be taken to sample not too sparsely to achieve a stable result, but—if runtime is of any concern—not too densely either.

Beam tracing (Heckbert and Hanrahan, 1984) is a similar method which replaces the infinitely thin rays with pyramidal volumes (see Figure 5.2). This is a useful modification due to spatial coherence: If a given ray emitted from a source can be traced to a receiver volume via a certain series of reflections, then the same is probably true for a ray emitted in a slightly different direction. For each beam, its entire volume is tested for intersections with reflective surfaces, such that all possible rays within it can be traced at once. This reduces the risk of sampling the space too sparsely or too densely, at the cost of more computational work that a beam tracer needs to do for each beam (compared to a ray tracer for each ray). Funkhouser *et al.* (2004) described a simulation system based on beam tracing, noting that it is fast enough for interactive use when combined with clever data structures of pre-computed geometrical information.

Sound particle simulation (*e.g.* Stephenson, 1990) is closely related to ray tracing. It draws on the idea of sound as an infinitesimally small object that is bounced off of walls and other surfaces. While the underlying geometrical considerations are the same as for the ray-tracing perspective, the particle-simulator view stresses temporal and energetic aspects of sound propagation; instead of intersecting a line with a surface and inferring path lengths and delays, these algorithms will move particles in space in discrete timesteps and detect when collisions occur. A recent implementation based on this model view has been created by Picaut and Fortin (2012) under the auspices of the French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR). They call the method SPPS and distribute it in an open-source program named i-Simpa.

■ **Image-source model.** In terms of propagation delay and geometric attenuation, the specular reflection (*i.e.*, without regard for scattering) of a sound source at a surface behaves like another sound source placed behind that surface. This imaginary *image source* is located at the same distance and along the same normal (straight line perpendicular to the surface) as the true sound source, *i.e.*, it is mirrored. It is referred to as a first-order image source, as it was found by reflecting the sound source one time. Unlike an ideal mirror, a typical reflective surface absorbs sound energy, which is accounted for by processing the sound from the image source with appropriate digital filters. In a straightforward implementation of the image-source model (also called simply the "image model", especially by early developers such as Gibbs and Jones, 1972), these processes of mirroring and filtering

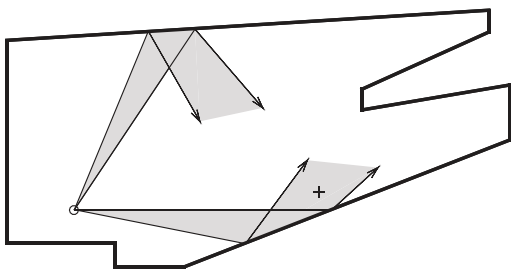

Figure 5.2 Illustration of beam tracing. This method functions very similarly to classical ray tracing, but it assigns a non-infinitesimal volume to each "ray". Reproduced from Savioja and Svensson (2015) in accordance with the Creative Commons Attribution 3.0 Unported License.

are simply iterated for every image source of the $(n-1)$-th order, and every face of every object in a room, to obtain all the reflections of the n-th order, up to a predefined maximum n. Figure 5.3 illustrates the idea visually in two dimensions.

When allowing the simulated room to have an arbitrary geometry—to be made up of an arbitrary set of surfaces (Borish, 1984; Santon, 1976)—ray-tracing methods have been found to come to a result faster than is possible by dealing with image sources (Stephenson, 1990). This is due to exponential growth of the number of image sources as the order increases, as well as the computationally expensive checks whether a mirror image of a sound source with regard to a surface is actually visible from the receiver's position. One way to make the image-source model more tractable is to only consider sequences of surface reflections which are detected by a ray tracer (Vorländer, 1989). Another is to restrict the simulated room to be cuboid and to disallow any additional surfaces within it, in which case the growth of the image-source count becomes quadratic (Grünbaum, 1994) and many geometrical considerations can be avoided (*e.g.*, due to the absence of protruding walls, no attention needs to be paid to any occlusion effects). Such a simplified geometry is commonly known as a *shoebox room* and allows the calculation of image source positions with a closed-form equation (Allen and Berkley, 1979) and efficient computational techniques (McGovern, 2009).

As opposed to ray tracing (Kuttruff, 1971; Mehta and Mulholland, 1976), the image-source approach can only accurately determine specular reflections, which are those for which the law of reflection holds: The angle of incidence is required to be equal to the angle of reflection. Since sound waves undergo scattering at real-world surfaces, a plausible simulation of room acoustics has to consider the resulting diffuse reflections too (Dalenbäck *et al.*, 1994; Hodgson, 1990). This motivates the popular combination of an image-source model with another simulation technique in hybrid approaches (Vorländer, 2008), as described later. There are also methods to simulate diffusivity "within" an image-source model by extending it with filters for temporal smearing (Buttler *et al.*, 2018; Siltanen, Lokki, Tervo, *et al.*, 2012).

A modern, purely image-source-based room-acoustical simulation program has been developed at TU München as the core software component of the "real-time Simulated Open Field Environment" (RTSOFE; Seeber and Clapp, 2017). This environment comprises a purpose-built anechoic

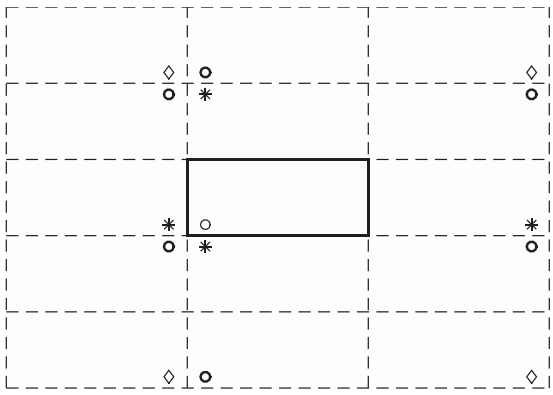

Figure 5.3 Illustration of the image-source model. The rectangle with the solid lines represents the room whose acoustics are to be simulated; the dashed rectangles are its mirror images. The sound source location is given by the thin circle. In this two-dimensional room, there are then four first-order image sources, one for each reflection at the four walls (denoted by the asterisks). Eight second-order image sources are generated by reflecting the first-order image sources at all four walls again and discarding duplicates; six of these are shown here (denoted by thick circles). The lozenges represent some third-order image sources. Reproduced from Savioja and Svensson (2015) in accordance with the Creative Commons Attribution 3.0 Unported License.

chamber with a loudspeaker array for research in psychophysics along with custom software to render sound in simulated enclosed spaces. The simulation program relies on highly efficient vectorised operations available in the instruction sets of modern central processing units and can thus rapidly generate great numbers of image sources at high orders for arbitrary room geometries.

■ **Radiosity.** In thermal engineering, the transfer of heat can be modelled with a partial differential equation and solved numerically using the *finite element method*: subdividing the space of an objects of interest into a mesh; estimating the energy contribution of each patch in the mesh to every other to obtain a system of linear equations; and finding the solution to this system (Sparrow and Cess, 2018). Goral *et al.* (1984) discovered the similarity of this problem to that of global illumination in computer graphics and named their application in this domain the *radiosity* method. They noted its utility in dealing with purely diffuse reflectors, which require special consideration in other models such as ray tracing.

G. R. Moore (1984) formulated it again in room acoustics and provided a computer implementation which uses the results of the radiosity calculations as an input for subsequent processing of specular reflections with an image-source model. In general, the special importance of specular reflections in room acoustics is widely recognised by authors who study radiosity in this context, as evidenced by the popularity of combining it with image-source models or (derivatives of) ray tracing (Koutsouris *et al.*, 2013; Lewers, 1993; Tsingos and Gascuel, 1997). Another major aspect that sets apart acoustic radiosity is that it needs to express energy as a function not only of place, but also of time; this is not necessary in computer graphics, where for all practical purposes, thanks to the high speed of light, the illumination of a scene reaches a steady state instantly.

The radiosity method has been found to deliver results that agree well with reality in predicting the overall acoustical properties of a room, though its results are lacking when it comes to the precise determination of individual reflections (Hodgson and Nosal, 2006). As supplementations with other models are possible, it might be surprising that radiosity appears to be a fringe method in acoustics. Its long-standing limitation to convex rooms and the computational effort it requires are possible causes (Nosal *et al.*, 2004).

■ **Hybrid approaches.** As already suggested in the discussion of radiosity, current computer programs for the simulation of room acoustics often use not just one of the mentioned algorithms, but combine two or more of them to benefit from the strengths of each. Many more examples of this are found in commercial software, typically aimed at an audience of practitioners in the field of acoustical engineering. The combination of an image-source model with a ray-tracing method is particularly popular. Such products include

ODEON (Naylor, 1993): ODEON A/S, Copenhagen; ray tracing and an image-source model with simplifications at higher orders;

CATT-Acoustic (Dalenbäck, 1995): CATT Computer Aided Theatre Technique, Gothenburg; an image-source model and ray/cone tracing (van Maercke, 1986);

EASE (Ahnert and Feistel, 1993; Schmitz *et al.*, 2001): AFMG Technologies GmbH, Berlin; image-source model and ray tracing with diffuse rain (a derivate of radiosity; Heinz, 1993).

A notable example of such a hybrid from research is RAVEN (Schröder and Vorländer, 2011), created at RWTH in Aachen, Germany for use in their virtual reality setup at the Institute of Technical Acoustics. The program is described in detail by its original developer in his doctoral thesis (Schröder, 2011). It was developed with a focus on efficiency and interactive control, topics which are discussed in the following section. Many physical phenomena that are commonly handled only by wave-based simulation methods, such as diffraction (the bending of sound waves at the edges of an obstacle) and transmission (the passage of sound waves from one medium to another), are handled by the program in addition to the hybrid ray-tracing/image-source model.

Another hybrid method is implemented in RAZR (Wendt *et al.*, 2014) from the University of Oldenburg, Germany. Compared to the above models, it is highly simplified and approximates arbitrary room geometries by shoebox rooms to allow the efficient calculation of early reflections with an image-source model at low orders. The effects of scattering due to objects within the room and rough wall surfaces are simulated in the time domain with sparse infinite impulse-response filters (Buttler *et al.*, 2018). Late reverberation, where the generation of image sources becomes computationally expensive, is handled by a spatially mapped network of delay lines and feedback connections (*feedback-delay network* for short; Jot and Chaigne, 1991). Despite these simplifications, the model has been confirmed to be highly perceptually valid in a subjective study with expert listeners where RAZR was tested alongside some of the highly complex models described above (Brinkmann *et al.*, 2019).

TASCAR (Grimm, Luberadzka, *et al.*, 2019) is another framework from the University of Oldenburg. The name is short for "toolbox for acoustic scene creation and rendering" and already reveals that the scope of the software is broader than just room-acoustical simulation. For the purposes of research in audiology, it allows the creation of multiple virtual sound sources and receivers in space, and almost arbitrary interactive changes to the scene while the program is running. TASCAR has included an image-source model from an early version, and has recently (in early 2020) gained a simple feedback-delay network too.

5.1.2 Real-time applications

In general terms, for a computer system to operate in *real time* means that given an input x at time t_1, it generates the correct output $f(x)$ no later than at the *deadline* time $t_2 = t_1 + \Delta t$ (Ben-Ari, 2006), where $\Delta t > 0$ is the acceptable latency. In audio applications, x is one audio signal or several (from prior recordings, synthesis, or from a live signal from a microphone) together with a problem-specific set of parameters (such as gains, filter settings, or in the case of room-acoustical simulation, some geometrical information about the simulated scene), and $f(x)$ is the signal to be emitted by loudspeakers or headphones. The acceptable latency, then, is dependent on the desired overall delay between an interaction with the system and its response in the form of an acoustical signal. Δt is typically required to be some tens of milliseconds.

In the context of computational room acoustics, it is useful to draw a distinction between auralisation *vs.* simulation in real time. The former term places the real-time constraint only on the processing of sound, not on that of geometry (*i.e.*, x is only the audio signal): The acoustics of the room can thus be calculated in advance, and the real-time part of the system only needs to be concerned with applying these acoustics to a stream of sound. The latter concept considers geometrical information as part of the input data. This is required to allow movement within a scene, for example by head tracking, or by input from a device such as a keyboard, mouse or joystick. A system which fulfills the real-time constraint only for the auralisation step gives rise to a static VAE (called thus because the geometry is fixed), which stands in contrast to a more flexible dynamic VAE.

■ **Static virtual acoustic environments.** A highly useful and popular (see *e.g.* Kleiner *et al.*, 1993; Vorländer, 2008) tool for real-time auralisation are real-time finite impulse response (FIR) filters based on fast convolution, see sections 1.3.3 and 1.3.4. These can apply the acoustics of a room (simulated or measured with microphones), represented as a finite impulse response, to an arbitrary incoming audio signal at a very low latency. Fast convolution is required because direct convolution would be become prohibitive for FIR lengths of just a few milliseconds, whereas the perceptible part of the impulse response of a typical room is typically hundreds of milliseconds long. While fast convolution it is more efficient for long FIRs, it tends to be slower for shorter ones (see Strum and Kirk, 1988). Moreover, it introduces additional latency, as its efficiency benefits depend on the collection of a sufficiently large amount of input data before starting a processing cycle.

This issue of additional latency can be solved by *block*, *partitioning* or *sectioned* convolution (all synonyms). This class of methods splits up the FIR into at least two segments, and a convolution (in the time or in the frequency domain) is calculated for a chunk of the input signal with each of these segments. The two classical algorithms of this kind are called *overlap–add* and *overlap–save* (see Rabiner and Gold, 1975). An extension, which uses non-uniform segment sizes, combines time-domain convolution for early (latency-critical) segments of the impulse response with the frequency-domain algorithm for the later segments, and makes sure that the computational load remains even over time, became known as the Gardner scheme (W. G. Gardner, 1995). A number of partitioning schemes, filter structures and other optimisations has been developed since (*e.g.* Battenberg and Avižienis, 2011; García, 2002). A comprehensive review is given in the dissertation of Wefers (2014).

Partitioning convolvers are widely implemented in free and commercial software. On modern computers, these programs can filter multiple channels with several second-long FIRs, without adding any latency beyond the delays introduced by the hardware and the operating system's audio processing stack. All the room-acoustical simulation programs described in the previous section are able to calculate finite impulse responses for a given room, source and listener orientations, and a variety of reproduction setups. Consequently, each combination of a simulation software and a suitable convolver can be considered a real-time auralisation system. Such a system cannot by itself adapt its output to changes in the scene, at least not without a perceptible interruption while the convolver is restarted. Due to this restriction, this approach produces a *static virtual acoustic environment* (Xie, 2013). Accounting for interactive changes of simulation parameters requires additional work.

■ **Dynamic virtual acoustic environments.** Following the deliberations above, an obvious implementation of a real-time system which can accommodate interactive changes may be built on the basis of a convolver which supports time-varying impulse responses. In fact, this technique can be applied in a way that is agnostic to the choice of room-acoustical simulation model and software. Before the auralisation, one can pre-calculate or measure room impulse responses for a grid of parameters of interest (*e.g.*, the possible positions and orientations of a listener inside a room) in order to build a database. Theoretically, this may be understood as sampling impulse responses from a function of space, sometimes called the *plenacoustic function* (Ajdler *et al.*, 2006). The real-time part of the system then only needs to estimate the true value of this function corresponding to the momentary parameters, by interpolating between the appropriate sampling points, and to convolve an acoustic signal with this reconstructed impulse response. The design of algorithms for the interpolation step is an ongoing field of research (*e.g.* García-Gómez and López, 2018; Kearney *et al.*, 2009; Samarasinghe *et al.*, 2015; Zhang *et al.*, 2019).

Clearly, this approach is characterised by a trade-off between producing an accurate auralisation and keeping the database small: To get an acceptably small interpolation error, a sufficiently high number of impulse responses must be calculated ahead of time. For example, considering horizontal listener rotation alone, an angular resolution of 2° (*i.e.*, 180 sampling points to cover the full 360°) have been recommended to avoid audible interpolation artifacts (Lindau, Maempel, *et al.*, 2008). Depending on the parameter values that must be expected to occur during the auralisation, the parameter space—and hence the impulse response database—can become prohibitively large. This problem may be avoided when the room-acoustical simulation itself is sufficiently fast to provide updated impulse responses in real time. This is indeed a common mode of operation: It is, for example, the principle behind RAVEN (Schröder, 2011) and the system by Funkhouser *et al.* (2004). RTSOFE (Seeber and Clapp, 2017) also adheres to this scheme. It can generate impulse responses for changed geometries hundreds of times per second with less than a millisecond of latency. The RTSOFE system also comes with a custom-built partitioning convolver that further reduces output latency by performing the early part of the convolution (up to approximately 100 ms on suitable hardware) in the time domain, which it achieves by considering the sparse structure of the early part of a room impulse response.

In some way, the generation of finite room impulse responses can be considered a detour: Rather than immediately processing the desired sound according to the acoustics of a simulated room, the calculations are first performed for a brief impulse, and the sound is only brought into the system in the convolution step. This does not come without issues. Most notably, abruptly switching from one impulse response to another tends to produce audible artifacts, such that some kind of transition is required whenever the geometry changes. Probably the most common transition is a cross-fade from the signal convolved with the old impulse response into the one convolved with the new. This implies that within such a transition period, the presented impulse response is not accurate for either geometry. If, for example, a late echo from a far-away surface changes in timing because the listener moved after the sound was emitted, but before the echo arrived at their ear, the cross-fading approach may generate two fainter echoes, one of which is physically incorrect entirely, while the other one has an incorrect amplitude.

Depending on the chosen room-acoustical simulation model and its concrete implementation in software, the generation of an impulse response can be bypassed, thereby avoiding the necessity of fading and the problems it brings. An alternative is to directly apply delays, gains, filters *etc.* with time-varying parameters to the input sound. The "Digital Interactive Virtual Acoustics" (DIVA) platform from Helsinki University of Technology (Lokki, 2002; Savioja, Huopaniemi, *et al.*, 1999) functions in this way, as does the TASCAR framework (Grimm, Luberadzka, *et al.*, 2019). The model of RAZR (Wendt *et al.*, 2014) also readily lends itself to this treatment, although its reference implementation in MATLAB has so far been limited to the generation of impulse responses.

■ **Latency considerations.** Psychophysical studies give some indications as to how fast a dynamic VAE needs to operate for it to be considered plausible. Listeners have been found to perform well in sound localisation tasks even when there are delays of 150–250 ms between rotating their heads and receiving updated signals via headphones which reflect the rotation (Sandvad, 1996; Wenzel, 1999). For immersion in a virtual acoustic environment, however, it should be taken into account they can perceive much smaller latencies in the range of 55–75 ms (Brungart, Kordik, *et al.*, 2006; Lindau, 2009; Mackensen, 2004; Yairi and Iwaya, 2006). Brungart, Kordik, *et al.* (2006) even argue that in situations where a VAE is overlaid on real-world sounds, as in augmented-reality applications, the target latency should be at most 30 ms.

If the subject's own voice is part of the simulation, such as in a virtual echo-acoustic environments (Flanagin *et al.*, 2017; Schörnich *et al.*, 2012; Wallmeier, Geßele, *et al.*, 2013; Wallmeier, Kish, *et al.*, 2015; Wallmeier and Wiegrebe, 2014a,b), even smaller latencies are required: For a physically accurate simulation of reflections, the latency must not be larger than the time taken by a sound to travel to the nearest virtual reflective surface and back to the listener who emitted it; this time is just 5.8 ms for a simulated reflector 1 m away.

5.1.3 Motivation

Given the existence of a number of dynamic VAEs which let listeners and sound sources move interactively within a simulated room, it might not be immediately obvious why another system of this kind might be desirable. I originally decided to pursue the present project of suitably extending RAZR in view of planned studies of human echolocation in virtual rooms. Such experiments necessitate very low latencies not only for the simulation of reflected sounds from nearby walls, but also to obtain an accurate stimulation with echoes even in the presence of spontaneous self-motion, which has been shown to be very important for human echolocators (Milne *et al.*, 2014; Tonelli *et al.*, 2018; Wallmeier and Wiegrebe, 2014b). At the same time, a VAE for this purpose should already be well-evaluated for the accuracy and *plausibility* (Lindau and Weinzierl, 2012) of its simulation results.

RAZR is an attractive basis for such a real-time system not only because subjective listening tests have already demonstrated the fidelity of its outputs (*e.g.* Wendt *et al.*, 2014, 2016), but also because

the simplicity of the model allows the effects of interactive geometrical updates to be calculated very quickly. For this reason, a real-time implementation of RAZR can be expected to support operations within very low latency thresholds. Taken together, these considerations should make this new software very useful for many more applications than merely for echolocation experiments.

5.2 The RAZR model

This section describes the signal processing components of the RAZR model as originally described by Wendt *et al.* (2014) together with some extensions. The original MATLAB version of RAZR is freely available from the web at `http://www.razrengine.com/`. Based on a static configuration, it generates image sources (section 5.2.1) and collects their individual contributions to the simulated acoustics of a shoebox room either into an overall two-channel FIR for headphone presentation (called a *binaural room impulse response* or BRIR), or into a multi-channel FIR for loudspeaker arrays. Some image source outputs are fed via a geometry-based channel mapping (section 5.2.3) into a feedback delay network (section 5.2.2), whose output channels are also integrated into the output FIR much like the image sources. The remainder of this section will introduce these three main components in detail.

5.2.1 The image-source model (ISM)

The ISM component of RAZR is currently restricted to empty cuboid (shoebox) rooms. An extension to arbitrary geometries is in development.

Image sources for specular reflections are generated up to a specified order, by adding impulses with the appropriate amplitudes (according to the $\frac{1}{r}$ distance law) at the appropriate time points of the room impulse response. The maximum reflection order is typically low; by default, it is 3. Each impulse individually undergoes several steps of filtering, summarised in Figure 5.4.

■ **Reflection filters.** The reflective properties of the walls are specified in frequency-band reflection coefficients (or alternatively by using a materials database, which maps human-readable names to these coefficients). A range of methods is implemented to fit IIR filter coefficients to this specification. The default is to use "composed parametric equalisers": In a first step, frequency bands with similar reflection coefficients are merged. Subsequently, second-order low-shelf and high-shelf filters are designed, using the method of Holters and Zölzer (2006), to yield the desired gains at the edges of the frequency range of interest. Finally, if there are more than two frequency bands to be considered,

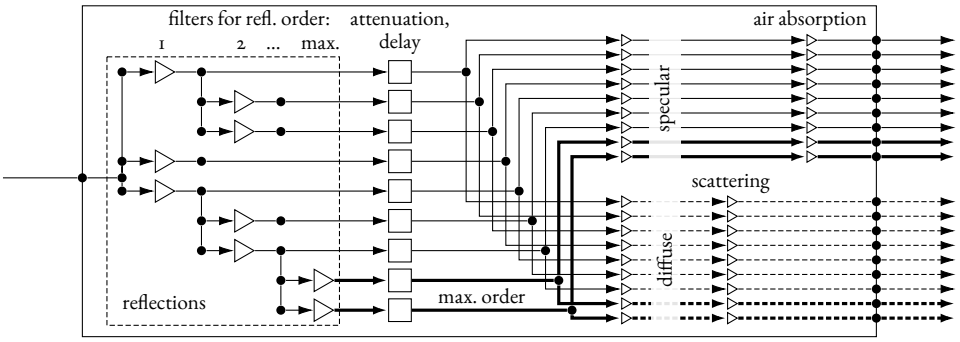

Figure 5.4 Block diagram of the signal processing related to the image-source model (ISM) implemented in RAZR. Diffuse paths are dashed and maximum-order paths are highlighted in bold; these contribute to the input of the feedback-delay network (FDN). Triangles represent IIR filters; the ones within the box labelled "reflections" contain reflection, source directivity and smearing filters. Squares represent the gains and delays due to sound propagation.

peak filters are calculated following the same approach. Their order depends on the bandwidth (2 for an octave). The shelving and peak filters are then composed in order to obtain one IIR filter per wall. For every image source, these filters are applied to the signal as often as the sound is reflected at the corresponding wall.

■ **Air absorption.** Following Grimm, Wendt, *et al.* (2014), to account for the stronger attenuation of high-frequency sound by air, a simple IIR filter is employed with the normalised feedforward coefficient

$$b_0 = \exp\left(-\frac{rf_S}{c\alpha}\right)$$

(with the distance r, the sampling rate f_S, the speed of sound c and the empirical constant $\alpha = 7782$), and the normalised feedback coefficient

$$a_1 = b_0 - 1.$$

This filter is applied once to each image source as well as to the direct sound path.

■ **Source directivity.** Real-world sound sources do not radiate acoustic energy equally in all directions. Instead, sound typically spreads in a frequency-dependent spatial pattern depending on the size and shape of the sound source, with low frequencies radiated in an omnidirectional manner and high frequencies focused towards a certain direction. RAZR includes both FIR-based (Blau *et al.*, in print) and IIR-based (Steffens *et al.*, 2019, and in revision) filters to model this effect for virtual human speakers. The FIR approach is based on measurements with microphones placed at various angles around an emitter of sound (analogously to how a HRTF database models the receiver directivity of a human head). The IIR approach follows C. P. Brown and Duda (1998): It sets

$$\omega_0 = \frac{c}{af_S},$$
$$\alpha = 1 + \frac{\alpha_{\min}}{2} + \left(1 - \frac{\alpha_{\min}}{2}\right)\cos\frac{\theta}{\theta_0\pi}$$

and the normalised feedforward and feedback coefficients

$$b_0 = \frac{\alpha + \omega_0}{1 + \omega_0}, \qquad b_1 = -\frac{\alpha + \omega_0}{1 + \omega_0}, \qquad a_1 = -\frac{1 - \omega_0}{1 + \omega_0}.$$

θ refers to the azimuthal angle for which the filter should be computed, θ_0 is the angle at which the overall attenuation is maximal (set to π, *i.e.*, behind the speaker), $\alpha_{\min} = 0.05$, and a is the radius of the spherical head implied by this model. f_S and c are as above.

This filter is applied once for each image source.

■ **Spatial mapping.** For each image source, RAZR calculates the azimuth and elevation relative to the receiver. This information is used according to the spatialisation mode selected by the user. Each image source can be included into the room impulse response

- by writing it out diotically, ignoring the spatial information,
- by applying broadband interaural level differences,
- by filtering its output according to an HRTF database,
- by processing it using IIR filters derived from an extension (including elevation-dependent filters and an altered azimuth dependency to better account for ILDs; Buttler, 2018) of a spherical head model (C. P. Brown and Duda, 1998), similar to the approach described above for source directivity, or
- rendered onto a loudspeaker array using vector-base amplitude panning (Pulkki, 1997).

■ **Surface and object scattering.** The image-source model, let alone one limited to an empty shoebox room, does not by itself give rise to scattering phenomena at walls and interior objects. RAZR therefore approximates these effects in the time domain through IIR filters (Buttler *et al.*, 2018) based on Schroeder's (1962) all-pass reverberators. These filters account for "local reverberation" (Siltanen, Lokki, Tervo, *et al.*, 2012) produced by scattered reflections at each of the six walls of the shoebox (surface scattering) as well as for multiple scattered reflections on interior objects for sound travelling through the room resulting in an temporal spread or smearing (object scattering).

For object scattering, RAZR associates with each side wall X of the room a cascade of four all-pass filters in series with an overall group delay of

$$\tau_X = R_X \frac{s d_X}{c},$$

where s is a user-configurable factor (0.05 by default), d_X is the dimension of the room along the axis normal to the wall, c is the speed of sound, and R_X is a random number from the uniform distribution on $[0.9, 1.1)$. The order of the kth IIR all-pass filter in the cascade ($k = 0, 1, 2, 3$) is

$$n_{X,k} = \left\lfloor \frac{\lfloor \tau_X f_S \rceil}{\pi^k} \right\rfloor,$$

where $\lfloor \rceil$ denotes rounding to the nearest integer and f_S is the sampling rate. Multiple options are available to compute the filter coefficients. As with the reflection filters, each of the cascades is applied to each image-source signal as often as the sound was reflected at the corresponding wall.

For a perceptually plausible simulation of diffuse reflections at the walls, the output of each image source is passed through a surface scattering pipeline. At the level of the ISM, this pipeline starts with a separation of the sound into a specular and a diffuse component, after which only the specular component is directly passed on to to the spatialiser, whereas the diffuse component is *only* forwarded to the FDN. The specular component is derived by processing the signal with a bi-quadratic IIR low-shelf filter, whereas the filter for the diffuse component has complementary (*i.e.*, high-pass) characteristics. The surface scattering pipeline for the diffuse outputs of the ISM continues in the feedback-delay network. This is the case for all image-source orders, as opposed to the specular outputs, of which only the maximal-order ones are passed on to the FDN.

5.2.2 The feedback-delay network (FDN)

By default, the FDN consists of 12 channels, with an overall architecture as described by Jot and Chaigne (1991). It receives input from the image-source model as described in section 5.2.3, maintaining the specular–diffuse split generated by the surface scattering mechanism in the ISM. These two

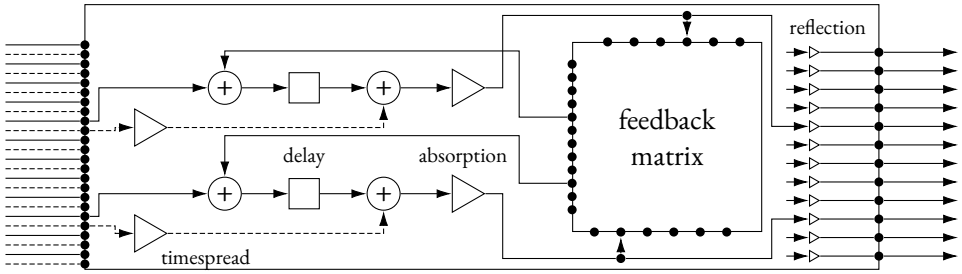

Figure 5.5 Block diagram of the feedback-delay network (FDN) implemented in RAZR. The inputs (left) are generated by the mapping from ISM to FDN, to be described in section 5.2.3; solid lines correspond to the specular and dashed lines to the diffuse components of the image-source outputs. Triangles represent IIR filters. Squares represent the delays.

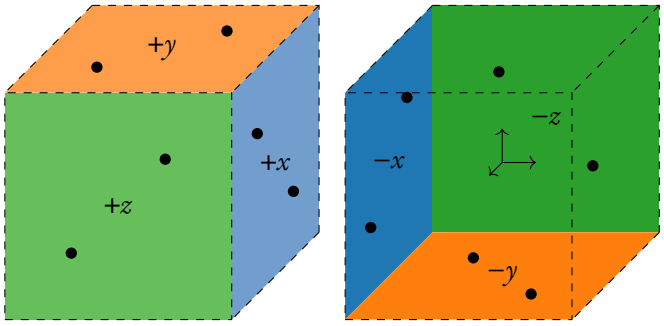

Figure 5.6 Positions assigned to the twelve default FDN channels in RAZR. **Left:** The 2 × 3 channels on the diagonals of "positive" walls $+x$, $+y$ and $+z$. **Right:** The 2 × 3 channels on the opposite diagonals of "negative" walls $-x$, $-y$ and $-z$. The cube is centered around the receiver, represented by the axes.

components per channel are mixed together in the FDN with appropriate timing. Figure 5.5 presents a block diagram of the current implementation.

RAZR assigns a spatial location to each channel such that on a room-aligned cube centered at the receiver's position, two channels each are mapped to every face of the cube; the two positions on each face lie on one of its diagonals; and the opposite diagonals are used on each pair of parallel faces of the cube. As such, there are four channels located on two y–z planes, two of which share the same low x coordinate (denoted here as the $-x$ channels), and the other two of which have the same high x coordinate (the $+x$ channels). Together, these make up the four x channels. Analogously, there are $(\pm)y$ and $(\pm)z$ channels on the other coordinate planes. See Figure 5.6 for an illustration.

■ **Delays.** Given a shoebox room with the dimensions (d_x, d_y, d_z) and the average edge length $D = \frac{1}{3}\left(d_x + d_y + d_z\right)$, the delay lengths $\tau_{X,i}$ of each of the four FDN channels $i \in \{1, 2, 3, 4\}$ associated with each dimension $X \in \{x, y, z\}$ are randomly chosen from a uniform distribution over an open interval:

$$\tau_{X,i} \in \left(\frac{d_X - 0.1D}{c}, \frac{d_X + 0.1D}{c}\right),$$

such that the expected value for every dimension is simply the time taken by sound to travel over the whole length of the corresponding edge, $\frac{1}{c}d_X$. The random jitter of up to ±10% of the average edge length serves to avoid the exact same delay lengths being assigned to multiple FDN channels.

The delays are applied to the appropriate specular inputs near the entrance of the FDN module, just after the corresponding result of the feedback matrix operation is added to each input sample.

■ **Feedback matrix.** In its preset configuration, RAZR randomly generates an orthogonal feedback matrix $A \in \mathrm{O}(12)$ by filling a precursor $A' \in [-1, 1)^{12\times12}$ with uniformly distributed random numbers and processing it with the Gram–Schmidt orthogonalisation algorithm. A is used as a linear transformation on the 12 input channels (after combining the specular and diffuse inputs, and after absorption filtering) to produce 12 feedback channels which are mixed back into the corresponding input channels. This leads to a stochastic distribution of energy that helps simulate diffuse reflections.

■ **Absorption and reflection filtering.** Absorption filters are required to model reverberation decay; without them, the total energy within the FDN module would increase unboundedly over time. Their transfer functions $H^{\mathrm{a}}_{X,i}(f)$ satisfy the equation

$$20\log_{10}\left|H^{\mathrm{a}}_{X,i}(f)\right| = -\frac{60\tau_{X,i}}{T_{60}(f)},$$

where $T_{60}(f)$ is the reverberation time estimated using Eyring's (1930) formula

$$T_{60}(f) = \frac{24 \ln 10}{c} \cdot \frac{V}{-S \cdot \ln\left(1 - \overline{\alpha}(f)\right)}$$

from the room volume $V = \prod_X d_X$, the total wall surface area $S = \sum_{X \in \{x,y,z\}} \prod_{Y \neq X} d_Y$, and the mean absorption coefficient[1] $\overline{\alpha}(f) = \frac{1}{6} \sum_{X \in \{x,y,z\}} \left(\alpha_{+X}(f) + \alpha_{-X}(f)\right)$.

Outside of the feedback loop, before the FDN-processed signals undergo spatial rendering, reflection filters are applied to each channel. The reflection filter for each FDN channel is identical to the ISM reflection filter computed for the corresponding wall.

■ **Surface scattering.** As described in the section on the ISM, image-source outputs may be split by a pair of filters into a specular part damping high frequencies, and a diffuse part damping low frequencies. These diffuse outputs (regardless of the order of the image-source they are associated with) enter the FDN without the delay that is applied to the specular components, but undergo timespread filtering using an all-pass IIR filter cascade. This cascade is very similar to the one described for object scattering in section 5.2.1, but its parameters are chosen slightly differently.

■ **Spatial mapping.** The output channels of the feedback-delay network are passed through the same spatialisation process as the image sources, using the relative virtual source positions illustrated in Figure 5.6.

5.2.3 The mapping from ISM to FDN

A crucial part of RAZR is the way in which the physically exact output of the image-source model is linked to the input of the feedback-delay network. Each image source of maximal order is not only output directly to the simulated room impulse response, but also enters one or more channels of the FDN, which is meant to approximate diffuse reflections for late reverberation (while bypassing the computational difficulties of generating image sources for high reflection orders). Moreover, if scattering is enabled in the image-source model, diffuse outputs of *all* orders are passed into the feedback-delay network as well. A schematic of this mapping process is given in Figure 5.7.

[1] Note that Eyring's original formulation weights the absorption coefficients by surface area.

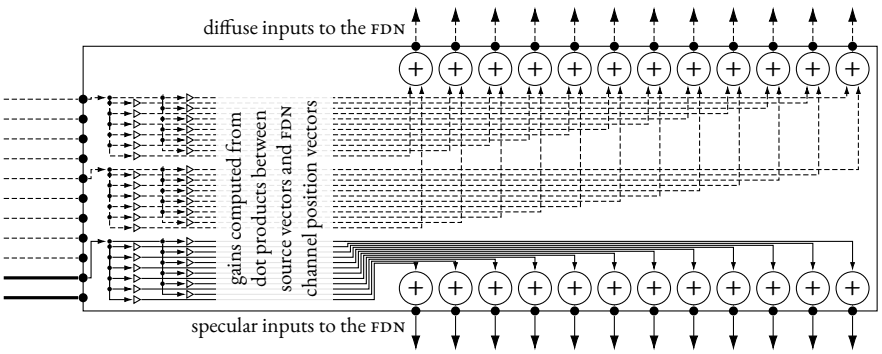

Figure 5.7 Block diagram of the mapping from ISM to FDN. Diffuse image-source outputs of all reflection orders (dashed lines) are processed separately from specular outputs of maximal reflection orders (solid lines). For both signal components, the output from each image source is mixed into each FDN input channel with an appropriate gain. This is illustrated here for two exemplary diffuse streams and one specular stream. The gain values depend on ISM and FDN geometry.

■ **Smart channel mapping.** In the original version of the simulation program (Wendt *et al.*, 2014), each maximal-order image source output was assigned quasi-randomly to exactly one feedback-delay network channel. Essentially, this treated the spatial origin of the simulated low-order reflections as irrelevant, while still assigning the whole energy to an arbitrary channel with a precisely defined location. The current version of RAZR uses a "smart" mapping instead: For each image source position s_i and FDN channel position c_j (where both vectors are relative to the receiver position), the negative dot product $g_{i,j} = -s_i \cdot c_j$ is calculated. The *specular* output of the i-th maximal-order image source is fed into the j-th channel with a normalised gain of

$$\hat{g}_{i,j} = \frac{g_{i,j}}{\sum_J \left(\max\left\{g_{i,J}, 0\right\}\right)^2}$$

if $g_{i,j} > 0$. This method splits up the energy emitted from each maximal-order image source across all FDN channels which are located in an opposing direction, thus roughly approximating the next reflection from the other side of the room.

If surface scattering is enabled in the ISM, the *diffuse* outputs of the image source *of all orders* are mapped using a similar procedure, but with $g_{i,j} = +s_i \cdot c_j$ (*i.e.*, with opposite sign). Each diffuse signal is thus rendered to emanate most strongly from FDN channels in the vicinity *of the image source that produced it* (so it should be heard at the same time as the image source, consequently bypassing the input delay in the FDN module), whereas for a specular signal, the selected channels are close to where the image source *of the next-higher order* would lie (so it is appropriately delayed).

5.3 liveRAZR: A real-time implementation of RAZR

As part of this dissertation, liveRAZR was created by the author as a re-implementation of the RAZR model in C++, with a focus on real-time geometry and signal processing. The source code is compliant with the C++17 standard and has been successfully built with the Visual C++ compiler on Windows 10 (both Microsoft, Redmond, WA) and with GCC (GNU Project) on Debian GNU/Linux 11. The core differences of this new code in comparison to the existing MATLAB implementation are described in this section. Figure 5.8 shows a high-level overview of the building blocks of liveRAZR. Programmers can use liveRAZR as a library, *i.e.*, run the room simulation and auralisation from their own code. For example, a program controlling a psychophysical experiment can make appropriate calls to liveRAZR functions to adjust stimulus properties according to prior results, subject motion, *etc*. For ease of use, liveRAZR is also provided as a standalone program, together with a server through which it accepts commands. Other programs, running on the same or even on a different computer, can control some selected simulation parameters through this interface (see section 5.3.8).

It must be noted that liveRAZR does not yet include most filter design functionality present in RAZR, as MATLAB is indeed the more suitable tool for this task, and the provision of all the prerequisites in liveRAZR was not considered a priority. Instead, liveRAZR currently comes with a script which runs the appropriate RAZR routines to calculate the necessary filter coefficients, delays, *etc.*, and generates a liveRAZR configuration file which includes all required precomputed parameters. In this sense, liveRAZR is not yet a fully-featured room-acoustical simulation framework, and should instead be seen as an optional real-time component of the RAZR system. This status, however, is due to change in the future, as the implementation of all currently MATLAB-only features in C++ is desired.

5.3.1 Buffer-by-buffer processing

RAZR builds up a finite impulse response with a duration of a few seconds at most. For simplicity, it holds the entire result buffer in memory throughout its runtime. As liveRAZR directly processes a given input signal which might, such as in the case of microphone input, be arbitrary long, this is not a viable strategy. This new real-time implementation instead processes in a loop short input buffers

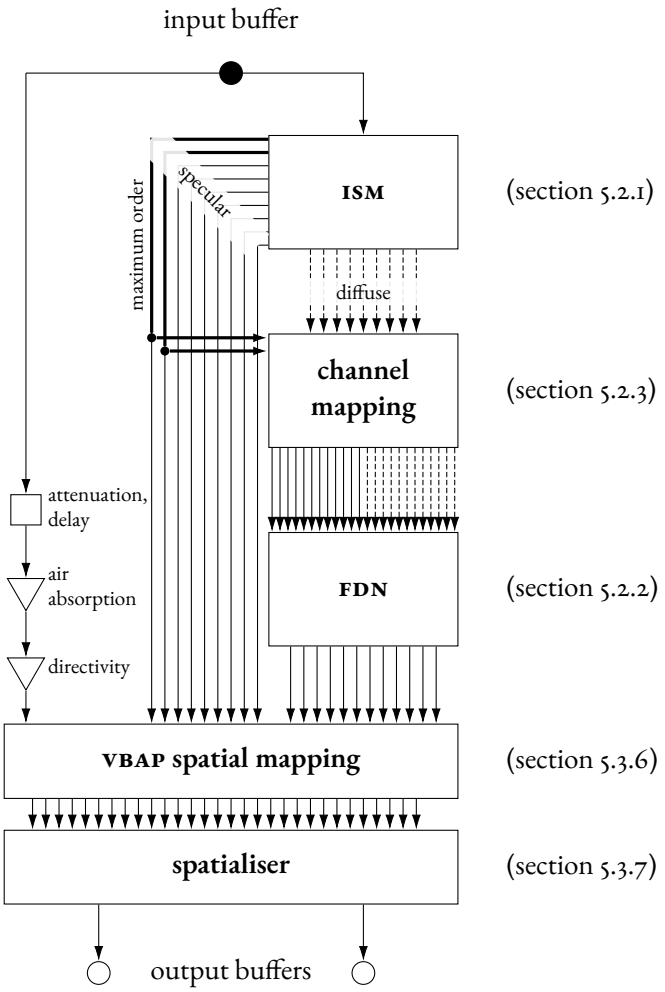

Figure 5.8 Synoptic block diagram illustrating the core of the main liveRAZR signal processing loop. ISM, ISM-to-FDN channel mapping, and FDN operate much like in the MATLAB implementation of RAZR, but directly process an input signal rather than turning a single impulse into an impulse response: The shown signal processing steps are repeatedly performed for small buffers of an input signal, generating equally-sized buffers of a multi-channel output signal. Finer-grained block diagrams for these modules can be found in the referenced sections. The spatial mapping and spatialiser modules are liveRAZR-specific designs. This diagram does not show the geometry-updating routine which runs parallelly in the background and affects the operating state of all signal processing modules.

with a fixed, small number of samples which, taken together in sequence, constitute the complete input signal (from a file or microphone). In each iteration of this main signal-processing loop, an output buffer (of equal length as the input buffer) is written and then immediately emitted to a file or sound card (see section 5.3.3).

All the required storage space for input, output and intermediate signal is preallocated at the start of liveRAZR, as this is essential to make a signal-processing operation with predictable runtime properties possible. These internal buffers are cyclically overwritten as *ring buffers*, such that audio samples in one processing module which are no longer required by any downstream module are automatically discarded and memory is thus efficiently reused.

5.3.2 Object-oriented architecture

liveRAZR uses a modular approach by encapsulating its functionality into appropriate classes which can either already be put together as needed, or facilitate the straightforward integration of projected additional features. Some examples for the benefits of this modularity are:

- Modular WAVE and ASIO interfaces allow input signals to be read either from files or from the input channels of a sound card, and the result of the simulation to be written either to a file or to the output channels of a sound card. This is described in some more detail in section 5.3.3.
- Multiple choices of spatialiser make it possible to create auralisations for loudspeaker playback or for binaural reproduction via headphones, see section 5.3.7.
- The direct-sound path, ISM, channel mapping, and FDN modules are encapsulated in a pipeline object which makes it easy to simulate the effects of the acoustics of a room on multiple sound sources at once, such as more than one human talker or virtual loudspeaker.
- Geometry code is decoupled from signal-processing code and can therefore loop at a rate different to the main signal-processing loop. A background geometry processing module maintains a queue of trajectory sampling points (consisting of source and receiver positions and orientations) which are to be reached in the future, taken either from a configuration file or from interactive submissions via a network-based interface (see section 5.3.8). As soon as each sampling point becomes available, this background loop calls geometry code for ISM position updates (currently limited to shoebox geometries, but an extension to more general geometries with axis-aligned quadrilateral faces is in development), for the channel mapping between the ISM and FDN, for the FDN itself, and for spatial rendering via the spatial mapper and spatialiser.

  The results for each of these modules is held in memory. For each output buffer, a momentary geometrical state is computed by linear interpolation between the states previously calculated for the two enclosing sampling points, and all modules are updated atomically (*i.e.*, such that the geometry in the ISM is always consistent with that in the ISM-to-FDN mapper, that in the spatial mapper, *etc.*).

5.3.3 WAVE and ASIO input/output interfaces

Via the command line interface of liveRAZR, reading sound source signals and writing the resulting auralisations is supported using the Waveform Audio File Format (WAVE) standard specified by IBM (Armonk, NY) and Microsoft (Redmond, WA). By means of the public-domain dr_wav library developed by David Reid, Australia, a wide range of digital sample representations are supported for the (single-channel) input signals, such as linear pulse code modulation (8, 12, 16, 24, and 32 bits per sample) and single-precision or double-precision IEEE floating point. Output signals can have an arbitrary number of channels (two for binaural output, or the number of loudspeaker channels for array auralisation); samples are written as single-precision floating point numbers.

On operating systems for which the Audio Stream Input/Output protocol (ASIO, Steinberg Media Technologies GmbH, Hamburg) is provided, liveRAZR can use it to interface with supported sound cards at a low latency. This modular feature is missing on platforms for which ASIO is not available, such as Linux-based operating systems. Where it is available, users can freely opt to use ASIO input channels in place of WAVE input files, ASIO output channels in place of WAVE output files, or both.

5.3.4 Efficient and numerically stable filters

RAZR uses IIR filters of relatively high orders to simulate the absorption effects at surfaces. For reasons of numerical stability, most of the IIR filters in liveRAZR are implemented as second-order sections (see section 1.3.2). The coefficients of the filters designed by RAZR can be transformed to those for biquadratic filter cascades by finding and grouping the poles and zeroes of the transfer function; this is done automatically by the MATLAB script which converts a RAZR setup to a liveRAZR configuration file. This is not ideal; future versions of RAZR will be able to design second-order sections directly.

5.3.5 Time-varying filters for smearing

As described in section 5.2.1, "Surface and object scattering", RAZR designs a cascade of all-pass filters with a fixed, high order and with fixed, sparse coefficients and uses it within the ISM to account for effects of reflections at object boundaries. This feature has been slightly refined in liveRAZR:

- Instead of one all-pass cascade per wall, liveRAZR uses one per image source. Rather than running the per-wall filters potentially multiple times as RAZR does, this filter is always applied once per image source.
- The group delay-related time constants τ_X are consequently replaced by one τ_i per image source i, where
$$\tau_i = s\frac{d_i}{c},$$
d_i is the distance between the receiver and the image source, c is the speed of sound, and s is a factor which can be set in the configuration file (0 by default, *i.e.*, no smearing).
- The filter orders and coefficients are calculated as explained previously, but they can change at every geometry update in the image source model, based on the current value of τ_i calculated at the beginning of each new audio buffer.

In the context of its planned extension to handle arbitrary room geometries in the image source model, RAZR will eventually adopt these modifications to the model, as there is no clear replacement for the calculation of per-wall τ_X values when the concept of axis dimensions ceases to be meaningful.

5.3.6 The VBAP spatial mapper

Vector-base amplitude panning (VBAP, Pulkki, 1997) is a method to find appropriate gain coefficients for loudspeakers in an array, such that an arbitrary virtual sound source position can be simulated even if no physical loudspeaker exists at that exact point. RAZR uses this technique in its array rendering mode, and in order to interpolate between sampling points in an HRTF database. For this purpose, a three-dimensional loudspeaker array or HRTF database is represented as a polyhedron with triangular faces (VBAP *triplets*), referred to as the VBAP *mesh* here. The corners of the mesh, each representing a single loudspeaker or an HRTF, are *vertices*. One such triplet of vertices is selected to render the sound of each virtual source.

liveRAZR operates similarly to RAZR in this regard, but extends its VBAP implementation to better deal with virtual sound source and receiver positions that may vary at runtime. After such a change

to the simulated scene, it must provide an updated output signal sufficiently quickly to fulfill the real-time constraints. For relatively fine-grained meshes as are typical for HRTF databases, the computational expense of VBAP becomes an issue: Naive implementation, such as the one included in RAZR, test every triplet in the mesh against every virtual sound source by means of a matrix–vector multiplication. When there are many virtual sources and/or points in the VBAP mesh, this can become prohibitively slow in an interactive setting.

The three-dimensional VBAP implementation in liveRAZR therefore includes a space-partitioning k-d tree (for $k = 3$; Bentley, 1975). Using this data structure, a subset of candidate VBAP triplets can be determined for every given virtual sound source, such that triplets in a very different region of space need not be tested. The k-d tree is built up once, at program startup: Initially, the tree has one leaf node containing the entire mesh. Then, a yz plane is chosen heuristically such that approximately half of the mesh is on one side of the plane, and the other half is on the other side. The two sub-meshes are stored as leaf nodes of the tree, whereas the x coordinate resulting in the division is stored in a new non-leaf root node. This process is then repeated recursively for each of the two leaves, next along an appropriate xz plane, then an xy plane, then a yz plane again, *etc.*, until the tree contains six layers of non-leaf nodes.

As opposed to a standard k-d tree, the purpose-built implementation in liveRAZR does not store points, but VBAP triplets which together make up a triangulation on a sphere. A triplet is included in a leaf node if the smaller spherical cap obtained by intersecting the unit sphere with the plane described by the three points intersects the space of the node. The leaf node obtained by traversing the spatial partition for a virtual source position is thus guaranteed to contain all triplets for which VBAP will succeed.

Two-dimensional VBAP is also implemented as a special case for horizontal loudspeaker arrays. This geometrically straightforward case does not require an acceleration data structure.

5.3.7 Spatial rendering

The application of VBAP described in the previous section results in an assignment of n gain coefficients to each virtual source (direct sound, image source or FDN channel), where n is the number of vertices in the HRTF or loudspeaker array mesh.[2] It is the role of a *spatialiser* to render the individual signals accordingly. Currently, there are three spatialiser implementations.

■ **Simple spatialiser.** The simple approach is applicable for both array and binaural rendering if FIR filtering is not required. In the loudspeaker array mode, it simply applies the VBAP gains to each source by multiplying the samples from the virtual sources with the coefficients, and adding up all contributing virtual sources in each of the n output channels. Optionally, gains and delays can be specified for each channel, which allows for the compensation of geometrical differences between the loudspeaker array and a sphere (to delay and attenuate the signals from loudspeakers which are closer to the listener position than others).

When generating a binaural auralisation, this procedure is performed twice, separately for the left and for the right ear. This produces $2n$ intermediate output channels, one for each vertex, which are merged into just 2 at the output of the spatialiser. Again, the configuration of optional gains and delays is possible, and these can be set separately for the left and right sides; this mechanism can be used to generate interaural time and level differences.

If the loudspeakers in an array require the application of a compensation impulse response, or if HRTFs must be applied in binaural reproduction, one of the following two spatialisers is used instead.

[2] Most of these gain coefficients will be zero, as VBAP will identify 3 points in the mesh as a valid triplet, or 2 as a valid pair in the two-dimensional case.

■ **Array spatialiser.** This implementation is used when rendering to a loudspeaker array, and a compensation impulse response should be applied at every output channel, *e.g.* to correct for different loudspeaker frequency responses.

At the core of the array spatialiser, there is a multi-threaded, lock-free processor for fast FIR filtering. To initialise this frequency-domain convolver, for each vertex,

- the associated finite impulse response is partitioned, with part sizes b, b, $2b$, $2b$, $4b$, $4b$, ..., where b is the size of the audio buffer size, with the last part right-padded with zeroes;
- each of the parts is doubled in size (by padding with zeroes at the end) and transformed with an FFT for real-valued signals; and
- a processing plan is created out of a limited range of primitives (FFT, inverse FFT, vector addition, elementwise vector multiplication) as a list of tasks, where tasks can depend on each other.

The processing plan implements the scheme by W. G. Gardner (1995). Characteristically for the scheme, each processed audio buffer is assigned to a cycle; in each cycle, only a certain set of impulse response parts are actually convolved, which ensures a relatively even computational load throughout the runtime of the convolver. The first two parts, each of length b, undergo fast convolution in every cycle; they correspond to the earliest part of the impulse response, and so this part of the convolution is needed immediately. A part of length kb for integer $k > 1$, however, is only processed in every kth cycle. A cycle which does not handle a part of length kb will instead append its input signal to the buffer of the next cycle which does.

Within this scheme, the fast convolver implemented in liveRAZR also takes care to minimise the number of fast Fourier transforms that have to be calculated. This especially concerns the handling of aliasing in the outputs of the partial convolutions: Some internal buffers correspond to identical points in the output stream, *i.e.*, they overlap completely. For example, the convolution of the first b samples of the input signal with the *second* part of the impulse response in one cycle is always perfectly aligned with the convolution of the next b samples of the input signal with the *first* part of the impulse response in the following cycle. The array spatialiser recognises these aliased chunks of signal and sums up their spectra before running an inverse FFT (instead of running multiple inverse FFTs and summing the time-domain results).

Figure 5.9 illustrates the entire procedure schematically.

■ **Binaural ("merging") spatialiser.** The fundamental principle behind the merging spatialiser is the same as for the array spatialiser, with a crucial difference: Whereas the latter performs one fast convolution for each vertex, the former runs two (one for the left-ear, one for the right-ear output channel) for every virtual sound source (direct sound, image source or FDN channel). The frequency-domain results of all the partial convolutions for the left-ear and the right-ear channels are added up ("merged") *before* the inverse fast Fourier transforms.

This different mode of operation implies that the finite impulse responses (HRTFs) are variable at the runtime of the convolver, as every virtual source may change its position relative to the listener during the auralisation. The VBAP gain coefficients are therefore used to explicitly interpolate between HRTFs, as opposed to the array spatialiser, which uses them to calculate the contributions of each virtual source to each output channel (with a constant impulse response). The HRTFs for each virtual source are currently computed by straightforward linear interpolation, as this can be done efficiently in the frequency domain. This simple procedure may lead to audible artifacts when one interpolated HRTF is replaced with another, especially due to the ambiguity of periodic quantity of phase. An appropriate implementation of phase unwrapping (*e.g.* Kaplan and Ulrych, 2007; Karam and Oppenheim, 2007; McGowan and Kuc, 1982; Al-Nashi, 1989; Steiglitz and Dickinson, 1982; Tribolet, 1977) will alleviate this issue in future versions of liveRAZR. If the spatial updates are sufficiently smooth to avoid large jumps within the HRTF grid, however, the problem is negligible.

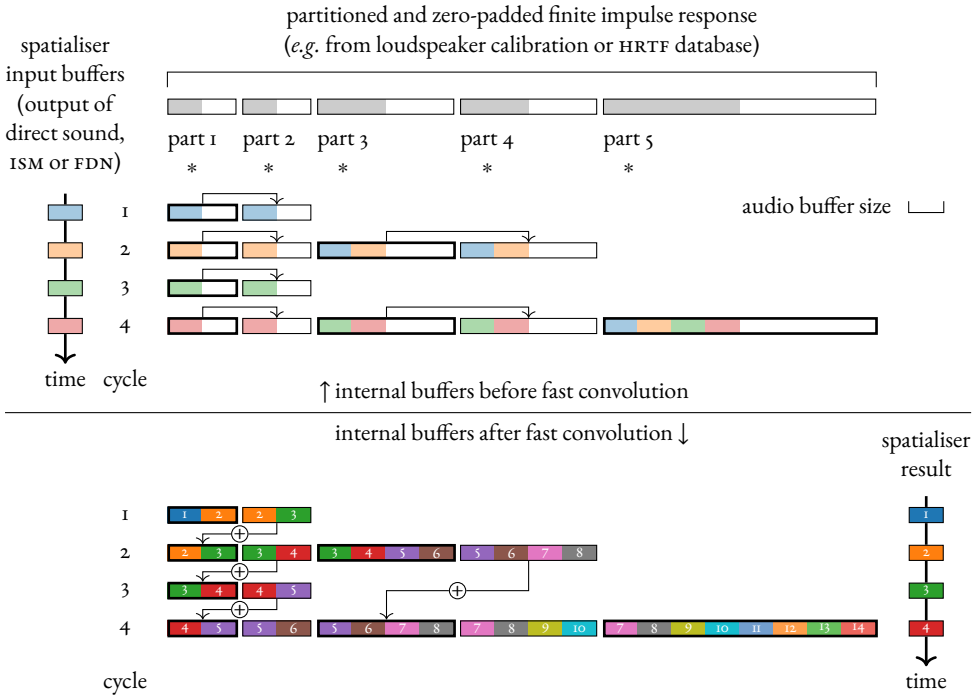

Figure 5.9 Chart of the fast convolution process based on W. G. Gardner (1995), as implemented in the array and merging spatialisers in liveRAZR. **Above the horizontal line:** In each cycle, the current block of output signal (timeline on the left) is copied to the internal buffers (middle) in the current cycle (for the first two parts of the partitioned impulse response) and, if the current cycle does not contain all parts, to the appropriate locations in future cycles as well (see matching colours). The bold-framed internal buffers need to be processed with FFTs; the spectra for two internal buffers with the same size within each cycle are identical and can be copied after the transformation (arrows). White areas in the internal buffers indicate zero padding. **Below the horizontal line:** After frequency-domain elementwise multiplication of each internal buffer with the corresponding part of the FIR (very top of figure) and inverse FFT, all internal buffers (middle) contain chunks of signal which correspond to more than one block of output signal (timeline on the right). The colours and white numbers indicate the target cycles: All parts of the internal buffers with the same target cycle are added up. As an optimisation, the addition is done before the inverse FFT for internal buffers with the same pattern of target cycles (arrows), such that only the bold-framed buffers have to be transformed.

5.3.8 Interactive control

Beside the ability to process signals in real time, the core motivation to create an implementation of the RAZR model in C++ was to allow sound sources and/or receivers to move dynamically during the simulation. This can be achieved by using the application programming interface of the liveRAZR library, *i.e.*, by integrating it into another purpose-built program which provides appropriate geometrical data. Unfortunately, this type of integration is not always easy to achieve: It requires the user to possess programming skills in a compatible language; it can be difficult to combine liveRAZR with certain large software frameworks such as game engines; and it requires extra care to keep the simulation loop running in real time.

liveRAZR thus provides an interface for interactive control even in its standalone mode. This facility is based on the networked Open Sound Control (OSC) protocol. The liveRAZR executable accepts messages which allow external software

- to turn each sound source on or off;

- to load WAVE files from the file system into memory;
- to switch the signal emitted by each sound source with such a signal previously loaded from the file system, or with an input channel from an ASIO device;
- to change the position and orientation of each sound source; and
- to change the position and orientation of the receiver.

The latter two types of message can be furnished with a target timestamp. Until this timestamp is reached, liveRAZR will calculate intermediate positions by linearly interpolating between the two most recently received position updates, and by spherically-linearly interpolating ("slerping"; Shoemake, 1985) between the corresponding orientation updates.

5.3.9 Fractional-delay filtering

Because the receiver and any sound sources in the scene may move, the sound propagation delays implied by the source-to-receiver distance can obviously change as well during an ongoing simulation. This poses a difficulty because liveRAZR has to work on a digital signal which is necessarily sampled, *i.e.*, its value is only available at certain discrete points in time. The variable delay values do not adhere to these timepoints, and it is problematic to snap them to the nearest available sample, as this would give rise to potential discontinuities: An image-source processor would eventually skip a sample when the receiver moves towards towards the source position, or it might jump backward across a sample binary if the receiver retreats. A fractional-delay filter is required to effectively interpolate between samples. Conveniently, it also simulates the Doppler effect (Strauss, 1998, cited by Välimäki and Laakso, 2001).

The unpredictability of changes in delay can be most easily supported with a random-access FIR filter. In liveRAZR, this uses a windowed-sinc design as described by Cain *et al.* (1995). To keep the memory access effort within reasonable bounds, the filter was limited to 5 coefficients, taken from the nearest matching row in a look-up table for 1024 fractional values between 0 and 0.999023. liveRAZR employs such a structure both in the image-source model and in the spatialiser.

5.4 Verification of liveRAZR

To test the basic functionality of liveRAZR, its output for an impulse input signal in a static scene with a single sound source was compared to the equivalent impulse response generated by RAZR using the same parameters. Two such tests were conducted, one with the "laboratory" room and one with the "aula" room which are included as examples in RAZR. See Table 5.1 for reference.

For simplicity, signal processing stages which are known to differ between RAZR and liveRAZR were disabled in both programs in order to allow a sample-by-sample comparison of the resulting impulse responses. Beside some details of smearing and source directivity filtering, this concerns any rendering to headphones or loudspeaker arrays. Due to substantial differences in the way that RAZR and liveRAZR implement their respective spatialisation functionalities, a comparison of these processing modules is a major task which I have left to a future undertaking.

To identify remaining differences within the two major building blocks of the RAZR model, the comparison was based on a separate comparison of three output signals: **ISM,** the sum of the impulse responses from all image sources; **FDN,** the sum of the twelve output channels from the feedback-delay network; and **entire IR,** the sum of ISM, FDN, and the direct sound. The image-source and direct-sound IRs contain spatial information encoded in their geometric attenuations and delays, but due to the disabled spatialisers, each of these three signals comprises only one channel and consequently does not include any binaural information (*e.g.*, no ITDs, no ILDs and no HRTFs).

The resulting impulse responses from both programs are plotted in Figure 5.10. In fact, only the RAZR outputs are shown, because the impulse responses generated by liveRAZR are visually indistinguishable. A dashed red curve in each panel shows the relative error $e(i)$ calculated with a sliding

	Laboratory	Aula
Dimensions	4.97 m × 4.12 m × 3 m	12 m × 30 m × 10 m
Absorpt. coeff., $-z$ wall	0.06, 0.15, 0.40, 0.60, 0.60	0.05, 0.10, 0.13, 0.16, 0.22
$-y$ wall	0.10, 0.05, 0.04, 0.07, 0.10	
$-x$ wall	0.30, 0.10, 0.10, 0.10, 0.10	
$+x$ wall	0.20, 0.20, 0.10. 0.07, 0.04	
$+y$ wall	0.70, 0.60, 0.70, 0.70, 0.50	
$+z$ wall	0.01, 0.02, 0.02, 0.02, 0.03	
Receiver position	(2.77 m, 1.30 m, 1.51 m)	(6.70 m, 25.30 m, 1.20 m)
Receiver orientation	(90°, 0°)	(95°, 0°)
Source position	(1.40 m, 3.02 m, 1.36 m)	(6.30 m, 28.00 m, 1.20 m)

Table 5.1 Specifications of the rooms which were used for the verification of liveRAZR based on a comparison to RAZR. Absorption coefficients ("Absorpt. coeff.") are given in octave bands centred at 0.25 kHz, 0.5 kHz, 1 kHz, 2 kHz and 4 kHz; in the aula, they are identical for all walls. Orientations are given in azimuth and elevation, where (0°, 0°) points towards the $+x$ wall.

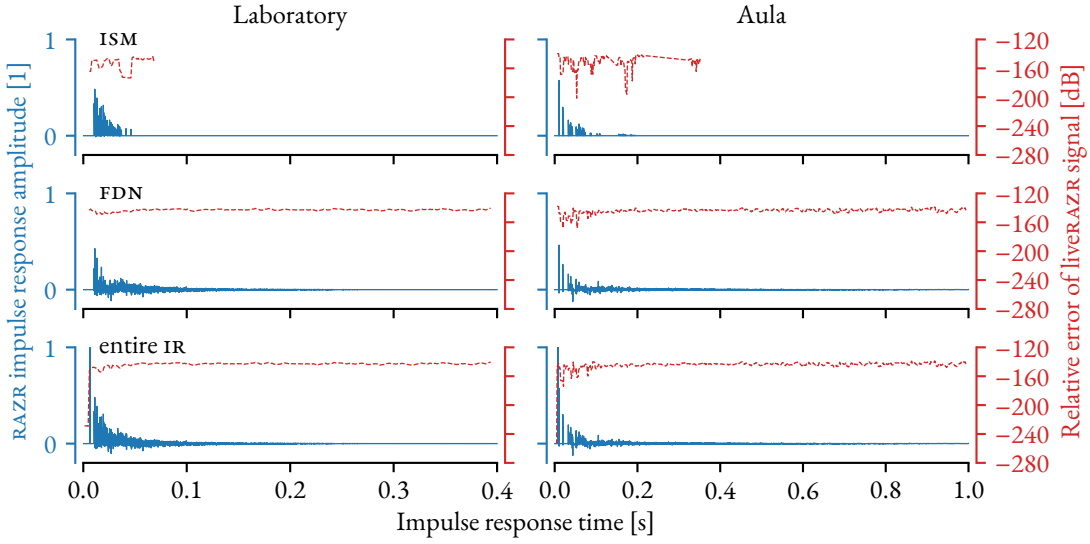

Figure 5.10 Comparison between the impulse responses generated by RAZR and liveRAZR for two identical configurations. The impulse response calculated by the original MATLAB version of RAZR is drawn in blue. The dashed red line represents the error of the equivalent liveRAZR output in decibels. **Left and right columns:** Results for the laboratory and aula room configurations, respectively. **Top row:** Partial IR containing only the 62 image-source signals. **Middle row:** Partial IR containing only the 12 feedback-delay network channels. **Bottom row:** Sum of all image sources, FDN channels, and the direct sound, a total of 75 contributing virtual sources.

rectangular window (5.8 ms in width), akin to a "signal-to-noise" ratio where the liveRAZR output is the "signal" $s(i)$ and the RAZR output is the reference "noise" $r(i)$,

$$e(i) = 10 \log_{10} \frac{\sum_{k \in \{0,1,\ldots,255\}} (s(i+k) - r(i+k))^2}{\sum_{k \in \{0,1,\ldots,255\}} (r(i+k))^2}$$

This error ratio never exceeded −135 dB relative to the RAZR signal, which indicates merely minor numerical errors. Indeed, no audible differences could be detected in an informal listening test based on the impulse responses.

5.5 Runtime analysis of liveRAZR

As a real-time system, the runtime of the signal processing pipeline in liveRAZR will clearly be a concern for most future applications. I acquired such performance data on a computer with an Intel CORE i7-9850H central processing unit (six physical cores, 2.60 GHz clock speed) on a pre-release Debian 11 "Bullseye" system with a Linux 5.7.10 kernel (released on 22 July 2020). The liveRAZR source code was translated to machine code with the GNU C++ compiler, from version 10.2 of the GNU Compiler Collection, using -O3 and link-time optimisation. None but essential system processes were running besides liveRAZR at the time of the measurements. The input and output signals were held in pre-allocated memory; the measured runtime values do not include any reading and writing from a storage medium, or any communication with audio hardware.

The input signal was 10 s long at 44.1 kHz, and processing was stopped immediately after 441 000 samples of input signal were run through the pipeline. The internal audio buffer size was set to 64 samples, corresponding to an added input–output latency of 1.45 ms if the process were running as an on-line audio processor.

A variety of configuration files was generated to study the effects of several parameters, namely

- the maximum image source order, set to either 1, 2 or 3;
- whether or not scattering was applied;
- whether or not fractional delays were applied in the image-source model;
- whether or not the image sources were modelled as directional;
- whether or not the feedback-delay network was running;
- whether or not a spatialiser was running, and if it was,
  - whether it was the array or the binaural spatialiser;
  - whether the spatial mapper and spatialiser were running with a mesh of 103 (in 202 triangles; "sparse") or 187 vertices (in 370 triangles; "dense"); and
  - whether or not each output channel (in the case of the array spatialiser) or each virtual source (in the case of the binaural spatialiser) was convolved with a 444-tap finite impulse response.

All meaningful combinations of these parameter values were tested, resulting in 1632 different liveRAZR configurations. The runtime measurements were taken five times, for a total of 8160 program runs and a total time of approximately 4 h and 5 min (including measurement instrumentation and program startup overheads).

Figure 5.11 shows a summary of the measurement results, separating only the effects of rendering target (array *vs.* binaural; columns), overall number of virtual sound sources (horizontal axes), and VBAP mesh density/FIR filtering in the spatialiser (colours). These correspond to configuration parameters which exhibited particularly relevant effects on runtime. Within each group, runtime variations are caused by the scattering, fractional delay, and source directivity settings, as well as random differences between the five runs. The mean runtime, averaged across these variables, is indicated with a horizontal bar, and the within-group distributions of runtimes are displayed as kernel density plots.

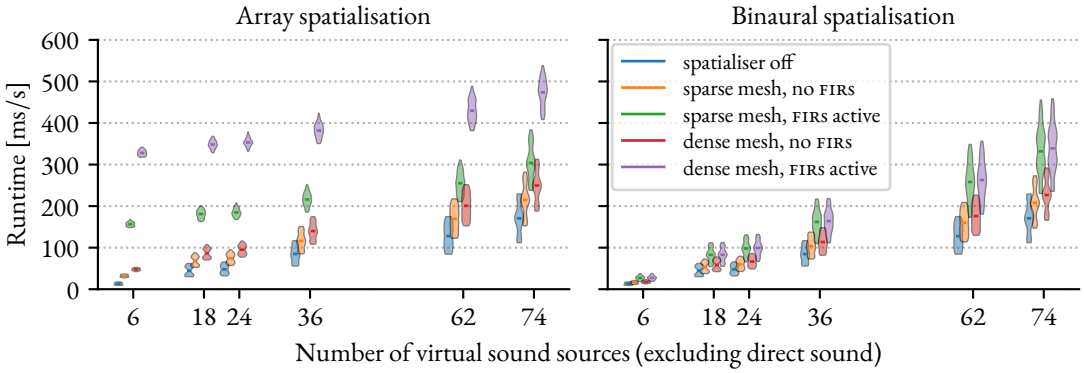

Figure 5.11 Runtimes of liveRAZR relative to 1 s of input signal as a function of the number of virtual sound sources, for various spatialiser configurations. The data are summarised using kernel density estimates, with the mean highlighted with a short horizontal bar. **Left:** Results for the array spatialiser. **Right:** Results for the binaural (merging) spatialiser. **Blue:** Values when spatialisation is disabled entirely; included as a reference in both columns. **Orange and red:** Values for a spatialiser without convolution, on a sparse and dense VBAP mesh, respectively. **Green and purple:** Values for a spatialiser with (*e.g.*, HRTF) convolution, on a sparse and dense VBAP mesh, respectively.

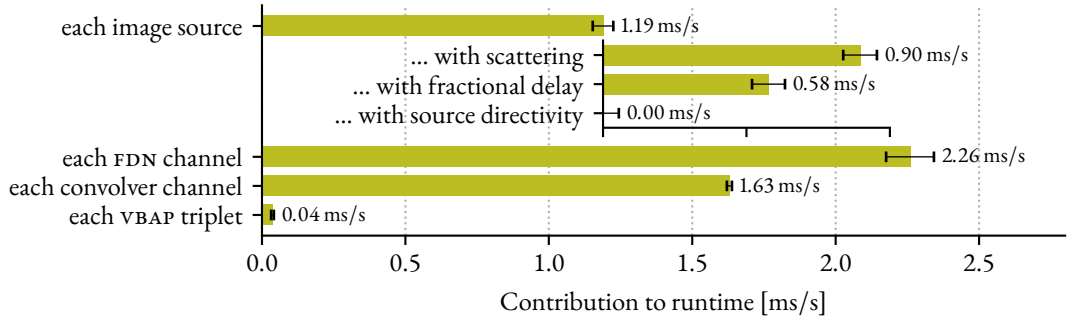

Figure 5.12 Results of a linear model fitted to the runtime data presented in Figure 5.11. Each bar shows the estimated runtime contribution of a liveRAZR processing stage in milliseconds of added runtime per second of input signal. The error bars show standard errors.

5.5.1 Runtime model

With the aim of providing a more comprehensive interpretation of the parameters which influence the performance of liveRAZR, a linear model was fitted to the measurement data. The seven independent variables were chosen as

- the number of image sources,
- the number of image sources processed with scattering filters,
- the number of image sources processed with fractional delay filters,
- the number of image sources processed with source directivity filters,
- the number of FDN channels,[3]
- the number of simultaneous convolution channels in the spatialiser, and
- the number of triplets in the spatialiser VBAP mesh.

[3]Even though the number of FDN channels is a constant in liveRAZR (the feedback-delay network can either be turned on or off, not changed in size), this binary independent variable was chosen to take the values 0 and 12 rather than 0 and 1, such that the magnitudes of the resulting regression coefficients for image sources and FDN channels would be comparable.

To separate the effects of upstream on downstream processing steps, a number of interaction terms was also included in the model, namely (a) between the number of image sources and each of the numbers of image sources processed with further filters, (b) between the number of image sources and the number of FDN channels, and (c) between the number of total virtual sources and the spatialiser mesh parameters. This yielded a model with $R^2 = 0.95$.

The regression coefficients are plotted in Figure 5.12. Due to the intact scales of the independent and dependent variables, they may be read as the extra runtime caused when each of the listed parameters is increased by 1. Note that the interactions were omitted from the figure for simplicity.

5.6 An application of liveRAZR: Looming in rooms

As introduced in section 1.2.4, changes in sound level at a human listener's ear due to the motion of a sound source are subject to a bias: Listeners are more sensitive to increasing than to decreasing levels (Neuhoff, 1998). Past investigations of this effect have been confined to anechoic space, where the absence of reflections makes the localisation of sound sources in depth particularly difficult (see Kolarik, B. C. Moore, *et al.*, 2016). This was also the case in the experiment from Chapter 4, which investigated the reactions of human subjects to the simulated warning sounds of approaching rattlesnakes. But while a hypothetical encounter with an animal lends itself well to an experiment that disregards the acoustics of enclosed spaces, human hearing frequently takes place in rooms, and the lack of experiments which focus on this aspect appears to me as an oversight. The few publications which address this issue at all either only consider ground reflections (Bach, Neuhoff, *et al.*, 2009; Neuhoff *et al.*, 2009) or are limited to variations of looming sounds, without stationary or receding conditions for comparison (Wilkie and Stockman, 2020).

The pilot experiment described in this section was therefore devised to study the detection of differences in the motion profiles of frontally approaching *vs.* receding sound sources outside (where only relative comparisons between stimuli are possible) *vs.* inside of rooms (where absolute cues for sound source distance are available; see section 1.2.3). The question under investigation was how human listeners' abilities to discriminate the distances covered by two moving sound sources are affected by the availability of room cues in a task which can be assumed to be subject to the looming effect. One could reasonably expect that an enclosed space makes the task either easier (by providing room-related cues, specifically DRRs) or more difficult (by compressing loudness differences due to the presence of reflections).

Because of the requirement of presenting moving sound sources in room-acoustical conditions, this experiment was additionally suited as a "real-life test" of liveRAZR.

5.6.1 Methods

■ **Listeners, task and conditions.** In each trial of a 2-alternative, 2-interval forced-choice (2-AFC) experiment, 6 normal-hearing listeners (23–33 years of age, 4 female) were asked to identify the interval in which the movement of a solitary virtual sound source in front of them, auralised via a horizontal loudspeaker array, covered a larger distance. In every pair of intervals, the sound source was in uniform horizontal motion of the same direction, either toward or away from the listener. It emitted a click sound every 0.1 s for a total duration of 2 s per interval and was otherwise silent. The midpoint of each trajectory was 4 m away from the listener, and the sum of the motion distances of the two intervals in each trial was always 8 m. The two intervals in each trial were presented either both anechoically, or both echoically with the acoustics of a liveRAZR-simulated room.

The difference between the motion distances of the first and second interval in each trial varied by $d \in \{0.2\,\text{m}, 0.8\,\text{m}, 1.4\,\text{m}, 2.0\,\text{m}, 2.6\,\text{m}, 3.2\,\text{m}, 3.8\,\text{m}\}$, such that the sound source in the correct interval travelled $D^+ = 4\,\text{m} + \frac{1}{2}d$, *vs.* $D^- = 4\,\text{m} - \frac{1}{2}d$ in the incorrect interval. Hence, at the two extreme points of the range, a listener had to auditorily discriminate between two virtual motion

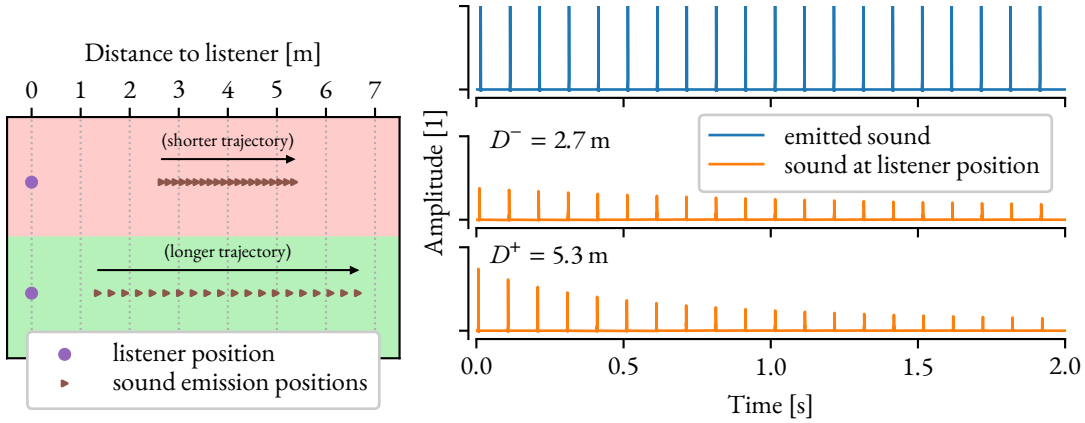

Figure 5.13 Stimulation in the "looming in rooms" 2-AFC experiment. **Left:** Schematic of one trial in the anechoic–receding condition (distance difference $d = 2.6$ m), with the listener location represented by a purple circle, and the simulated locations where the virtual moving sound source emitted clicks represented by brown triangles. The listener had to indicate in which interval the virtual source covered a larger distance (green background) than the other interval (red background). **Right:** Waveforms of the click train emitted by the virtual sound source in each 2 s interval (top frame, blue) and of the stimuli generated thereof by liveRAZR, including distance-based attenuation, as played back through a loudspeaker in front of the listener (other two frames, orange, aligned with the intervals in the trial schematic).

distances of 3.9 m *vs.* 4.1 m (most difficult) and between 2.1 m *vs.* 5.9 m (easiest). The left column of Figure 5.13 illustrates the trajectories of an exemplary trial in the receding condition.

80 repetitions of 2-AFC responses were acquired for each listener and each of the 7 distance differences, 2 source travel directions, and 2 acoustical conditions. Listeners were given an acoustic feedback after each response to let them know if their decision was correct. The order in which the trials were presented was fully randomised. Each listener completed the experiment in 8 sessions of about 45 min each.

■ **Stimulus and reproduction setup.** In every interval, the virtual sound source emitted twenty 5.8 ms clicks with a f^{-1} (pink) power spectrum. This output signal remained constant within each trial, *i.e.*, differences in level and in any other sound property at the listener's ear were only due to the simulated movements of the virtual sound source, not because this source itself changed its emission (see the right column of Figure 5.13). A transient signal was chosen in order to allow the reverberant tails to be clearly audible in the echoic trials. This is in contrast to typical stimuli from other experiments on the auditory looming effect, which frequently use signals from stationary processes, such as ramped sinusoids, white noise, or steady-state vowels (beginning with Neuhoff, 1998). Thanks to fMRI studies (Bach, Schachinger, *et al.*, 2008; Seifritz *et al.*, 2002) where pulsed stimuli are convenient, however, amplitude-modulated sounds are known to elicit the same bias. The pink spectral colouration was intended to provide for a less artificial-sounding stimulation, as entirely flat spectra are rare in natural environments (see *e.g.* Ewert, 2020).

liveRAZR was employed to synthesise the acoustic scene of each interval at a sampling rate of 44.1 kHz. It was configured to generate a 36-channel output signal via the array spatialiser for the virtual source, either without reflections (ISM and FDN turned off) or, in the echoic condition, with the simulated wall reflections from a shoebox room with the dimensions of 8 m × 12 m × 3.5 m (broadband $RT_{60} = 430$ ms). The listener was positioned in this room with one of the longer walls 2 m to their right, one of the shorter walls 1 m behind them, the head at a height of 1.6 m above the floor, and the virtual source directly in front. The direct sound was thus always mapped to just one loud-

speaker at 0° azimuth, whereas each image source and each FDN channel was panned between a pair of neighbouring loudspeakers using 2-dimensional VBAP.

The stimuli were presented to the listeners in an anechoic chamber with a 2 m × 2 m base and a height of 2.2 m via 36 loudspeakers (Plus XS.2, CANTON Elektronik, Weilrod, Germany) corresponding to the liveRAZR output channels. The loudspeakers were mounted on the wall of the chamber around the listener's head, with an azimuthal difference of 10° between them, such that they evenly covered the horizontal plane. The ASIO interface of liveRAZR sent the synthesised signals to two audio interfaces with 24 channels each (24I/O, MOTU, Cambridge MA, USA) that were connected to the loudspeakers via four 12-channel power amplifiers (CI 9120, NAD Electronics International, Pickering ON, Canada).

■ **Data analysis.** For each listener, four psychometric curves were fitted to the numbers of correct and incorrect responses calculated from the binary 2-AFC data, with distance difference as the independent variable: One curve each for the anechoic–approaching, echoic–approaching, anechoic–receding and echoic–receding conditions. The fits were acquired using a maximum-likelihood estimation method with the Python version of the *psignifit* toolbox (Wichmann and Hill, 2001a). Threshold values for the just-noticeable distance difference were extracted from these curves, with confidence intervals extracted by bootstrapping (Wichmann and Hill, 2001b).

A separate statistical analysis was conducted for an overview across listeners. This was done due to the low number of listeners in this pilot study and the high uncertainty of the psychometric outcome measures due to low performance of some listeners in some conditions (see below). As a preferable alternative to fitting psychometric curves to pooled data, which would not control for inter-individual differences, the analysis was instead based on a logistic mixed model on response correctness as a binomially-distributed response variable. The independent variables were distance difference as an interval-scaled covariate; acoustic condition (anechoic/echoic), movement direction (approaching/receding) and global gain (0 dB/+12 dB) as categorical covariates; and listener identity as a random factor (allowing for between-listener variance in intercept as well as distance-difference slope). Interactions between room condition and movement direction were also considered.

As there is no mathematically straightforward way to describe the distribution of the variances (and thus to obtain p-values) of coefficients in mixed logistic regression models when sample sizes are small, the following section will instead report fitted logit coefficient values x and their standard errors σ (in the format $x \pm \sigma$), along with $z = \frac{x}{\sigma}$. For orientation, if z were normally distributed, then two-sided $p < 0.05$ would be reached for $|z| > 1.96$, and $p < 0.01$ for $|z| > 2.58$.

5.6.2 Results

Panels #1 to #6 of Figure 5.14 show the ratios of correct responses for all listeners and conditions, and the four psychometric curves fitted to each listener's data. Just-noticeable differences in distance difference, defined as the value of the independent parameter where the psychometric curve exceeded a score of 67% correct responses, are also reported. This unusually low threshold for a 2-AFC experiment was chosen to ensure that the value lies within the observed range for almost all listeners and conditions. At 67%, only listener #6 fell short of this criterion in the anechoic–receding condition.

Thresholds were lower for approaching sound sources in all listeners. The logistic model analysis confirmed this difference, with a large slope of 0.70 ± 0.05 ($z = 12.64$) for the binary receding–approaching covariate. The overall difference between the echoic and anechoic conditions was clearly insignificant with a coefficient value of 0.07 ± 0.06 ($z = 1.20$). The value of 0.18 ± 0.08 ($z = 2.34$), however, points at a likely interaction between acoustic condition and movement direction. Indeed, estimated marginal means (bottom right in Figure 5.14) suggest that room acoustics provided a benefit especially in the receding trials. The estimated slope associated with distance difference was doubtlessly significantly different from 0 at 0.46 ± 0.03 ($z = 18.08$).

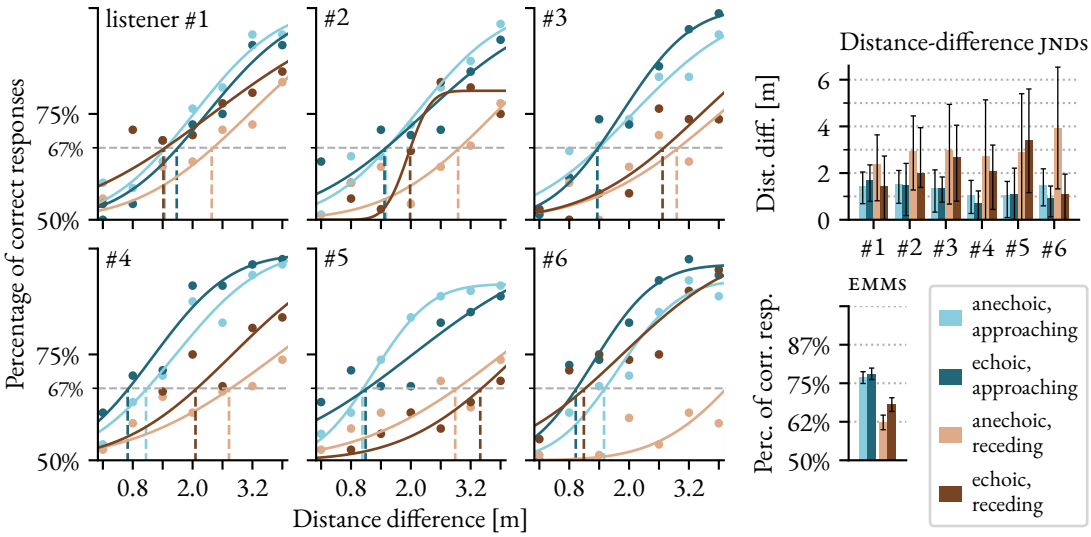

Figure 5.14 Results from the "looming in rooms" experiment. *#1–#6:* Per-listener performance data (as a percentage of trials answered correctly) and psychometric curves, separately for each of the four combinations of two acoustic conditions × two sound-source movement direction. **Distance-difference JNDs:** Just-noticeable differences in distance difference for a threshold of 67% correct responses, corresponding to the dashed vertical lines in panels #1–#6. The error bars show 95% confidence intervals based on a bootstrap analysis. **EMMS:** Estimated marginal means of performance scores for the four combined conditions across distance differences, corrected for variations in intercept and distance difference-related slopes between listeners. The error bars show asymptotic 95% confidence intervals.

The logistic model fit the data with $R^2 = 0.76$.

5.6.3 Discussion of the experiment

The finding that the subjects performed worse for receding than for approaching sounds is consistent with the existing literature on the looming bias, and in particular with Neuhoff (2016). This study reported that listeners' perceptions of the speeds of approaching virtual sound sources were more precise than when the virtual sources moved away from them, in the sense that their numerical speed estimates for the three true speeds of 15 m/s, 20 m/s and 25 m/s were not significantly different from each other. In other words, a speed ratio greater than than 167 : 100 would have been required for successful discrimination. Because the stimulus duration was held constant in the present experiment while the travel distance varied, comparisons of perceived speed would be a viable strategy to solve the task. Ratios of up to 281:100 were presented here. 167:100 would correspond to a distance difference of 2 m, which was indeed below the individual discrimination threshold for receding sources for most listeners. Similarly, a speed ratio of 125 : 100 was sufficient for approaching sources in Neuhoff (2016). This would be a distance difference of 0.89 m, which was slightly below the corresponding thresholds determined here. It does appear likely that the task at hand would be somewhat more difficult, as in a comparison of the two experiments, only Neuhoff's listeners had access to potentially highly useful interaural cues: These stimuli were "bypass trajectories", moving on a line parallel to the interaural axis either between azimuthal angles of 87.6° and 45.0°, or between 88.1° and 82.4°.

In the introductory paragraphs of this section, I hypothesised that the presence of room acoustics could lead to an improvement just as well as a degradation of listener performance in this task. The psychometric data give some indication of the former, in that for five out of six subjects, just-noticeable differences in motion distance were decreased in the echoic–receding condition compared

to the anechoic–receding trials (and EMM performance was consequently increased). In fact, for one subject, the availability of room cues brought only the echoic–receding performance in line with the approaching conditions. In contrast, there is no evidence for any detrimental effects of the loudness compression caused by reverberation. On the contrary, an analysis of the pairs of stimuli with the loudness model by B. C. Moore, Glasberg, *et al.* (2016) suggests that the significantly greater ($p = 0.0005$, Mann-Whitney U test) between-interval differences of within-interval loudness change in the anechoic intervals (median difference of 5.7 phon) do not positively influence the ratio of correct responses compared to the echoic intervals (median difference of only 2.6 phon). Taken together, it can be speculated that room-related cues improve the auditory discrimination of motion distances, to some extent counteracting the looming bias. More data is clearly required to support or refute this conjecture.

Beside dealing with the limitations of this pilot experiment such as the low number of subjects and the small range of tested distance differences, a full investigation will also need to include some control conditions to ascertain that any observed effects between the anechoic and echoic conditions are really due to room-acoustical effects. To investigate the impact of spatial cues, the echoic–anechoic pair of conditions could be augmented with a non-spatial echoic condition in which the direct sound and all reflections are played back from one loudspeaker. This might make the compression of loudness differences more disruptive by hampering the segregation of the direct sound and its reflections. Furthermore, the role of the DRR could be isolated by making it vary, *e.g.*, by artificially scaling the reverberant energy after every click to keep a constant ratio to the energy of the direct sound.

5.7 Discussion

This chapter introduced liveRAZR, a new real-time room-acoustical simulation system which I have developed based on the perceptually-validated RAZR model (Wendt *et al.*, 2014), with a macroscopic description of its architecture and of several implementation details. While liveRAZR does not yet contain all the features offered by the original RAZR, its present feature set already allowed its successful use in a psychoacoustic experiment. For configurations where their feature sets overlap, RAZR and liveRAZR generate practically identical output. Moreover, an analysis of program runtimes has shown the system to be capable of real-time performance in a variety of scenarios.

This analysis also resulted in a runtime model which can inform potential users about what liveRAZR currently can and cannot achieve under real-time constraints, but also highlights limitations which arise directly from particular design choices. One such example is the interaction between the type of spatial rendering (to a loudspeaker array *vs.* binaurally to headphones) and spatial precision (quantified in spatialiser vertices, *i.e.*, the number of loudspeakers in an array or points in an HRTF database). Namely, the influence of the number of vertices in the mesh is large for the array spatialiser, but very small for the binaural (merging) spatialiser (*cf.* the green and purple data across both columns in Figure 5.11). This is caused by the crucial implementation difference outlined in section 5.3.7: In the array spatialiser, there are as many simultaneous convolution channels as there are vertices; in the merging spatialiser, there are twice as many as there are virtual sound sources. As the speed of the array spatialiser is less dependent on the number of sources, this implies a trade-off: When the simulation generates many sources relative to the number of spatialiser vertices (sampling points of an HRTF database), we can expect that the array spatialiser would lead to better performance. liveRAZR currently always chooses the merging spatialiser for binaural synthesis, because the threshold at which the array spatialiser becomes faster is not straightforward to determine. Future revisions of the software may try to detect such cases, or feature a configuration toggle to let the user explicitly choose one over the other implementation.

A second limitation also concerns the spatial rendering module: It is known from psychophysics that sound sources near the listener can be more accurately localised in depth when they are located at an azimuth other than 0° (Kopčo and Shinn-Cunningham, 2011). This is partly due to interaural level differences which, due to near-field effects observed where the wavelengths of sound are roughly equal or larger than the distance from source to receiver, decrease with increasing distance until they reach a constant azimuth-dependent level at 1 m (Brungart and Rabinowitz, 1999; Coleman, 1963; Stewart, 1911). None of the liveRAZR spatialisers can currently account for this. Moreover, it is also not yet possible to pick HRTFs corresponding to different azimuth and elevation angles for the left and right ears, despite the perceptual differences that this parallax phenomenon causes in real-world close-up sources (H.-Y. Kim *et al.*, 2001). These omissions currently make liveRAZR an unsuitable choice if an accurate simulation of sound sources in a listener's peripersonal space is required.

The most significant limitation in practice, however, is the restriction of liveRAZR—as well as RAZR—to cuboid room geometries. This has been a deliberate design choice to keep the underlying model (and its software implementations) simpler than many of the other programs mentioned in section 5.1. However, it also greatly limits the scope of possible applications; for this reason, an extension of liveRAZR to more general room layouts is already underway.

CHAPTER 6

Listeners & sources with moving ears

This chapter is a reproduction of an open-access, peer-reviewed publication which has appeared in *Frontiers of Ecology and Evolution* on 12 April 2019. It can be accessed through the DOI 10.3389/fevo.2019.00116, or on the publisher's website under the URI `https://www.frontiersin.org/articles/10.3389/fevo.2019.00116/full`. The full citation is

> Ella Z. Lattenkamp, Stephanie M. Shields, Michael Schutte, Jassica Richter, Meike Linnenschmidt, Sonja C. Vernes, and Lutz Wiegrebe (2019). "The vocal repertoire of pale spear-nosed bats in a social roosting context". In: *Frontiers in Ecology and Evolution* 7, 116.

This publication has already been included in an earlier collection of works, namely in

> Ella Z. Lattenkamp (2020). "Vocal learning in the pale spear-nosed bat, *Phyllostomus discolor*". Dissertation. Nijmegen: Radboud University. ISBN: 978-94-92910-10-3

6.0 Author contributions

L.W., M.L., S.C.V., and E.Z.L. conceived and supervised the study. S.M.S. recorded the data. S.M.S., E.Z.L., and M.L. developed the classification key. L.W. wrote the syllable detection and analysis program. E.Z.L. and S.M.S. performed the syllable classification. M.S. conducted the statistical analyses and data presentation. J.R. rated the behavioral context. E.Z.L. wrote the first draft of the manuscript. All authors contributed to the writing, editing, and revising of the final paper.

The Vocal Repertoire of Pale Spear-Nosed Bats in a Social Roosting Context

Ella Z. Lattenkamp [1,2*†], ***Stephanie M. Shields*** [1†], ***Michael Schutte*** [1], ***Jassica Richter*** [1], ***Meike Linnenschmidt*** [1], ***Sonja C. Vernes*** [2,3] ***and Lutz Wiegrebe*** [1]

[1] *AG Wiegrebe, Department Biology II, Ludwig-Maximilians University Munich, Martinsried, Germany,* [2] *Neurogenetics of Vocal Communication Group, Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands,* [3] *Donders Institute for Brain, Cognition and Behaviour, Nijmegen, Netherlands*

Commonly known for their ability to echolocate, bats also use a wide variety of social vocalizations to communicate with one another. However, the full vocal repertoires of relatively few bat species have been studied thus far. The present study examined the vocal repertoire of the pale spear-nosed bat, *Phyllostomus discolor*, in a social roosting context. Based on visual examination of spectrograms and subsequent quantitative analysis of syllables, eight distinct syllable classes were defined, and their prevalence in different behavioral contexts was examined. Four more syllable classes were observed in low numbers and are described here as well. These results show that *P. discolor* possesses a rich vocal repertoire, which includes vocalizations comparable to previously reported repertoires of other bat species as well as vocalizations previously undescribed. Our data provide detailed information about the temporal and spectral characteristics of syllables emitted by *P. discolor*, allowing for a better understanding of the communicative system and related behaviors of this species. Furthermore, this vocal repertoire will serve as a basis for future research using *P. discolor* as a model organism for vocal communication and vocal learning and it will allow for comparative studies between bat species.

Keywords: vocal communication, *Phyllostomus discolor*, syllable classes, vocal repertoire, social behavior

OPEN ACCESS

Edited by:
Claudia Fichtel, Deutsches Primatenzentrum, Germany

Reviewed by:
Marco Gamba, University of Turin, Italy
Genevieve Spanjer Wright, University of Maryland, College Park, United States

****Correspondence:***
Ella Z. Lattenkamp
ella.lattenkamp@evobio.eu

[†] *These authors share first authorship*

Specialty section:
This article was submitted to Behavioral and Evolutionary Ecology, a section of the journal Frontiers in Ecology and Evolution

Received: *11 January 2019*
Accepted: *21 March 2019*
Published: *12 April 2019*

Citation:
Lattenkamp EZ, Shields SM, Schutte M, Richter J, Linnenschmidt M, Vernes SC and Wiegrebe L (2019) The Vocal Repertoire of Pale Spear-Nosed Bats in a Social Roosting Context. Front. Ecol. Evol. 7:116. doi: 10.3389/fevo.2019.00116

INTRODUCTION

Bats are highly gregarious mammals that have been extensively studied for their ability to echolocate (i.e., gain spatial information from the echoes of prior emitted ultrasonic calls). However, bats also emit social vocalizations to communicate with conspecifics and some bat species have been shown to possess rich vocal repertoires (e.g., Kanwal et al., 1994; Ma et al., 2006; Bohn et al., 2008), supporting intricate social interactions (Wilkinson, 1995, 2003). Current literature on vocal communication in bats illustrates that social vocalizations can be very complex, are highly important for bat sociality, and often vary notably between species. However, research in this field has only been scratching the surface; there is still much to learn about social communication in bats. Relative to the total number of bat species (being the second richest order of mammals with over 1,300 species), very few species have been studied, and even fewer have had their vocal repertoires described.

Research on social communication in bats generally focuses on studying a specific subset of vocalizations in a species repertoire—such as neonatal calls (Gould, 1975), calls produced during ontogeny (Knörnschild et al., 2006, 2010a), mother-infant calls (Esser and Schmidt, 1989), male song (Davidson and Wilkinson, 2004)—or more commonly on studying only one particular type of vocalization—such as distress calls (Russ et al., 2004; Hechavarría et al., 2016) or aggressive calls (Bastian and Schmidt, 2008). Fewer studies have sought to describe the repertoire of a species more comprehensively, defining several types of syllables emitted often in specific behavioral contexts (e.g., Behr, 2006; Knörnschild et al., 2010b; Wright et al., 2013). Even fewer have investigated the occurrence of syllable combination and temporal emission patterns (e.g., Kanwal et al., 1994; Bohn et al., 2008). These studies have reported a great deal of vocal diversity, ranging from 2 to 22 described vocalization types per species.

The pale spear-nosed bat, *Phyllostomus discolor,* has been in the focus of scientific attention for several years and has been investigated in a variety of psychophysical and neurophysiological studies (e.g., Firzlaff et al., 2006; Hoffmann et al., 2008; Heinrich and Wiegrebe, 2013) and, more recently, neurogenetics studies (Rodenas-Cuadrado et al., 2015, 2018). *P. discolor* is a scientifically particularly interesting species as it belongs to the handful of bat species for which evidence of vocal learning (i.e., the ability to produce new or strongly modified vocalizations according to auditory experiences) has been presented (Esser, 1994; Knörnschild, 2014; Lattenkamp et al., 2018). Social vocalizations of *P. discolor* are thus especially intriguing as these bats are a valuable system for the study of vocal learning that will help deepen our understanding of this phenomenon (Lattenkamp and Vernes, 2018). However, previous studies of social vocalizations in *P. discolor* have mainly focused on mother-infant communication (Esser and Schmidt, 1989; Esser, 1994; Esser and Schubert, 1998; Luo et al., 2017).

The current study is the first to assess the vocal communicative repertoire of *P. discolor* in an undisturbed social roosting context, which covers about 80% of their daily activity (La Val, 1970). Pairs and groups of three, four, and six pale spear-nosed bats were repeatedly recorded with a high resolution ultrasonic microphone array under anechoic conditions. Following the methodology of Kanwal et al. (1994), vocalizations were initially classified by two independent human raters and the classifications were subsequently statistically verified based on a fixed set of 19 automatically extracted spectral and temporal vocalization parameters. Eight distinct syllable classes were identified, and four additional, infrequently emitted classes were observed, suggesting that *P. discolor* possesses a diverse vocal repertoire. For the eight distinct syllable classes, the behavioral context at the time of emission was analyzed. The combined results present an extensive assessment of the vocal repertoire of the pale spear-nosed bat, *P. discolor*, in a social roosting context.

MATERIALS AND METHODS

Terminology

We follow previous literature in defining syllables as continuous vocal emissions surrounded by periods of silence (Kanwal et al., 1994; Doupe and Kuhl, 1999; Behr and Von Helversen, 2004; Bohn et al., 2008; Gadziola et al., 2012). By this definition, syllables are the smallest, independent acoustic unit of a vocalization. A call can consist of a single or multiple syllables (Gadziola et al., 2012). For clarity, we specifically focused on studying individual syllables rather than the less objective entity of a call. Syllable classes are used to describe groups of statistically different syllables (cf. Gadziola et al., 2012; Hechavarría et al., 2016), which are assigned depending on the outcome of the classification process described below. We follow the definitions of syllable train and phrase used by Kanwal et al. (1994) (cf. "simple phrase" and "combination phrase" used by Ma et al. (2006). The term syllable train describes a combination of two or more syllables from the same class, while a phrase describes a combination of syllables from at least two different classes. The silent period between any two syllables in a train or phrase is roughly similar and may be longer than the duration of any one syllable (Kanwal et al., 1994).

Animals

Six adult pale spear-nosed bats, *P. discolor*, were recorded in pairs or groups of three, four, and six. Recordings were done between January and March 2018 for 5 days per week. The animals recorded in this experiment originated from a breeding colony at Ludwig Maximilian University of Munich, where they were born and housed together throughout their lives. The sex ratio between the bats was equal. One male and one female were approximately 1 year old, while the other bats were between 6 and 9 years old. The bats were provided with a species specific diet (fruits, supplements, and meal worms) and had *ad libitum* access to water during and outside of the experiment. This experiment was conducted under the principles of laboratory animal care and the regulations of the German Law on Animal Protection. The license to keep and breed *P. discolor* as well as all experimental protocols were approved by the German Regierung von Oberbayern (approval 55.2-1-54-2532-34-2015).

Recording Setup

The recording setup was mounted in a sound-insulated chamber (2.24 × 1.27 × 2.24 m^3; L × W × H; **Figure 1A**) and consisted of a box containing recording equipment and space for the bats to roost (**Figure 1**). The instrumented box was mounted 1.5 meters above the ground, allowing the bats to fly in and out as they pleased. The ceiling light was only turned on when the experimenter was in the room. Otherwise, the chamber was only dimly illuminated by a small lamp, encouraging the bats to remain in the darker roosting area inside the box. During experimental sessions, the chamber was monitored via an infrared CCD camera (Renkforce CMOS, Conrad Electronic, Hirschau, Germany). Temperature and humidity were monitored from outside the chamber.

Vocalizations and behaviors were recorded with both high temporal and spatial resolution via a custom-built acoustic camera. This acoustic camera consisted of a 16-unit ultrasonic microphone array (custom-made on basis of SPU0410LR5H, Knowles Corporation, Itasca, IL, USA) and a high resolution infrared video camera (Point Gray Research Grasshopper3 GS3-U3-41C6NIR; FLIR Integrated Imaging Solutions, Inc.,

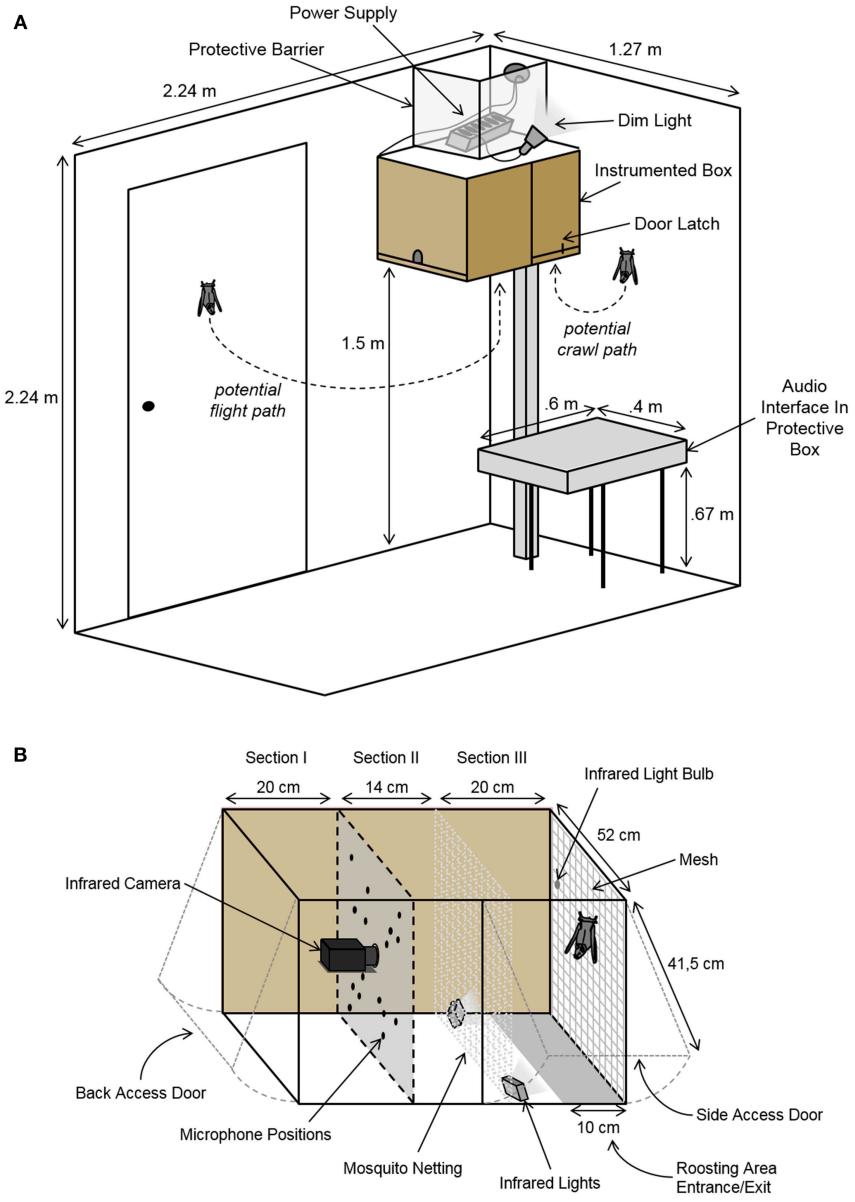

FIGURE 1 | Schematic of the setup**. (A)** The sound-insulated chamber containing the instrumented box and the audio interface. The instrumented box was mounted to the wall allowing a free flight path in and out of the box. Dim illumination encouraged the bats to remain in the darker roosting area inside the box. The chamber was monitored via an infrared camera. **(B)** Detailed schematic of the instrumented box containing the acoustic camera. The microphone array and camera faced the freely accessible roosting space. The box was divided into three sections: (I) an area containing the equipment and corresponding cords, (II) a secured space between the bats' roosting area and the panel with the recording equipment, and (III) a section for the bats. The bats' roosting area was illuminated with two infrared lights.

Richmond, BC, Canada) controlled and synchronized via a custom-written MATLAB (R2015a, MathWorks, Cambridge, MA, USA) script. By comparing time-of-arrival differences between all microphones of the array, the acoustic camera allows to determine the exact location of a sound source in the recorded video. The camera and microphones were mounted inside of the instrumented box (54 × 52 × 41.5 cm^3; L × W × H; **Figure 1**), which was lined with acoustic foam. The bats could enter or exit through a 10 cm wide opening along the bottom of the backside of the box (cf. section Results, **Figure 1B**). Two additional doors with latches allowed the experimenter to access the bats and the equipment independently (**Figure 1**). The back wall of the bats' roosting space was lined with mesh for the bats to hang from and crawl on. Two small infrared lights were mounted in the

lower corners of the bats' area, illuminating the back wall. An additional infrared light bulb was hung from the mesh mounted on the back wall. This infrared light was used to synchronize the recorded video with the recorded audio. Audio data was recorded via a Horus audio interface (Merging Technologies SA, Puidoux, Switzerland) placed next to the instrumented box in the experimental chamber (**Figure 1A**).

Recording Procedure

The six bats were observed in the recording chamber for 47 sessions (either 1.5 or 3 h long), amounting to a total of 96 h of observation. All 15 possible pair combinations between the six bats were observed for 1.5 h each. On these pair-recording days, the remaining four bats were added into the recording chamber after the first 1.5 h and all six bats were subsequently observed for another 1.5 h. In two additional sessions, first all males and then all females were observed together for 3 h each. Next, all 15 possible combinations of four bats were observed for 3 h as well.

During the recording sessions, the bats were monitored in real-time. The recording of audio and visual data was manually triggered by an experimenter from outside the chamber, when social vocalizations were emitted in the chamber. Ultrasonic vocalizations were made audible for the experimenter via real-time heterodyning of two of the 16 microphone channels and presented via headphones. The data acquisition was controlled via a custom-written MATLAB script, which saved a 10 s audio ring buffer synchronously for all 16 microphones (sampling rate: 192; microphone gain: 18 dB). The corresponding 10 s long video files were recorded synchronously via StreamPix 6 Single-Camera (NorPix, Inc., Montreal, QC, Canada) (frame rate: 100/s; shutter speed: 9.711 ms). The video files were compressed using the Norpix Motion-JPEG Encoder AVI Video Codec.

Acoustic Analysis

For the acoustic analysis, we detected and extracted all vocalizations surrounded by silence via a custom-written MATLAB script. Syllable detection was based on amplitude peaks identified in the recordings, which were at least 20 dB louder than the background noise and were separated in time from previously detected peaks by at least 5 ms. For each identified syllable, the recording from the microphone that picked up the loudest signal was used for analysis. Nineteen acoustic parameters were extracted or calculated for each detected syllable: (1) Syllable duration and (2) maximum syllable amplitude were calculated. To represent the overall frequency content of the syllable, 5 parameters were calculated: (3) spectral centroid frequency (SCF; i.e., weighted mean of the frequencies contained in a syllable), (4) peak frequency (PF; i.e., the frequency with the most energy content), (5) minimum frequency, (6) maximum frequency, and (7) overall syllable bandwidth. The fundamental frequency ($f0$) contour of each syllable was detected using the YIN algorithm (de Cheveigné and Kawahara, 2002), and six parameters describing this $f0$ contour were then extracted: (8) mean $f0$, (9) minimum $f0$, (10) maximum $f0$, and (11) starting $f0$ at the syllable onset. Seven additional parameters describing the $f0$ contour were extracted: (12, 13) the coefficients of the best-fitting linear (degree 1) polynomial and (14, 15, 16) quadratic (degree 2) polynomial to the raw contour of the $f0$. (17, 18) Furthermore, the root-mean-square errors (RMSE) between the fitted polynomials and the $f0$ contours were calculated (19). Lastly, the aperiodicity of syllables was also calculated via the YIN algorithm. It represents how noisy a signal is and functions as a proxy for entropic state of the vocalization (i.e., an aperiodicity of $\geq$0.1 indicates high entropy). The YIN algorithm first assesses the degree of aperiodicity of a recorded call and then tries to assign a fundamental frequency to those call segments where aperiodicity is low enough to do so. In the analyses of some quite complex syllables (see below), the fundamental frequency estimate may jump very quickly between quite different values.

Syllable Classification

Qualitative Categorization

Following Kanwal et al. (1994) and Ma et al. (2006), a preliminary classification key consisting of 20 vocalization classes was generated based on the spectrograms of a subset of recordings and previous literature (Kanwal et al., 1994; Ma et al., 2006). Subsequently, two independent raters visually assessed the spectrograms and waveforms of the extracted syllables based on their duration and frequency information, such as spectral contour, aperiodicity, or suppression of frequencies. The syllables were presented to the raters in four different ways: (1) the waveform of the syllable; (2) the spectrogram of the extracted syllable; (3) the spectrogram of the extracted syllable scaled to a fixed 100 ms window; (4) the spectrogram in a 100 ms context window, which displayed the recording 50 ms before and after the extracted syllable. This way of displaying the data allowed the raters to determine whether the syllable was extracted well or erroneously. Syllables were either sorted into syllable classes defined in the preliminary classification key, or they were marked as unsuitable for analysis due to low quality (e.g., because of spectral smear, syllable overlap, or incorrect extraction). A few vocalizations were marked as not matching any of the syllable classes present in the preliminary key. These potentially novel syllable classes were later reexamined, and two additional syllable classes were suggested as a result.

Quantitative Categorization

For the quantitative categorization only high quality recordings of social syllables that were classified identically by both raters were used. Only classes containing at least 50 detected syllables were analyzed. The separability of the classes based on the 19 extracted spectro-temporal parameters was verified and refined based on a 5-fold cross validation procedure (Hastie et al., 2009). The dataset was stratified prior to splitting into folds to avoid empty classes and reduce variance (Forman and Scholz, 2010). In each fold, ~80% of the data for each class were employed to fit a linear discriminant analysis (LDA) classifier (Hastie et al., 2009), and this classifier was used to predict the classes of the remaining 20% of the calls. Each call was used in the test dataset exactly once. A mean confusion matrix was computed from the ground-truth labels assigned by the human raters and the labels predicted by the LDA classifier. The confusion matrix was normalized by multiplying each row vector with a constant factor to have row sums of 1. The normalized confusion matrix guided the

refinement of the preliminary labels obtained from the qualitative categorization. As the ultimate goal of the classification process was the development of an automatic classifier, which renders human raters redundant in the future, an algorithmically greedy procedure was used to merge the pair of classes with the highest off-diagonal normalized confusion score. This procedure was done with the input of the human raters, confirming the reasonableness of the merge. The LDA analysis was then rerun on the altered dataset and this algorithm was iterated as long as the human raters agreed that the two candidate classes for merging were non-trivial to separate by their spectrograms. The merging was continued, until a 60% overlap of the human raters and LDA classification was reached.

Behavioral Video Analysis

We assessed the behavioral context observed during the emission of syllables belonging to the previously established classes. For that reason, an ethogram containing 56 detailed behaviors for *P. discolor* was generated based on personal observations (ML, SS, EL). More specifically, the ethogram encompassed 20 behaviors observed in neutral contexts, 18 in prosocial, and 18 in antagonistic behavioral contexts. This ethogram was used by a naïve rater to score the behaviors observed in the video files. The rater was blinded to the emitted syllables contained in the videos. The behavioral scoring was done in the Behavioral Observation Research Interactive Software (BORIS) (Friard and Gamba, 2016), and the behavior that occurred at the time of syllable emission was extracted.

RESULTS

Within the 96 h of observation 1,434 recordings were made. The automatic syllable finder identified 57,955 vocalizations in these recordings, which were assessed by the two independent raters. The majority of these vocalizations were excluded from the subsequent quantitative analyses for several reasons: 56% (n = 32,551) were excluded, because one or both raters marked them as unsuitable for the classification (due to syllable overlap or low recording quality occurring when vocalizations were emitted outside the instrumented box) or because the two independent raters disagreed on their classification; 2% (n = 1,115) of the recorded sounds were excluded as they presented no vocalizations, but rather scratching noises produced by the bats brachiating on the back wall of the box; and 10% (n = 5,630) of the data were eventually excluded, because not all 19 spectro-temporal syllable parameters could fully be extracted. The remaining 32% (n = 18,658) of the vocalizations represented conservatively selected, high quality syllables classified identically by both independent raters. These syllables were qualitatively and quantitatively assessed as belonging to 13 syllable classes. Of these 13 classes eight were represented by more than 50 syllables and thus evaluated as commonly occurring in this social roosting context (n = 6,162) and four classes were represented by <50 syllables and are thus reported as rarely occurring (n = 81). The largest class (n = 12,416) was comprised of calls with a suppressed fundamental frequency (SF class) and is reported separately below.

For the 19 extracted spectro-temporal parameters, the 25th, 50th, and 75th percentiles (i.e., first, second, and third quartiles) are reported below to represent data distribution. These values are presented as follows: Q50 [Q25 Q75]. Additionally, all quartiles for each parameter are listed in **Supplementary Table S1** for each common syllable class and in **Supplementary Table S2** for each rare syllable class and the suppressed fundamental frequency class. An example of all commonly occurring syllables is given in **Figure 2**, while the variation within these classes is illustrated in the Supplementary Material (**Supplementary Figure S1**).

Common Syllable Classes

High Entropy (HE) Vocalizations

The majority of high quality, commonly emitted social syllables belong to the high entropy (HE) class (n = 3,860; 63% of all syllables in the commonly occurring classes). HE syllables were termed according to their appearance in the spectrogram (i.e., smeared along the frequency axis), and can generally be described as noisy or screechy vocalizations (**Figure 2A**). They can still retain some degree of harmonicity, similar to synthesized tonal noises (iterated rippled noises) (Yost, 1996), and if the residual tonality was strong enough, modulations of the fundamental frequency (typically sinusoidal) could be observed (**Supplementary Figure S1**). As expected, HE syllables displayed a very high degree of aperiodicity (0.42 [0.34 0.48]; cf. Q50 [Q25 Q75], **Figure 3**). The short average duration of HE syllables (6.24 [4.74 10.27] ms) can be explained by our definition of syllable: The raters observed that long HE calls are often composed of several HE syllables (cf. **Figure 9A**), which were analyzed individually, if the call was strongly amplitude modulated and the modulation period longer than 5 ms (cf. 5 ms criterion for syllable separation).

Linearly Downward Frequency Modulated (lDFM) Vocalizations

Seven hundred and twenty-seven syllables (12%) are composed of linear downward frequency modulations (lDFM) of the fundamental frequency (**Figure 2B**). Linearly DFM syllables are usually relatively short (6.74 [5.35 8.78] ms). They have a steep downward slope (−1.70 [−2.06 −1.41] kHz/ms) and the highest mean fundamental frequency (17.27 [15.83 18.65] kHz; **Figure 3**) of all commonly occurring syllables.

Non-linearly Downward Frequency Modulated (nlDFM) Vocalizations

Non-linearly downward frequency modulated (nlDFM) syllables (n = 562; 9%) also sweep downward, but they have a curved shape, or an irregular offset including small constant frequency or upward frequency modulated components (**Figure 2C**). These nlDFM syllables are generally longer than lDFM syllables (17.10 [13.72 20.13] ms; **Figure 3**) and have a lower mean $f0$ (14.72 [13.56 15.71] kHz; **Figure 3**). While lDFM and nlDFM syllables have a comparable bandwidth (lDFM: 28.50 [23.25 33.75] kHz; nlDFM: 28.50 [23.25 33.00] kHz), the slope of nlDFM syllables is less steep on average (−0.74 [−1.07 −0.53] kHz/ms).

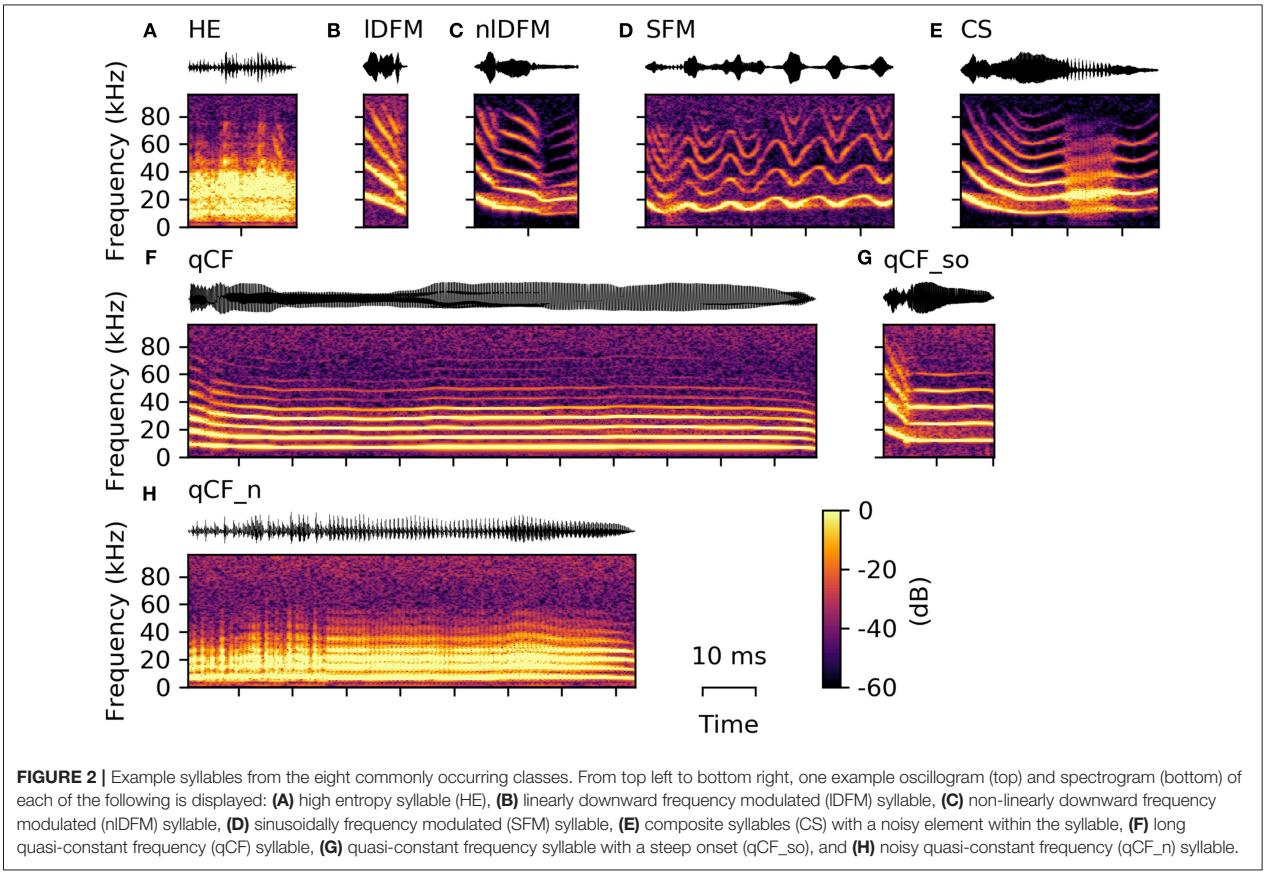

FIGURE 2 | Example syllables from the eight commonly occurring classes. From top left to bottom right, one example oscillogram (top) and spectrogram (bottom) of each of the following is displayed: **(A)** high entropy syllable (HE), **(B)** linearly downward frequency modulated (lDFM) syllable, **(C)** non-linearly downward frequency modulated (nlDFM) syllable, **(D)** sinusoidally frequency modulated (SFM) syllable, **(E)** composite syllables (CS) with a noisy element within the syllable, **(F)** long quasi-constant frequency (qCF) syllable, **(G)** quasi-constant frequency syllable with a steep onset (qCF_so), and **(H)** noisy quasi-constant frequency (qCF_n) syllable.

Sinusoidally Frequency Modulated (SFM) Vocalizations

Also frequently occurring were syllables with a sinusoidal $f0$ contour (SFM) ($n = 445$; 7%). SFM syllables have a stable sinusoidal frequency modulation with small overall variation in modulation depth and modulation frequency, and they generally do not have an onset that notably exceeds the first frequency modulation (**Figure 2D**). However, SFM syllables can also have a steep linear downward sweep onset and a horizontal, ascending, or descending SFM tail (cf. **Supplementary Figure S1**). Irregular SFM syllables are also emitted and consist of inconsistent sinusoidal frequency modulations. SFM syllables can vary in both the rate and depth of oscillations. Similar to HE syllables, SFM vocalizations are often strongly amplitude modulated and our definition of syllables thus determines the rather short average durations of the SFM syllables (5.51 [4.66 7.90] ms; **Figure 3**).

Composite (CS) Vocalizations

Composite syllables (CS; $n = 286$; 5%) contain both tonal and noisy elements. Frequently, the syllable begins with a tonal, downward frequency-modulated sweep and then ends with a HE element. One or more HE elements can also occur within syllables (**Figure 2E**). In most cases, a CS is a SFM syllable that is interrupted by one or more HE elements. These syllables had the third highest average aperiodicity (0.09 [0.06 0.12]; **Figure 3**) of the commonly emitted syllables.

Quasi-Constant Frequency (qCF) Vocalizations

Quasi-constant frequency (qCF) syllables ($n = 67$; 1%) have a near constant fundamental frequency for the duration of the entire syllable (**Figure 2F**). qCF syllables are tonal and have no specific onset, but rather start immediately with the constant frequency element. Overall, syllables in the qCF class tended to have low mean $f0$s (7.19 [6.15 9.87] kHz; **Figure 3**).

Quasi-Constant Frequency Vocalizations With a Steep Onset (qCF_so)

Tonal qCF syllables can also have a steep downward frequency modulated onset (qCF_so; $n = 89$; 1%; **Figure 2G**). A separate class was created for those qCF_so syllables as they necessarily differ in many parameters from pure qCF syllables, which lack such a clear onset. For example, qCF_so syllables have stronger negative $f0$ slopes than the qCF syllables, because of the added onset (qCF_so: −0.40 [−0.51 −0.28] kHz/ms; qCF: −0.05 [−0.20 0.01] kHz/ms). For the same reason, the qCF_so syllables are generally longer (qCF_so: 21.03 [17.98 24.33] ms; qCF: 10.64 [7.20 20.52] ms).

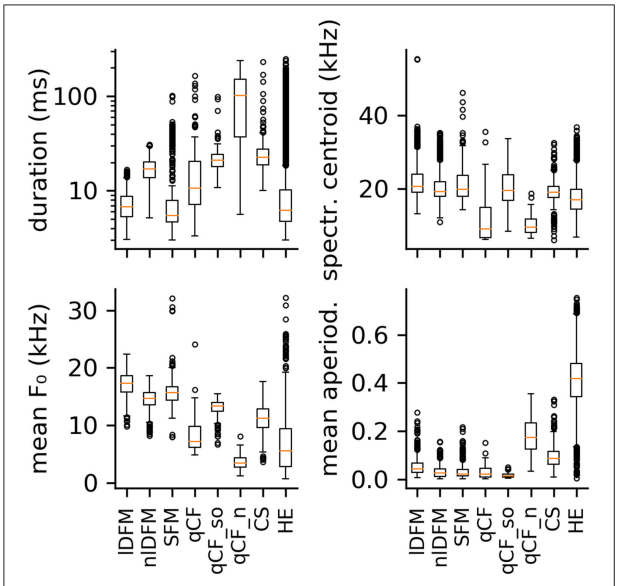

FIGURE 3 | Boxplots of four selected spectral and temporal parameters. From top left to bottom right: syllable duration, spectral centroid frequency, mean fundamental frequency, and mean aperiodicity. Distributions are shown for the eight commonly occurring syllable classes: linearly downward frequency modulated (lDFM); non-linearly downward frequency modulated (nlDFM), sinusoidally frequency modulated (SFM), quasi-constant frequency (qCF), quasi-constant frequency with a steep onset (qCF_so), noisy quasi-constant frequency (qCF_n), composite syllables (CS), and high entropy syllables (HE).

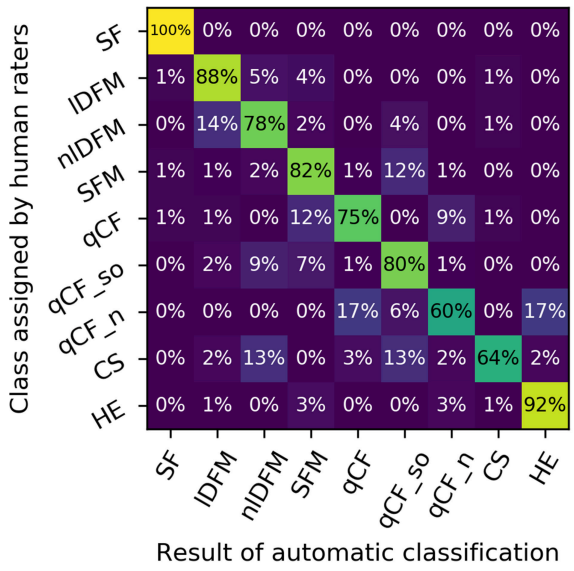

FIGURE 4 | Confusion matrix depicting the distinguishability of the suppressed fundamental frequency (SF) class and the eight commonly occurring syllable classes: linearly downward frequency modulated (lDFM); non-linearly downward frequency modulated (nlDFM), sinusoidally frequency modulated (SFM), quasi-constant frequency (qCF), quasi-constant frequency with a steep onset (qCF_so), noisy quasi-constant frequency (qCF_n), composite syllables (CS), and high entropy syllables (HE). Rows: classes as specified by human raters. Columns: class labels as predicted by an automatic LDA classifier. Rows are normalized to a sum of 100%.

Noisy Quasi-Constant Frequency (qCF_n) Vocalizations

Noisy quasi-constant syllables (qCF_n) are essentially high entropy versions of the tonal qCF syllables (n = 126; 2%; **Figure 2H**). They also did not start with a frequency modulated onset. Of all syllable classes, qCF_n syllables had the longest average durations (102.20 [37.13 151.90] ms), lowest mean f0s (3.45 [2.71 4.38] kHz), and lowest spectral centroids (9.51 [8.07 11.78] kHz). They had the second highest average aperiodicity (0.17 [0.12 0.23]; **Figure 3**).

In the quantitative analysis, the LDA classifier performed with an overall accuracy of 87% over the eight classes described above (chance level: 12.5%) (**Figure 4**). The mean overall precision score was 89%, mean overall recall 87%, mean per-class precision 67%, and mean per-class recall 76%. **Figure 4** reproduces the row-normalized confusion matrix, i.e., each cell shows which percentage of calls of a specific human-rated class is assigned to a specific class label by the automatic classifier. The confusion matrix shows that particularly high recall scores are attained for lDFM and HE calls, which also separate comparatively well univariately (based on mean f0 and mean aperiodicity, respectively).

Behavioral Context of the Common Syllable Classes

For each of the eight commonly occurring syllable classes, 20 videos were scored for the behaviors displayed by the bats during syllable emission. For the lDFM and nlDFM classes only 19 instances could successfully be scored as the behavior for one instance was performed outside the field of view of the camera. From the ethogram of 56 detailed behaviors, only 23 behaviors were observed during syllable emission (**Supplementary Table S3**). Only one single observation was ever made, where a vocalization was emitted in a neutral behavioral context (**Supplementary Table S3**; **Figure 5A**). More specifically, a single HE syllable was emitted in a context scored as "brachiating on walls or ceiling." Other than that, syllables were always emitted either in a prosocial or an antagonistic behavioral context.

The behavioral analyses show that the HE syllables are emitted 95% of the time in antagonistic encounters (**Supplementary Table S3**). One exception is the above mentioned single observation of a HE syllable emitted in a neutral context. All other syllables were, with varying prevalence, emitted in both, prosocial and antagonistic contexts (**Supplementary Table S3**; **Figure 5A**). Syllables from the qCF, SFM, and nlDFM classes were emitted in prosocial behavioral contexts in 75–85% of the scored videos (**Supplementary Table S3**; **Figure 5A**). CS, lDFM, and qCF_so syllables were emitted slightly more often in prosocial than antagonistic contexts (in 55–63% of the videos, **Supplementary Table S3**). Noisy qCF syllables (qCF_n) were emitted in antagonistic behavioral contexts in 40% of

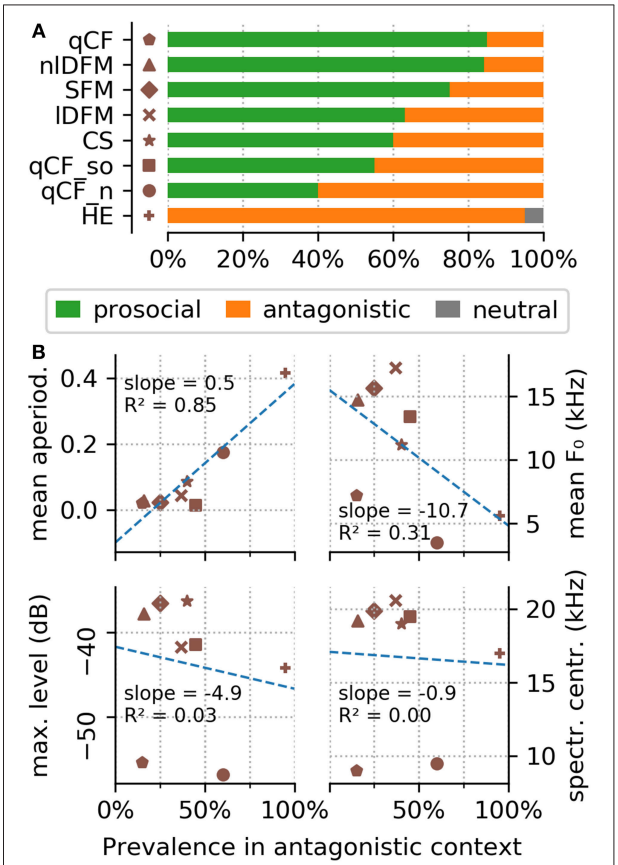

FIGURE 5 | Behavioral context of the common syllable classes. **(A)** The top panel shows the behavioral context scored for the eight commonly occurring syllable classes. **(B)** The bottom panel shows the correlation of four syllable parameters (top left to bottom right: mean aperiodicity, mean fundamental frequency, maximum level, and aperiodicity) in dependence of their prevalence in antagonistic encounters.

the scored videos. Stable correlations were found between some acoustic parameters and the behavioral context in which a syllable was emitted: Specifically, the measured aperiodicity of the syllables is strongly positively correlated with their prevalence in antagonistic encounters (**Figure 5B**). Also syllable $f0$s are lower during antagonistic behaviors (**Figure 5B**).

Rare Syllable Classes

In addition to the commonly occurring syllable classes, several vocalizations were repeatedly, but extremely infrequently emitted. Specifically, out of the total of 18,658 high quality recordings fewer than 50 vocalizations per rare syllable class were recorded. Thus, not enough data are available to include these vocalizations in the statistical analysis. They are described in the following as purely observational and should be considered as rarely emitted, at least in a social roosting context.

Puffs

During the recording sessions, the bats repeatedly emitted air puffs ($n = 42$), which appeared to result from bats forcefully expelling air through their nostrils. These sounds are not necessarily to be considered sneezing, but are rather short nasal exhalation potentially used to clean the nostrils. The spectrograms of puffs appear to be noisy sound clouds with a sharp onset (**Figure 6A**). As the puffs did not contain a tonal component, the mean aperiodicity and bandwidth of these puffs were the highest of all recorded vocalizations (aperiodicity: 0.43 [0.40 0.47] and bandwidth: 45.75 [42.00 48.75] kHz).

V-Shaped Vocalizations

Syllables from this class ($n = 30$) consisted of a downward frequency modulated onset and a subsequent upward sweep, resulting in a characteristic "V"-shaped frequency contour (**Figure 6B**). Vocalizations in the V-shaped class are in shape comparable to the sinusoidal vocalizations, but always end within the first modulation.

Noisy Quasi-Constant Frequency Vocalizations With Steep Onset (qCF_nso)

The qCF_nso syllables were recorded only five times and were a combination of the qCF_n and the qCF_so syllable classes (**Figure 6C**). They also consist of a steep downward frequency modulated onset followed by a quasi-constant syllable element. However, they were emitted with higher sound pressure levels than qCF_n and higher aperiodicity than qCF_so syllables (**Supplementary Tables S1**, **S2**), resulting in a noisy version of the qCF_so syllable type.

Hooked Frequency Modulated (hFM) Vocalizations

Upward- or downward-hooked frequency modulated (hFM) syllables ($n = 4$) are characterized by the similarity between the shape of the vocalization displayed in the spectrogram and a hook. These syllables are typically short and can appear in either an upward-hooked (**Figure 6D**) or a downward-hooked (**Figure 6E**) form. These two hFM syllable types were the least abundant (upward-hooked: $n = 1$; downward-hooked: $n = 3$). HFM syllables had the highest average spectral centroid aside from syllables with a suppressed fundamental frequency (27.08 [21.71 33.09] kHz). However, comparative results should be taken with care, as the quantitative characteristics of this class are not well-supported due to the small number of syllables detected.

Suppressed Fundamental (SF) Class

The vast majority of recorded syllables belonged to the suppressed fundamental (SF) class ($n = 12{,}416$; 66% of the high quality, uniformly rated syllables). This syllable class can easily be distinguished from all other recorded syllables by its high spectral centroid (**Figure 7**). In fact, the spectral centroid frequency is a parameter showing a clear bimodal distribution of the data, splitting SF syllables and syllables of all other classes (**Figure 7**).

Syllables in the SF class have either a fully or partially suppressed fundamental frequency, and the dominant harmonic is instead the second or even third harmonic (**Figure 8**). SF syllables typically had short durations (4.07 [3.46 5.04] ms,

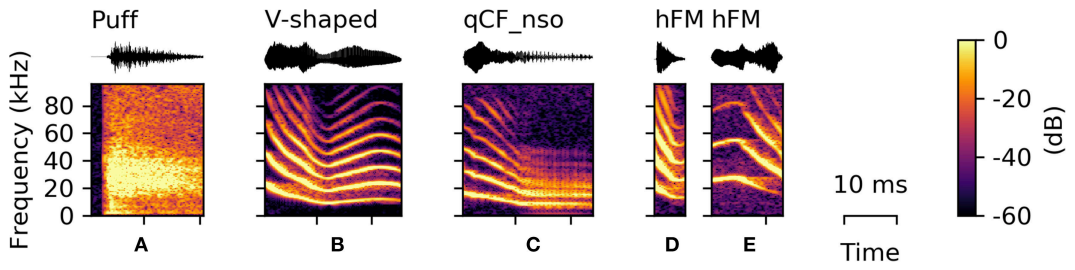

FIGURE 6 | Examples of syllables from rarely occurring classes. **(A)** Puff sound, **(B)** V-shaped, **(C)** noisy quasi-constant frequency syllables with a steep onset (qCF_nso), **(D)** upward-hooked, and **(E)** downward-hooked frequency modulated (hFM) syllables.

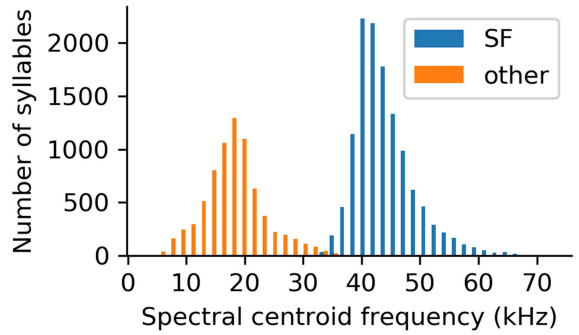

FIGURE 7 | Spectral centroid frequencies of all analyzed syllables, separated based on whether or not the syllable was classified as belonging to the suppressed fundamental frequency class.

Supplementary Table S2) and high spectral centroids (43.05 [40.65 46.51] kHz, **Supplementary Table S2**). Especially the very short durations indicate that this syllable class includes the species-specific echolocation calls, which typically range in duration between 0.3 and 2.5 ms (Rother and Schmidt, 1985; Kwiecinski, 2006; Luo et al., 2015). However, the SF class also included syllables, which structurally resembled syllables from other commonly occurring syllables classes with the only decisive difference that the fundamental frequency was fully or partially suppressed (**Figure 8**). Based on these strong characteristics and the varying shape of the SF syllables, this class can be easily separated from the other classes, but should rather be regarded as a meta-class, containing versions with suppressed fundamental frequency of most other syllable types. The function of these SF calls is currently uncertain and might or might not vary from the normal context of the syllable type with expressed fundamental frequency.

Syllable Combinations: Trains and Phrases

Very few studies have investigated temporal emission patterns of syllables and the existence of consistently-occurring syllable combinations (e.g., Kanwal et al., 1994; Bohn et al., 2008; Knörnschild et al., 2014; Smotherman et al., 2016). Previous literature shows, however, that for certain bat species the temporal emission pattern of social vocalizations can be highly complex. *Phyllostomus discolor* also emits combinations of syllables in a standardized order and with constant temporal emission patterns. Temporal relationships between syllables were not analyzed in the current work, thus we cannot draw qualitative conclusions about this aspect of the vocalizations. However, during syllable classification we observed several syllable combinations of varying length, complexity, and number of contained syllables (**Figure 9**).

Observed syllable trains consist of multiple syllables from the same class repeated with roughly the same temporal distance, whereby the silent interval can be longer than the preceding syllable (**Figures 9B,C**). Syllable trains can be of varying overall length, depending on the number of contained syllables. Phrases consist of syllables from two or more classes (**Figures 9D–F**), which can be repeated several times (usually in a fixed temporal distance). We found eight different types of syllable combination, which were repeatedly recorded over the duration of the experiment. The behavioral purpose of syllable trains and phrases is thus far purely speculative. A repetitive emission of phrases might serve to emphasize the transmitted information, but the number of phrase repetitions could also carry information by itself. Though the function and magnitude of syllable trains and phrases in these bats is currently unknown, we want to report our observation of them to encourage further research in this direction.

DISCUSSION

Vocalizations of *P. discolor*: Known and Novel

Here we present an extensive assessment of the vocal repertoire of the pale spear-nosed bat, *P. discolor*. As we recorded vocalizations in a social roosting context, which is the main pastime of *P. discolor* (Kwiecinski, 2006), we are confident that we identified the majority of social vocalizations emitted by this species. From 18,658 high-quality syllable recordings, we were able to define eight distinct classes, uniquely different from each other in their spectro-temporal parameters. We were also able to support the acoustic analysis with a detailed assessment of the behavioral contexts in which these eight syllable classes are generally emitted (**Supplementary Table S3**; **Figure 5**). Furthermore, we describe

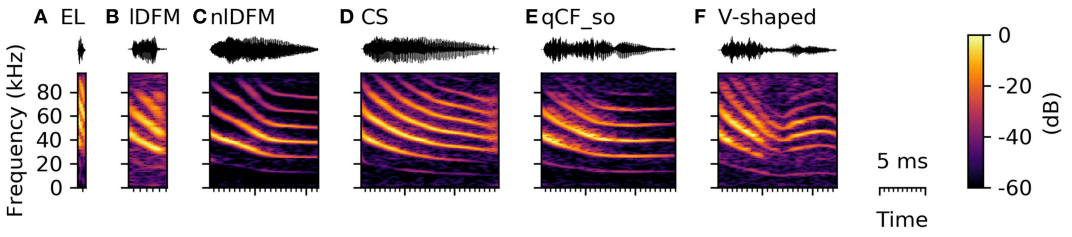

FIGURE 8 | Spectrograms of syllables with a suppressed fundamental frequency resembling syllables from other classes. **(A)** species-typical echolocation call (EL), **(B)** linearly downward frequency modulated syllable (lDFM), **(C)** non-linearly downward frequency modulated syllable (nlDFM), **(D)** composite syllable, **(E)** quasi-constant frequency syllable with a steep onset (qCF_so), and **(F)** V-shaped syllable.

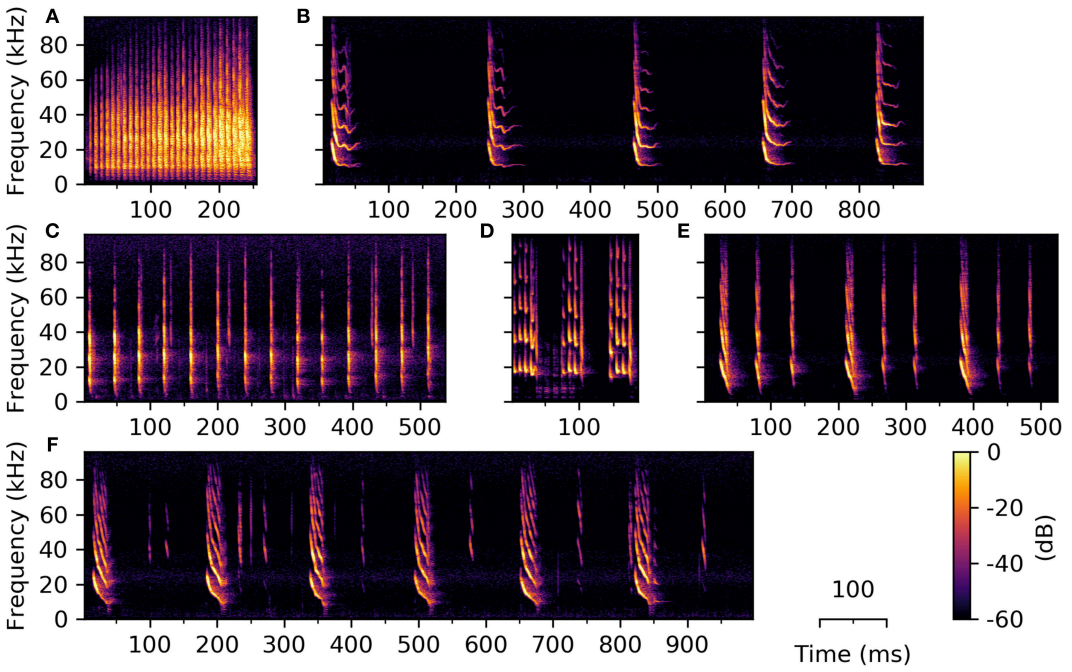

FIGURE 9 | Example spectrograms of three syllable trains **(A–C)** and three phrases **(D–F)** emitted by *P. discolor*. Syllable trains consist of several repetitions of one syllable type, e.g., **(A)** high entropy syllables, **(B)** sinusoidally frequency modulated, or **(C)** simple linearly frequency modulated syllables. **(D–F)** Phrases consist of syllables from at least two different classes; Phrases are depicted three **(D,E)** or six **(F)** times in order to show the temporal relationship between repetitions of the different phrases.

four additional call classes, which were only infrequently emitted by the bats and are thus described here, but not analyzed on the basis of their spectro-temporal characteristics.

Most syllable classes described in the present study have never before been observed for this species. Especially the quasi-constant frequency modulated (qCF) class and classes containing qCF elements (i.e., qCF_so and qCF_n) have hitherto not been reported for *P. discolor*. From our behavioral observations (**Supplementary Table S3**) it becomes apparent that all three classes containing syllables with a qCF element are used in very versatile behavioral contexts. This could indicate a loose behavioral association with the syllable structure and one could speculate about a behaviorally more meaningful variation of these syllables in their specific context (e.g., duration of qCF element could indicate special emphasis on a particular meaning). However, such speculations await experimental confirmation.

Sinusoidally frequency modulated (SFM) syllables have received considerable attention in previous literature. In *P. discolor*, SFM syllables were found to be used in mother-infant communication (as e.g., maternal directive calls and late forms of infant isolation calls) and can encode individual signatures, and even vocal dialects (Gould, 1975; Esser and Schmidt, 1989; Esser and Lud, 1997; Esser and Schubert, 1998). We can

confirm that the majority of the analyzed SFM syllables were emitted in the behavioral contexts "attention seeking" or "vocal contact," which are both in line with previous observations (**Supplementary Table S3**). In addition to the usage of SFM syllables in these contexts, we also demonstrated their emission in antagonistic encounters (**Supplementary Table S3**; **Figure 5A**). Emission of one syllable type in a variety of different behavioral contexts (cf. **Supplementary Table S3**; **Figure 5A**) suggests complex communicative function or purpose. Thus, our results support previous findings, which advocate syllable subgroups, in which vocalizations with very similar acoustic parameters can be further split up based on associated behaviors (Bohn et al., 2008; Kanwal, 2009). As described above, the syllable classification here presented is based purely on spectrogram shape and the extracted syllable parameters. This allows us to present mathematically distinct syllable classes and validates our first, subjective classification scheme. Nevertheless, the established classes may be further differentiated according to their behavioral contexts. Our behavioral assessments show that syllables from a single class with very similar acoustic characteristics can be used in up to 10 different behavioral contexts (**Supplementary Table S3**). The establishment of syllable subgroups (i.e., splitting of the presented syllable classes) based on their contextual usage would require extensive, detailed behavioral observations and ideally confirmation via playback experiments. We also want to highlight the possibility that additional syllable classes might be contained in the *P. discolor* repertoire, which were not emitted in the here reported social roosting context.

Comparison to the Closely Related Species (*P. hastatus*): Emerging Vocal Complexity

The number of distinct syllable classes assessed in this study (eight) is comparable to vocal repertoire descriptions of other bat species, which also found between 2 and 10 syllable types (e.g., Nelson, 1964; Gould, 1975; Barclay et al., 1979; Kanwal et al., 1994; Pfalzer and Kusch, 2003; Bohn et al., 2004; Wright et al., 2013; Knörnschild et al., 2014). When comparing the vocal repertoire of *P. discolor* to a closely related species (*P. hastatus*, which lives under essentially identical social and ecological conditions), it is noticeable, that the vocal repertoire of *P. hastatus* is less expansive. In addition to their echolocation calls, only two types of social calls are reported for *P. hastatus*, namely group-specific foraging calls, so-called screech calls, and infant isolation calls (Bohn et al., 2004). The screech calls of *P. hastatus* were shown to be used for the recognition of social group members during foraging, while infant isolation calls help mothers to recognize offspring (Boughman, 1997; Boughman and Wilkinson, 1998; Wilkinson and Boughman, 1998). Vocalizations reported as infant isolation calls are distinctly different between *P. discolor* and *P. hastatus*, with the former using single, clearly sinusoidally frequency modulated calls (Esser and Schmidt, 1989) and the latter typically using a pair of linear or bent frequency modulated calls (Bohn et al., 2007). The broadband, noisy screech calls of *P. hastatus* are similar in their spectral characteristics to the here defined high entropy (HE) syllables (Boughman, 1997), they are, however, used for the coordination of foraging activities and are not emitted in antagonistic behaviors contexts as observed in this study (**Supplementary Table S3**). The surprising difference in the size of the vocal repertoires of these closely related species, which are so similar in their ecology and lifestyle, only highlights the value of *P. discolor* as a model species for vocal communication and vocal learning. The vocal repertoires of the other members of the genus (*P. elongatus* and *P. latifolius*) are still unknown. Uncovering the evolutionary background of the emergence of such differences in vocal complexity in closely related species might help us to shed light on the evolution of communicative systems and the capacity for vocal learning in bats.

Similarities to Distantly Related Species: Acoustic Universals

A number of distantly related bat species were reported to emit high entropy calls during aggressive encounters (e.g., Russ et al., 2004; Hechavarría et al., 2016; Prat et al., 2016). It has been hypothesized that aggressive vocalizations tend to always be long, rough, and lower in frequency (Briefer, 2012). We confirmed a strongly positive correlation between the mean syllable class aperiodicity and its prevalence in antagonistic confrontations (**Figure 5B**). We also detected a negative correlation between the mean fundamental frequency of a syllable class and its occurrence during aggressive encounters. Overall, these findings support the idea of shared characteristics of mammalian vocalizations in strongly emotional behavioral contexts and provide further evidence for acoustic universals and potential for interspecies communication (Filippi, 2016; Filippi et al., 2017).

Temporal Emission Patterns: Evidence for Higher Order Vocal Constructs

Previous studies suggest that syllable sequences such as trains or phrases can encode combinational meaning or emphasis, thus increasing the available vocal complexity for a given bat species (e.g., Behr and Von Helversen, 2004; Bohn et al., 2008; Smotherman et al., 2016; Knörnschild et al., 2017). Sequences of syllables, which present higher order vocal constructs, have been described for a few bat species (for review see Smotherman et al., 2016). However, for the family Phyllostomidae, which is a very ecologically diverse and speciose bat family [i.e., >140 described species within 56 genera (Wetterer et al., 2000)], to date there have been only two published observations of the use of such hetero-syllabic constructs. Specifically, only for Seba's short-tailed bat (*Carollia perspicillata*) and the buffy flower bat (*Erophylla sezekorni*) descriptions of syllable combinations (i.e., simple trains and phrases) are available (Murray and Fleming, 2008; Knörnschild et al., 2014). Here we provide further evidence for syntax usage in a phyllostomid bat, which opens this family up for future in-depth research on this topic.

CONCLUSIONS

In the framework of this study, 18,658 high-quality social vocalizations of the pale spear-nosed bat, *P. discolor*, were

recorded under laboratory conditions. From 6,162 of these, it was possible to define eight robust syllable classes, including some vocalizations not previously known to be produced by these bats. Furthermore, we were also able to assess the behavioral contexts in which these syllable classes are generally emitted, and could show that e.g., high entropy syllables are exclusively emitted in aggressive encounters. We also describe four additional, rarely occurring syllable classes (i.e., 81 recordings in total). The majority of recorded syllables ($n = 12{,}416$) present evidence for a meta-class of vocalizations, i.e., syllables from different classes with the joint characteristic of having a suppressed fundamental frequency. Finally, we present tentative evidence for emission of syllable trains and phrases in this Neo-tropical bat species', highlighting the described complexity of *P. discolor* vocalizations. Together, these results present an extensive assessment of the vocal repertoire of *P. discolor* in a social roosting context and the associated behavioral contexts.

AUTHOR CONTRIBUTIONS

LW, ML, SV, and EL conceived and supervised the study. SS recorded the data. SS, EL, and ML developed the classification key. LW wrote the syllable detection and analysis program. EL and SS performed the syllable classification. MS conducted the statistical analyses and data presentation. JR rated the behavioral context. EL wrote the first draft of the manuscript. All authors contributed to the writing, editing, and revising of the final paper.

FUNDING

The research was funded by a Human Frontiers Science Program Research Grant (RGP0058/2016) awarded to LW and SV. SS was funded by the Fulbright U.S. Student Program.

ACKNOWLEDGMENTS

The authors want to thank Mirjam Knörnschild for helpful discussions concerning this manuscript and the LMU workshop for their help in constructing the recording setup. We would also like to thank the two reviewers for their helpful and constructive feedback on earlier versions of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fevo.2019.00116/full#supplementary-material

Supplementary Figure S1 | Syllable diversity in the commonly occurring syllable classes. The different commonly occurring classes contain syllables with some structural variety. Here we want to give an impression about the different shapes syllables from any class can take. From top left to bottom right, example oscillograms (top) and spectrograms (bottom) of the following are displayed: **(A)** noisy, **(B)** long, and **(C)** short high entropy syllables (HE), **(D)** linearly downward frequency modulated (lDFM) syllable, **(E,F)** non-linearly downward frequency modulated (nlDFM) syllables, **(G)** regular sinusoidally frequency modulated (SFM) syllable, **(H)** SFM syllables with a downward-frequency modulated onset, **(I)** ascending and **(J)** short SFM syllables, composite syllables (CS) with a noisy element **(K)** at the end or **(L)** within the syllable, **(M)** short, and **(N)** long quasi-constant frequency (qCF) syllables, **(O)** quasi-constant frequency syllable with a steep onset (qCF_so), and **(P)** noisy quasi-constant frequency (qCF_n) syllable.

Supplementary Table S1 | Measured and calculated acoustic parameters of the common syllable classes. For the 19 extracted spectro-temporal parameters, the 25th (Q25), 50th (Q50), and 75th (Q75) percentiles (i.e., first, second, and third quartiles) are reported to represent data distribution.

Supplementary Table S2 | Measured and calculated acoustic parameters of the rare syllable classes and the suppressed fundamental frequency class (SF). For the 19 extracted spectro-temporal parameters, the 25th (Q25), 50th (Q50), and 75th (Q75) percentiles (i.e., first, second, and third quartiles) are reported to represent data distribution.

Supplementary Table S3 | Behavioral contexts scored for 20 syllables per class. For each of the eight commonly occurring syllable classes, 20 videos were scored for the behaviors displayed by the bats during syllable emission. For the lDFM and nlDFM syllable classes only 19 instances could successfully be scored as the behavior during syllable emission was performed outside the field of view of the camera in the remaining two cases. A single vocalization from the HE class was emitted in a neutral behavioral context, which was scored as "brachiating on walls or ceiling". Other than that, all's syllables were emitted either in a prosocial or an antagonistic behavioral context. From the ethogram of 56 detailed behaviors, which was used for the behavioral scoring, only 23 behaviors were observed during syllable emission.

REFERENCES

Barclay, R. M. R., Fenton, M. B., and Thomas, D. W. (1979). Social behavior of the little brown bat, *Myotis lucifugus*. *Behav. Ecol. Sociobiol.* 6, 137–146

Bastian, A., and Schmidt, S. (2008). Affect cues in vocalizations of the bat, *Megaderma lyra*, during agonistic interactions. *J. Acoust. Soc. Am.* 124, 598–608. doi: 10.1121/1.2924123

Behr, O. (2006). The vocal repertoire of the sac-winged bat, *Saccopteryx bilineata*. Doctoral thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg.

Behr, O., and Von Helversen, O. (2004). Bat serenades - Complex courtship songs of the sac-winged bat (*Saccopteryx bilineata*). *Behav. Ecol. Sociobiol.* 56, 106–115. doi: 10.1007/s00265-004-0768-7

Bohn, K. M., Boughman, J. W., Wilkinson, G. S., and Moss, C. F. (2004). Auditory sensitivity and frequency selectivity in greater spear-nosed bats suggest specializations for acoustic communication. *J. Comp. Physiol. A Sens. Neural Behav. Physiol.* 190, 185–192. doi: 10.1007/s00359-003-0485-0

Bohn, K. M., Schmidt-French, B., Ma, S. T., and Pollak, G. D. (2008). Syllable acoustics, temporal patterns, and call composition vary with behavioral context in Mexican free-tailed bats. *J. Acoust. Soc. Am.* 124, 1838–1848. doi: 10.1121/1.2953314

Bohn, K. M., Wilkinson, G. S., and Moss, C. F. (2007). Discrimination of infant isolation calls by female greater spear-nosed bats, *Phyllostomus hastatus*. *Anim. Behav.* 73, 423–432. doi: 10.1016/j.anbehav.2006.09.003

Boughman, J. W. (1997). Greater spear-nosed bats give group-distinctive calls. *Behav. Ecol. Sociobiol.* 40, 61–70. doi: 10.1007/s002650050316

Boughman, J. W., and Wilkinson, G. S. (1998). Greater spear-nosed bats discriminate group mates by vocalizations. *Anim. Behav.* 55, 1717–1732. doi: 10.1006/anbe.1997.0721

Briefer, E. F. (2012). Vocal expression of emotions in mammals: mechanisms of production and evidence. *J. Zool.* 288, 1–20. doi: 10.1111/j.1469-7998.2012.00920.x

Davidson, S. M., and Wilkinson, G. S. (2004). Function of male song in the greater white-lined bat, *Saccopteryx bilineata*. *Anim. Behav.* 67, 883–891. doi: 10.1016/j.anbehav.2003.06.016

de Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* 111, 1917–1930. doi: 10.1121/1.1458024

Doupe, A. J., and Kuhl, P. K. (1999). Birdsong and human speech: common themes and mechanisms. *Annu. Rev. Neurosci.* 22, 567–631.

Esser, K. H. (1994). Audio-vocal learning in a non-human mammal: the lesser spear-nosed bat *Phyllostomus discolor*. *Neuroreport* 5, 1718–1720

Esser, K. H., and Lud, B. (1997). Discrimination of sinusoidally frequency modulated sound signals mimicking species specific communication calls in the FM bat *Phyllostomus discolor*. *J. Comp. Physiol. A* 180, 513–522.

Esser, K. H., and Schmidt, U. (1989). Mother-infant communication in the lesser spear-nosed bat *Phyllostomus discolor* (Chiroptera, Phyllostomidae) - evidence for acoustic learning. *Ethology* 82, 156–168.

Esser, K. H., and Schubert, J. (1998). Vocal dialects in the lesser spear-nosed bat *Phyllostomus discolor*. *Naturwissenschaften* 85, 347–349. doi: 10.1007/s001140050513

Filippi, P. (2016). Emotional and interactional prosody across animal communication systems: a comparative approach to the emergence of language. *Front. Psychol.* 7:1393.doi: 10.3389/fpsyg.2016.01393

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., et al. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proc. R. Soc. B Biol. Sci.* 284, 1–9. doi: 10.1098/rspb.2017.0990

Firzlaff, U., Schörnich, S., Hoffmann, S., Schuller, G., and Wiegrebe, L. (2006). A neural correlate of stochastic echo imaging. *J. Neurosci.* 26, 785–791. doi: 10.1523/JNEUROSCI.3478-05.2006

Forman, G., and Scholz, M. (2010). Apples-to-apples in cross-validation studies. *ACM SIGKDD Explor. Newsl.* 12:49. doi: 10.1145/1882471.1882479

Friard, O., and Gamba, M. (2016). BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods Ecol. Evol.* 7, 1325–1330. doi: 10.1111/2041-210X.12584

Gadziola, M. A., Grimsley, J. M. S. S., Faure, P. A., and Wenstrup, J. J. (2012). Social vocalizations of big brown bats vary with behavioral context. *PLoS ONE* 7:e44550. doi: 10.1371/journal.pone.0044550

Gould, E. (1975). Neonatal vocalizations in bats of eight genera. *J. Mammal.* 56, 15–29. doi: 10.2307/1379603

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Edn*. New York, NY: Springer New York.

Hechavarría, J. C., Beetz, M. J., Macias, S., and Kössl, M. (2016). Distress vocalization sequences broadcasted by bats carry redundant information. *J. Comp. Physiol. A* 202, 503–515. doi: 10.1007/s00359-016-1099-7

Heinrich, M., and Wiegrebe, L. (2013). Size constancy in bat biosonar? Perceptual interaction of object aperture and distance. *PLoS ONE* 8:e61577. doi: 10.1371/journal.pone.0061577

Hoffmann, S., Baier, L., Borina, F., Schuller, G., Wiegrebe, L., and Firzlaff, U. (2008). Psychophysical and neurophysiological hearing thresholds in the bat *Phyllostomus discolor*. *J. Comp. Physiol. A* 194, 39–47. doi: 10.1007/s00359-007-0288-9

Kanwal, J. S. (2009). "Audiovocal communication in bats," in *Encyclopedia of Neurosciences*, ed L. R. Squire (Oxford: Academic Press), 681–690.

Kanwal, J. S., Matsumura, S., Ohlemiller, K., and Suga, N. (1994). Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats. *J. Acoust. Soc. Am.* 96, 1229–1254. doi: 10.1121/1.410273

Knörnschild, M. (2014). Vocal production learning in bats. *Curr. Opin. Neurobiol.* 28, 80–85. doi: 10.1016/j.conb.2014.06.014

Knörnschild, M., Behr, O., and von Helversen, O. (2006). Babbling behavior in the sac-winged bat (*Saccopteryx bilineata*). *Naturwissenschaften* 93, 451–454. doi: 10.1007/s00114-006-0127-9

Knörnschild, M., Blüml, S., Steidl, P., Eckenweber, M., and Nagy, M. (2017). Bat songs as acoustic beacons - Male territorial songs attract dispersing females. *Sci. Rep.* 7, 1–11. doi: 10.1038/s41598-017-14434-5

Knörnschild, M., Feifel, M., and Kalko, E. K. V. (2014). Male courtship displays and vocal communication in the polygynous bat *Carollia perspicillata*. *Behaviour* 151, 781–798. doi: 10.1163/1568539X-00003171

Knörnschild, M., Glöckner, V., and Von Helversen, O. (2010b). The vocal repertoire of two sympatric species of nectar-feeding bats (*Glossophaga soricina* and *G. commissarisi*). *Acta Chiropterol.* 12, 205–215. doi: 10.3161/150811010X504707

Knörnschild, M., Nagy, M., Metz, M., Mayer, F., and Von Helversen, O. (2010a). Complex vocal imitation during ontogeny in a bat. *Biol. Lett.* 6, 156–159. doi: 10.1098/rsbl.2009.0685

Kwiecinski, G. G. (2006). *Phyllostomus discolor*. *Mamm. Species* 1–11. doi: 10.1644/801.1

La Val, R. K. (1970). Banding patterns and activity periods of some costa Rican Bats. *Southwest. Nat.* 15, 1–10.

Lattenkamp, E. Z., and Vernes, S. C. (2018). Vocal learning: a language-relevant trait in need of a broad cross-species approach. *Curr. Opin. Behav. Sci.* 21, 209–215. doi: 10.1016/j.cobeha.2018.04.007

Lattenkamp, E. Z., Vernes, S. C., and Wiegrebe, L. (2018). Volitional control of social vocalisations and vocal usage learning in bats. *J. Exp. Biol.* 221:jeb180729. doi: 10.1242/jeb.180729

Luo, J., Goerlitz, H. R., Brumm, H., and Wiegrebe, L. (2015). Linking the sender to the receiver: vocal adjustments by bats to maintain signal detection in noise. *Sci. Rep.* 5, 1–11. doi: 10.1038/srep18556

Luo, J., Lingner, A., Firzlaff, U., and Wiegrebe, L. (2017). The Lombard effect emerges early in young bats: implications for the development of audio-vocal integration. *J. Exp. Biol.* 220, 1032–1037. doi: 10.1242/jeb.151050

Ma, J., Kobayasi, K., Zhang, S., and Metzner, W. (2006). Vocal communication in adult greater horseshoe bats, *Rhinolophus ferrumequinum*. *J. Comp. Physiol. A* 192, 535–550. doi: 10.1007/s00359-006-0094-9

Murray, K. L., and Fleming, T. H. (2008). Social structure and mating system of the buffy flower bat, *Erophylla sezekorni* (Chiroptera, Phyllostomidae). *J. Mammal.* 89, 1391–1400. doi: 10.1644/08-MAMM-S-068.1

Nelson, J. E. (1964). Vocal communication in Australian flying foxes (Pteropodidae; Megachiroptera). *Z. Tierpsychol.* 21, 857–870. doi: 10.1111/j.1439-0310.1964.tb01224.x

Pfalzer, G., and Kusch, J. J. (2003). Structure and variability of bat social calls: implications for specificity and individual recognition. *J. Zool.* 261, 21–33. doi: 10.1017/S0952836903003935

Prat, Y., Taub, M., and Yovel, Y. (2016). Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. *Sci. Rep.* 6:39419. doi: 10.1038/srep39419

Rodenas-Cuadrado, P., Chen, X. S., Wiegrebe, L., Firzlaff, U., and Vernes, S. C. (2015). A novel approach identifies the first transcriptome networks in bats: a new genetic model for vocal communication. *BMC Genomics* 16, 1–18. doi: 10.1186/s12864-015-2068-1

Rodenas-Cuadrado, P. M., Mengede, J., Baas, L., Devanna, P., Schmid, T. A., Yartsev, M., et al. (2018). Mapping the distribution of language related genes FoxP1, FoxP2, and CntnaP2 in the brains of vocal learning bat species. *J. Comp. Neurol.* 526, 1235–1266. doi: 10.1002/cne.24385

Rother, G., and Schmidt, U. (1985). Die ontogenetische Entwicklung der Vokalisation bei *Phyllostomus discolor* (Chiroptera). *Z. Säugetierkd.* 50, 17–26. doi: 10.1017/CBO9781107415324.004

Russ, J. M., Jones, G., Mackie, I. J., and Racey, P. A. (2004). Interspecific responses to distress calls in bats (Chiroptera: Vespertilionidae): a function for convergence in call design? *Anim. Behav.* 67, 1005–1014. doi: 10.1016/j.anbehav.2003.09.003

Smotherman, M., Knörnschild, M., Smarsh, G., and Bohn, K. (2016). The origins and diversity of bat songs. *J. Comp. Physiol. A* 202, 535–554. doi: 10.1007/s00359-016-1105-0

Wetterer, A., Rockman, M. V., and Simmons, N. B. (2000). Phylogeny of phyllostomid bats (Mammalia: Chiroptera): data from diverse morphological systems, sex chromosomes, and restriction sites. *Bull. Am. Museum Nat. Hist.* 248, 1–200. doi: 10.1206/0003-0090(2000)248<0001:POPBMC>2.0.CO;2

Wilkinson, G. S. (1995). Information transfer in bats. in "*Ecology, Evolution and Behaviour Bats*" eds P. A. Racey and S. M. Swift *Symp. Zool. Soc. London* 67, 345–360.

Wilkinson, G. S. (2003). "Social and vocal complexity in bats," in *Animal Social Complexity: Intelligence, Culture and Individualize Societies*, Chapter 12, eds

F. B. M. de Waal, and P. L. Tyack (Cambridge, MA: Harvard University Press), 322–341.

Wilkinson, G. S., and Boughman, J. W. (1998). Social calls coordinate foraging in greater spear-nosed bats. *Anim. Behav.* 55, 337–350. doi: 10.1006/anbe.1997.0557

Wright, G. S., Chiu, C., Xian, W., Wilkinson, G. S., Moss, C. F., Gadziola, M., et al. (2013). Social calls of flying big brown bats (*Eptesicus fuscus*). *Front. Physiol.* 4:214. doi: 10.3389/fphys.2013.00214

Yost, W. A. (1996). Pitch strength of iterated rippled noise. *J. Acoust. Soc. Am.* 100, 3329–3335.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6.9 Supplementary material

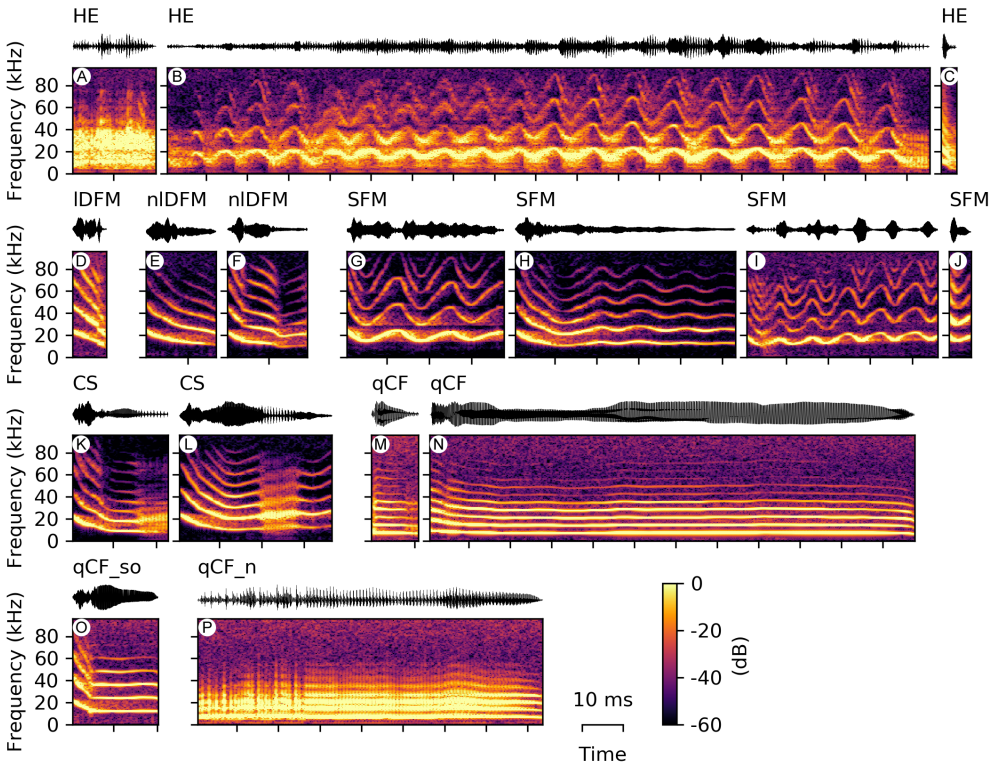

Figure 6.S1 Syllable diversity in the commonly occurring syllable classes. The different commonly occurring classes contain syllables with some structural variety. Here we want to give an impression about the different shapes syllables from any class can take. From top left to bottom right, example oscillograms (top) and spectrograms (bottom) of the following are displayed: **(A)** noisy, **(B)** long, and **(C)** short high entropy syllables (HE), **(D)** linearly downward frequency modulated (lDFM) syllable, **(E, F)** non-linearly downward frequency modulated (nlDFM) syllables, **(G)** regular sinusoidally frequency modulated (SFM) syllable, **(H)** SFM syllables with a downward-frequency modulated onset, **(I)** ascending and **(J)** short SFM syllables, composite syllables (CS) with a noisy element **(K)** at the end or **(L)** within the syllable, **(M)** short, and **(N)** long quasi-constant frequency (qCF) syllables, **(O)** quasi-constant frequency syllable with a steep onset (qCF_so), and **(P)** noisy quasi-constant frequency (qCF_n) syllable.

		HE (3860)			lDFM (727)			nlDFM (562)			SFM (445)		
		Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75
duration	(ms)	4.74	6.24	10.27	5.35	6.74	8.78	13.72	17.10	20.13	4.66	5.51	7.90
max. level	(dB)	−50.1	−44.2	−39.3	−47.3	−41.8	−36.9	−43.5	−37.8	−33.6	−40.7	−36.6	−31.3
spectral centroid	(kHz)	14.43	16.99	19.80	19.01	20.59	23.93	17.93	19.19	21.91	18.02	19.89	23.69
peak frequency	(kHz)	9.75	14.25	19.50	17.25	19.50	21.00	15.00	16.50	20.25	15.75	17.25	21.75
min. frequency	(kHz)	4.50	6.00	7.50	10.50	12.00	14.25	9.00	11.25	12.75	12.75	13.50	15.00
max. frequency	(kHz)	31.50	35.25	39.75	36.00	40.50	45.00	35.25	39.00	42.75	34.50	38.25	48.00
bandwidth	(kHz)	25.50	29.25	33.75	23.25	28.50	33.75	23.25	28.50	33.00	20.25	25.50	33.75
mean f_0	(kHz)	2.83	5.58	9.43	15.83	17.27	18.65	13.56	14.72	15.71	14.38	15.66	16.71
min. f_0	(kHz)	0.75	1.06	5.06	10.54	13.11	15.60	8.54	10.21	11.39	12.54	14.41	15.73
max. f_0	(kHz)	8.35	12.00	16.15	20.45	21.43	22.23	21.06	22.43	23.90	16.99	18.08	20.60
f_0 onset	(kHz)	1.66	9.89	15.01	20.45	21.43	22.23	21.06	22.43	23.90	16.65	17.78	19.82
f_0, LR slope	(kHz/ms)	−1.33	−0.11	0.11	−2.06	−1.70	−1.41	−1.07	−0.74	−0.53	−0.38	−0.15	0.05
f_0, LR intercept	(kHz)	2.06	7.27	13.64	21.73	22.66	23.43	19.72	20.83	21.55	15.02	16.19	17.63
f_0, QR quadr. coeff.	(kHz/ms^2)	−0.02	0.09	0.84	−0.11	−0.02	0.03	0.03	0.04	0.06	0.26	0.65	0.90
f_0, QR lin. coeff.	(kHz/ms)	−4.97	−0.92	0.18	−1.97	−1.48	−0.97	−1.80	−1.47	−1.12	−4.01	−2.81	−1.21
f_0, QR const. coeff.	(kHz)	2.69	8.64	16.67	21.41	22.52	23.30	21.72	22.68	23.46	17.72	19.13	20.61
f_0, LR error	(kHz)	0.59	1.94	3.29	0.19	0.26	0.35	0.62	0.87	1.07	0.36	0.73	1.36
f_0, QR error	(kHz)	0.43	1.56	2.69	0.14	0.20	0.27	0.33	0.43	0.58	0.10	0.19	0.57
mean aperiodicity	(1)	0.34	0.42	0.48	0.03	0.04	0.07	0.01	0.03	0.04	0.01	0.02	0.04

		CS (286)			qCF (67)			qCF_so (89)			qCF_n (126)		
		Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75
duration	(ms)	18.68	22.60	27.63	7.20	10.64	20.52	17.98	21.03	24.33	37.13	102.2	151.9
max. level	(dB)	−41.2	−36.3	−32.3	−60.7	−55.3	−52.5	−47.8	−41.5	−37.7	−60.1	−56.8	−54.1
spectral centroid	(kHz)	17.66	19.01	20.63	6.75	9.04	14.92	16.88	19.47	23.82	8.07	9.51	11.78
peak frequency	(kHz)	14.25	17.25	20.25	6.00	7.50	13.50	12.00	14.25	23.25	6.00	6.75	7.50
min. frequency	(kHz)	7.50	9.00	10.50	5.25	6.00	9.38	10.50	11.25	12.00	4.50	4.50	5.25
max. frequency	(kHz)	33.75	38.25	43.50	12.38	19.50	31.12	36.00	39.75	43.50	18.75	22.50	26.25
bandwidth	(kHz)	24.75	29.25	34.31	5.25	13.50	23.25	26.25	28.50	32.25	14.25	18.00	22.31
mean f_0	(kHz)	9.64	11.20	12.89	6.15	7.19	9.87	12.47	13.38	14.00	2.71	3.45	4.38
min. f_0	(kHz)	0.77	1.16	1.50	5.51	6.27	7.68	10.14	11.43	11.98	1.00	1.00	1.03
max. f_0	(kHz)	20.93	21.87	22.72	6.83	8.07	11.79	20.18	21.34	22.38	7.38	8.01	8.73
f_0 onset	(kHz)	20.86	21.78	22.65	6.37	7.62	9.97	20.18	21.34	22.38	4.03	5.39	6.96
f_0, LR slope	(kHz/ms)	−0.92	−0.75	−0.49	−0.20	−0.05	0.01	−0.51	−0.40	−0.28	−0.01	0.00	0.03
f_0, LR intercept	(kHz)	17.17	19.24	20.71	6.40	7.74	9.87	15.86	17.25	18.06	1.85	3.10	4.30
f_0, QR quadr. coeff.	(kHz/ms^2)	0.00	0.02	0.03	−0.01	0.00	0.02	0.03	0.05	0.06	0.00	0.00	0.00
f_0, QR lin. coeff.	(kHz/ms)	−1.50	−1.07	−0.61	−0.26	−0.10	0.05	−1.68	−1.43	−1.09	−0.12	−0.01	0.04
f_0, QR const. coeff.	(kHz)	19.18	20.75	22.20	6.39	7.84	10.37	19.52	20.98	21.95	2.07	3.94	5.19
f_0, LR error	(kHz)	1.52	1.97	2.53	0.10	0.19	0.42	1.33	1.54	1.73	2.13	2.43	2.67
f_0, QR error	(kHz)	1.19	1.78	2.31	0.07	0.13	0.28	0.50	0.70	0.85	1.89	2.25	2.57
mean aperiodicity	(1)	0.06	0.09	0.12	0.01	0.02	0.05	0.01	0.01	0.02	0.12	0.17	0.23

Table 6.S1 Measured and calculated acoustic parameters of the common syllable classes. For the 19 extracted spectro-temporal parameters, the 25th (Q25), 50th (Q50), and 75th (Q75) percentiles (*i.e.*, first, second, and third quartiles) are reported to represent data distribution. f_0: Fundamental frequency. LR: Linear regression. QR: Quadratic regression.

		Puff (42)			**V-shaped (30)**			**qCF_nso (5)**			**hFM (4)**			**SF (12 416)**		
		Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75	Q25	Q50	Q75
duration	(ms)	19.26	20.81	21.84	19.20	21.99	25.13	24.72	29.79	54.05	11.42	13.43	14.92	3.46	4.07	5.04
max. level	(dB)	−68.4	−66.4	−64.7	−38.8	−36.6	−34.8	−56.1	−55.3	−45.3	−48.0	−44.2	−38.4	−61.2	−57.2	−52.3
spectral centroid	(kHz)	25.57	27.63	30.22	18.84	20.00	21.93	12.13	14.45	16.71	21.71	27.08	33.09	40.65	43.05	46.51
peak frequency	(kHz)	23.44	25.12	26.81	15.00	19.50	21.56	7.50	11.25	14.25	18.94	23.25	30.00	37.50	39.75	40.50
min. frequency	(kHz)	1.50	1.50	2.25	9.19	9.75	10.50	5.25	5.25	6.00	10.50	12.00	15.38	27.75	29.25	30.75
max. frequency	(kHz)	44.25	47.25	51.94	36.75	40.88	43.88	31.50	33.75	38.25	39.19	48.38	58.88	64.50	72.00	78.75
bandwidth	(kHz)	42.00	45.75	48.75	27.00	30.00	33.56	24.00	28.50	33.00	29.06	31.50	38.25	34.50	43.50	50.25
mean f_0	(kHz)	14.45	25.35	26.28	12.49	12.93	13.48	4.68	5.72	6.68	14.98	15.52	17.50	18.73	19.84	34.85
min. f_0	(kHz)	1.04	9.56	23.01	9.37	10.02	10.54	1.03	1.06	1.09	9.56	11.66	13.60	16.29	17.78	32.22
max. f_0	(kHz)	25.60	27.73	32.00	20.62	21.53	22.21	16.07	16.15	16.53	19.33	20.06	21.97	21.09	23.62	37.77
f_0 onset	(kHz)	17.97	23.90	26.67	20.62	21.53	22.21	16.07	16.15	16.53	18.38	19.50	21.56	21.05	23.54	37.66
f_0, LR slope	(kHz/ms)	−0.21	−0.07	0.07	−0.45	−0.29	−0.20	−0.65	−0.58	−0.13	−1.36	−1.09	−0.98	−3.70	−2.38	−1.80
f_0, LR intercept	(kHz)	15.05	25.73	27.71	15.44	16.08	17.07	7.97	11.17	14.98	20.62	21.38	24.04	22.30	24.09	40.46
f_0, QR quadr. coeff.	(kHz/ms^2)	−0.04	−0.02	0.00	0.04	0.06	0.08	0.01	0.02	0.06	−0.10	−0.05	0.11	−0.31	0.08	0.72
f_0, QR lin. coeff.	(kHz/ms)	−0.09	0.19	0.97	−1.98	−1.56	−1.23	−2.22	−1.31	−0.80	−1.47	−0.19	0.20	−5.26	−2.54	−1.17
f_0, QR const. coeff.	(kHz)	14.86	24.12	27.15	19.86	21.02	22.23	14.21	18.56	20.63	20.24	22.16	23.96	22.40	25.58	41.08
f_0, LR error	(kHz)	1.33	1.97	3.87	1.98	2.10	2.19	2.58	3.56	3.58	0.84	0.96	1.13	0.16	0.27	0.47
f_0, QR error	(kHz)	1.23	1.89	3.35	0.59	0.79	1.13	1.71	2.14	2.31	0.41	0.65	0.83	0.10	0.17	0.31
mean aperiodicity	(1)	0.40	0.43	0.47	0.02	0.03	0.03	0.13	0.13	0.15	0.03	0.06	0.10	0.11	0.16	0.24

Table 6.S2 Measured and calculated acoustic parameters of the rare syllable classes and the suppressed fundamental frequency class (SF). For the 19 extracted spectro-temporal parameters, the 25th (Q25), 50th (Q50), and 75th (Q75) percentiles (*i.e.*, first, second, and third quartiles) are reported to represent data distribution. f_0: Fundamental frequency. LR: Linear regression. QR: Quadratic regression.

Syllable class	%	Prosocial behaviour	*N*	%	Antagonistic behaviour	*N*
HE	0			95	mock biting	1
					defense of roostposition	1
					mate defense	4
					vocal protest	3
					vocal admonishment	7
					turning/movement towards	1
					physical escape from aggression	1
					physical avoidance of aggression	1
lDFM	63	vocal contact	8	37	vocal protest	4
		mate guarding	1		vocal admonishment	1
		attention seeking	1		defense of roost position	1
		general approach	2		mate defense	1
nlDFM	84	vocal contact	8	16	vocal admonishment	2
		looking at/turning towards	2		lunging	1
		grooming	2			
		attention seeking	2			
		general approach	2			
SFM	75	vocal contact	4	25	vocal admonishment	1
		attention seeking	11		lunging	1
					rejection of advances	1
					vocal protest	2
CS	60	attention seeking	5	40	vocal admonishment	7
		vocal contact	7		turning/movement towards	1
qCF	85	vocal contact	14	15	vocal protest	1
		grooming	1		turning/movement towards	1
		vocal protest	1		mate defense	1
		attention seeking	1			
qCF_so	55	vocal contact	6	45	mock biting	3
		mate approach	1		turning/Movement towards	1
		attention seeking	1		mock biting	3
		cuddling	1		mate defense	2
		sniffing other bat	1			
		grooming	1			
qCF_n	40	attention seeking	4	60	vocal admonishment	6
		vocal contact	2		defense of roost position	3
		general approach	1		vocal protest	2
		looking at/turning towards	1		mate defense	1

Table 6.S3 Behavioral contexts scored for 20 syllables per class. For each of the eight commonly occurring syllable classes, 20 videos were scored for the behaviors displayed by the bats during syllable emission. For the lDFM and nlDFM syllable classes only 19 instances could successfully be scored as the behavior during syllable emission was performed outside the field of view of the camera in the remaining two cases. A single vocalization from the HE class was emitted in a neutral behavioral context, which was scored as "brachiating on walls or ceiling". Other than that, all syllables were emitted either in a prosocial or an antagonistic behavioral context. From the ethogram of 56 detailed behaviors, which was used for the behavioral scoring, only 23 behaviors were observed during syllable emission. %: Percentage of videos scored in either a prosocial or antagonistic behavioural context. *N*: Number of videos scored for the corresponding behaviour.

CHAPTER 7

Discussion

7.1 Vision and the perception of room acoustics

There is evidence that in the adaptation of the auditory system to reflections (as observed in the buildup of the precedence effect, see section 1.4.1), a persistent temporal and spatial pattern of the reflected sound is important (Clifton and Freyman, 1997; Clifton, Freyman, Litovsky, *et al.*, 1994). Keen and Freyman (2009)[1] suggested that "[r]*eflected sound* [...] *is analyzed to form a model of the auditory space*", a hypothetical process they termed *"room-acoustics model"*. The concept of *"room learning"* (proposed by Seeber, Müller, *et al.*, 2016; see also Seeber and Clapp, 2020) is similar, but explicitly addresses the fact that when a listener behaves naturally inside a room, the delays and directions of the reflections vary with his or her ever-changing position and orientation.

Theoretically, the visual image of the room could contribute to the formation of any such models. Pictures and videos have already been shown to inform human subjects about expected reverberation characteristics (McCreery and Calamia, 2006; Valente and Braasch, 2008). Inspired by evidence from the literature that there are adaptive suppressive processes that deal with reverberation (Brandewie and Zahorik, 2010; Watkins, 2005a,b; see section 1.4.2), the research question underlying Chapter 2 is whether additional visual stimulation can thus alter the perception of reverberation of a sound.

In this experiment, I presented speech in virtual reverberant environments together with either a matching, a mismatching, or no visual stimulus, and asked listeners to judge the magnitude of perceived reverberation on a scale from 1 to 10. If visual stimulation had an influence, ratings should have decreased in audiovisually congruent and increased in incongruent conditions, analogous to how the fusion of echoes in the precedence effect builds up with a sustained spatial arrangement of leading and lagging sound and breaks down when the pattern changes (Clifton, 1987). It is important to note, however, that the experimental design did not preclude any other possible effects of the visual stimulation: Alternatively, one might reasonably assume that an integration of matching auditory and visual stimuli might make reverberation more easily detectable (much like a sound presented in synchrony can enhance the visibility of a light; *e.g.* Bolognini *et al.*, 2005), or even that the listeners' ratings of a trial could be dominated by the identity of the visual stimulus (as in the ventriloquism effect, see section 1.5).

7.1.1 Summary of the findings

The data presented in Chapter 2 do not lend support to any such hypothesis of audiovisual integration in the perception of room acoustics. While listeners were able to meaningfully rate reverberation on a numerical scale across different acoustic scenes, a modification of the visual stimulus alone did not produce significant changes of their responses. Specifically, I neither observed differences in rating between room-acoustically identical audio-only and audiovisually congruent trials, nor between pairs of audiovisual trials that only deviated from each other in one visual aspect (namely the visually presented room size, source distance, or source azimuth).

The only conspicuous result was apparent between the ratings of room-acoustically identical trials with congruent visual stimuli *vs.* visual stimuli with a mismatch of sound source azimuth, which

[1] These are the same authors; Rachel Keen published under the name Rachel K. Clifton until the year 2002.

were slightly (though not significantly at $\alpha = 0.05$) lowered with angular disparities of 60° and 90°. As the Discussion section of Chapter 2 already suggested, this is probably not an indication for a multisensory effect, but rather an incidental finding related to the head-orientation benefit (HOB) to speech intelligibility: Kock (1950) and, more recently, Grange and Culling (2016) concordantly reported that the ability of listeners to understand speech in the presence of a single source of noise is improved when they turn their heads so as to create a difference in the binaural patterns (ITDs and ILDs) of the signal *vs.* the noise. When the distractor is not a noise source from a single location in space, but reverberant sound arriving from all spatial directions, it should simply be most effective to point one ear directly at the sound source, as this maximises signal level (and consequently the ratio of signal level to noise level) at that ear. The strategy spontaneously employed by hearing-impaired listeners with their better ear to maximise speech understanding, who preferentially pointed their nose 60° away from the sound source they were attending to (Brimijoin, McShefferty, *et al.*, 2012), agrees well with this consideration. It is easy to imagine that the perception of room acoustics is subject to a similar HOB: If one ear is acoustically favoured by yielding a better signal-to-noise ratio, judgments of reverberation are probably based on this ear too. While I did not explicitly test this hypothesis, listeners in the experiment were asked to look at the visual source when visual stimuli were presented, and reverberation ratings were indeed most strongly decreased in conditions with a 60° disparity between visual and acoustic source azimuth.

7.1.2 Non-dominance of vision in an audiovisual task

As pointed out in section 1.5, vision has frequently been found to dominate over audition in audiovisual tasks. This holds especially when they relate to some aspect of spatial perception, where the ventriloquism effect (Howard and Templeton, 1966) serves as the canonical example—though others are easy to find, such as in distance perception (*e.g.* Mendonça *et al.*, 2016), recalibration of auditory space after a shift of visual space (*e.g.* Recanzone, 1998), or room size estimation (Maempel and Jentsch, 2013). The counterexample presented in Chapter 2 might seem like an unexpected departure from this rule. However, the modern view holds that *"it is reliability, not vision, that captures auditory localization"* (Witten and Knudsen, 2005), and it can be assumed that hearing allows one to be a more reliable judge of reverberation than seeing does: After all, there are many complex factors that influence the acoustical properties of an enclosed space, and the appearance of a room is not necessarily a good predictor of what it sounds like.

In this view, the dominance of audition in the present experiment does not appear so unusual. In fact it is well in line with some other recent studies which investigated the audiovisual perception of attributes that relate to the listening environment: Gil-Carvajal *et al.* (2016) found that visual awareness of a room (in which the location of the sound source was not apparent) does not affect assessments of source azimuth and "compactness". In a manuscript which has not yet been subject to peer review, Libesman *et al.* (preprint) reported that listeners did not take visually mediated sound source distance information into account when judging the loudness of a sound. Postma and Katz (2017a) tested the influence of visual stimulation on plausibility, apparent source width, listener envelopment and source distance estimates in an auralisation of a theatre, but only found an effect on distance. Most recently, Salmon *et al.* (2020) had listeners rate the similarity of pairs of sound stimuli with varying visual stimulation (which was, as in the Chapter 2 study), presented in virtual reality via a head-mounted display). Their results (discussed below) also suggest that the subjects payed no regard to the visuals.

7.1.3 Alternative explanations for the results

Rather than immediately accepting the negative results in Chapter 2 as proof of the absence of a visual influence on the perception of reverberation, one needs to consider that the experiment may simply

have been unsuitable to reveal the hypothesised effect. For example, if the effect existed but were very small, the 1–10 rating scale might have been too coarse for listeners to express any differences they perceived; or the immersion of the listeners in the virtual environment might have been insufficient for multisensory integration to occur; or perhaps shortcomings of the simulation of the VR audiovisual scenes forbid an interpretation of the results in a real-world context (*cf.* Maempel and Horn, 2018). Given the relatively long stimuli which were presented to the listeners, it is also possible that purely auditory processes of reverberation suppression (see section 1.4.2) were already active and shadowed any possible influence of the visual modality.

There is also a chance that the instructions (clearly pointing to an auditory task) and the familiarisation protocol (which was based on purely auditory stimuli) might have focused the attention of the listeners in the main experiment on their sense of hearing more than it would be the case in a natural listening situation. As the three simulated room environments were very different in their reverberation times, it is even conceivable that some subjects might have quantified the duration of the reverberant tail (*e.g.*, by counting the seconds until they could no longer hear it), even though reverberation time was never explicitly mentioned in the instructions. Moreover, it is known that aspects of attention can enhance or reduce the precedence effect (London *et al.*, 2012; Wallmeier, Geßele, *et al.*, 2013), which could translate to the perception of reverberation.

Some of these hypotheticals may be discounted by looking at the results in the context of similar findings. The study by Salmon *et al.* (2020) in particular is comparable in its underlying idea, but it also differs in some important ways: These authors asked subjects for direct pairwise perceptual comparisons of audiovisual stimuli, using dissimilarity ratings on an unmarked, continuous scale. This protocol could be expected to bring subtle perceptual differences to light more effectively than the 1–10 scheme in my experiment could, and to avoid an inordinate reliance of listeners on one perceptual attribute like reverberation time. They also used recorded sounds and visuals of real rooms rather than synthetic ones, precluding concerns that a low quality of simulations could restrict the interpretability of their findings. Their findings were nonetheless consistent with those under discussion here: High dissimilarity scores could be explained almost completely by acoustical differences, whereas different visual stimuli generated near-identical judgments given the same sound stimuli.

In fact, to my knowledge, no study until now has shown any effect of visual stimulation on auditory perceptual attributes unless they are related to the spatial location of the source. The possibility remains, however, that the visual impression of a room might modulate processes other than conscious judgments of room acoustics: For example, recently published data showed that ratings of perceived reverberance are not necessarily a good proxy measure for performance in a speech understanding task (Ellis and Zahorik, 2019). This implies that a more complete investigation of audiovisual integration would have to include measurements of more than one dependent variable.

7.2 Translational motion and the auditory perception of distance

As an individual moves through space along a line different from the line which passes through two objects, the azimuthal angle between him or her and the nearer object changes faster than that between him or her and the more distant object. This well-known phenomenon is called *motion parallax* and has been shown to aid visual depth perception in animals, for example in locusts (Sobel, 1990), in pigeons (Xiao and Frost, 2013), and in humans (Rogers and Graham, 1979). Past studies of a possible exploitation of motion parallax in human hearing have looked at absolute distance perception (see section 1.5.3). This is remarkable considering that the derivation of an absolute distance measure from a parallax angle requires a comparison to a reference, which is nontrivial even in vision (*cf.* Hagen and M. Teghtsoonian, 1981; Ono *et al.*, 1986).

In the study presented here in Chapter 3, listeners were confronted with two sound sources which were positioned at different distances at 0° both in azimuth and elevation, whose emissions did not

overlap in time, and which differed in fundamental frequency (and consequently in pitch). The listeners had to detect which of the two sound sources in each trial was closer to them, without having to determine its distance in absolute terms. It seems clear this relative discrimination task could be solved with a more straightforward use of motion parallax information.

7.2.1 Summary of the findings

Listeners performed poorly when they were sitting still. This was expected, as sound level would constitute the only reliable cue for distance in an anechoic environment such as the one in the experiment, yet this parameter was deliberately randomised for each sound source in each trial. With a trained left–right swaying movement of the shoulders and head, however, performance improved in every single subject, especially as the distance between the two sound sources was increased. In the aggregate, with this motion, the listeners could already solve the problem of relative distance discrimination with an accuracy of 75 % when one speaker was 56 cm and the other 40 cm away, the most difficult pair of distances; when they did not move, they never reached this threshold even in the easiest condition of 98 cm *vs.* 30 cm.

A second experiment in virtual acoustic space, based on a horizontal loudspeaker array and a motion platform, made it possible to additionally move listeners passively (such that vestibular self-motion signals were available to the listeners) and purely virtually (by only moving the sound sources in VAS, but keeping the listeners stationary) while always presenting the same auditory cues. Performance was best with active motion, worst with pure sound source movement, and intermediate when listeners were moved by the platform. This suggests that a precise self-motion trajectory, which arises more readily from combined proprioceptive–vestibular information than from the vestibular system alone (see Mergner and Rosemeier, 1998), is beneficial for the correct interpretation of auditory motion parallax.

7.2.2 Results in context of related research

Simpson and Stanton (1973) already studied auditory distance perception by listeners whose heads were stationary *vs.* allowed to move, but it is not entirely clear whether any motion occurred that was suitable to elicit a motion parallax. In one of the two experiments described in the paper, they wrote that *"subjects were told of the advantage of head movement, given demonstrations, and were encouraged and reminded (but not forced) during the session to make whatever head movements they thought helpful"*; elsewere, they elaborated that these demonstrations were of *"rotate, tip and pivot movements"*, consistent with the terminology of Thurlow and Runge (1967) for purely rotational motion along the three main head axes. Simpson and Stanton (1973) did not report which kinds of movement they actually observed in their listeners, but they presumably did not include any displacements of the head.

The first examiners whose listeners definitely had the opportunity to exploit auditory cues which varied with translational self-motion thus appear to be Speigle and Loomis (1993) and Ashmead *et al.* (1995). Speigle and Loomis (1993) had listeners approach a sound source while blindfolded in an outdoors environment (*i.e.*, with few auditory distance cues) and turned the sound off after they were still some distance away. They then had to walk to the place where they thought the sound had come from. Their accuracy in this task hardly differed from one where they listened from a stationary position. In contrast to this finding, Ashmead *et al.* (1995), in a very similar experiment, reported that relative errors of listeners' distance estimates fell by about 45 % in a walking *vs.* stationary condition. One notable difference is that Speigle and Loomis (1993) kept the level of the sound constant at the source position, whereas Ashmead *et al.* (1995) varied it randomly. The latter argued that familiarity with the sound levels associated with various distances might have offered Speigle and Loomis's (1993) subjects a cue that was reliable enough such that the dynamic changes provided no additional benefit.

The Chapter 3 study is aligned with Ashmead *et al.*'s (1995) both in that motion-independent auditory distance cues are minimised, and in that motion was shown to be beneficial under such circumstances. However, the nature of the motion-based cues that the listeners utilised were probably quite different. In Ashmead *et al.* (1995), the walking trajectories always brought them closer to the source. In some conditions, they walked directly towards it, so that motion parallax would not have occurred at all; in others, the source was not at 0° in azimuth, but sound level still systematically increased with every step the listeners took. This situation is similar to that in time-to-contact studies (*e.g.* Rosenblum *et al.*, 1993; Shaw *et al.*, 1991), and based on the ecological importance of looming sounds (Neuhoff, 1998), it is likely that dynamic changes in level at the ear are a more salient cue than motion parallax.

In the two present experiments, it was mostly the azimuthal angles of the two sound sources which varied with the listeners' (or, in one condition, the sources') side-to-side swaying motion. At a glance, this may seem unusual and perhaps difficult to exploit; after all, binaural cues are most effective for azimuthal information, and those are not substantially dependent on distance. In vision, on the other hand, the evaluation of the angular discrepancies due to the slightly different viewpoints provided by the two eyes (*i.e.*, binocular parallax) is entirely natural (stereopsis; *e.g.* Blakemore, 1970). Notably, for distances up to about 1 m, even the auditory system is able to estimate the distance of a sound source based on the differences in the incidence angles at the two ears, presumably by comparing azimuth information given by the two monaural spectral (HRTF) cues (H.-Y. Kim *et al.*, 2001). This is prior evidence that the computation of depth from two different azimuthal angles at two different points in space is a familiar task for the sound localisation system.

Another interesting parallel can be drawn to studies of auditory motion speed discrimination. Altman and Viskov (1977) and Grantham (1986) found JNDs of approximately 10 °/s. In the at-threshold active-motion condition of the first experiment from Chapter 3, the two sources moved at rates of 29.9 °/s and 22.3 °/s relative to the listener, a difference of 7.6 °/s, which is arguably of the same order as these JNDs. Of course, direct comparisons are difficult, as the stimuli in these earlier experiments were different in spectral content, duration, and the ratio of speed difference to absolute speed. Nevertheless, if the mechanisms that the listeners used to solve the tasks were similar in nature, the second experiment from Chapter 3 would suggest that active motion reduces such speed discrimination thresholds.

The apparent differences in performance between active and passive motion in this second (VAS) experiment are also interesting in that they suggest an influence of proprioception (sensory inputs from the musculoskeletal system) and/or motor commands (the signals emitted by the brain to move the upper body) on sound localisation. It is conceivable that proprioceptive signals/efference copies (duplicates of those motor signals which remain in the central nervous system) are integrated with auditory spatial processing, such that the efferent signals give rise to more reliable head position estimates which may aid the evaluation of the dynamically changing binaural cues. Such circuits have been proposed before; in fact, the general idea can be traced back to Wallach (1939, 1940). Recently, Genzel, Firzlaff, *et al.* (2016) have drawn on this hypothetical integration process to explain performance differences in sound localisation by listeners who turned their head actively, passively, or had their active head rotations counteracted by a rotating chair. Pastore *et al.* (2020) have pointed out that while the concept is well-established in vision (see *e.g.* Furman and Gur, 2012), physiological evidence for it is still lacking when it comes to the auditory system. The present findings regarding auditory motion parallax, however, are an addition to an already substantial basis of behavioural evidence (*e.g.* Brimijoin and Akeroyd, 2012, 2014; Freeman, Culling, *et al.*, 2017).

Finally, while the listeners had no difficulty in using auditory motion parallax in these two experiments, it is not clear whether they would naturally do so in everyday life. That listeners actively and intuitively seek out dynamic cues to aid distance perception was previously observed by Rébillat *et al.*

(2012): They remarked that when subjects were asked to estimate the distance to an object in a virtual environment, they consistently walked without being explicitly instructed to.

7.3 Stimulus characteristics and the auditory perception of looming sounds

Past investigations of the perception of looming sounds (see section 1.2.4) have often been based on very simplistic stimuli, such as pure tones, noises or vowels which were straightforwardly varied in amplitude. Based on just such a study, for example, Neuhoff (1998) famously proposed that the auditory system is biased towards overestimating approaching (compared to receding) sources. In the real world, of course, the physical processes implied by sound source motion alter sound in a more complex pattern. This was recognised by Neuhoff *et al.* (2009) and Bach, Neuhoff, *et al.* (2009), whose stimuli incorporated *"absolute decay* [...]*, Doppler shift, atmospheric filtering, gain attenuation due to atmospheric spreading, ground reflection attenuation, and head-related transfer function"* (Bach, Neuhoff, *et al.*, 2009). Even in those experiments, however, alterations of the stimulus due to time-varying changes of the emission, or due to the presence of enclosing surfaces, were not considered.

In Chapter 4 and section 5.6, I looked into the effects of two very different stimulus manipulations: The first is a modulation of a repetition rate, based on a behaviour observed in rattlesnakes, which were showed to increase the rate with which they flick their tails as an object approaches them. The second is an inclusion of the sound reflections that are characteristic of an enclosed space.

7.3.1 Summary of the findings

In Chapter 4, I compared human distance perception based on looming sounds which emitted repetitive rattle-like bursts of clicks at a constant rate *vs.* ones whose rate was modulated by distance—specifically, increased with decreasing distance, as evidenced in behavioural experiments with western diamondback rattlesnakes. The human psychoacoustic experiment showed that the listeners estimated the sound source to be just 1 m away sooner (*i.e.*, at a greater actual remaining distance) in the rate-modulated condition. I regard this as evidence for an underestimation of distance caused by faster rattling. The distance-to-rate mapping notably included a discontinuity at 4 m, where (again consistently with rattlesnake behaviour) the rattling rate increased abruptly at the moment where the distance fell below this threshold, and which appeared to have an additional startling effect which often seemed to suggest imminent contact.

Loudness is well-known to be integrated over time, effectively by summing up momentary impressions over approximately 0.1–0.2 s (see Buus *et al.*, 1997 for a review). In Chapter 4, it was suggested that this phenomenon may be part of the explanation for why the listeners' auditory distance percepts were compressed in the rate-modulated condition—via an increase in loudness despite the constant level of each rattling sound. This analysis, however, should be taken with a grain of salt: The loudness predictions of the model used (B. C. Moore, Glasberg, *et al.*, 2016) do not consider the looming bias. I am, in fact, not aware of any published model of perceived loudness which could reproduce even the pioneering finding of Neuhoff (1998) that stimuli which increase in level lead to greater changes in loudness than their time-reversed counterparts. Despite this, it is clearly appropriate to assume that loudness integration continues to occur when a sound source moves closer; in fact, if the looming bias is mediated by loudness, its effects may well be amplified.

The paradigm employed in section 5.6 was one of motion distance discrimination in looming *vs.* receding and anechoic *vs.* echoic conditions. Since the duration of all presented stimuli was the same, it could equivalently be described as motion speed discrimination (*cf.* Neuhoff, 2016). Consistently with this reference, discrimination performance was better for approaching than for receding trajectories; in fact, some listeners failed to discriminate a linear receding movement over 2.1 m from one over 5.9 m, while they could all do this near-perfectly with approaching sources. The data from 6 listeners was insufficient to draw definitive conclusions on differences between the anechoic and echoic

conditions, but they lend support to the hypothesis that the presence of reflections (possibly via the availability of an additional DRR cue) moves the performance for receding stimuli somewhat towards the—still significantly better—results attained in the looming condition.

7.3.2 Remarks on the rate-modulating behaviour of rattlesnakes

The rattle of rattlesnakes is purported to have *"evolved via elaboration of a simple behavior"* (Allf *et al.*, 2016): Many snakes, especially those in the viper family which includes the Crotalus and Sisturus genera which together make up the rattlesnakes, shake their tails when agitated (B. A. Young, 2003). Out of this behaviour, which could be aposematic by itself, the rattle at the end of the tail might have emerged by natural selection for resilience against biting attacks and for a more imposing acoustic signal (see Reiserer and Schuett, 2016 for a review).

The rate modulation of rattlesnake tail vibrations with distance (Chapter 4), or possibly more generally with perceived level of threat, have not been previously described, but it might be an additional stage in this hypothetical evolutionary process. In the absence of a rattle, this adaptive behaviour would probably be much less effective. To support the view that it has arisen as an interspecies acoustic communication system which can convey distance information, it could be valuable to confirm that snakes which vibrate their tails but do not have a rattle do not show this behaviour. It might be especially interesting to look for the presence of the two modes that were referred to as low-frequency (LF) and high-frequency (HF) rattling in Chapter 4 in different species: According to B. A. Young and I. P. Brown (1995), the former is associated with front–back vibrations and the latter with vibrations from side to side, a nuance of behaviour that appears to be of little use without a rattle.

7.3.3 Remarks on the looming bias in the context of room acoustics

A first study which considered room-acoustical cues in the context of auditory looming perception was published very recently (Wilkie and Stockman, 2020), after the data described in section 5.6 were acquired. Wilkie and Stockman (2020) collected time-to-impact estimates from listeners in various conditions, where one factor was the type of cue that was used to indicate approach (only level cues, only binaural cues, only DRR cues, any pair of these cues, and all three cues at once). As an inevitable consequence of the chosen time-to-impact paradigm, this experiment did not include a receding condition for comparison; moreover, the description of the data analysis unfortunately suggests a certain lack of rigour. In any event, the results clearly show that changes in level were necessary for the looming bias: Without them, the expected underestimation of the time to impact did not occur. The data also suggest that additional DRR cues further increased this bias, while binaural cues—unsurprisingly, given that all stimuli were presented at 0° in azimuth—had no influence.

Due to differences in the dependent variables and stimuli, it is difficult to compare my results from section 5.6 with these findings. The drawn conclusions, however, are quite different: Wilkie and Stockman (2020) reasoned that additional auditory spatial cues lead to a stronger manifestation of the looming bias. On the other hand, I suggested that they would facilitate more accurate distance judgments, particularly for receding stimuli, in spite of the bias. These ideas are not necessarily contradictory, but a unified experiment would be needed to test them both.

In spite of the scarcity of the available literature and the limitations of the present experiment, it is apparent that the size of any effect of room-acoustical cues on the perception of looming *vs.* receding sounds must be minor. Specifically, such an interaction clearly does not outweigh the strong effects of the looming bias.

7.4 Social communication in bats and the study of vocal learning

There are compelling arguments to closely consider bats in the investigation of mammalian vocal learning (Knörnschild, 2014; Lattenkamp and Vernes, 2018). When looking for evidence for the capability of an animal species to acquire new vocalisations based on auditory input, it is generally highly useful to know which vocalisations are commonly emitted by its individuals. Chapter 6 compiled such a repertoire for the pale spear-nosed bat, Phyllostomus discolor, for which indications of vocal learning have already been published (Esser, 1994; Esser and U. Schmidt, 1989).

7.4.1 Summary of the findings

From 18 658 vocalisations, recorded from six adult pale spear-nosed bats in social settings, eight distinct and common syllable classes could be extracted based on a combination of a visual classification of spectrograms by human raters and an automatic classification of computer-extracted acoustical parameters. In addition, there were four further distinct, but rare syllable classes, and apparent combinations of syllables to form trains (repetitions of syllables in the same class) and phrases (composed of more than one syllable class).

In principle, it is conceivable to develop the automatic classifier further, in a way that would allow it to be used to quickly identify candidate observations of novel vocalisations in the sense of vocal learning. I have already made such an attempt in the context of an upcoming publication involving a longitudinal survey of some bats, based on their vocalisations as juveniles and as adults, but ultimately abandoned it because a suitable method to properly preprocess the data (in a way that is robust to the effects of different recording environments) proved elusive. In all likelihood, the design of a classifier that would be useful outside of the specific work presented in Chapter 6 would require major additional work.

Lattenkamp, Vernes, and Wiegrebe (2020) have since provided additional evidence for some capacity of vocal learning in these bats: They demonstrated that at least one individual altered the fundamental frequency of its calls so as to match an acoustic template.

7.5 Room-acoustical simulation in real time and a brief outlook

In Chapter 5, I have described the conception and development of liveRAZR, a new program to present virtual sound sources and their reflections from the walls of simulated rooms with a straightforward geometry, in real-time scenarios where both the listener and the virtual sources are allowed to rotate and move through space. For computational efficiency, it combines an image source model for shoebox rooms (Allen and Berkley, 1979) with a feedback-delay network (Jot and Chaigne, 1991). This implementation is designed to be identical to the RAZR model, which in its original incarnation (Wendt *et al.*, 2014) could only simulate impulse responses for single sound sources in stationary source–receiver geometries.

Contrary to a number of previous related works, liveRAZR avoids the usual two-step approach in which the room-acoustical simulation generates a set of finite impulse responses which is then handed over to an auralisation engine (*e.g.* Funkhouser *et al.*, 2004; Schröder, 2011; Seeber and Clapp, 2017), and instead applies an appropriate sequence of processing operations directly on a digital sound signal. While this makes the implementation considerably more challenging, it should prevent a large part of artifacts which arise from the transitions, which are inevitable in FIR-based approaches, between the snapshots of scene geometries at different points in time.

liveRAZR—like the original RAZR—strives to be suited to a range of reproduction hardware, instead of being tailored to a specific hardware setup. It is therefore conceivable to use this software for many different purposes, ranging from computer games to immersive presentations of vocal or musical performances in virtual reality. I will continue to use it in psychoacoustical research to investigate

a range of topics related to spatial perception, including the design of auditory stimuli to help people orient themselves in and navigate through space when vision is impaired (in the manner of Massiceti *et al.*, 2018), as well as echolocation in fully virtual environments (in loose continuation of the work by Wallmeier and Wiegrebe, 2014b). One aim will be to manipulate auditory cues in order to see how their distortion, or their complete absence, affects performance in a variety of tasks. On top of allowing such low-level experimental control, related to the classical understanding of psychophysics going back to Fechner (1860), I also hope that liveRAZR will further encourage and facilitate experiments which deliberately approximate sensory inputs that might plausibly occur in real life, the "naturalistic stimuli" which are becoming increasingly common in cognitive neuroscience research (see Sonkusare *et al.*, 2019).

Of course, the development of liveRAZR is far from finished. Important matters for its progress include the implementation of arbitrary room geometries (and not just six walls that are pairwise parallel), a model of sound diffraction, greater ease of use including the deployment to a wider audience, and of course further optimisations of computing time in order to support more complex scenes and a greater diversity of hardware. From a scientific point of view, these advances will also allow the program to be used for novel experiments in the auditory perception of space.

Bibliography

Ahnert, Wolfgang and Rainer Feistel (1993). "EARS auralization software". In: *Journal of the Audio Engineering Society* 41.11, pp. 894–904 (cit. on p. 65).

Ajdler, Thibaut, Luciano Sbaiz, and Martin Vetterli (2006). "The plenacoustic function and its sampling". In: *IEEE Transactions on Signal Processing* 54.10, pp. 3790–3804 (cit. on p. 67).

Alais, David and David Burr (2004). "The ventriloquist effect results from near-optimal bimodal integration". In: *Current Biology* 14.3, pp. 257–262 (cit. on p. 19).

Alexander, Richard D. (1962). "Evolutionary change in cricket acoustical communication". In: *Evolution* 16.4, pp. 443–467 (cit. on p. 22).

Allen, Jont B. and David A. Berkley (1979). "Image method for efficiently simulating small-room acoustics". In: *The Journal of the Acoustical Society of America* 65.4, pp. 943–950 (cit. on pp. 64, 118).

Allf, Bradley C., Paul A.P. Durst, and David W. Pfennig (2016). "Behavioral plasticity and the origins of novelty: the evolution of the rattlesnake rattle". In: *The American Naturalist* 188.4, pp. 475–483 (cit. on p. 117).

Altman, J.A., V.P. Romanov, and I.P. Pavlov (1988). "Psychophysical characteristics of the auditory image movement perception during dichotic stimulation". In: *International Journal of Neuroscience* 38.3-4, pp. 369–379 (cit. on p. 10).

Altman, J.A. and O.V. Viskov (1977). "Discrimination of perceived movement velocity for fused auditory image in dichotic stimulation". In: *The Journal of the Acoustical Society of America* 61.3, pp. 816–819 (cit. on pp. 10, 115).

Altmann, Christian F., Esther Wilczek, and Jochen Kaiser (2009). "Processing of auditory location changes after horizontal head rotation". In: *Journal of Neuroscience* 29.41, pp. 13074–13078 (cit. on p. 21).

Anderson, Paul W. and Pavel Zahorik (2014). "Auditory/visual distance estimation: accuracy and variability". In: *Frontiers in Psychology* 5, 1097 (cit. on pp. 10, 20).

Ashihara, Kaoru (2007). "Hearing thresholds for pure tones above 16 kHz". In: *The Journal of the Acoustical Society of America* 122.3, EL52–EL57 (cit. on p. 2).

Ashmead, Daniel H., DeFord L. Davis, and Anna Northington (1995). "Contribution of listeners' approaching motion to auditory distance perception." In: *Journal of Experimental Psychology: Human Perception and Performance* 21.2, p. 239 (cit. on pp. 21, 114, 115).

Bach, Dominik R., John G. Neuhoff, Walter Perrig, and Erich Seifritz (2009). "Looming sounds as warning signals: The function of motion cues". In: *International Journal of Psychophysiology* 74.1, pp. 28–33 (cit. on pp. 85, 116).

Bach, Dominik R., Hartmut Schachinger, John G. Neuhoff, Fabrizio Esposito, Franceso di Salle, Christoph Lehmann, Marcus Herdener, Klaus Scheffler, and Erich Seifritz (2008). "Rising Sound Intensity: An Intrinsic Warning Cue Activating the Amygdala". In: *Cerebral Cortex* 18.1, pp. 145–150 (cit. on pp. 11, 86).

Backer, Kristina C., Kevin T. Hill, Antoine J. Shahin, and Lee M. Miller (2010). "Neural time course of echo suppression in humans". In: *Journal of Neuroscience* 30.5, pp. 1905–1913 (cit. on p. 16).

Baird, Spencer Fullerton and Charles Girard (1853). *Catalogue of North American reptiles in the Museum of the Smithsonian Institution*. Smithsonian Institution (cit. on pp. 47, 53).

Barclay, Robert M.R., M. Brock Fenton, and Donald W. Thomas (1979). "Social behavior of the little brown bat, *Myotis lucifugus*". In: *Behavioral Ecology and Sociobiology* 6.2, pp. 137–146 (cit. on p. 102).

Bass, Henry E., Louis C. Sutherland, Allen J. Zuckerwar, David T. Blackstock, and D.M. Hester (1995). "Atmospheric absorption of sound: Further developments". In: *The Journal of the Acoustical Society of America* 97.1, pp. 680–683 (cit. on p. 2).

Bastian, Anna and Sabine Schmidt (2008). "Affect cues in vocalizations of the bat, *Megaderma lyra*, during agonistic interactions". In: *The Journal of the Acoustical Society of America* 124.1, pp. 598–608 (cit. on p. 93).

Batteau, Dwight W. (1967). "The role of the pinna in human localization". In: *Proceedings of the Royal Society B: Biological Sciences* 168.1011, pp. 158–180 (cit. on p. 8).

Battenberg, Eric and Rimas Avižienis (2011). "Implementing real-time partitioned convolution algorithms on conventional operating systems". In: *Proceedings of the 14th International Conference on Digital Audio Effects*. Paris, pp. 248–235 (cit. on p. 67).

Begault, Durand R. (1994). *3-D sound for Virtual Reality and Multimedia*. Boston. ISBN: 978-0-12-084735-8 (cit. on p. 14).

Behr, Oliver (2006). "The vocal repertoire of the sac-winged bat, *Saccopteryx bilineata*". Dissertation. Erlangen-Nürnberg: Friedrich-Alexander-Universität (cit. on p. 93).

Behr, Oliver and Otto von Helversen (2004). "Bat serenades—complex courtship songs of the sac-winged bat *(Saccopteryx bilineata)*". In: *Behavioral Ecology and Sociobiology* 56.2, pp. 106–115 (cit. on pp. 93, 102).

von Békésy, György (1938). "Über die Entstehung der Entfernungsempfindung beim Hören". In: *Akustische Zeitschrift* 3, pp. 21–31 (cit. on pp. 8, 36).

— (1949). "The moon illusion and similar auditory phenomena". In: *The American Journal of Psychology* 62.4, pp. 540–552 (cit. on p. 8).

Ben-Ari, Mordechai (2006). *Principles of Concurrent and Distributed Programming*. Second edition. Essex: Pearson Education. ISBN: 978-0-321-31283-9 (cit. on p. 66).

Bentley, Jon Louis (1975). "Multidimensional binary search trees used for associative searching". In: *Communications of the ACM* 18.9, pp. 509–517 (cit. on p. 78).

Bertelson, Paul and Beatrice de Gelder (2004). "The Psychology of Multimodal Perception". In: *Crossmodal Space and Crossmodal Attention*. Ed. by Charles Spence and Jon Driver. Oxford: Oxford University Press, pp. 141–177. ISBN: 978-0-19-852487-8 (cit. on p. 19).

Bianchi, Federica, Sarah Verhulst, and Torsten Dau (2013). "Experimental evidence for a cochlear source of the precedence effect". In: *Journal of the Association for Research in Otolaryngology* 14.5, pp. 767–779 (cit. on p. 15).

Bishop, Christopher W., Sam London, and Lee M. Miller (2011). "Visual influences on echo suppression". In: *Current Biology* 21.3, pp. 221–225 (cit. on pp. 20, 29).

— (2012). "Neural time course of visually enhanced echo suppression". In: *Journal of Neurophysiology* 108.7, pp. 1869–1883 (cit. on p. 29).

Bishop, Christopher W., Deepak Yadav, Sam London, and Lee M. Miller (2014). "The effects of preceding lead-alone and lag-alone click trains on the buildup of echo suppression". In: *The Journal of the Acoustical Society of America* 136.2, pp. 803–817 (cit. on p. 28).

Bizley, Jennifer K. and Yihan Dai (2020). "Non-auditory processing in the central auditory pathway". In: *Current Opinion in Physiology* 18, pp. 100–105 (cit. on p. 21).

Blakemore, Colin (1970). "The range and scope of binocular depth discrimination in man". In: *The Journal of Physiology* 211.3, pp. 599–622 (cit. on p. 115).

Blau, Matthias, Armin Budnik, Mina Fallahi, Henning Steffens, Stephan D. Ewert, and Steven van de Par (in print). "Toward realistic binaural auralizations—perceptual comparison between measurement- and simulation-based auralizations and the real room for a classroom scenario". In: *Acta Acustica united with Acustica* (cit. on p. 70).

Blauert, Jens (1969). "Sound localization in the median plane". In: *Acta Acustica united with Acustica* 22.4, pp. 205–213 (cit. on p. 8).

— (1997). *Spatial Hearing: The Psychophysics of Human Sound Localization*. Revised edition. Cambridge, Mass.: MIT Press. ISBN: 978-0-262-02413-6 (cit. on pp. 6, 8, 28, 36).

Blauert, Jens and Pierre L. Divenyi (1988). "Spectral selectivity in binaural contralateral inhibition". In: *Acustica* 66.5, pp. 267–274 (cit. on p. 15).

Bohn, Kirsten M., Janette Wenrick Boughman, Gerald S. Wilkinson, and Cynthia F. Moss (2004). "Auditory sensitivity and frequency selectivity in greater spear-nosed bats suggest specializations for acoustic communication". In: *Journal of Comparative Physiology A* 190.3, pp. 185–192 (cit. on p. 102).

Bohn, Kirsten M., Barbara Schmidt-French, Sean T. Ma, and George D. Pollak (2008). "Syllable acoustics, temporal patterns, and call composition vary with behavioral context in Mexican free-tailed bats". In: *The Journal of the Acoustical Society of America* 124.3, pp. 1838–1848 (cit. on pp. 92, 93, 100, 102).

Bohn, Kirsten M., Gerald S. Wilkinson, and Cynthia F. Moss (2007). "Discrimination of infant isolation calls by female greater spear-nosed bats, *Phyllostomus hastatus*". In: *Animal Behaviour* 73.3, pp. 423–432 (cit. on p. 102).

Bolognini, Nadia, Francesca Frassinetti, Andrea Serino, and Elisabetta Làdavas (2005). "'Acoustical vision' of below threshold stimuli: interaction among spatially converging audiovisual inputs". In: *Experimental Brain Research* 160.3, pp. 273–282 (cit. on p. 111).

Borish, Jeffrey (1984). "Extension of the image model to arbitrary polyhedra". In: *The Journal of the Acoustical Society of America* 75.6, pp. 1827–1836 (cit. on p. 64).

Boudreau, James C. and Chiyeko Tsuchitani (1968). "Binaural interaction in the cat superior olive s segment." In: *Journal of Neurophysiology* 31.3, pp. 442–454 (cit. on p. 7).

Boughman, Janette Wenrick (1997). "Greater spear-nosed bats give group-distinctive calls". In: *Behavioral Ecology and Sociobiology* 40.1, pp. 61–70 (cit. on p. 102).

— (1998). "Vocal learning by greater spear–nosed bats". In: *Proceedings of the Royal Society B: Biological Sciences* 265.1392, pp. 227–233 (cit. on p. 23).

Boughman, Janette Wenrick and Gerald S. Wilkinson (1998). "Greater spear-nosed bats discriminate group mates by vocalizations". In: *Animal Behaviour* 55.6, pp. 1717–1732 (cit. on p. 102).

Bradbury, Jack W. and Sandra L. Vehrencamp (2011). *Principles of Animal Communication*. Second edition. Sunderland, Mass.: Sinauer. ISBN: 978-0-87893-045-6 (cit. on pp. 22, 23).

Brandewie, Eugene J. and Pavel Zahorik (2010). "Prior listening in rooms improves speech intelligibility". In: *The Journal of the Acoustical Society of America* 128.1, pp. 291–299 (cit. on pp. 17, 28, 111).

— (2018). "Speech intelligibility in rooms: Disrupting the effect of prior listening exposure". In: *The Journal of the Acoustical Society of America* 143.5, pp. 3068–3078 (cit. on p. 18).

Bregman, Albert S. (1994). *Auditory Scene Analysis. The Perceptual Organization of Sound*. Cambridge, Mass.: MIT press. ISBN: 978-0-262-52195-6 (cit. on pp. 1, 153).

Briefer, Elodie F. (2012). "Vocal expression of emotions in mammals: mechanisms of production and evidence". In: *Journal of Zoology* 288.1, pp. 1–20 (cit. on p. 102).

Brimijoin, W. Owen and Michael A. Akeroyd (2012). "The role of head movements and signal spectrum in an auditory front/back illusion". In: *i-Perception* 3.3, pp. 179–182 (cit. on p. 115).

— (2014). "The moving minimum audible angle is smaller during self motion than during source motion". In: *Frontiers in Neuroscience* 8, 273 (cit. on pp. 21, 39, 115).

Brimijoin, W. Owen, David McShefferty, and Michael A. Akeroyd (2012). "Undirected head movements of listeners with asymmetrical hearing impairment during a speech-in-noise task". In: *Hearing Research* 283.1-2, pp. 162–168 (cit. on p. 112).

Brinkmann, Fabian (2019). "Binaural processing for the evaluation of acoustical environments". Dissertation. Berlin: TU Berlin (cit. on p. 1).

Brinkmann, Fabian, Lukas Aspöck, David Ackermann, Steffen Lepa, Michael Vorländer, and Stefan Weinzierl (2019). "A round robin on room acoustical simulation and auralization". In: *The Journal of the Acoustical Society of America* 145.4, pp. 2746–2760 (cit. on p. 66).

Bronkhorst, Adelbert W. (1995). "Localization of real and virtual sound sources". In: *The Journal of the Acoustical Society of America* 98.5, pp. 2542–2553 (cit. on p. 21).

— (2000). "The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions". In: *Acta Acustica united with Acustica* 86.1, pp. 117–128 (cit. on p. 5).

Bronkhorst, Adelbert W. and Tammo Houtgast (1999). "Auditory distance perception in rooms". In: *Nature* 397.6719, pp. 517–520 (cit. on pp. 9, 36).

Brown, Andrew D., G. Christopher Stecker, and Daniel J. Tollin (2015). "The precedence effect in sound localization". In: *Journal of the Association for Research in Otolaryngology* 16.1, pp. 1–28 (cit. on p. 28).

Brown, C. Phillip and Richard O. Duda (1998). "A structural model for binaural sound synthesis". In: *IEEE Transactions on Speech and Audio Processing* 6.5, pp. 476–488 (cit. on p. 70).

Brown, Sarah G., George H. Boettner, and Jayne E. Yack (2007). "Clicking caterpillars: acoustic aposematism in Antheraea polyphemus and other Bombycoidea". In: *Journal of Experimental Biology* 210.6, pp. 993–1005 (cit. on p. 23).

Bruce, Vicki, Patrick R. Green, and Mark A. Georgeson (2003). *Visual perception: Physiology, Psychology, & Ecology*. 4th edition. New York: Psychology Press. ISBN: 978-1-84169-237-1 (cit. on p. 5).

Brungart, Douglas S., Alexander J. Kordik, and Brian D. Simpson (2006). "Effects of headtracker latency in virtual audio displays". In: *Journal of the Audio Engineering Society* 45.1/2, pp. 32–44 (cit. on p. 68).

Brungart, Douglas S. and William M. Rabinowitz (1999). "Auditory localization of nearby sources. Head-related transfer functions". In: *The Journal of the Acoustical Society of America* 106.3, pp. 1465–1479 (cit. on p. 90).

Butler, Robert A., Elena T. Levy, and William D. Neff (1980). "Apparent distance of sounds recorded in echoic and anechoic chambers." In: *Journal of Experimental Psychology: Human Perception and Performance* 6.4, p. 745 (cit. on p. 8).

Buttler, Oliver (2018). "Optimierung und Erweiterung einer effizienten Methode zur Synthese binauraler Raumimpulsantworten". MA thesis. Oldenburg: Carl-von-Ossietzky-Universität (cit. on p. 70).

Buttler, Oliver, Torben Wendt, Steven van de Par, and Stephan D. Ewert (2018). "Perceptually plausible room acoustics simulation including diffuse reflections". In: *The Journal of the Acoustical Society of America* 143.3, pp. 1829–1830 (cit. on pp. 64, 66, 71).

Buus, Sören, Mary Florentine, and Torben Poulsen (1997). "Temporal integration of loudness, loudness discrimination, and the form of the loudness function". In: *The Journal of the Acoustical Society of America* 101.2, pp. 669–680 (cit. on pp. 5, 116).

Cain, Gerald D., Anush Yardim, and P. Henry (1995). "Offset windowing for FIR fractional-sample delay". In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 2. IEEE, pp. 1276–1279 (cit. on p. 81).

Calcagno, Esteban R., Ezequiel L. Abregu, Manuel C. Eguía, and Ramiro Vergara (2012). "The role of vision in auditory distance perception". In: *Perception* 41.2, pp. 175–192 (cit. on p. 20).

Campos, Jennifer, Robert Ramkhalawansingh, and M. Kathleen Pichora-Fuller (2018). "Hearing, self-motion perception, mobility, and aging". In: *Hearing Research* 369, pp. 42–55 (cit. on p. 22).
Canévet, Georges, Bertram Scharf, and Marie-Claire Botte (1983). "Loudness adaptation, when induced, is real". In: *British Journal of Audiology* 17.1, pp. 49–57 (cit. on p. 5).
Cant, Nell Beatty and John H. Casseday (1986). "Projections from the anteroventral cochlear nucleus to the lateral and medial superior olivary nuclei". In: *Journal of Comparative Neurology* 247.4, pp. 457–476 (cit. on p. 7).
Cant, Nell Beatty and Richard L. Hyson (1992). "Projections from the lateral nucleus of the trapezoid body to the medial superior olivary nucleus in the gerbil". In: *Hearing Research* 58.1, pp. 26–34 (cit. on p. 6).
Carlile, Simon and Johahn Leung (2016). "The perception of auditory motion". In: *Trends in Hearing* 20, 2331216516644254 (cit. on p. 10).
Chadwick, L.E. and Hermann Rahn (1954). "Temperature dependence of rattling frequency in the rattlesnake, *Crotalus v. viridis*". In: *Science* 119.3092, pp. 442–443 (cit. on pp. 24, 47).
Chittka, Lars and Axel Brockmann (2005). "Perception space—the final frontier". In: *PLOS Biology* 3.4, e137 (cit. on pp. 4, 153).
Clifton, Rachel K. (1987). "Breakdown of echo suppression in the precedence effect". In: *The Journal of the Acoustical Society of America* 82.5, pp. 1834–1835 (cit. on pp. 15, 17, 28, 111).
Clifton, Rachel K. and Richard L. Freyman (1989). "Effect of click rate and delay on breakdown of the precedence effect". In: *Perception & Psychophysics* 46.2, pp. 139–145 (cit. on pp. 15, 28).
— (1997). "The precedence effect: Beyond echo suppression". In: *Binaural and Spatial Hearing in Real and Virtual Environments*. Ed. by Robert Gilkey and Timothy R. Anderson, pp. 233–256 (cit. on p. 111).
Clifton, Rachel K., Richard L. Freyman, Ruth Y. Litovsky, and Daniel McCall (1994). "Listeners' expectations about echoes can raise or lower echo threshold". In: *The Journal of the Acoustical Society of America* 95.3, pp. 1525–1533 (cit. on p. 111).
Clifton, Rachel K., Richard L. Freyman, and Jennifer Meo (2002). "What the precedence effect tells us about room acoustics". In: *Perception & Psychophysics* 64.2, pp. 180–188 (cit. on p. 28).
Coleman, Paul D. (1962). "Failure to localize the source distance of an unfamiliar sound". In: *The Journal of the Acoustical Society of America* 34.3, pp. 345–346 (cit. on p. 36).
— (1963). "An analysis of cues to auditory depth perception in free space". In: *Psychological Bulletin* 60.3, p. 302 (cit. on p. 90).
— (1968). "Dual role of frequency spectrum in determination of auditory distance". In: *The Journal of the Acoustical Society of America* 44.2, pp. 631–632 (cit. on p. 8).
Cooley, James W. and John W. Tukey (1965). "An algorithm for the machine calculation of complex Fourier series". In: *Mathematics of Computation* 19.90, pp. 297–301 (cit. on p. 14).
Corcoran, Aaron J., Jesse R. Barber, and William E. Conner (2009). "Tiger moth jams bat sonar". In: *Science* 325.5938, pp. 325–327 (cit. on p. 23).
Dalenbäck, Bengt-Inge (1995). "A New Model for Room Acoustic Prediction and Auralization". Dissertation. Gothenburg: Chalmers University of Technology. ISBN: 978-91-7197-200-2 (cit. on p. 65).
Dalenbäck, Bengt-Inge, Mendel Kleiner, and Peter Svensson (1994). "A macroscopic view of diffuse reflection". In: *Journal of the Audio Engineering Society* 42.10, pp. 793–807 (cit. on p. 64).
Damaschke, Jörg, Helmut Riedel, and Birger Kollmeier (2005). "Neural correlates of the precedence effect in auditory evoked potentials". In: *Hearing Research* 205.1-2, pp. 157–171 (cit. on p. 16).
Daniel, Peter M. and David Whitteridge (1961). "The representation of the visual field on the cerebral cortex in monkeys". In: *The Journal of Physiology* 159.2, p. 203 (cit. on p. 5).
Davidson, Susan M. and Gerald S. Wilkinson (2004). "Function of male song in the greater white-lined bat, *Saccopteryx bilineata*". In: *Animal Behaviour* 67.5, pp. 883–891 (cit. on p. 93).

Davis, Kevin A., Ramnarayan Ramachandran, and Bradford J. May (2003). "Auditory processing of spectral cues for sound localization in the inferior colliculus". In: *Journal of the Association for Research in Otolaryngology* 4.2, pp. 148–163 (cit. on p. 7).

Dawkins, Richard and John R. Krebs (1978). "Animal Signals: Information or Manipulation?" In: *Behavioural Ecology: An Evolutionary Approach*. Ed. by John R. Krebs and Nicholas B. Davies. Blackwell Scientific Publications, pp. 282–309. ISBN: 978-0-632-00285-6 (cit. on p. 22).

De Cheveigné, Alain and Hideki Kawahara (2002). "YIN, a fundamental frequency estimator for speech and music". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1917–1930 (cit. on p. 95).

DeAngelis, Gregory C. and Dora E. Angelaki (2012). "Visual–vestibular integration for self-motion perception". In: *The Neural Bases of Multisensory Processes*. Ed. by Micah M. Murray and Mark T. Wallace. Boca Raton: CRC Press, Taylor & Francis. ISBN: 978-1-4398-1217-4 (cit. on p. 21).

Devore, Sasha, Antje Ihlefeld, Kenneth Hancock, Barbara Shinn-Cunningham, and Bertrand Delgutte (2009). "Accurate sound localization in reverberant environments is mediated by robust encoding of spatial cues in the auditory midbrain". In: *Neuron* 62.1, pp. 123–134 (cit. on p. 16).

Djelani, Thomas and Jens Blauert (2001). "Investigations into the build-up and breakdown of the precedence effect". In: *Acta Acustica united with Acustica* 87.2, pp. 253–261 (cit. on p. 16).

Dokka, Kalpana, Paul R. MacNeilage, Gregory C. DeAngelis, and Dora E. Angelaki (2011). "Estimating distance during self-motion: A role for visual–vestibular interactions". In: *Journal of Vision* 11.13, p. 2 (cit. on p. 36).

Doupe, Allison J. and Patricia K. Kuhl (1999). "Birdsong and human speech: common themes and mechanisms". In: *Annual Review of Neuroscience* 22.1, pp. 567–631 (cit. on p. 93).

Driver, Jon and Toemme Noesselt (2008). "Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments". In: *Neuron* 57.1, pp. 11–23 (cit. on p. 19).

Dürer, Albrecht (1525). *Underweysung der messung mit dem zirckel und richtscheyt in linien ebnen unnd gantzen corporen*. Nürnberg. Reprinted in Nördlingen (1983): Uhl (cit. on p. 62).

Ekdale, Eric G. (2016). "Form and function of the mammalian inner ear". In: *Journal of Anatomy* 228.2, pp. 324–337 (cit. on p. 3).

Ellard, Colin G., Melvyn A. Goodale, and Brian Timney (1984). "Distance estimation in the mongolian gerbil: The role of dynamic depth cues". In: *Behavioural Brain Research* 14.1, pp. 29–39 (cit. on p. 39).

Ellis, Gregory M. and Pavel Zahorik (2019). "A dissociation between speech understanding and perceived reverberation". In: *Hearing Research* 379, pp. 52–58 (cit. on p. 113).

Englitz, Bernhard, Sandra Tolnai, Marei Typlt, Jürgen Jost, and Rudolf Rübsamen (2009). "Reliability of synaptic transmission at the synapses of Held in vivo under acoustic stimulation". In: *PLOS ONE* 4.10, e7014 (cit. on p. 6).

Epstein, Michael and Jeremy Marozeau (2010). "Loudness and Intensity Coding". In: *The Oxford Handbook of Auditory Science: Hearing*. Ed. by Christopher J. Plack. Vol. 3, pp. 45–69. ISBN: 978-0-19-923355-7 (cit. on p. 4).

Ernst, Marc O. and Martin S. Banks (2002). "Humans integrate visual and haptic information in a statistically optimal fashion". In: *Nature* 415.6870, pp. 429–433 (cit. on p. 40).

Esser, Karl-Heinz (1994). "Audio-vocal learning in a non-human mammal: the lesser spear-nosed bat *Phyllostomus discolor*". In: *NeuroReport* 5.14, pp. 1718–1720 (cit. on pp. 23, 93, 118).

Esser, Karl-Heinz and B. Lud (1997). "Discrimination of sinusoidally frequency-modulated sound signals mimicking species-specific communication calls in the FM-bat *Phyllostomus discolor*". In: *Journal of Comparative Physiology A* 180.5, pp. 513–522 (cit. on p. 101).

Esser, Karl-Heinz and Uwe Schmidt (1989). "Mother-infant communication in the lesser spear-nosed bat *Phyllostomus discolor* (Chiroptera, Phyllostomidae)—evidence for acoustic learning". In: *Ethology* 82.2, pp. 156–168 (cit. on pp. 23, 93, 101, 102, 118).

Esser, Karl-Heinz and Julia Schubert (1998). "Vocal dialects in the lesser spear-nosed bat *Phyllostomus discolor*". In: *Naturwissenschaften* 85.7, pp. 347–349 (cit. on pp. 93, 101).

Ewert, Stephan D. (2020). "Defining the proper stimulus and its ecology—mammals". In: *The Senses: A Comprehensive Reference*. Ed. by Bernd Fritzsch and Benedikt Grothe. Second edition. Vol. 2. Elsevier, Academic Press, pp. 187–206 (cit. on p. 86).

Eyring, Carl F. (1930). "Reverberation time in 'dead' rooms". In: *The Journal of the Acoustical Society of America* 1.2A, pp. 217–241 (cit. on pp. 3, 73).

Fechner, Gustav Theodor (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel (cit. on pp. 5, 119).

Feddersen, W. E., Thomas T. Sandel, Donald C. Teas, and Lloyd A. Jeffress (1957). "Localization of high-frequency tones". In: *the Journal of the Acoustical Society of America* 29.9, pp. 988–991 (cit. on p. 7).

Fenton, M. Brock and Lawrence E. Licht (1990). "Why rattle snake?" In: *Journal of Herpetology*, pp. 274–279 (cit. on pp. 24, 46, 56).

Fichtel, Claudia (2004). "Reciprocal recognition of sifaka *(Propithecus verreauxi verreauxi)* and red-fronted lemur *(Eulemur fulvus rufus)* alarm calls". In: *Animal Cognition* 7.1, pp. 45–52 (cit. on p. 23).

Filippi, Piera (2016). "Emotional and interactional prosody across animal communication systems: A comparative approach to the emergence of language". In: *Frontiers in Psychology* 7, 1393 (cit. on p. 102).

Filippi, Piera, Jenna V. Congdon, John Hoang, Daniel L. Bowling, Stephan A. Reber, Andrius Pašukonis, Marisa Hoeschele, Sebastian Ocklenburg, Bart De Boer, Christopher B. Sturdy, Albert Newen, and Onur Güntürkün (2017). "Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals". In: *Proceedings of the Royal Society B: Biological Sciences* 284.1859, 20170990 (cit. on p. 102).

Firzlaff, Uwe, Sven Schörnich, Susanne Hoffmann, Gerd Schuller, and Lutz Wiegrebe (2006). "A neural correlate of stochastic echo imaging". In: *Journal of Neuroscience* 26.3, pp. 785–791 (cit. on p. 93).

Fisher, H. Geoffrey and Sanford J. Freedman (1968). "Localization of sound during simulated unilateral conductive hearing loss". In: *Acta Oto-Laryngologica* 66.1-6, pp. 213–220 (cit. on p. 8).

Fitch, W. Tecumseh (2000). "The evolution of speech: A comparative review". In: *Trends in Cognitive Sciences* 4.7, pp. 258–267 (cit. on p. 22).

Flanagin, Virginia L., Sven Schörnich, Michael Schranner, Nadine Hummel, Ludwig Wallmeier, Magnus Wahlberg, Thomas Stephan, and Lutz Wiegrebe (2017). "Human exploration of enclosed spaces through echolocation". In: *Journal of Neuroscience* 37.6, pp. 1614–1627 (cit. on pp. 18, 68).

Forman, George and Martin Scholz (2010). "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement". In: *ACM SIGKDD Explorations Newsletter* 12.1, pp. 49–57 (cit. on p. 95).

Forsthofer, Michael, Michael Schutte, Harald Luksch, Tobias Kohl, Lutz Wiegrebe, and Boris P. Chagnaud (2021). "Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans". In: *Current Biology*. Published online ahead of print: `https://doi.org/10.1016/j.cub.2021.07.018` (cit. on pp. 24, 45).

Freeman, Tom C.A., Rebecca A. Champion, and Paul A. Warren (2010). "A Bayesian model of perceived head-centered velocity during smooth pursuit eye movement". In: *Current Biology* 20.8, pp. 757–762 (cit. on p. 39).

Freeman, Tom C.A., John F. Culling, Michael A. Akeroyd, and W. Owen Brimijoin (2017). "Auditory compensation for head rotation is incomplete". In: *Journal of Experimental Psychology: Human Perception and Performance* 43.2, pp. 371–380 (cit. on pp. 21, 39, 115).

Freyman, Richard L., Rachel K. Clifton, and Ruth Y. Litovsky (1991). "Dynamic processes in the precedence effect". In: *The Journal of the Acoustical Society of America* 90.2, pp. 874–884 (cit. on pp. 15, 17).

Friard, Olivier and Marco Gamba (2016). "BORIS: A free, versatile open-source event-logging software for video/audio coding and live observations". In: *Methods in Ecology and Evolution* 7.11, pp. 1325–1330 (cit. on p. 96).

Fullard, James H., M. Brock Fenton, and James A. Simmons (1979). "Jamming bat echolocation: the clicks of arctiid moths". In: *Canadian Journal of Zoology* 57.3, pp. 647–649 (cit. on p. 23).

Funkhouser, Thomas, Nicolas Tsingos, Ingrid Carlbom, Gary W. Elko, Mohan Sondhi, James E. West, Gopal Pingali, Patrick Min, and Addy Ngan (2004). "A beam tracing method for interactive architectural acoustics". In: *The Journal of the Acoustical Society of America* 115.2, pp. 739–756 (cit. on pp. 63, 67, 118).

Furman, Moran and Moshe Gur (2012). "And yet it moves: Perceptual illusions and neural mechanisms of pursuit compensation during smooth pursuit eye movements". In: *Neuroscience & Biobehavioral Reviews* 36.1, pp. 143–151 (cit. on pp. 39, 115).

Gadziola, Marie A., Jasmine M.S. Grimsley, Paul A. Faure, and Jeffrey J. Wenstrup (2012). "Social vocalizations of big brown bats vary with behavioral context". In: *PLOS ONE* 7.9, e44550 (cit. on p. 93).

Galambos, Robert and Hallowell Davis (1943). "The response of single auditory-nerve fibers to acoustic stimulation". In: *Journal of Neurophysiology* 6.1, pp. 39–57 (cit. on p. 6).

Gamble, Eleanor A. (1909). "Minor studies from the psychological laboratory of Wellesley College: Intensity as a criterion in estimating the distance of sounds." In: *Psychological Review* 16.6, p. 416 (cit. on p. 8).

García, Guillermo (2002). "Optimal filter partition for efficient convolution with short input/output delay". In: *Audio Engineering Society Convention 113*. Paper no. 5660. Audio Engineering Society. Los Angeles (cit. on p. 67).

García-Gómez, Victor and Jose J. López (2018). "Binaural room impulse responses interpolation for multimedia real-time applications". In: *Audio Engineering Society Convention 144*. Paper no. 9962. Audio Engineering Society. Milan (cit. on p. 67).

Gardner, Mark B. (1968a). "Historical background of the Haas and/or precedence effect". In: *The Journal of the Acoustical Society of America* 43.6, pp. 1243–1248 (cit. on p. 15).

— (1968b). "Proximity image effect in sound localization". In: *The Journal of the Acoustical Society of America* 43.1, p. 163 (cit. on pp. 19, 20).

Gardner, William G. (1995). "Efficient convolution without input–output delay". In: *Journal of the Audio Engineering Society* 43.3, pp. 127–136 (cit. on pp. 67, 79, 80).

Garrido-Jurado, Sergio, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez (2014). "Automatic generation and detection of highly reliable fiducial markers under occlusion". In: *Pattern Recognition* 47.6, pp. 2280–2292 (cit. on p. 42).

Gates, George A. and John H. Mills (2005). "Presbycusis". In: *The Lancet* 366.9491, pp. 1111–1120 (cit. on p. 2).

Gelfand, Stanley A and Shlomo Silman (1979). "Effects of small room reverberation upon the recognition of some consonant features". In: *The Journal of the Acoustical Society of America* 66.1, pp. 22–29 (cit. on p. 9).

Genzel, Daria, Uwe Firzlaff, Lutz Wiegrebe, and Paul R. MacNeilage (2016). "Dependence of auditory spatial updating on vestibular, proprioceptive, and efference copy signals". In: *Journal of Neurophysiology* 116.2, pp. 765–775 (cit. on pp. 21, 115).

Genzel, Daria, Michael Schutte, W. Owen Brimijoin, Paul R. MacNeilage, and Lutz Wiegrebe (2018). "Psychophysical evidence for auditory motion parallax". In: *Proceedings of the National Academy of Sciences* 115.16, pp. 4264–4269 (cit. on pp. 24, 35).

Ghahramani, Zoubin, Daniel M. Wolptrt, and Michael I. Jordan (1997). "Computational models of sensorimotor integration". In: *Advances in Psychology*. Vol. 119. Elsevier, pp. 117–147 (cit. on p. 19).

Gibbs, Barry M. and D.K. Jones (1972). "A simple image method for calculating the distribution of sound pressure levels within an enclosure". In: *Acustica* 26.1, pp. 24–32 (cit. on p. 63).

Gil-Carvajal, Juan C., Jens Cubick, Sébastien Santurette, and Torsten Dau (2016). "Spatial hearing with incongruent visual or auditory room cues". In: *Scientific Reports* 6, 37342 (cit. on pp. 32, 112).

Glaudas, Xavier, Terence M. Farrell, and Peter G. May (2005). "Defensive behavior of free-ranging pygmy rattlesnakes *(Sistrurus miliarius)*". In: *Copeia* 2005.1, pp. 196–200 (cit. on p. 47).

Goldberg, Jay M. and Paul B. Brown (1969). "Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization." In: *Journal of Neurophysiology* 32.4, pp. 613–636 (cit. on p. 6).

Goldstein, E. Bruce (1997). *Wahrnehmungspsychologie*. Heidelberg: Spektrum Akademischer Verlag. ISBN: 978-3-8274-0189-2 (cit. on p. 153).

Goodall, Jane (1986). *The Chimpanzees of Gombe: Patterns of Behavior*. Cambridge, Mass.: Harvard University Press. ISBN: 978-0-674-11649-8 (cit. on p. 22).

Goral, Cindy M., Kenneth E. Torrance, Donald P. Greenberg, and Bennett Battaile (1984). "Modeling the interaction of light between diffuse surfaces". In: *SIGGRAPH Computer Graphics* 18.3, pp. 213–222 (cit. on p. 65).

Gould, Edwin (1975). "Neonatal vocalizations in bats of eight genera". In: *Journal of Mammalogy* 56.1, pp. 15–29 (cit. on pp. 93, 101, 102).

Grange, Jacques A. and John F. Culling (2016). "The benefit of head orientation to speech intelligibility in noise". In: *The Journal of the Acoustical Society of America* 139.2, pp. 703–712 (cit. on pp. 33, 112).

Grantham, D. Wesley (1986). "Detection and discrimination of simulated motion of auditory targets in the horizontal plane". In: *The Journal of the Acoustical Society of America* 79.6, pp. 1939–1949 (cit. on p. 115).

Greene, Harry W. (1969). "Antipredator Mechanisms in Reptiles". In: *Biology of the Reptilia, Volume 16, Ecology B: Defense and Life History*. Ed. by Carl Gans and Raymond B. Huey. New York: Alan R. Liss, pp. 1–152. ISBN: 978-0-8451-4402-2 (cit. on p. 23).

Grimm, Giso, Joanna Luberadzka, and Volker Hohmann (2019). "A toolbox for rendering virtual acoustic environments in the context of audiology". In: *Acta Acustica united with Acustica* 105.3, pp. 566–578 (cit. on pp. 66, 68).

Grimm, Giso, Torben Wendt, Volker Hohmann, and Stephan D. Ewert (2014). "Implementation and perceptual evaluation of a simulation method for coupled rooms in higher order ambisonics". In: *Proceedings of the EAA Joint Symposium on Auralization and Ambisonics 2014*. Berlin, pp. 27–32 (cit. on p. 70).

Grothe, Benedikt, Michael Pecka, and David McAlpine (2010). "Mechanisms of sound localization in mammals". In: *Physiological Reviews* 90.3, pp. 983–1012 (cit. on pp. 5–7).

Grünbaum, Branko (1994). "Uniform tilings of 3-space". In: *Geombinatorics* 4.2, pp. 49–56 (cit. on p. 64).

Haas, Helmut (1951). "Über den Einfluß eines Einfachechos auf die Hörsamkeit von Sprache". In: *Acustica* 1.2, pp. 49–58 (cit. on p. 15).

Hagen, Margaret A. and Martha Teghtsoonian (1981). "The effects of binocular and motion-generated information on the perception of depth and height". In: *Perception & Psychophysics* 30.3, pp. 257–265 (cit. on p. 113).

Hall, Deborah A. and David R. Moore (2003). "Auditory neuroscience: The salience of looming sounds". In: *Current Biology* 13.3, R91–R93 (cit. on p. 11).

Hameed, Sharaf, Jyri Pakarinen, Kari Valde, and Ville Pulkki (2004). "Psychoacoustic cues in room size perception". In: *Audio Engineering Society Convention 116*. Paper no. 6084. Audio Engineering Society. Berlin (cit. on p. 18).

Hartmann, William M (1983). "Localization of sound in rooms". In: *The Journal of the Acoustical Society of America* 74.5, pp. 1380–1391 (cit. on p. 9).

Hartung, Klaus and Constantine Trahiotis (2001). "Peripheral auditory processing and investigations of the 'precedence effect' which utilize successive transient stimuli". In: *The Journal of the Acoustical Society of America* 110.3, pp. 1505–1513 (cit. on p. 28).

Hastie, Trevor, Jerome Friedman, and Robert Tibshirani (2009). *The Elements of Statistical Learning*. Vol. 1. Springer Series in Statistics 10. New York: Springer. ISBN: 978-0-387-84858-7 (cit. on p. 95).

Hauser, Marc D., Noam Chomsky, and W. Tecumseh Fitch (2002). "The faculty of language: What is it, who has it, and how did it evolve?" In: *science* 298.5598, pp. 1569–1579 (cit. on p. 22).

Hechavarría, Julio C., M. Jerome Beetz, Silvio Macias, and Manfred Kössl (2016). "Distress vocalization sequences broadcasted by bats carry redundant information". In: *Journal of Comparative Physiology A* 202.7, pp. 503–515 (cit. on pp. 93, 102).

Heckbert, Paul S. and Pat Hanrahan (1984). "Beam tracing polygonal objects". In: *SIGGRAPH '84: Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques*. Minneapolis: ACM Press, pp. 119–127 (cit. on p. 63).

Heinrich, Melina and Lutz Wiegrebe (2013). "Size constancy in bat biosonar? Perceptual interaction of object aperture and distance". In: *PLOS ONE* 8.4, e61577 (cit. on p. 93).

Heinz, Renate (1993). "Binaural room simulation based on an image source model with addition of statistical methods to include the diffuse sound scattering of walls and to predict the reverberant tail". In: *Applied Acoustics* 38.2-4, pp. 145–159 (cit. on p. 65).

von Helmholtz, Hermann (1925). *Treatise on Physiological Optics*. New York (cit. on p. 36).

Hendrickx, Etienne, Mathieu Paquier, Vincent Koehl, and Julian Palacino (2015). "Ventriloquism effect with sound stimuli varying in both azimuth and elevation". In: *The Journal of the Acoustical Society of America* 138.6, pp. 3686–3697 (cit. on p. 33).

Henry, Joseph (1851). *Scientific Writings of Joseph Henry*. Vol. II. Smithsonian Institution, Washington, D.C. (cit. on p. 15).

Hirsh, Ira J. and W. Dixon Ward (1952). "Recovery of the auditory threshold after strong acoustic stimulation". In: *The Journal of the Acoustical Society of America* 24.2, pp. 131–141 (cit. on p. 5).

Hodgson, Murray (1990). "Evidence of diffuse surface reflection in rooms". In: *The Journal of the Acoustical Society of America* 88.S1, S185–S185 (cit. on p. 64).

Hodgson, Murray and Eva-Marie Nosal (2006). "Experimental evaluation of radiosity for room sound-field prediction". In: *The Journal of the Acoustical Society of America* 120.2, pp. 808–819 (cit. on p. 65).

Hoffmann, Susanne, Leonie Baier, Frank Borina, Gerd Schuller, Lutz Wiegrebe, and Uwe Firzlaff (2008). "Psychophysical and neurophysiological hearing thresholds in the bat *Phyllostomus discolor*". In: *Journal of Comparative Physiology A* 194.1, pp. 39–47 (cit. on p. 93).

Hofmann, Georg Rainer (1990). "Who invented ray tracing?" In: *The Visual Computer* 6.3, pp. 120–124 (cit. on p. 62).

Holmes, Gordon (1918). "Disturbances of vision by cerebral lesions". In: *The British Journal of Ophthalmology* 2.7, p. 353 (cit. on p. 5).

Holters, Martin and Udo Zölzer (2006). "Parametric higher-order shelving filters". In: *14th European Signal Processing Conference*. IEEE, pp. 1–4 (cit. on p. 69).

Howard, Ian P. and Brian J. Rogers (1995). *Binocular Vision and Stereopsis*. New York: Oxford University Press. ISBN: 978-0-19-508476-4 (cit. on p. 36).

— (2002). *Seeing in Depth: Depth Perception*. Vol. 2. Toronto: University of Toronto Press/I. Porteous. ISBN: 978-0-9730873-1-4 (cit. on p. 36).

Howard, Ian P. and William B. Templeton (1966). *Human Spatial Orientation*. London: John Wiley & Sons (cit. on pp. 19, 112).

Hristov, Nickolay I. and William E. Conner (2005). "Sound strategy: acoustic aposematism in the bat–tiger moth arms race". In: *Naturwissenschaften* 92.4, pp. 164–169 (cit. on p. 23).

Inductiveload on Wikimedia Commons (2009). *Anatomy of the Human Ear.svg*. Downloaded on 12 November 2020. URL: `https://commons.wikimedia.org/w/index.php?title=File:Anatomy_of_the_Human_Ear.svg%5C&oldid=453296791` (cit. on pp. 4, 153).

International Organization for Standardization (1993). *ISO 9613-1: Attenuation of Sound During Propagation Outdoors. Part 1: Calculation of the Absorption of Sound by the Atmosphere* (cit. on p. 2).

— (2016). *ISO 532-2: Acoustics–Methods for calculating loudness–Part 2: Moore–Glasberg method* (cit. on p. 5).

Janik, Vincent M. and Peter J.B. Slater (1997). "Vocal Learning in Mammals". In: *Advances in the Study of Behaviour*. Ed. by Peter J.B. Slater, Jay S. Rosenblatt, Charles T. Snowdon, and Manfred Milinski. Vol. 26. New York: Academic Press, pp. 59–100. ISBN: 978-0-12-004526-6 (cit. on p. 22).

Jongkees, Leonard B.W. and J.J. Groen (1946). "On directional hearing". In: *The Journal of Laryngology & Otology* 61.9, pp. 494–504 (cit. on p. 8).

Joris, Philip X., Laurel H. Carney, Philip H. Smith, and Tom C. Yin (1994). "Enhancement of neural synchronization in the anteroventral cochlear nucleus. I. Responses to tones at the characteristic frequency". In: *Journal of Neurophysiology* 71.3, pp. 1022–1036 (cit. on p. 6).

Jot, Jean-Marc and Antoine Chaigne (1991). "Digital delay networks for designing artificial reverberators". In: *Audio Engineering Society Convention 90*. Paper no. 3030. Audio Engineering Society. Paris (cit. on pp. 66, 71, 118).

Kanwal, Jagmeet S. (2009). "Audiovocal Communication in Bats". In: *Encyclopedia of Neurosciences*. Ed. by Squire Larry R. ISBN: 978-0-08-045046-9 (cit. on p. 102).

Kanwal, Jagmeet S., Sumiko Matsumura, Kevin Ohlemiller, and Nobuo Suga (1994). "Analysis of acoustic elements and syntax in communication sounds emitted by mustached bats". In: *The Journal of the Acoustical Society of America* 96.3, pp. 1229–1254 (cit. on pp. 92, 93, 95, 100, 102).

Kaplan, Sam T. and Tadeusz J. Ulrych (2007). "Phase unwrapping: a review of methods and a novel technique". In: *Let it Flow – 2007 CSPG CSEG Convention*. Canadian Society of Petroleum Geologists/Canadian Society of Exploration Geophysicists, pp. 534–537 (cit. on p. 79).

Kaplanis, Neofytos, Sören Bech, Sören Holdt Jensen, and Toon van Waterschoot (2014). "Perception of reverberation in small rooms: a literature study". In: *Proceedings of the AES 55th International Conference: Spatial Audio*. Audio Engineering Society. Helsinki (cit. on p. 18).

Karam, Zahi N. and Alan V. Oppenheim (2007). "Computation of the one-dimensional unwrapped phase". In: *15th International Conference on Digital Signal Processing*. IEEE, pp. 304–307 (cit. on p. 79).

Kearney, Gavin, Claire Masterson, Stephen Adams, and Frank Boland (2009). "Dynamic time warping for acoustic response interpolation: Possibilities and limitations". In: *2009 17th European Signal Processing Conference*. IEEE, pp. 705–709 (cit. on p. 67).

Keen, Rachel and Richard L. Freyman (2009). "Release and re-buildup of listeners' models of auditory space". In: *The Journal of the Acoustical Society of America* 125.5, pp. 3243–3252 (cit. on p. 111).

Kim, Duck O., Brian B. Bishop, and Shigeyuki Kuwada (2010). "Acoustic cues for sound source distance and azimuth in rabbits, a racquetball and a rigid spherical model". In: *Journal of the Association for Research in Otolaryngology* 11.4, pp. 541–557 (cit. on p. 40).

Kim, Duck O., Pavel Zahorik, Laurel H. Carney, Brian B. Bishop, and Shigeyuki Kuwada (2015). "Auditory distance coding in rabbit midbrain neurons and human perception: Monaural amplitude modulation depth as a cue". In: *Journal of Neuroscience* 35.13, pp. 5360–5372 (cit. on p. 9).

Kim, Hae-Young, Yôiti Suzuki, Shouichi Takane, and Toshio Sone (2001). "Control of auditory distance perception based on the auditory parallax model". In: *Applied Acoustics* 62.3, pp. 245–270 (cit. on pp. 90, 115).

Kirchner, W.H. and J. Röschard (1999). "Hissing in bumblebees: An interspecific defence signal". In: *Insectes sociaux* 46.3, pp. 239–243 (cit. on p. 23).

Kirkup, Stephen (2019). "The boundary element method in acoustics: A survey". In: *Applied Sciences* 9.8, p. 1642 (cit. on p. 62).

Kissner, Kelley J., Mark R. Forbes, and Diane M. Secoy (1997). "Rattling behavior of prairie rattlesnakes (*Crotalus viridis viridis*, Viperidae) in relation to sex, reproductive status, body size, and body temperature". In: *Ethology* 103.12, pp. 1042–1050 (cit. on p. 47).

Klatzky, Roberta L. (1998). "Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections". In: *Spatial Cognition*. Ed. by Christian Freksa, Christopher Habel, and Karl F. Wender. Vol. 1404. Lecture Notes in Computer Science. Berlin, Heidelberg, pp. 1–17. ISBN: 978-3-540-64603-7 (cit. on p. 21).

Kleiner, Mendel, Bengt-Inge Dalenbäck, and Peter Svensson (1993). "Auralization—An Overview". In: *Journal of the Audio Engineering Society* 41.11, pp. 861–875 (cit. on pp. 61, 66).

Klumpp, R.G. and H.R. Eady (1956). "Some measurements of interaural time difference thresholds". In: *The Journal of the Acoustical Society of America* 28.5, pp. 859–860 (cit. on p. 6).

Kneip, Laurent and Claude Baumann (2008). "Binaural model for artificial spatial sound localization based on interaural time delays and movements of the interaural axis". In: *The Journal of the Acoustical Society of America* 124.5, pp. 3108–3119 (cit. on p. 36).

Knörnschild, Mirjam (2014). "Vocal production learning in bats". In: *Current Opinion in Neurobiology* 28, pp. 80–85 (cit. on pp. 23, 93, 118).

Knörnschild, Mirjam, Oliver Behr, and Otto von Helversen (2006). "Babbling behavior in the sac-winged bat *(Saccopteryx bilineata)*". In: *Naturwissenschaften* 93.9, pp. 451–454 (cit. on p. 93).

Knörnschild, Mirjam, Simone Blüml, Patrick Steidl, Maria Eckenweber, and Martina Nagy (2017). "Bat songs as acoustic beacons—male territorial songs attract dispersing females". In: *Scientific Reports* 7, 13918 (cit. on p. 102).

Knörnschild, Mirjam, Marion Feifel, and Elisabeth K.V. Kalko (2014). "Male courtship displays and vocal communication in the polygynous bat *Carollia perspicillata*". In: *Behaviour* 151.6, pp. 781–798 (cit. on pp. 100, 102).

Knörnschild, Mirjam, Volker Glöckner, and Otto von Helversen (2010). "The vocal repertoire of two sympatric species of nectar-feeding bats (*Glossophaga soricina* and *G. commissarisi*)". In: *Acta Chiropterologica* 12.1, pp. 205–215 (cit. on p. 93).

Knörnschild, Mirjam, Martina Nagy, Markus Metz, Frieder Mayer, and Otto von Helversen (2010). "Complex vocal imitation during ontogeny in a bat". In: *Biology Letters* 6.2, pp. 156–159 (cit. on p. 93).

Kock, Winston E. (1950). "Binaural localization and masking". In: *The Journal of the Acoustical Society of America* 22.6, pp. 801–804 (cit. on p. 112).

Kohl, Tobias and Bruce A. Young (2011). "Electrophysiology of the snake retina". In: *Annual Meeting of the SICB*. Vol. 51. Society for Integrative & Comparative Biology, E71 (cit. on p. 53).

Kolarik, Andrew J., Silvia Cirstea, and Shahina Pardhan (2013). "Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues". In: *The Journal of the Acoustical Society of America* 134.5, pp. 3395–3398 (cit. on p. 9).

Kolarik, Andrew J., Brian C.J. Moore, Pavel Zahorik, Silvia Cirstea, and Shahina Pardhan (2016). "Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss". In: *Attention, Perception, & Psychophysics* 78.2, pp. 373–395 (cit. on pp. 8, 36, 85).

Kopčo, Norbert, Keerthi Kumar Doreswamy, Samantha Huang, Stephanie Rossi, and Jyrki Ahveninen (2020). "Cortical auditory distance representation based on direct-to-reverberant energy ratio". In: *NeuroImage* 208, 116436 (cit. on p. 10).

Kopčo, Norbert, Samantha Huang, John W. Belliveau, Tommi Raij, Chinmayi Tengshe, and Jyrki Ahveninen (2012). "Neuronal representations of distance in human auditory cortex". In: *Proceedings of the National Academy of Sciences* 109.27, pp. 11019–11024 (cit. on pp. 9, 40).

Kopčo, Norbert and Barbara G. Shinn-Cunningham (2011). "Effect of stimulus spectrum on distance perception for nearby sources". In: *The Journal of the Acoustical Society of America* 130.3, pp. 1530–1541 (cit. on pp. 9, 90).

Kopp-Scheinpflug, Cornelia, William R. Lippe, Gerd J. Dörrscheidt, and Rudolf Rübsamen (2003). "The medial nucleus of the trapezoid body in the gerbil is more than a relay: comparison of pre- and postsynaptic activity". In: *Journal of the Association for Research in Otolaryngology* 4.1, pp. 1–23 (cit. on p. 6).

Koutsouris, Georgios I., Jonas Brunskog, Cheol-Ho Jeong, and Finn Jacobsen (2013). "Combination of acoustical radiosity and the image source method". In: *The Journal of the Acoustical Society of America* 133.6, pp. 3963–3974 (cit. on p. 65).

Krauth, Ekkehard and Roland Bücklein (1962). "Neuere Ergebnisse raumakustischer Modellversuche". In: *Elektroakustik II*. Ed. by Johannes Wosnik. Vol. 26. Nachrichtentechnische Fachberichte. Springer, pp. 53–56. ISBN: 978-3-663-00472-1 (cit. on p. 61).

Krokstad, Asbjørn, Staffan Strom, and Svein Sørsdal (1968). "Calculating the acoustical room response by the use of a ray tracing technique". In: *Journal of Sound and Vibration* 8.1, pp. 118–125 (cit. on p. 62).

Kuhl, Patricia K. and Andrew N. Meltzoff (1996). "Infant vocalizations in response to speech: Vocal imitation and developmental change". In: *The Journal of the Acoustical Society of America* 100.4, pp. 2425–2438 (cit. on p. 22).

Kulowski, Andrzej (1982). "Error investigation for the ray tracing technique". In: *Applied Acoustics* 15.4, pp. 263–274 (cit. on p. 63).

Kuttruff, Heinrich (1971). "Simulierte Nachhallkurven in Rechteckräumen mit diffusem Schallfeld". In: *Acustica* 25.6, pp. 333–342 (cit. on p. 64).

— (2007). *Acoustics: An introduction*. New York: Taylor & Francis. ISBN: 978-0-415-38679-1 (cit. on p. 1).

— (2016). *Room Acoustics*. 6th edition. Boca Raton: CRC Press. ISBN: 978-1-4822-6043-4 (cit. on p. 62).

Kuwabara, Nobuyuki and John M. Zook (1992). "Projections to the medial superior olive from the medial and lateral nuclei of the trapezoid body in rodents and bats". In: *Journal of Comparative Neurology* 324.4, pp. 522–538 (cit. on p. 6).

Kuwada, Clinton A., Brian B. Bishop, Shigeyuki Kuwada, and Duck O. Kim (2010). "Acoustic recordings in human ear canals to sounds at different locations". In: *Otolaryngology—Head and Neck Surgery* 142.4, pp. 615–617 (cit. on p. 40).

Kwiecinski, Gary G. (2006). "*Phyllostomus discolor*". In: *Mammalian Species* 801, pp. 1–11 (cit. on p. 100).

Lappin, Joseph S., Duje Tadin, Jeffrey B. Nyquist, and Anne L. Corn (2009). "Spatial and temporal limits of motion perception across variations in speed, eccentricity, and low vision". In: *Journal of Vision* 9.1, pp. 30–30 (cit. on p. 39).

Larsen, Erik, Nandini Iyer, Charissa R. Lansing, and Albert S. Feng (2008). "On the minimum audible difference in direct-to-reverberant energy ratio". In: *The Journal of the Acoustical Society of America* 124.1, pp. 450–461 (cit. on pp. 10, 18).

Lattenkamp, Ella Z. (2020). "Vocal learning in the pale spear-nosed bat, *Phyllostomus discolor*". Dissertation. Nijmegen: Radboud University. ISBN: 978-94-92910-10-3 (cit. on pp. 23, 91).

Lattenkamp, Ella Z., Stephanie M. Shields, Michael Schutte, Jassica Richter, Meike Linnenschmidt, Sonja C. Vernes, and Lutz Wiegrebe (2019). "The vocal repertoire of pale spear-nosed bats in a social roosting context". In: *Frontiers in Ecology and Evolution* 7, 116 (cit. on pp. 25, 91).

Lattenkamp, Ella Z. and Sonja C. Vernes (2018). "Vocal learning: a language-relevant trait in need of a broad cross-species approach". In: *Current Opinion in Behavioral Sciences* 21, pp. 209–215 (cit. on pp. 23, 93, 118).

Lattenkamp, Ella Z., Sonja C. Vernes, and Lutz Wiegrebe (2018). "Volitional control of social vocalisations and vocal usage learning in bats". In: *Journal of Experimental Biology* 221.14 (cit. on p. 93).

— (2020). "Vocal production learning in the pale spear-nosed bat, *Phyllostomus discolor*". In: *Biology Letters* 16.4, p. 20190928 (cit. on p. 118).

Lewers, Tim (1993). "A combined beam tracing and radiant exchange computer model of room acoustics". In: *Applied Acoustics* 38.2-4, pp. 161–178 (cit. on p. 65).

Libesman, Sol, Thomas J. Whitford, and Damien J. Mannion (preprint). "Loudness judgements are not necessarily affected by visual cues to sound source distance". Downloaded on 20 October 2020. URL: https://psyarxiv.com/kjb78 (cit. on p. 112).

Lindau, Alexander (2009). "The perception of system latency in dynamic binaural synthesis". In: *Proceedings of the International Conference on Acoustics NAG/DAGA 2009*. Rotterdam, pp. 1063–1066 (cit. on p. 68).

Lindau, Alexander, Vera Erbes, Steffen Lepa, Hans-Joachim Maempel, Fabian Brinkmann, and Stefan Weinzierl (2014). "A spatial audio quality inventory (SAQI) ". In: *Acta Acustica united with Acustica* 100.5, pp. 984–994 (cit. on p. 29).

Lindau, Alexander, Hans-Joachim Maempel, and Stefan Weinzierl (2008). "Minimum BRIR grid resolution for dynamic binaural synthesis". In: *Journal of the Acoustical Society of America* 123.5, p. 3498 (cit. on p. 67).

Lindau, Alexander and Stefan Weinzierl (2012). "Assessing the plausibility of virtual acoustic environments". In: *Acta Acustica united with Acustica* 98.5, pp. 804–810 (cit. on p. 68).

Litovsky, Ruth Y., H. Steven Colburn, William A. Yost, and Sandra J. Guzman (1999). "The precedence effect". In: *The Journal of the Acoustical Society of America* 106.4, pp. 1633–1654 (cit. on pp. 15, 28).

Litovsky, Ruth Y., Brian J. Fligor, and Mark J. Tramo (2002). "Functional role of the human inferior colliculus in binaural hearing". In: *Hearing Research* 165.1-2, pp. 177–188 (cit. on p. 16).

Little, Alex D., Donald H. Mershon, and Patrick H. Cox (1992). "Spectral content as a cue to perceived auditory distance". In: *Perception* 21.3, pp. 405–416 (cit. on p. 9).

Lochner, J.P.A. and J.F. Burger (1958). "The subjective masking of short time delayed echoes by their primary sounds and their contribution to the intelligibility of speech". In: *Acustica* 8.1, pp. 1–10 (cit. on p. 15).

Lokki, Tapio (2002). "Physically-based Auralization—Design, Implementation, and Evaluation". Dissertation. Espoo: Helsinki University of Technology. ISBN: 978-951-22-6157-4 (cit. on p. 68).

London, Sam, Christopher W. Bishop, and Lee M. Miller (2012). "Spatial attention modulates the precedence effect." In: *Journal of Experimental Psychology: Human Perception and Performance* 38.6, pp. 1371–1379 (cit. on p. 113).

Loomis, Jack M., Roberta L. Klatzky, John W. Philbeck, and Reginald G. Golledge (1998). "Assessing auditory distance perception using perceptually directed action". In: *Perception & Psychophysics* 60.6, pp. 966–980 (cit. on pp. 32, 36, 39).

Lu, Thomas, Li Liang, and Xiaoqin Wang (2001). "Neural representations of temporally asymmetric stimuli in the auditory cortex of awake primates". In: *Journal of Neurophysiology* 85.6, pp. 2364–2380 (cit. on p. 11).

Luo, Jinhong, Holger R. Goerlitz, Henrik Brumm, and Lutz Wiegrebe (2015). "Linking the sender to the receiver: vocal adjustments by bats to maintain signal detection in noise". In: *Scientific Reports* 5, 18556 (cit. on p. 100).

Luo, Jinhong, Andrea Lingner, Uwe Firzlaff, and Lutz Wiegrebe (2017). "The Lombard effect emerges early in young bats: Implications for the development of audio-vocal integration". In: *Journal of Experimental Biology* 220.6, pp. 1032–1037 (cit. on p. 93).

Lutfi, Robert A. and Wen Wang (1999). "Correlational analysis of acoustic cues for the discrimination of auditory motion". In: *The Journal of the Acoustical Society of America* 106.2, pp. 919–928 (cit. on p. 10).

Ma, Jie, Kohta Kobayasi, Shuyi Zhang, and Walter Metzner (2006). "Vocal communication in adult greater horseshoe bats, *Rhinolophus ferrumequinum*". In: *Journal of Comparative Physiology A* 192.5, pp. 535–550 (cit. on pp. 92, 93, 95).

Mackensen, Philip (2004). "Auditive Localization. Head movements, an additional cue in Localization". Dissertation. Berlin: Technische Universität Berlin (cit. on p. 68).

Maempel, Hans-Joachim and Michael Horn (2018). "Audiovisual perception of real and virtual rooms". In: *Journal of Virtual Reality and Broadcasting* 14.5 (cit. on pp. 20, 113).

Maempel, Hans-Joachim and Matthias Jentsch (2013). "Auditory and visual contribution to egocentric distance and room size perception". In: *Building Acoustics* 20.4, pp. 383–401 (cit. on pp. 20, 112).

van Maercke, Dirk (1986). "Simulation of sound fields in time and frequency domain using a geometrical model". In: *Proceedings of the 12th International Congress on Acoustics* 2, E11–7 (cit. on p. 65).

Malacarne, Giorgio and Cristina Giacoma (1986). "Chemical signals in European newt courtship". In: *Italian Journal of Zoology* 53.1, pp. 79–83 (cit. on p. 22).

Malavasi, Rachele and Ludwig Huber (2016). "Evidence of heterospecific referential communication from domestic horses *(Equus caballus)* to humans". In: *Animal Cognition* 19.5, pp. 899–909 (cit. on p. 23).

Maldonado, Héctor (1970). "The deimatic reaction in the praying mantis *Stagmatoptera biocellata*". In: *Zeitschrift für vergleichende Physiologie* 68.1, pp. 60–71 (cit. on p. 23).

Mann, Zoe F. and Matthew W. Kelley (2011). "Development of tonotopy in the auditory periphery". In: *Hearing Research* 276.1-2, pp. 2–15 (cit. on p. 4).

Markl, Hubert (1965). "Stridulation in leaf-cutting ants". In: *Science* 149.3690, pp. 1392–1393 (cit. on p. 22).

Marozeau, Jeremy, Michael Epstein, Mary Florentine, and Becky Daley (2006). "A test of the binaural equal-loudness-ratio hypothesis for tones". In: *The Journal of the Acoustical Society of America* 120.6, pp. 3870–3877 (cit. on p. 5).

Martin, James H. and Roland M. Bagby (1972). "Temperature-frequency relationship of the rattlesnake rattle". In: *Copeia*, pp. 482–485 (cit. on pp. 24, 46, 47).

— (1973). "Properties of rattlesnake shaker muscle". In: *Journal of Experimental Zoology* 185.3, pp. 293–300 (cit. on p. 24).

Mason, Matthew J. (2016). "Structure and function of the mammalian middle ear. II: Inferring function from structure". In: *Journal of Anatomy* 228.2, pp. 300–312 (cit. on p. 4).

Massiceti, Daniela, Stephen Lloyd Hicks, and Joram Jacob van Rheede (2018). "Stereosonic vision: Exploring visual-to-auditory sensory substitution mappings in an immersive virtual reality navigation paradigm". In: *PLOS ONE* 13.7, e0199389 (cit. on p. 119).

Maynard Smith, John and David Harper (2003). *Animal Signals*. Oxford Series in Ecology and Evolution. Oxford: Oxford University Press. ISBN: 978-0-19-852684-1 (cit. on p. 22).

McCreery, Anthony and Paul Calamia (2006). "Cross-modal perception of room acoustics". In: *The Journal of the Acoustical Society of America* 120.5, pp. 3150–3150 (cit. on pp. 20, 29, 111).

McGovern, Stephen G. (2009). "Fast image method for impulse response calculations of box-shaped rooms". In: *Applied Acoustics* 70.1, pp. 182–189 (cit. on p. 64).

McGowan, Richard S. and Roman Kuc (1982). "A direct relation between a signal time series and its unwrapped phase". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30.5, pp. 719–726 (cit. on p. 79).

McKee, Suzanne P. and Douglas G. Taylor (2010). "The precision of binocular and monocular depth judgments in natural settings". In: *Journal of Vision* 10.10, pp. 5–5 (cit. on p. 39).

Mehta, Madan L. and K.A. Mulholland (1976). "Effect of non-uniform distribution of absorption on reverberation time". In: *Journal of Sound and Vibration* 46.2, pp. 209–224 (cit. on p. 64).

Mendonça, Catarina, Pietro Mandelli, and Ville Pulkki (2016). "Modeling the perception of audiovisual distance: Bayesian causal inference and other models". In: *PLOS ONE* 11.12, e0165391 (cit. on p. 112).

Meredith, M. Alex and Barry E. Stein (1983). "Interactions among converging sensory inputs in the superior colliculus". In: *Science* 221.4608, pp. 389–391 (cit. on p. 19).

Mergner, Thomas and Thomas Rosemeier (1998). "Interaction of vestibular, somatosensory and visual signals for postural control and motion perception under terrestrial and microgravity conditions—a conceptual model". In: *Brain Research Reviews* 28.1-2, pp. 118–135 (cit. on p. 114).

Mershon, Donald H., William L. Ballenger, Alex D. Little, Patrick L. McMurtry, and Judith L. Buchanan (1989). "Effects of room reflectance and background noise on perceived auditory distance". In: *Perception* 18.3, pp. 403–416 (cit. on p. 9).

Mershon, Donald H. and John N. Bowers (1979). "Absolute and relative cues for the auditory perception of egocentric distance". In: *Perception* 8.3, pp. 311–322 (cit. on p. 9).

Mershon, Donald H., Douglas H. Desaulniers, Thomas L. Amerson, and Stephan A. Kiefer (1980). "Visual capture in auditory distance perception: Proximity image effect reconsidered." In: *Journal of Auditory Research* 20.2, pp. 129–136 (cit. on p. 20).

Mershon, Donald H. and L. Edward King (1975). "Intensity and reverberation as factors in the auditory perception of egocentric distance". In: *Perception & Psychophysics* 18.6, pp. 409–415 (cit. on p. 8).

Meyer, Jens and Gary W. Elko (2004). "Spherical microphone arrays for 3D sound recording". In: *Audio Signal Processing for Next-Generation Multimedia Communication Systems*. Ed. by Yiteng Huang and Jacob Benesty. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers, pp. 67–89. ISBN: 978-1-4020-7768-5 (cit. on p. 14).

Michael, R.R. and Julius A. Rudinsky (1972). "Sound production in Scolytidae: specificity in male Dendroctonus beetles". In: *Journal of Insect Physiology* 18.11, pp. 2189–2201 (cit. on p. 22).

Middlebrooks, John C. and David M. Green (1991). "Sound localization by human listeners". In: *Annual Review of Psychology* 42.1, pp. 135–159 (cit. on pp. 5, 7, 51).

Middlebrooks, John C., James C. Makous, and David M. Green (1989). "Directional sensitivity of sound-pressure levels in the human ear canal". In: *The Journal of the Acoustical Society of America* 86.1, pp. 89–108 (cit. on p. 7).

Miller, George A. (1947). "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness". In: *The Journal of the Acoustical Society of America* 19.4, pp. 609–619 (cit. on p. 10).

Mills, Allen W. (1958). "On the minimum audible angle". In: *The Journal of the Acoustical Society of America* 30.4, pp. 237–246 (cit. on p. 6).

— (1960). "Lateralization of high-frequency tones". In: *The Journal of the Acoustical Society of America* 32.1, pp. 132–134 (cit. on p. 7).

Milne, Jennifer L., Melvyn A. Goodale, and Lore Thaler (2014). "The role of head movements in the discrimination of 2-D shape by blind echolocation experts". In: *Attention, Perception, & Psychophysics* 76.6, pp. 1828–1837 (cit. on p. 68).

Moore, Brian C.J. (2014). "Development and current status of the 'Cambridge' loudness models". In: *Trends in Hearing* 18, 2331216514550620 (cit. on p. 5).

Moore, Brian C.J. and Brian R. Glasberg (2007). "Modeling binaural loudness". In: *The Journal of the Acoustical Society of America* 121.3, pp. 1604–1612 (cit. on p. 5).

Moore, Brian C.J., Brian R. Glasberg, Ajanth Varathanathan, and Josef Schlittenlacher (2016). "A loudness model for time-varying sounds incorporating binaural inhibition". In: *Trends in hearing* 20, p. 2331216516682698 (cit. on pp. 5, 51, 89, 116).

Moore, Gregg R. (1984). "An approach to the analysis of sound in auditoria: Model design and computer implementation". Dissertation. Cambridge: University of Cambridge (cit. on p. 65).

Moore, Maurus J. and Donald M. Caspary (1983). "Strychnine blocks binaural inhibition in lateral superior olivary neurons". In: *Journal of Neuroscience* 3.1, pp. 237–242 (cit. on p. 7).

Morein-Zamir, Sharon, Salvador Soto-Faraco, and Alan Kingstone (2003). "Auditory capture of vision: examining temporal ventriloquism". In: *Cognitive Brain Research* 17.1, pp. 154–163 (cit. on p. 19).

Moscatelli, Alessandro, Vincent Hayward, Mark Wexler, and Marc O. Ernst (2015). "Illusory tactile motion perception: An analog of the visual Filehne illusion". In: *Scientific Reports* 5, 14584 (cit. on p. 39).

Murray, Kevin L. and Theodore H. Fleming (2008). "Social structure and mating system of the buffy flower bat, Erophylla sezekorni (Chiroptera, Phyllostomidae)". In: *Journal of Mammalogy* 89.6, pp. 1391–1400 (cit. on p. 102).

Náběl ek, Anna K. and Paul A. Dagenais (1986). "Vowel errors in noise and in reverberation by hearing-impaired listeners". In: *The Journal of the Acoustical Society of America* 80.3, pp. 741–748 (cit. on p. 9).

Naji, Jenny J. and Tom C.A. Freeman (2004). "Perceiving depth order during pursuit eye movement". In: *Vision Research* 44.26, pp. 3025–3034 (cit. on p. 39).

Al-Nashi, Hamid (1989). "Phase unwrapping of digital signals". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.11, pp. 1693–1702 (cit. on p. 79).

Nawrot, Mark and Keith Stroyan (2009). "The motion/pursuit law for visual depth perception from motion parallax". In: *Vision Research* 49.15, pp. 1969–1978 (cit. on p. 40).

Naylor, Graham M (1993). "ODEON—Another hybrid room acoustical model". In: *Applied Acoustics* 38.2-4, pp. 131–143 (cit. on p. 65).

Nelson, John E. (1964). "Vocal communication in Australian flying foxes (Pteropodidae; Megachiroptera)". In: *Zeitschrift für Tierpsychologie* 21.7, pp. 857–870 (cit. on p. 102).

Neuhoff, John G. (1998). "Perceptual bias for rising tones". In: *Nature* 395.6698, pp. 123–124 (cit. on pp. 10, 11, 51, 85, 86, 115, 116).

— (2001). "An adaptive bias in the perception of looming auditory motion". In: *Ecological Psychology* 13.2, pp. 87–110 (cit. on pp. 10, 11, 51).

— (2016). "Looming sounds are perceived as faster than receding sounds". In: *Cognitive Research: Principles and Implications* 1.1, pp. 1–9 (cit. on pp. 11, 88, 116).

Neuhoff, John G., Rianna Planisek, and Erich Seifritz (2009). "Adaptive sex differences in auditory motion perception: Looming sounds are special." In: *Journal of Experimental Psychology: Human Perception and Performance* 35.1, p. 225 (cit. on pp. 85, 116).

Nielsen, Jens Bo and Torsten Dau (2010). "Revisiting perceptual compensation for effects of reverberation in speech identification". In: *The Journal of the Acoustical Society of America* 128.5, pp. 3088–3094 (cit. on pp. 16, 17).

Nosal, Eva-Marie, Murray Hodgson, and Ian Ashdown (2004). "Improved algorithms and methods for room sound-field prediction by acoustical radiosity in arbitrary polyhedral rooms". In: *The Journal of the Acoustical Society of America* 116.2, pp. 970–980 (cit. on p. 65).

Olsen, Kirk N., Catherine J. Stevens, and Julien Tardieu (2010). "Loudness change in response to dynamic acoustic intensity." In: *Journal of Experimental Psychology: Human Perception and Performance* 36.6, p. 1631 (cit. on p. 11).

Ono, Mika E., Josée Rivest, and Hiroshi Ono (1986). "Depth perception as a function of motion parallax and absolute-distance information." In: *Journal of Experimental Psychology: Human Perception and Performance* 12.3, p. 331 (cit. on p. 113).

Palmer, Alan R. and Ian J. Russell (1986). "Phase-locking in the cochlear nerve of the guinea-pig and its relation to the receptor potential of inner hair-cells". In: *Hearing Research* 24.1, pp. 1–15 (cit. on p. 6).

Panerai, Francesco, Valérie Cornilleau-Pérès, and Jacques Droulez (2002). "Contribution of extraretinal signals to the scaling of object distance during self-motion". In: *Perception & Psychophysics* 64.5, pp. 717–731 (cit. on p. 39).

Pastore, M. Torben, Yi Zhou, and William A. Yost (2020). "Cross-modal and cognitive processes in sound localization". In: *The Technology of Binaural Understanding*. Springer, pp. 315–350. ISBN: 978-3-030-00385-2 (cit. on p. 115).

Paul, Stephan (2009). "Binaural recording technology: A historical review and possible future developments". In: *Acta Acustica united with Acustica* 95.5, pp. 767–788 (cit. on p. 14).

Payne, Roger S. (1971). "Acoustic location of prey by barn owls *(Tyto alba)*". In: *Journal of Experimental Biology* 54.3, pp. 535–573 (cit. on p. 5).

Pecka, Michael, Thomas P. Zahn, Bernadette Saunier-Rebori, Ida Siveke, Felix Felmy, Lutz Wiegrebe, Achim Klug, George D. Pollak, and Benedikt Grothe (2007). "Inhibiting the inhibition: a neuronal network for sound localization in reverberant environments". In: *Journal of Neuroscience* 27.7, pp. 1782–1790 (cit. on p. 16).

Pepperberg, Irene M. (2002). "Cognitive and communicative abilities of grey parrots". In: *Current Directions in Psychological Science* 11.3, pp. 83–87 (cit. on p. 23).

Perrett, Stephen and William Noble (1997a). "The contribution of head motion cues to localization of low-pass noise". In: *Perception & Psychophysics* 59.7, pp. 1018–1026 (cit. on p. 21).

— (1997b). "The effect of head rotations on vertical plane sound localization". In: *The Journal of the Acoustical Society of America* 102.4, pp. 2325–2332 (cit. on p. 21).

Perrott, David R., Thomas Z. Strybel, and Carol L. Manligas (1987). "Conditions under which the Haas precedence effect may or may not occur." In: *The Journal of Auditory Research* 27.1, pp. 59–72 (cit. on p. 15).

Pfalzer, Guido and Jürgen Kusch (2003). "Structure and variability of bat social calls: implications for specificity and individual recognition". In: *Journal of Zoology* 261.1, pp. 21–33 (cit. on p. 102).

Philbeck, John W. and Donald H. Mershon (2002). "Knowledge about typical source output influences perceived auditory distance". In: *The Journal of the Acoustical Society of America* 111.5, pp. 1980–1983 (cit. on p. 8).

Picaut, Judicaël and Nicolas Fortin (2012). "SPPS, a particle-tracing numerical code for indoor and outdoor sound propagation prediction". In: *Proceedings of the Acoustics 2012 Nantes Conference*, pp. 1417–1422 (cit. on p. 63).

Pick, Herbert L., David H. Warren, and John C. Hay (1969). "Sensory conflict in judgments of spatial direction". In: *Perception & Psychophysics* 6.4, pp. 203–205 (cit. on p. 19).

Pilley, John W. and Alliston K. Reid (2011). "Border collie comprehends object names as verbal referents". In: *Behavioural processes* 86.2, pp. 184–195 (cit. on p. 23).

Place, Aaron J. and Charles I. Abramson (2008). "Habituation of the rattle response in Western Diamondback rattlesnakes, *Crotalus atrox*". In: *Copeia* 2008.4, pp. 835–843 (cit. on pp. 48, 54).

Ploog, Detlev W. (1992). "The Evolution of Vocal Communication". In: *Nonverbal Vocal Communication: Comparative and Developmental Approaches*. Ed. by Hanuš Papoušek, Uwe Jürgens, and

Mechthild Papoušek. Vol. 6. Cambridge & Paris: Cambridge University Press & Editions de la Maison des Sciences de l'Homme. ISBN: 978-0-521-41265-0 (cit. on p. 22).

Pollack, Gerald S. (2014). "Neurobiology of acoustically mediated predator detection". In: *Journal of Comparative Physiology A* 201.1, pp. 99–109 (cit. on p. 5).

Poole, Joyce H., Peter L. Tyack, Angela S. Stoeger-Horwath, and Stephanie Watwood (2005). "Elephants are capable of vocal learning". In: *Nature* 434.7032, pp. 455–456 (cit. on p. 23).

Postma, Barteld N.J. and Brian F.G. Katz (2017a). "Influence of visual rendering on the acoustic judgements of a theater auralization". In: *Proceedings of Meetings on Acoustics 173*. Vol. 30. 1. Acoustical Society of America, 015008 (cit. on p. 112).

— (2017b). "The influence of visual distance on the room-acoustic experience of auralizations". In: *The Journal of the Acoustical Society of America* 142.5, pp. 3035–3046 (cit. on p. 28).

Poulton, Edward Bagnall (1890). *The Colours of Animals: Their Meaning and Use Especially Considered in the Case of Insects*. New York: D. Appleton and Company (cit. on p. 23).

Prat, Yosef, Mor Taub, and Yossi Yovel (2016). "Everyday bat vocalizations contain information about emitter, addressee, context, and behavior". In: *Scientific Reports* 6, 39419 (cit. on p. 102).

Prior, Kent A. and Patrick J. Weatherhead (1994). "Response of free-ranging eastern massasauga rattlesnakes to human disturbance". In: *Journal of Herpetology* 28.2, pp. 255–257 (cit. on p. 47).

Pulkki, Ville (1997). "Virtual sound source positioning using vector base amplitude panning". In: *Journal of the Audio Engineering Society* 45.6, pp. 456–466 (cit. on pp. 30, 58, 70, 77).

Qian, Ning (1997). "Binocular disparity and the perception of depth". In: *Neuron* 18.3, pp. 359–368 (cit. on p. 36).

Rabiner, Lawrence R. and Bernard Gold (1975). *Theory and application of digital signal processing*. Englewood Cliffs: Prentice-Hall. ISBN: 978-0-13-914101-0 (cit. on p. 67).

Rakerd, Brad, William M. Hartmann, and Joy Hsu (2000). "Echo suppression in the horizontal and median sagittal planes". In: *The Journal of the Acoustical Society of America* 107.2, pp. 1061–1064 (cit. on p. 15).

Rauschecker, Josef P. and Biao Tian (2000). "Mechanisms and streams for processing of 'what' and 'where' in auditory cortex". In: *Proceedings of the National Academy of Sciences* 97.22, pp. 11800–11806 (cit. on p. 10).

Rébillat, Marc, Xavier Boutillon, Étienne Corteel, and Brian F.G. Katz (2012). "Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments". In: *ACM Transactions on Applied Perception* 9.4, pp. 1–17 (cit. on p. 115).

Recanzone, Gregg H. (1998). "Rapidly induced auditory plasticity: the ventriloquism aftereffect". In: *Proceedings of the National Academy of Sciences* 95.3, pp. 869–875 (cit. on p. 112).

Reinert, Howard K., David Cundall, and Lauretta M. Bushar (1984). "Foraging behavior of the timber rattlesnake, *Crotalus horridus*". In: *Copeia*, pp. 976–981 (cit. on p. 55).

Reiserer, Randall S. and Gordon W. Schuett (2016). "The Origin and Evolution of the Rattlesnake Rattle: Misdirection, Clarification, Theory, and Progress". In: *Rattlesnakes of Arizona*. Ed. by Gordon W. Schuett, Martin J. Feldner, Charles F. Smith, and Randall S. Reiserer. Vol. 2 (cit. on pp. 46, 117).

Reiss, Diana and Brenda McCowan (1993). "Spontaneous vocal mimicry and production by bottlenose dolphins *(Tursiops truncatus)*: evidence for vocal learning." In: *Journal of Comparative Psychology* 107.3, p. 301 (cit. on p. 23).

Rhode, William S. and Steven Greenberg (1992). "Physiology of the cochlear nuclei". In: *The mammalian auditory pathway: Neurophysiology*. Springer, pp. 94–152 (cit. on p. 4).

Rindel, Jens H. (2002). "Modelling in auditorium acoustics. From ripple tank and scale models to computer simulations". In: *Revista de Acústica* 33.3-4, pp. 31–35 (cit. on pp. 61, 62).

Ritsma, Roelof J. (1967). "Frequencies dominant in the perception of the pitch of complex sounds". In: *The Journal of the Acoustical Society of America* 42.1, pp. 191–198 (cit. on p. 42).

Rodenas-Cuadrado, Pedro, Xiaowei Sylvia Chen, Lutz Wiegrebe, Uwe Firzlaff, and Sonja C. Vernes (2015). "A novel approach identifies the first transcriptome networks in bats: A new genetic model for vocal communication". In: *BMC Genomics* 16.1, pp. 1–18 (cit. on p. 93).

Roffler, Suzanne K. and Robert A. Butler (1968). "Factors that influence the localization of sound in the vertical plane". In: *The Journal of the Acoustical Society of America* 43.6, pp. 1255–1259 (cit. on p. 8).

Rogers, Brian J. and Maureen Graham (1979). "Motion parallax as an independent cue for depth perception". In: *Perception* 8.2, pp. 125–134 (cit. on pp. 36, 113).

Romani, Gian Luca, Samuel J. Williamson, and Lloyd Kaufman (1982). "Tonotopic organization of the human auditory cortex". In: *Science* 216.4552, pp. 1339–1340 (cit. on p. 4).

Rosenblum, Lawrence D., A. Paige Wuestefeld, and Helena M. Saldana (1993). "Auditory looming perception: Influences on anticipatory judgments". In: *Perception* 22.12, pp. 1467–1482 (cit. on p. 115).

Rother, G. and Uwe Schmidt (1985). "Die ontogenetische Entwicklung der Vokalisation bei *Phyllostomus discolor* (Chiroptera)". In: *Zeitschrift für Säugetierkunde* 50.1, pp. 17–26 (cit. on p. 100).

Rowe, Candy and T.I.M. Guilford (1999). "Novelty effects in a multimodal warning signal". In: *Animal Behaviour* 57.2, pp. 341–346 (cit. on p. 23).

Ruggero, Mario A. (1992). "Responses to sound of the basilar membrane of the mammalian cochlea". In: *Current Opinion in Neurobiology* 2.4, pp. 449–456 (cit. on p. 4).

Russ, Jon M., Gareth Jones, Iain J. Mackie, and Paula A. Racey (2004). "Interspecific responses to distress calls in bats (Chiroptera: Vespertilionidae): a function for convergence in call design?" In: *Animal Behaviour* 67.6, pp. 1005–1014 (cit. on pp. 93, 102).

Saberi, Kourosh and David R. Perrott (1990). "Minimum audible movement angles as a function of sound source trajectory". In: *The Journal of the Acoustical Society of America* 88.6, pp. 2639–2644 (cit. on p. 39).

Salmon, François, Étienne Hendrickx, Nicolas Épain, and Mathieu Paquier (2020). "The Influence of Vision on Perceived Differences Between Sound Spaces". In: *Journal of the Audio Engineering Society* 68.7/8, pp. 522–531 (cit. on pp. 112, 113).

Samarasinghe, Prasanga, Thushara Abhayapala, Mark Poletti, and Terence Betlehem (2015). "An efficient parameterization of the room transfer function". In: *IEEE/ACM Transactions on Audio, Speech and Language Processing* 23.12, pp. 2217–2227 (cit. on p. 67).

Sanders, Lisa D., Amy S. Joh, Rachel E. Keen, and Richard L. Freyman (2008). "One sound or two? Object-related negativity indexes echo perception". In: *Perception & Psychophysics* 70.8, pp. 1558–1570 (cit. on p. 16).

Sandvad, Jesper (1996). "Dynamic aspects of auditory virtual environments". In: *Audio Engineering Society Convention 100*. Paper no. 4226. Audio Engineering Society. Copenhagen (cit. on p. 68).

Santon, François (1976). "Numerical prediction of echograms and of the intelligibility of speech in rooms". In: *The Journal of the Acoustical Society of America* 59.6, pp. 1399–1405 (cit. on p. 64).

Satterthwaite, Franklin E. (1946). "An approximate distribution of estimates of variance components". In: *Biometrics Bulletin* 2.6, pp. 110–114 (cit. on p. 31).

Savioja, Lauri, Jyri Huopaniemi, Tapio Lokki, and Riitta Väänänen (1999). "Creating interactive virtual acoustic environments". In: *Journal of the Audio Engineering Society* 47.9, pp. 675–705 (cit. on pp. 14, 68).

Savioja, Lauri and U. Peter Svensson (2015). "Overview of geometrical room acoustic modeling techniques". In: *The Journal of the Acoustical Society of America* 138.2, pp. 708–730 (cit. on pp. 62–64, 153).

Schechtman, Eitan, Talia Shrem, and Leon Y. Deouell (2012). "Spatial localization of auditory stimuli in human auditory cortex is based on both head-independent and head-centered coordinate systems". In: *Journal of Neuroscience* 32.39, pp. 13501–13509 (cit. on p. 21).

Schlauch, Robert S., Dennis T. Ries, and Jeffrey J. DiGiovanni (2001). "Duration discrimination and subjective duration for ramped and damped sounds". In: *The Journal of the Acoustical Society of America* 109.6, pp. 2880–2887 (cit. on p. 11).

Schmidt, Justin O. (2004). "Venom and the good life in tarantula hawks (Hymenoptera: Pompilidae): How to eat, not be eaten, and live long". In: *Journal of the Kansas Entomological Society* 77.4, pp. 402–413 (cit. on p. 23).

Schmitz, Oliver, Michael Vorländer, Stefan Feistel, and Wolfgang Ahnert (2001). "Merging software for sound reinforcement systems and for room acoustics". In: *Audio Engineering Society Convention 110*. Paper no. 5352. Audio Engineering Society. Amsterdam (cit. on p. 65).

Schörnich, Sven, Andreas Nagy, and Lutz Wiegrebe (2012). "Discovering your inner bat: Echoacoustic target ranging in humans". In: *Journal of the Association for Research in Otolaryngology* 13.5, pp. 673–682 (cit. on p. 68).

Schröder, Dirk (2011). "Physically Based Real-Time Auralization of Interactive Virtual Environments". Dissertation. Aachen: RWTH. ISBN: 978-3-8325-3031-0 (cit. on pp. 65, 67, 118).

Schröder, Dirk and Michael Vorländer (2011). "RAVEN: A real-time framework for the auralization of interactive virtual environments". In: *Forum Acusticum 2011*. European Acoustics Association. Aalborg, pp. 1541–1546 (cit. on p. 65).

Schroeder, Manfred R. (1962). "Natural sounding artificial reverberation". In: *Journal of the Audio Engineering Society* 10.3, pp. 219–223 (cit. on pp. 62, 71).

— (1965). "New method of measuring reverberation time". In: *The Journal of the Acoustical Society of America* 37.6, pp. 1187–1188 (cit. on p. 3).

— (1969). "Digital simulation of sound transmission in reverberant spaces". In: *The Journal of the Acoustical Society of America* 47.2A, pp. 424–431 (cit. on p. 62).

Schroeder, Manfred R., Bishnu S. Atal, and Carol Bird (1962). "Digital computers in room acoustics". In: *Proceedings of the 4th International Congress on Acoustics*. Vol. 21. Copenhagen (cit. on p. 62).

Schutte, Michael, Stephan D. Ewert, and Lutz Wiegrebe (2019). "The percept of reverberation is not affected by visual room impression in virtual environments". In: *The Journal of the Acoustical Society of America* 145.3, EL229–EL235 (cit. on pp. 24, 27).

Scott-Phillips, Thomas C. (2008). "Defining biological communication". In: *Journal of Evolutionary Biology* 21.2, pp. 387–395 (cit. on p. 22).

Seebeck, August (1841). "Beobachtungen über einige Bedingungen der Entstehung von Tönen". In: *Annalen der Physik* 129.7, pp. 417–436 (cit. on p. 3).

Seeber, Bernhard U. and Samuel W. Clapp (2017). "Interactive simulation and free-field auralization of acoustic space with the rtSOFE". In: *The Journal of the Acoustical Society of America* 141.5, pp. 3974–3974 (cit. on pp. 64, 67, 118).

— (2020). "Auditory Room Learning and Adaptation to Sound Reflections". In: Springer, pp. 203–222. ISBN: 978-3-030-00385-2 (cit. on p. 111).

Seeber, Bernhard U., Matthias Müller, and Fritz Menzer (2016). "Does learning a room's reflections aid spatial hearing?" In: *Proceedings of the 22nd International Congress on Acoustics*. Paper no. 775. Buenos Aires (cit. on p. 111).

Seifritz, Erich, John G. Neuhoff, Deniz Bilecen, Klaus Scheffler, Henrietta Mustovic, Hartmut Schächinger, Raffaele Elefante, and Francesco Di Salle (2002). "Neural processing of auditory looming in the human brain". In: *Current Biology* 12.24, pp. 2147–2151 (cit. on pp. 11, 86).

Seraphim, Hans-Peter (1958). "Untersuchungen über die Unterschiedsschwelle exponentiellen Abklingens von Rauschbandimpulsen". In: *Acustica* 8.4, pp. 280–284 (cit. on p. 18).

Shams, Ladan, Yukiyasu Kamitani, and Shinsuke Shimojo (2000). "What you see is what you hear". In: *Nature* 408.6814, p. 788 (cit. on p. 19).

Shannon, Claude Elwood (1948). "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.3, pp. 379–423 (cit. on p. 11).

Shaw, Brian K., Richard S. McGowan, and Michael T. Turvey (1991). "An acoustic variable specifying time-to-contact". In: *Ecological Psychology* 3.3, pp. 253–261 (cit. on p. 115).

Shine, Richard, Li-Xin Sun, Mark Fitzgerald, and Michael Kearney (2002). "Antipredator responses of free-ranging pit vipers (*Gloydius shedaoensis*, Viperidae)". In: *Copeia* 2002.3, pp. 843–850 (cit. on p. 47).

Shinn-Cunningham, Barbara G., Patrick M. Zurek, and Nathaniel I. Durlach (1993). "Adjustment and discrimination measurements of the precedence effect". In: *The Journal of the Acoustical Society of America* 93.5, pp. 2923–2932 (cit. on p. 15).

Shoemake, Ken (1985). "Animating rotation with quaternion curves". In: *SIGGRAPH'85: Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques*. San Francisco, pp. 245–254 (cit. on p. 81).

Siltanen, Samuel, Tapio Lokki, and Lauri Savioja (2010). "Rays or waves? Understanding the strengths and weaknesses of computational room acoustics modeling techniques". In: *Proceedings of the International Symposium on Room Acoustics (ISRA)*. Melbourne (cit. on p. 62).

Siltanen, Samuel, Tapio Lokki, Sakari Tervo, and Lauri Savioja (2012). "Modeling incoherent reflections from rough room surfaces with image sources". In: *The Journal of the Acoustical Society of America* 131.6, pp. 4606–4614 (cit. on pp. 64, 71).

da Silva, José Aparecido (1985). "Scales for perceived egocentric distance in a large open field: Comparison of three psychophysical methods". In: *The American Journal of Psychology* 98.1, pp. 119–144 (cit. on p. 19).

Simpson, W.E. and Lee D. Stanton (1973). "Head movement does not facilitate perception of the distance of a source of sound". In: *The American Journal of Psychology* 86.1, p. 151 (cit. on pp. 21, 36, 114).

Smith, Philip H., Philip X. Joris, and Tom C. Yin (1998). "Anatomy and physiology of principal cells of the medial nucleus of the trapezoid body (MNTB) of the cat". In: *Journal of Neurophysiology* 79.6, pp. 3127–3142 (cit. on p. 6).

Smith, W. John (1977). *The Behavior of Communicating: An Ethological Approach*. Cambridge, Mass.: Harvard University Press. ISBN: 978-0-674-06466-9 (cit. on p. 22).

Smotherman, Michael, Mirjam Knörnschild, Grace Smarsh, and Kirsten Bohn (2016). "The origins and diversity of bat songs". In: *Journal of Comparative Physiology A* 202.8, pp. 535–554 (cit. on pp. 100, 102).

Sobel, Erik C. (1990). "The locust's use of motion parallax to measure distance". In: *Journal of Comparative Physiology A* 167.5, pp. 579–588 (cit. on p. 113).

Sonkusare, Saurabh, Michael Breakspear, and Christine Guo (2019). "Naturalistic stimuli in neuroscience: Critically acclaimed". In: *Trends in Cognitive Sciences* 23.8, pp. 699–714 (cit. on p. 119).

Spandöck, Friedrich (1934). "Akustische Modellversuche". In: *Annalen der Physik* 412.4, pp. 345–360 (cit. on p. 61).

Sparrow, Ephraim M. and Robert D. Cess (2018). *Radiation Heat Transfer*. Augmented edition. New York: Routledge. ISBN: 978-0-203-74138-2 (cit. on p. 65).

Speigle, Jon M. and Jack M. Loomis (1993). "Auditory distance perception by translating observers". In: *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*. IEEE. San Jose, pp. 92–99 (cit. on pp. 21, 36, 39, 114).

Spottiswoode, Claire N., Keith S. Begg, and Colleen M. Begg (2016). "Reciprocal signaling in honeyguide–human mutualism". In: *Science* 353.6297, pp. 387–389 (cit. on p. 23).

Stecker, G. Christopher and Ervin R. Hafter (2000). "An effect of temporal asymmetry on loudness". In: *The Journal of the Acoustical Society of America* 107.6, pp. 3358–3368 (cit. on p. 11).

Steffens, Henning, Steven van de Par, and Stephan D. Ewert (2019). "Perceptual relevance of speaker directivity modelling in virtual rooms". In: *Proceedings of the 23rd International Congress on Acoustics*. Aachen, pp. 2651–2658 (cit. on p. 70).

— (in revision). "The role of early and late reflections on perception of source orientation". In: *The Journal of the Acoustical Society of America* (cit. on p. 70).

Steiglitz, Kenneth and Bradley Dickinson (1982). "Phase unwrapping by factorization". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 30.6, pp. 984–991 (cit. on p. 79).

Stein, Barry E. and M. Alex Meredith (1993). *The Merging of the Senses*. Cambridge, Mass.: MIT Press. ISBN: 978-0-262-69301-1 (cit. on pp. 19, 28).

Stephenson, Uwe (1990). "Comparison of the mirror image source method and the sound particle simulation method". In: *Applied Acoustics* 29.1, pp. 35–72 (cit. on pp. 63, 64).

Stevens, Martin (2013). *Sensory Ecology, Behaviour, and Evolution*. Oxford: Oxford University Press. ISBN: 978-0-19-960178-3 (cit. on p. 22).

Stevens, Stanley Smith (1961). "To honor Fechner and repeal his law". In: *Science* 133.3446, pp. 80–86 (cit. on p. 5).

Stevens, Stanley Smith and Miguelina Guirao (1962). "Loudness, reciprocality, and partition scales". In: *The Journal of the Acoustical Society of America* 34.9B, pp. 1466–1471 (cit. on p. 49).

Stewart, George W. (1911). "The acoustic shadow of a rigid sphere, with certain applications in architectural acoustics and audition". In: *Physical Review (Series I)* 33.6, p. 467 (cit. on p. 90).

Stockham, Thomas G., Thomas M. Cannon, and Robert B. Ingebretsen (1975). "Blind deconvolution through digital signal processing". In: *Proceedings of the IEEE* 63.4, pp. 678–692 (cit. on p. 19).

Strauss, Holger (1998). "Implementing Doppler shifts for virtual auditory environments". In: *Audio Engineering Society Convention 104*. Paper no. 4687. Audio Engineering Society. Amsterdam (cit. on p. 81).

Strum, Robert D. and Donald E. Kirk (1988). *First Principles of Discrete Systems and Digital Signal Processing*. Reading: Addison-Wesley. ISBN: 978-0-201-09518-0 (cit. on p. 66).

Strutt, John William (1879). "XXXI. Investigations in optics, with special reference to the spectroscope". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 8.49, pp. 261–274 (cit. on p. 36).

— (1907). "On our perception of sound direction". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 13.74, pp. 214–232 (cit. on p. 6).

Strybel, Thomas Z., Carol L. Manllgas, and David R. Perrott (1992). "Minimum Audible Movement Angle as a Function of the Azimuth and Elevation of the Source". In: *Human Factors* 34.3, pp. 267–275 (cit. on p. 39).

Susini, Patrick, Stephen McAdams, and Bennett K. Smith (2007). "Loudness asymmetries for tones with increasing and decreasing levels using continuous and global ratings". In: *Acta Acustica united with Acustica* 93.4, pp. 623–631 (cit. on p. 11).

Susini, Patrick, Sabine Meunier, Régis Trapeau, and Jacques Chatron (2010). "End level bias on direct loudness ratings of increasing sounds". In: *The Journal of the Acoustical Society of America* 128.4, EL163–EL168 (cit. on p. 11).

Tajadura-Jiménez, Ana, Aleksander Väljamäe, Erkin Asutay, and Daniel Västfjäll (2010). "Embodied auditory perception: The emotional impact of approaching and receding sound sources." In: *Emotion* 10.2, pp. 216–229 (cit. on p. 11).

Teghtsoonian, Robert, Martha Teghtsoonian, and Georges Canévet (2005). "Sweep-induced acceleration in loudness change and the 'bias for rising intensities'". In: *Perception & Psychophysics* 67.4, pp. 699–712 (cit. on p. 11).

Teramoto, Wataru, Zhenglie Cui, Shuichi Sakamoto, and Jiro Gyoba (2014). "Distortion of auditory space during visually induced self-motion in depth". In: *Frontiers in Psychology* 5, 848 (cit. on pp. 22, 37).

Teramoto, Wataru, Shuichi Sakamoto, Fumimasa Furune, Jiro Gyoba, and Yôiti Suzuki (2012). "Compression of Auditory Space during Forward Self-Motion". In: *PLOS ONE* 7.6, e39402 (cit. on pp. 22, 37).

Teret, Elizabeth, M. Torben Pastore, and Jonas Braasch (2017). "The influence of signal type on perceived reverberance". In: *The Journal of the Acoustical Society of America* 141.3, pp. 1675–1682 (cit. on p. 18).

Thompson, Lonny L. (2006). "A review of finite-element methods for time-harmonic acoustics". In: *The Journal of the Acoustical Society of America* 119.3, pp. 1315–1330 (cit. on p. 62).

Thurlow, Willard R. and Charles E. Jack (1973). "Certain determinants of the 'ventriloquism effect'". In: *Perceptual and motor skills* 36.3, pp. 1171–1184 (cit. on p. 19).

Thurlow, Willard R. and Philip S. Runge (1967). "Effect of induced head movements on localization of direction of sounds". In: *The Journal of the Acoustical Society of America* 42.2, pp. 480–488 (cit. on p. 114).

Tolnai, Sandra, Rainer Beutelmann, and Georg M. Klump (2017). "Effect of preceding stimulation on sound localization and its representation in the auditory midbrain". In: *European Journal of Neuroscience* 45.3, pp. 460–471 (cit. on p. 16).

Tolnai, Sandra, Ruth Y. Litovsky, and Andrew J. King (2014). "The precedence effect and its buildup and breakdown in ferrets and humans". In: *The Journal of the Acoustical Society of America* 135.3, pp. 1406–1418 (cit. on p. 28).

Tonelli, Alessia, Claudio Campus, and Luca Brayda (2018). "How body motion influences echolocation while walking". In: *Scientific reports* 8.1, pp. 1–10 (cit. on p. 68).

Town, Stephen M., W. Owen Brimijoin, and Jennifer K. Bizley (2017). "Egocentric and allocentric representations in auditory cortex". In: *PLOS Biology* 15.6, e2001878 (cit. on p. 21).

Traer, James and Josh H. McDermott (2016). "Statistics of natural reverberation enable perceptual separation of sound and space". In: *Proceedings of the National Academy of Sciences* 113.48, E 7856–E 7865 (cit. on pp. 19, 28).

Tribolet, José (1977). "A new phase unwrapping algorithm". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.2, pp. 170–177 (cit. on p. 79).

Tsingos, Nicolas and Jean-Dominique Gascuel (1997). "A general model for the simulation of room acoustics based on hierarchical radiosity". In: *Visual Proceedings of Art Interdisciplinary Programs SIGGRAPH'97*. Los Angeles (cit. on p. 65).

Tsolias, Andreas, William J. Davies, *et al.* (2014). "Difference limen for reverberation time in auditoria". In: *Proceedings of Forum Acusticum 2014*. European Acoustics Association. Krakow (cit. on p. 18).

Umbers, Kate D.L., Jussi Lehtonen, and Johanna Mappes (2015). "Deimatic displays". In: *Current Biology* 25.2, R58–R59 (cit. on p. 23).

la Val, Richard K. (1970). "Banding returns and activity periods of some Costa Rican bats". In: *The Southwestern Naturalist*, pp. 1–10 (cit. on p. 93).

Valente, Daniel L. and Jonas Braasch (2008). "Subjective expectation adjustments of early-to-late reverberant energy ratio and reverberation time to match visual environmental cues of a musical performance". In: *Acta Acustica united with Acustica* 94.6, pp. 840–855 (cit. on pp. 20, 111).

— (2010). "Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues". In: *The Journal of the Acoustical Society of America* 128.4, pp. 1952–1964 (cit. on p. 20).

Välimäki, Vesa and Timo I. Laakso (2001). "Fractional Delay Filters—Design and Applications". In: *Nonuniform Sampling: Theory and Practice*. Ed. by Farokh Marvasti. Springer, pp. 835–895. ISBN: 978-1-46-135451-2 (cit. on p. 81).

Väljamäe, Aleksander (2009). "Auditorily-induced illusory self-motion: A review". In: *Brain Research Reviews* 61.2, pp. 240–255 (cit. on p. 22).

do Valle, Anderson Luis and Letícia de Almeida Leão-Vaz (2005). "The defensive reaction of rheas (Rhea americana) to a Rattlesnake Signal". In: *Revista de Etologia* 7.1, pp. 49–50 (cit. on p. 46).

Vernes, Sonja C. and Gerald S. Wilkinson (2020). "Behaviour, biology and evolution of vocal learning in bats". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 375.1789, 20190061 (cit. on p. 23).

Vorländer, Michael (1988). "Die Genauigkeit von Berechnungen mit dem raumakustischen Schallteilchenmodell und ihre Abhängigkeit von der Rechenzeit". In: *Acustica* 66.2, pp. 90–96 (cit. on p. 63).

— (1989). "Simulation of the transient and steady-state sound propagation in rooms using a new combined ray-tracing/image-source algorithm". In: *The Journal of the Acoustical Society of America* 86.1, pp. 172–178 (cit. on p. 64).

— (2008). *Auralization: Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*. RWTHedition. Berlin/Heidelberg: Springer. ISBN: 978-3-540-48829-3 (cit. on pp. 64, 66).

de Vries, Diemer, Edo M. Hulsebos, and Jan Baan (2001). "Spatial fluctuations in measures for spaciousness". In: *The Journal of the Acoustical Society of America* 110.2, pp. 947–954 (cit. on p. 18).

Wallach, Hans (1939). "On sound localization". In: *The Journal of the Acoustical Society of America* 10.4, pp. 270–274 (cit. on pp. 21, 115).

— (1940). "The role of head movements and vestibular and visual cues in sound localization." In: *Journal of Experimental Psychology* 27.4, p. 339 (cit. on pp. 21, 115).

Wallach, Hans, Edwin B. Newman, and Mark R. Rosenzweig (1949). "A precedence effect in sound localization". In: *The American Journal of Psychology* 62, pp. 315–336 (cit. on p. 14).

Wallmeier, Ludwig, Nikodemus Geßele, and Lutz Wiegrebe (2013). "Echolocation versus echo suppression in humans". In: *Proceedings of the Royal Society B: Biological Sciences* 280, 20131428 (cit. on pp. 16, 68, 113).

Wallmeier, Ludwig, Daniel Kish, Lutz Wiegrebe, and Virginia L. Flanagin (2015). "Aural localization of silent objects by active human biosonar: Neural representations of virtual echo-acoustic space". In: *European Journal of Neuroscience* 41.5, pp. 533–545 (cit. on p. 68).

Wallmeier, Ludwig and Lutz Wiegrebe (2014a). "Ranging in human sonar: effects of additional early reflections and exploratory head movements". In: *PLOS ONE* 9.12, e115363 (cit. on p. 68).

— (2014b). "Self-motion facilitates echo-acoustic orientation in humans". In: *Royal Society Open Science* 1.3, 140185 (cit. on pp. 68, 119).

Watkins, Anthony J. (2005a). "Perceptual compensation for effects of echo and of reverberation on speech identification". In: *Acta Acustica united with Acustica* 91.5, pp. 892–901 (cit. on pp. 16, 17, 111).

— (2005b). "Perceptual compensation for effects of reverberation in speech identification". In: *The Journal of the Acoustical Society of America* 118.1, pp. 249–262 (cit. on pp. 16, 17, 28, 111).

Watkins, Anthony J. and Andrew P. Raimond (2013). "Perceptual compensation when isolated test words are heard in room reverberation". In: *Basic Aspects of Hearing*. Springer, pp. 193–201 (cit. on p. 17).

Wefers, Frank (2014). "Partitioned convolution algorithms for real-time auralization". Dissertation. Aachen: RWTH. ISBN: 978-3-8325-3943-6 (cit. on pp. 14, 67).

Weinzierl, Stefan (2008). *Handbuch der Audiotechnik*. Berlin/Heidelberg: Springer. ISBN: 978-3-540-34300-4 (cit. on p. 40).

Weldon, Paul J. (2013). "Chemical aposematism". In: *Chemoecology* 23.4, pp. 201–202 (cit. on p. 23).

Wells, Kentwood D. and Joshua J. Schwartz (2007). "The behavioral ecology of anuran communication". In: *Hearing and sound communication in amphibians*. Ed. by Peter M. Narins, Albert S. Feng, Richard R. Fay, and Arthur N. Popper. Vol. 28. Springer Handbook of Auditory Research. Berlin/Heidelberg: Springer, pp. 44–86. ISBN: 978-0-387-32521-7 (cit. on p. 22).

Wendt, Torben, Steven van de Par, and Stephan D. Ewert (2014). "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse simulation". In: *Journal of the Audio Engineering Society* 62.11, pp. 748–766 (cit. on pp. 25, 66, 68, 69, 74, 89, 118).

— (2016). "Perceptually plausible acoustics simulation of single and coupled rooms". In: *The Journal of the Acoustical Society of America* 140.4, pp. 3178–3178 (cit. on pp. 29, 68).

Wenthold, Robert J., Diana Huie, Richard A. Altschuler, and Karen A. Reeks (1987). "Glycine immunoreactivity localized in the cochlear nucleus and superior olivary complex". In: *Neuroscience* 22.3, pp. 897–912 (cit. on p. 7).

Wenzel, Elizabeth M. (1992). "Localization in virtual acoustic displays". In: *Presence: Teleoperators & Virtual Environments* 1.1, pp. 80–107 (cit. on p. 14).

— (1999). "Effect of increasing system latency on localization of virtual sounds". In: *16th International Conference on Spatial Sound Reproduction*. Paper no. 16-004. Audio Engineering Society. Rovaniemi (cit. on p. 68).

Westheimer, Gerald (2005). "The resolving power of the eye". In: *Vision Research* 45.7, pp. 945–947 (cit. on p. 36).

Wetterer, Andrea L., Matthew V. Rockman, and Nancy B. Simmons (2000). "Phylogeny of phyllostomid bats (Mammalia: Chiroptera): data from diverse morphological systems, sex chromosomes, and restriction sites". In: *Bulletin of the American Museum of Natural History* 2000.248, pp. 1–200 (cit. on p. 102).

Wexler, Mark and Jeroen J.A. van Boxtel (2005). "Depth perception by the active observer". In: *Trends in Cognitive Sciences* 9.9, pp. 431–438 (cit. on p. 39).

Whittaker, Edmund Taylor (1915). "XVIII.—On the functions which are represented by the expansions of the interpolation-theory". In: *Proceedings of the Royal Society of Edinburgh* 35, pp. 181–194 (cit. on p. 11).

Whittle, L.S., S.J. Collins, and David W. Robinson (1972). "The audibility of low-frequency sounds". In: *Journal of Sound and Vibration* 21.4, pp. 431–448 (cit. on p. 2).

Wichmann, Felix A. and N. Jeremy Hill (2001a). "The psychometric function: I. Fitting, sampling, and goodness of fit". In: *Perception & Psychophysics* 63.8, pp. 1293–1313 (cit. on p. 87).

— (2001b). "The psychometric function: II. Bootstrap-based confidence intervals and sampling". In: *Perception & Psychophysics* 63.8, pp. 1314–1329 (cit. on p. 87).

Wigderson, Eyal, Israel Nelken, and Yosef Yarom (2016). "Early multisensory integration of self and source motion in the auditory system". In: *Proceedings of the National Academy of Sciences* 113.29, pp. 8308–8313 (cit. on p. 21).

Wightman, Frederic L. and Doris J. Kistler (1999). "Resolution of front–back ambiguity in spatial hearing by listener and source movement". In: *The Journal of the Acoustical Society of America* 105.5, pp. 2841–2853 (cit. on p. 21).

Wilkie, Sonia and Tony Stockman (2020). "The effect of audio cues and sound source stimuli on the perception of approaching objects". In: *Applied Acoustics* 167, p. 107388 (cit. on pp. 85, 117).

Wilkinson, Gerald S. (1995). "Information transfer in bats". In: *Ecology, Evolution and Behaviour of Bats*. Ed. by Paula A. Racey and Susan M. Swift. Vol. 67. Symposia of the Zoological Society of London. Clarendon Press, pp. 345–360. ISBN: 978-0-19-854945-1 (cit. on p. 92).

— (2003). "Social and Vocal Complexity in Bats". In: *Animal Social Complexity: Intelligence, Culture and Individualized Societies*. Ed. by Frans B.M. de Waal and Peter L. Tyack. Cambridge, Mass.: Harvard University Press. ISBN: 978-0-674-41913-1 (cit. on p. 92).

Wilkinson, Gerald S. and Janette Wenrick Boughman (1998). "Social calls coordinate foraging in greater spear-nosed bats". In: *Animal Behaviour* 55.2, pp. 337–350 (cit. on p. 102).

Wilson, Edward O. (2000). *Sociobiology: The New Synthesis*. Second, twenty-fifth anniversary edition. Cambridge, Mass.: Harvard University Press. ISBN: 978-0-674-00235-7 (cit. on p. 22).

Witten, Ilana B. and Eric I. Knudsen (2005). "Why seeing is believing: merging auditory and visual worlds". In: *Neuron* 48.3, pp. 489–496 (cit. on pp. 19, 112).

Wojtczak, Magdalena and Neal F. Viemeister (2005). "Forward masking of amplitude modulation: Basic characteristics". In: *The Journal of the Acoustical Society of America* 118.5, pp. 3198–3210 (cit. on p. 16).

Wright, Genevieve Spanjer, Chen Chiu, Wei Xian, Cynthia F. Moss, and Gerald S. Wilkinson (2013). "Social calls of flying big brown bats *(Eptesicus fuscus)*". In: *Frontiers in Physiology* 4, 214 (cit. on pp. 93, 102).

Xiao, Qian and Barrie J. Frost (2013). "Motion parallax processing in pigeon *(Columba livia)* pretectal neurons". In: *European Journal of Neuroscience* 37.7, pp. 1103–1111 (cit. on p. 113).

Xie, Bosun (2013). *Head-Related Transfer Function and Virtual Auditory Display*. Second edition. J. Ross Publishing. ISBN: 978-1-60427-070-9 (cit. on p. 67).

Yairi, Satoshi and Yukio Iwaya (2006). "Investigation of system latency detection threshold of virtual auditory display". In: *Proceedings of the 12th International Conference on Auditory Display*. International Community for Auditory Display. London, pp. 217–222 (cit. on p. 68).

Yang, Xuefeng and D. Wesley Grantham (1997). "Echo suppression and discrimination suppression aspects of the precedence effect". In: *Perception & Psychophysics* 59.7, pp. 1108–1117 (cit. on p. 15).

Yin, Tom C. (1994). "Physiological correlates of the precedence effect and summing localization in the inferior colliculus of the cat". In: *Journal of Neuroscience* 14.9, pp. 5170–5186 (cit. on p. 16).

Yost, William A. (1996). "Pitch strength of iterated rippled noise". In: *The Journal of the Acoustical Society of America* 100.5, pp. 3329–3335 (cit. on p. 96).

Yost, William A., Xuan Zhong, and Anbar Najam (2015). "Judging sound rotation when listeners and sounds rotate: Sound source localization is a multisystem process". In: *The Journal of the Acoustical Society of America* 138.5, pp. 3293–3310 (cit. on p. 21).

Young, Bruce A. (2003). "Snake bioacoustics: Toward a richer understanding of the behavioral ecology of snakes". In: *The Quarterly Review of Biology* 78.3, pp. 303–325 (cit. on p. 117).

Young, Bruce A. and Ilonna P. Brown (1993). "On the acoustic profile of the rattlesnake rattle". In: *Amphibia-Reptilia* 14.4, pp. 373–380 (cit. on p. 47).

— (1995). "The physical basis of the rattling sound in the rattlesnake *Crotalus viridis oreganus*". In: *Journal of Herpetology*, pp. 80–85 (cit. on p. 117).

Young, Eric D., George A. Spirou, John J. Rice, and Herbert F. Voigt (1992). "Neural organization and responses to complex stimuli in the dorsal cochlear nucleus". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 336.1278, pp. 407–413 (cit. on p. 7).

Zahorik, Pavel (2001). "Estimating sound source distance with and without vision". In: *Optometry and Vision Science* 78.5, pp. 270–275 (cit. on p. 20).

— (2002a). "Assessing auditory distance perception using virtual acoustics". In: *The Journal of the Acoustical Society of America* 111.4, pp. 1832–1846 (cit. on pp. 9, 10).

— (2002b). "Direct-to-reverberant energy ratio sensitivity". In: *The Journal of the Acoustical Society of America* 112.5, pp. 2110–2117 (cit. on pp. 10, 18, 36).

— (2009). "Perceptually relevant parameters for virtual listening simulation of small room acoustics". In: *The Journal of the Acoustical Society of America* 126.2, pp. 776–791 (cit. on p. 18).

Zahorik, Pavel and Eugene J. Brandewie (2016). "Speech intelligibility in rooms: Effect of prior listening exposure interacts with room acoustics". In: *The Journal of the Acoustical Society of America* 140.1, pp. 74–86 (cit. on p. 18).

Zahorik, Pavel, Douglas S. Brungart, and Adelbert W. Bronkhorst (2005). "Auditory distance perception in humans: A summary of past and present research". In: *Acta Acustica united with Acustica* 91.3, pp. 409–420 (cit. on pp. 8, 10).

Zhang, Wen, Prasanga N. Samarasinghe, and Thushara D. Abhayapala (2019). "Parameterization of the binaural room transfer function using modal decomposition". In: *The Journal of the Acoustical Society of America* 146.1, EL8–EL14 (cit. on p. 67).

Zölzer, Udo (2008). *Digital Audio Signal Processing*. Second edition. Chichester: Wiley. ISBN: 978-0-470-99785-7 (cit. on p. 1).

Zwicker, Eberhard, Georg Flottorp, and Stanley Smith Stevens (1957). "Critical bandwidth in loudness summation". In: *The Journal of the Acoustical Society of America* 29.5, pp. 548–557 (cit. on p. 5).

Zwislocki, Jozef and R.S. Feldman (1956). "Just noticeable differences in dichotic phase". In: *The Journal of the Acoustical Society of America* 28.5, pp. 860–864 (cit. on p. 6).

Acknowledgments

My biggest fortune in life has always been the support of **my parents Inge and Michael**. They have encouraged me in everything I have ever set out to do and stood by me in easy and in difficult times, so that at all the crossroads of my life so far, I felt empowered to set course for what fascinated me at the respective moment. And they have always done all that without making me feel even the tiniest amount of pressure. Danke für alles, was ihr mir ermöglicht habt.

In terms of my field of studies, I did not take the most obvious route to a Ph.D. in Systemic Neurosciences. Even though he is no longer with us, I would thus like to express my sincere appreciation for **Lutz Wiegrebe**, who was looking for a doctoral student at just the right time for me to pay attention. On one day in October 2016, I first contacted him via e-mail, somewhat apologetic for being in touch even though I was neither a biologist nor a psychologist in spite of what his call for applications said. He nevertheless asked me to stop by his office on the day after, and offered me the position on that same evening. I'll forever be grateful for his belief in me, for giving me the opportunity to enter such a fascinatingly different world from the one that I had come from—there are *bats* in the *basement* of a *faculty building?*—for his advice, and for being the most easygoing advisor any graduate student could ask for. "Why are you still here? Didn't you come at 8?" he asked at 5 p.m. "On a weekend? Don't even go there", he responded when I proposed on a Friday that I could revise a manuscript on the day after. I also owe him gratitude for encouraging me to let this dissertation become the hotchpotch that it is now; after all, as Lutz quipped, I could always just tie everything together with a title page that says "Something to do with Hearing". He had prepared me early for the possibility that I might not graduate under his supervision, but it was still a huge shock that he left us so soon. I miss him seriously and I miss him silly.

Thanks, too, to "Lutz' group" in the broadest sense, for the great research environment and for the good company: **Leonie Baier, Markus Drexl, Andrea Lingner, Meike Linnenschmidt, Sven Schörnich, Margarete Überfuhr, Ravi Umadi**—and particularly **Ella Lattenkamp**, whom I would have deeply regretted not meeting, and who also kindly agreed to proofread parts of this thesis, as did **Michael Pecka**. I am also indebted to the students who were involved in my projects, namely **Baccara Hizli, Vivien Lücke and Eva Mardus**. The Biozentrum was an enjoyable place to get up for in the morning thanks to all of you. Zooming out, this goes for everyone else at Neurobiology too; up until the day crowds started to feel a little bit dangerous about a year ago, the sprawling lunchtime chats with people too numerous to name were one of the highlights of most working days. On a more serious note, I am especially thankful for **Benedikt Grothe**, who agreed to supervise me as I brought my studies to their conclusion, and the others on my thesis advisory committee: **Christian Leibold**, **Bernhard Seeber** (TU München) and **Steffen Katzner**.

As with everything in life, there's no project described in this dissertation that I could have completed all by myself. On top of the names I have already mentioned in other contexts, it was a pleasure to work with my coauthors **Owen Brimijoin, Boris Chagnaud, Michael Forsthofer, Daria Genzel, Tobias Kohl, Harald Luksch, Paul MacNeilage, Jassica Richter, Stephanie Shields and Sonja Vernes**. And, of course, with **Stephan Ewert**, who invited me to Oldenburg for one very pleasant

spring, who has kindly continued to adopt me as a "virtual member" of his own working group, and who also gave me valuable feedback on a part of this dissertation.

I feel privileged to be a recipient of comfortable funding from the **Munich Center for Neuro-Sciences** and the **Deutsche Forschungsgemeinschaft**, and to have access to high-quality lab equipment provided by the **Bernstein Center for Computational Neuroscience**. As a lateral recruit to neurosciences, I am furthermore grateful for the learning and teaching opportunities that the **Graduate School of Systemic Neurosciences** has offered me, and especially for all the great people I had the pleasure to meet through it.

One nice thing about staying in the same city (inasmuch as Garching and Martinsried are practically both Munich) for a Ph.D. where one already did one's master's is that it's easy to stay in touch with older university friends besides making new ones. I'm very happy that **the Monday and HaPiNa folks** are still so present in my life—and that in the midst of a pandemic which sent social lives around the world into turmoil, it was possible to go to the swimming pool with some of you regardless of whether any pools were actually open. Even while being at opposite ends of the country, for that matter. As much as I like polarised light passed through liquid crystals, I can't wait to see you all again through pure atmosphere.

I'd also like to thank everyone who has lent me an ear in the course of my doctoral studies, who every once in a while agreed to just sit down and listen—**my experimental subjects**. Thank you for enduring the occasional tediousness of psychoacoustics for the benefit of science and myself, and thank you for only rarely dozing off in the VR chamber.

And very special thanks to one of them in particular, who has continued to lend me her ear on a regular basis, my partner **Narges**. For enduring yet another thesis within a few months, for letting me spontaneously redefine the word "home" in "working from home", and for simply making life more pleasant every day. (To be clear, we didn't meet because of psychophysics. No, we met because two calendar dates, one on a flight ticket and another in a tenancy contract, didn't quite match up.)

In this spirit: Thank you, **serendipity.**

List of publications

This is a list of peer-reviewed publications which I have co-authored, in reverse chronological order of their appearance. Asterisks indicate shared first authorship.

Michael Forsthofer*, **Michael Schutte***, Harald Luksch, Tobias Kohl, Lutz Wiegrebe, and Boris P. Chagnaud (2021). "Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans". In: *Current Biology*. Published online ahead of print: https://doi.org/10.1016/j.cub.2021.07.018.

Ella Z. Lattenkamp, Stephanie M. Shields, **Michael Schutte**, Jassica Richter, Meike Linnenschmidt, Sonja C. Vernes, and Lutz Wiegrebe (2019). "The vocal repertoire of pale spear-nosed bats in a social roosting context". In: *Frontiers in Ecology and Evolution* 7, 116.

Michael Schutte, Stephan D. Ewert, and Lutz Wiegrebe (2019). "The percept of reverberation is not affected by visual room impression in virtual environments". In: *The Journal of the Acoustical Society of America* 145.3, EL229–EL235.

Daria Genzel*, **Michael Schutte***, W. Owen Brimijoin, Paul R. MacNeilage, and Lutz Wiegrebe (2018). "Psychophysical evidence for auditory motion parallax". In: *Proceedings of the National Academy of Sciences* 115.16, pp. 4264–4269.

Michael Schutte and Matthias Dehmer (2014). "Large-Scale Analysis of Structural Branching Measures". In: *Journal of Mathematical Chemistry* 52.3, pp. 805–819.

Veronika Kraus, Matthias Dehmer, and **Michael Schutte** (2013). "On Sphere-Regular Graphs and the Extremality of Information-Theoretic Network Measures". MATCH Communications in Mathematical and in Computer Chemistry 70.3, pp. 885–900.

Copyright and terms of use

■ **Chapter 2.** The Acoustical Society of America (ASA) holds the copyright for the publication reproduced in this chapter. Any copying, redistribution, *etc.* of this chapter requires the permission of the ASA.

■ **Chapter 3.** Daria Genzel, Michael Schutte, W. Owen Brimijoin, Paul R. MacNeilage and Lutz Wiegrebe hold the copyright for the publication reproduced in this chapter. The publication rights lie exclusively with the National Academy of Sciences (NAS) of the USA. Any copying, redistribution, *etc.* of this chapter requires the permission of the NAS.

■ **Chapter 4.** Elsevier Inc. holds the copyright for the publication whose preprint is reproduced in this chapter. Any copying, redistribution, *etc.* of this chapter requires the permission of Elsevier Inc.

■ **Chapter 6.** Ella Z. Lattenkamp, Stephanie M. Shields, Michael Schutte, Jassica Richter, Meike Linnenschmidt, Sonja C. Vernes and Lutz Wiegrebe hold the copyright for the publication reproduced in this chapter. It may be distributed, transformed and built upon, as long as credit is given and any changes are indicated, under the terms of the Creative Commons Attribution 4.0 International License (see `https://creativecommons.org/licenses/by/4.0/legalcode` for the full terms and conditions). You can contact the corresponding author via e-mail to `ella.lattenkamp@evobio.eu` to request the source materials.

■ **Figure 1.1.** This illustration has been provided by Fabian Brinkmann and may be distributed, transformed and built upon, as long as credit is given and any changes are indicated, under the terms of the Creative Commons Attribution 4.0 International License (see `https://creativecommons.org/licenses/by/4.0/legalcode` for the full terms and conditions). It was inspired by Bregman (1994) and a figure which has appeared in Goldstein (1997). The artist has asked to remain anonymous.

■ **Figure 1.2.** Both parts of this figure are based on an illustration by Chittka and Brockmann (2005), which was modified and converted into vector graphics by Inductiveload on Wikimedia Commons (2009) and adapted for this dissertation (labelled on the left, used as the basis for a new illustration on the right) by Michael Schutte. All revisions of the figure, starting at Chittka and Brockmann's (2005) original and including the version in this dissertation, may be distributed, transformed and built upon, as long as credit is given and any changes are indicated, under the terms of the Creative Commons Attribution 2.5 Generic License (see `https://creativecommons.org/licenses/by/2.5/legalcode` for the full terms and conditions).

■ **Figures 5.1, 5.2 and 5.3.** These figures are unmodified reproductions of illustrations by Savioja and Svensson (2015). They may be distributed, transformed and built upon, as long as credit is given and any changes are indicated, under the terms of the Creative Commons Attribution 3.0 Unported License (see `https://creativecommons.org/licenses/by/3.0/legalcode` for the full terms and conditions).

COLOPHON

This dissertation was written between June 2020 and April 2021. It was typeset by the author with XeLaTeX. Diagrams and figures were created with the Ti*k*Z package, Inkscape, Python/matplotlib, and MATLAB.

The text is set in 11 pt EB Garamond, a free-software implementation by Georg Mayr-Duffner *et al.* of Claude Garamont's 16th-century typeface, with Italics based on a design by Robert Granjon.

Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation „Aspects of Room Acoustics, Vision and Motion in the Human Auditory Perception of Space“ selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation “Aspects of Room Acoustics, Vision and Motion in the Human Auditory Perception of Space” is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den 23. April 2021

Michael Schutte

Author contributions

■ **Chapter 2.** **Michael Schutte**, Stephan D. Ewert, and Lutz Wiegrebe (2019). "The percept of reverberation is not affected by visual room impression in virtual environments". In: *The Journal of the Acoustical Society of America* 145.3, EL229–EL235.

M.S., S.D.E. and L.W. designed the experiment. M.S. set up and performed the experiment, analyzed the data and prepared the figures. M.S., S.D.E. and L.W. interpreted the results. M.S. drafted the manuscript. M.S., S.D.E. and L.W. edited and revised the manuscript and approved its final version.

■ **Chapter 3.** Daria Genzel, **Michael Schutte**, W. Owen Brimijoin, Paul R. MacNeilage, and Lutz Wiegrebe (2018). "Psychophysical evidence for auditory motion parallax". In: *Proceedings of the National Academy of Sciences* 115.16, pp. 4264–4269.

D.G. and M.S. share the first authorship of this publication. M.S. and L.W. designed and set up Experiment I; D.G., W.O.B., P.R.M. and L.W. designed and set up Experiment II. M.S. performed Experiment I; D.G. and P.R.M. performed Experiment II. D.G., M.S., P.R.M. and L.W. analyzed the data and prepared the figures. All authors wrote and revised the manuscript and approved its final version.

■ **Chapter 4.** Michael Forsthofer, **Michael Schutte**, Harald Luksch, Tobias Kohl, Lutz Wiegrebe, and Boris P. Chagnaud (2021). "Frequency modulation of rattlesnake acoustic display affects acoustic distance perception in humans". In: *Current Biology*. Published online ahead of print: `https://doi.org/10.1016/j.cub.2021.07.018`.

M.F. and M.S. share the first authorship of this manuscript. B.P.C. planned and designed the study. M.F. and T.K. performed and M.F. analyzed the behavioral rattlesnake experiments. M.S. and L.W. designed, M.S. performed and M.S. and L.W. analyzed the psychophysical experiments. B.P.C., M.F. and M.S. wrote the paper. M.F., M.S., H.L., T.K. and B.P.C. edited and proofread the paper.

■ **Chapter 6.** Ella Z. Lattenkamp, Stephanie M. Shields, **Michael Schutte**, Jassica Richter, Meike Linnenschmidt, Sonja C. Vernes, and Lutz Wiegrebe (2019). "The vocal repertoire of pale spear-nosed bats in a social roosting context". In: *Frontiers in Ecology and Evolution* 7, 116.

L.W., M.L., S.C.V., and E.Z.L. conceived and supervised the study. S.M.S. recorded the data. S.M.S., E.Z.L., and M.L. developed the classification key. L.W. wrote the syllable detection and analysis program. E.Z.L. and S.M.S. performed the syllable classification. M.S. conducted the statistical analyses and data presentation. J.R. rated the behavioral context. E.Z.L. wrote the first draft of the manuscript. All authors contributed to the writing, editing, and revising of the final paper.

As (co-)first authors of the publications and manuscripts reproduced in this thesis, we confirm that the statements concerning the author contributions to our respective works are accurate.

Dr. Daria Genzel-Wehrfritz	Michael Forsthofer
Dr. Ella Z. Lattenkamp	Michael Schutte

As the supervisor of this dissertation, I confirm that these statements are accurate.

Prof. Dr. Benedikt Grothe