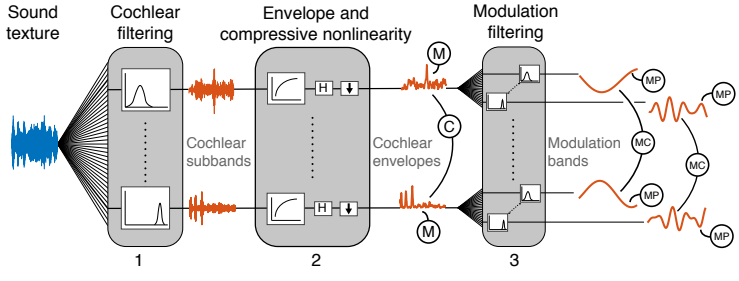
Technical University of Denmark

CONTRIBUTIONS TO
HEARING RESEARCH

Volume 28

Richard McWalter

Perceptual and Neural Response to Sound Texture



Hearing Systems
Department of Electrical Engineering

Perceptual and Neural Response to Sound Texture

PhD thesis by

Richard McWalter

Preliminary version: April 20, 2017

Technical University of Denmark

2017

Preprint version for the assessment committee.

Pagination will differ in the final published version.

This PhD dissertation is the result of a research project carried out at the Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark (Kgs. Lyngby, Denmark) and at the Danish Research Center for Magnetic Resonance Imaging (DRCMR, Hvidovre Hospital, Hvidovre, Denmark). Part of the research was carried out at Josh McDermott's Computation Audition Lab, Department of Brain and Cognitive Science, Massachusetts Institute of Technology (MIT, Cambridge, Massachusetts, USA).

The project was financed by the Technical University of Denmark. The external stay in Cambridge, MA was further supported by a scholarship from the Oticon Foundation.

Supervisor

Prof. Torsten Dau
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Collaborators

Asst. Prof. Josh H. McDermott
Lab for Computational Audition
Department of Brain and Cognitive Science
Massachusetts Institute of Technology
Cambridge, Massachusetts, USA

Jens Hjortkjear
Hearing Systems Group
Department of Electrical Engineering
Technical University of Denmark
Kgs. Lyngby, Denmark

Asst. Prof. Kristoffer Madsen
DTU Compute
Department of Appl. Math. and Comp. Sci.
Technical University of Denmark
Kgs. Lyngby, Denmark

Prof. Hartwig Siebner
Danish Research Center for Magnetic Resonance
Hvidovre Hospital
Copenhagen University Hospital
Hvidovre, Denmark

Abstract

The perception of natural sound develops in a cascade of brain regions along the auditory pathway. The peripheral regions respond more directly to the input acoustic waveform, whereas cortical regions appear to become more selective to particular features of natural stimuli. Although much is known about the early stages of the auditory system, the mechanisms and structure of the processing at later stages is more opaque. In this thesis, we employed an analysis-via-synthesis approach to probe the auditory system's response to naturalistic texture stimuli. Textures are ubiquitous in the natural world (e.g. rain, fire, stream) and their perceptual qualities can be captured by a set of time-averaged statistics measured at several stages of a standard auditory model. In the first study (chapter 2), we investigated the nature of the time-averaging mechanism that underlies sound texture perception. The second study (chapter 3) examined the neural locus that might host such a time-averaging mechanism using function magnetic resonance imaging (fMRI). The third study (chapter 4) investigated the role of envelope amplitude modulations in sound texture perception and we proposed a model to account for simple rhythmic structure. The fourth study (chapter 5), we proposed a texture statistics compensation strategy for the impaired auditory system. The projects expand our understanding of sound texture perception, the neural mechanisms that underpin such perception, and the utility of sound synthesis to characterize the auditory system.

Resumé

Opfattelsen af naturlig lyd udvikles i en kaskade i hjernens områder omkring lydvejen. De perifere områder svarer mere direkte til akustiske egenskaber, hvorimod de kortikale områder forekommer at være mere selektive til særlige egenskaber af naturlig stimuli. Selvom de tidlige stadier i det auditoriske system er velkendte, såer behandlingen af de senere stadiers mekanisme og struktur mere uklar. I denne tese, har vi benyttet en analyse-via-syntese fremgangsmåde til at undersøge det auditoriske systems svar til naturalistisk teksturstimuli. Teksturer er allestedsnærværende i den virkelige verden (f.eks. regn, ild, vandløb) og deres opfattelseskvaliteter kan blive opfanget som en mængde af tidsmidled statistik målt ved forskellige stadier af en standard auditorisk model. I det første projekt (kapitel 2) undersøgte vi den tidsmidlede mekanismes natur, som ligger til grund for opfattelse af lydtekstur. I det andet projekt (kapitel 3) blev det område i hjernen, hvor en sådan tidsmidlet mekanisme finder sted, undersøgt ved brug af funktionel magnetisk resonans scanning (fMRI). I det tredje projekt (kapitel 4) blev pakkeamplitudemodulation i opfattelsen af lydtekstur undersøgt, og vi fremlagde en model til at gøre rede for simple rytmiske strukturer. Til sidst (kapitel 5) begyndte vi at undersøge teksturopfattelse i det auditoriske system for hørehæmmede og fremlagde en kompensationsstrategi omkring teksturstatistik. Projekterne udviklede vores forståelse for opfattelsen af lydtekstur, det område i hjernen hvor mekanismer der understøtter en sådan opfattelse er, og ogsåmuligheden for anvendelse af sammenfatningen af lyd til at karakterisere det auditoriske system.

Related publications

Journal papers

- McWalter, R., McDermott, J.H., (**2017**). "Adaptive Time-Averaging of Auditory Scenes," in Review - submitted March 2017
- McWalter, R., Hjortkjear, J., Dau, T., Siebner, H., Madsen, K., (**2017**). "Response to the Statistical Structure of Texture in Auditory Cortex," in Prep.
- McWalter, R., Dau, T. (**2017**). "Amplitude Modulation Sensitivity in Sound Texture Perception," in Review - submitted March 2017

Conference papers

- McWalter, R., Dau, T., (**2015**). "Statistical Representation of Sound Textures in the Impaired Auditory System," Proceedings of International Symposium Auditory Audiological Research (ISAAR) 2015. p. 189-196.

Additional work

- Agerkvist, F.; Torras Rosell, A.; McWalter, R., (**2015**). "Improvements in Elimination of Loudspeaker Distortion in Acoustic Measurements," Proceedings of 138th International Audio Engineering Society (AES) Convention.
- Agerkvist, F.; Torras Rosell, A.; McWalter, R., (**2014**). "Eliminating Transducer Distortion in Acoustic Measurements," Proceedings of 137th International Audio Engineering Society (AES) Convention. p. 824-833 Convention Paper 9204.

Contents

1

Introduction

Human audition involves a network of brain regions spanning the auditory nerve to the auditory cortex and beyond. The input to the neural auditory system is through the cochlea, which is itself activated by pressure variations incident on the ear drum. These small vibrations are transformed into a neural code that preserves and highlights relevant sounds. Although much is known about the early stages of the auditory system, the mechanisms and structure of the processing at later stages is more opaque.

Our understanding of the auditory system is embedded in at least two main research areas: auditory perception and neural processing. Auditory perception reveals how the brain makes sense of the natural world. By examining how humans listen to controlled sounds, we can establish some of the basic attributes of the system, such as audio frequency selectivity and amplitude modulation frequency selectivity. The neural processing is, in essence, a vast network of neurons which facilitate this perception. The neurons can be examined in isolation, using methods such as patch clamp (Sakmann and Neher, 1984), or as larger populations, using methods such as functional magnetic resonance imaging (fMRI) (Ogawa et al., 1990), both of which offer insights into the structure and mechanisms underpinning sensory systems.

Sounds in the natural world span a broad range of levels and frequencies. However, it is the time-varying properties of sound that are arguably most critical for perception, since all natural sounds vary over time. Some sounds, such as a singing voice or a pianist playing a concerto, are comprised of complex patterns and variations which carry abundant time-varying information . Other sounds, such as a stream flowing or sparrows chirping, have more homogenous temporal properties (McDermott and Simoncelli, 2011). These sounds, ranging from highly variable to temporally homogenous, comprise the sounds of our natural environment.

Although the sounds that create our everyday auditory experience are rich and varied, many attempts to understand the processing mechanisms that underlie auditory perception are rooted in simple artificial stimuli (Theunissen and Elie, 2014). One possible issue with presenting simple artificial stimuli is that the response may not be universally reflected across all stimuli. In addition,

it may be the case that the auditory system has evolved to cope with natural sounds, whereas sounds generated with physically implausible features are represented in an unorthodox way by the system. Although natural sounds can be cumbersome to handle in an experimental context, their utility in characterizing the processes of the auditory system may be crucial (Theunissen and Elie, 2014).

Understanding the processing hierarchy of the auditory system can be undertaken with a multipronged approach. While auditory perception and neural processing represent two key aspects, modeling can also yield significant contributions to our understanding of the auditory system. In this thesis, my collaborators and I have attempted to combine perceptual observations, the statistics of natural sounds and model-based representations to further our understanding of the auditory system.

1.1 Auditory System

The auditory system is comprised of a cascade of processing layers that transform the pressure variations incident on the ears to neural representations that facilitate sound perception. Of the processing layers, much is known about the auditory periphery, but our understanding becomes opaque as we ascend the auditory pathway. For instance, the fine details of the processing of sound on the basilar membrane have been described for the mammalian auditory system, both in vivo and ex vivo (Ruggero, 1992a). Also the response of auditory-nerve fibers can be estimated with reasonable accuracy in non-human mammals. However, our ability to characterize the responses in the auditory cortex to complex natural stimuli is less well defined (Atiani et al., 2014; Miller et al., 2002a; Norman-Haignere et al., 2015a; Santoro et al., 2014; Theunissen et al., 2000). In this first section, some of the basic ideas of the auditory system will be summarized, which in turn will be the foundation for this project.

1.1.1 Outer-, Middle- and Inner-ear

Pressure variations in a given medium that are audible to the auditory system are referred to as sounds. For humans, sound typically operates on the frequency range between 20 Hz and 20 kHz (wavelengths of roughly 17m to 17mm). In terms of amplitude, humans can generally hear sounds as quiet as 0 dB SPL (sound pressure level) to sounds in excess of 120 dB SPL, which is a scale factor of about (Plack, 2005). Much of what dictates the range of sounds we hear can be attributed to the peripheral auditory system, or more specifically, the outer-, middle-, and inner-ear.

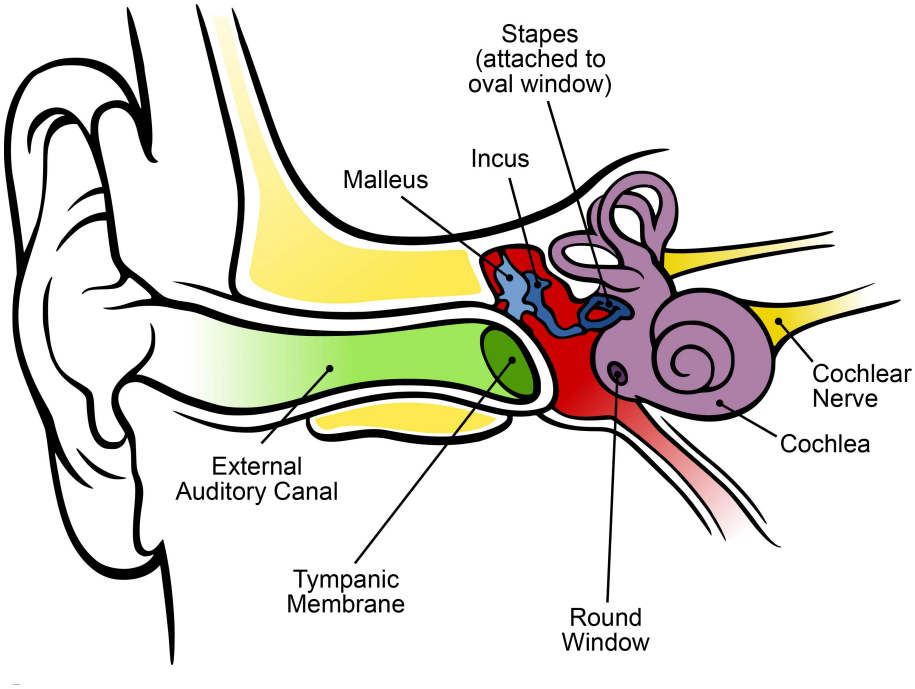


Figure 1.1: The peripheral auditory system, which constitutes the outer-, middle- and inner-ear. [Figure adapted from doi:10.1371/journal.pbio.0030137]

The outer-ear is typically defined as the pinna and ear canal, ending at the tympanic membrane. Both the pinna and ear canal act as acoustic filters that shape the frequency content of the waveform incident on the tympanic membrane. Broadly speaking, the ear canal acts as a Helmholtz resonator with a bandwidth between 1 to 6 kHz (Plack, 2005). The pinna aids in sound localization, primarily by shaping the spectral content for front and back sources and sources with varying elevation.

The middle-ear has two main functions: (1) to facilitate the transfer of pressure variations from air to liquid and (2) to attenuate certain incoming sounds. The middle-ear is composed of 3 small bones, malleus, incus and stapes (the smallest bones in the body) and the middle-ear reflex that connects the tympanic membrane of the outer ear to the oval window of the inner ear. The three bones act as an impedance match between the air-filled ear canal and the fluid-filled cochlear (inner-ear) (Rosowski and Relkin, 2001). The middle-ear reflex attenuates incoming sound that exceeds a moderate level (around 75dB SPL) and has a high-pass characteristic (Plack, 2005).

The inner-ear, or cochlea, is arguably the most important element of the peripheral auditory system, as it dictates the fundamental characteristics of the auditory system in general. The inner-ear is also where the transduction occurs from the acoustic vibrations of the oval window to the neural activity. Essentially, the cochlea is the gateway to the brain. One of the main features of the cochlear

is that it separates the incoming signal into a frequency tonotopy (Glasberg and Moore, 1990; Von Békésy and Wever, 1960). Anatomically, the cochlea can be separated into two compartments, the scala vistibuli and scala media, which are separated by the basilar membrane (Snell, 2010). The mechanical properties of the basilar membrane vary in stiffness and width. The membrane is stiff and narrow at the basal end (near the oval window) and wide a floppy at the apical end. These mechanical attributes translate to a frequency tonotopy, where high frequencies are represented at the oval window input and low frequencies are represented towards the apex. Pragmatically speaking, the cochlea operates as a mechanical frequency analyzer.

The cochlear is a non-linear system, compressing the input signal for a broad range of levels (Plack, 2005; Ruggero, 1992a). The compression arises from the active contribution of outer hair cells to the vibration of the basilar membrane. These efferent signals primarily operate between the input levels of 30 dB and 80 dB SPL (measured near the concha). Above and below the compressive region, the basilar membrane operates nearly linearly. The active mechanism facilitates the large operating dynamic range of the auditory system.

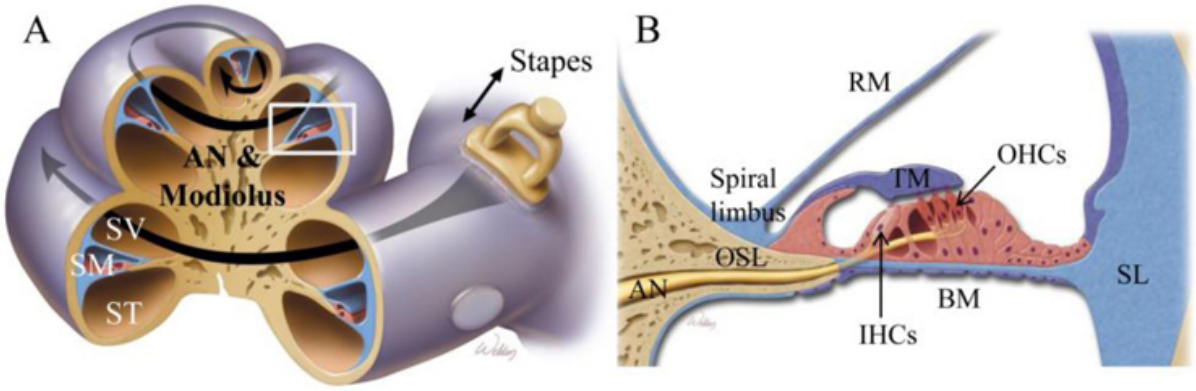


Figure 1.2: (A) Cochlear chambers (scala vestibuli (SV), scala media (SM), and scala tympani (ST)), auditory neurve fibers (AN). (B) The organ of Corti is comprised of outer hair cells (OHCs) and one row of inner hair cells (IHCs). The hair cells line the basilar membrane (BM). Hair cell stereocilia are deflected when forces develop from tectorial membrane (TM). Auditory nerve fibers connect the hair cells with the brainstem. [Figure taken from doi:10.1364/OE.19.015415]

The inner and outer hair cells play a role in the transduction from the mechanical vibrations of the basilar membrane to the auditory nerve. The basilar membrane vibration causes a depolarization of steriocilia (hair cell tips), which relays the signal to the auditory nerve fibers. The auditory nerve fibers code the acoustic information measured by the inner hair cells as a pattern of nerve impulses. The nerve impulses, or firing rate, vary depending on the acoustic signal characteristics, such as

increased firing rate relative to basilar membrane velocity. The pattern of vibration on the basilar membrane is thus transformed to a neural "rate-place code" in the auditory nerve (Lopez-Poveda and Meddis, 2001; Robles and Ruggero, 2001; Ruggero, 1992b).

1.1.2 Auditory Midbrain

The responses of auditory nerve fibers are transmitted to the cochlear nucleus (CN), located in the brainstem. The CN is divided into three main sections (anteroventral, posteroventral and dorsal) and is comprised of several types of neurons., many of which respond to low-level features of the acoustic input signal (e.g. frequency and onset). One prime utility of the CN cells is that they send their signals to several auditory nodes in the ascending pathway, such as superior olivary complex (SON), nuclei of the lateral lemniscus (LL), and inferior colliculus (IC) (Schnupp et al., 2011).

Another important processing node in the auditory system is the inferior colliculus (IC). Although the IC is a requisite processing stage in the afferent auditory pathway, there are still gaps regarding its general function in the auditory system. The IC facilitates some binaural processing as well as some reflexive movements (e.g. eye movements towards an unexpected sound (Schnupp et al., 2011). There is also recent work suggesting that neurons in the IC are noise invariant, in that they adapt to the mean level of the input (Dean et al., 2008; Kvale and Schreiner, 2004a). There is some evidence that neuronal responses in the IC are still rudimentary and that complex transformations of sensory input occurs between the IC and cortical processing levels (Atencio et al., 2012; Sharpee et al., 2011). Lastly, the IC has shown stimulus specific adaptation, which has been suggested to be indicative of a sensitivity to sound statistics (Ayala et al., 2013; Dean et al., 2005b).

The medial geniculate body (MGB) is located in the thalamus and connects the IC with the auditory cortex. Although the receptive fields of the thalamus have been difficult to identify, researchers have suggested that the thalamus facilitates the selectivity of envelope amplitude modulations of an acoustic signal. Amplitude modulation processing and perception in mammals has been an area of study for centuries, and arguably it is the amplitude modulations of acoustic signals that carry sound information (Joris et al., 2004a). Amplitude modulations also appear to change in the ascending auditory pathway from temporal coding at the periphery to rate coding at later stages. Most notably, the thalamus is the auditory processing node that appears to host amplitude modulation selective neurons, which is a known contributor to many aspects of auditory perception (Dau et al., 1997; Ewert et al., 2002; Joris et al., 2004a; Preuss and Möller-Preuss, 1990).

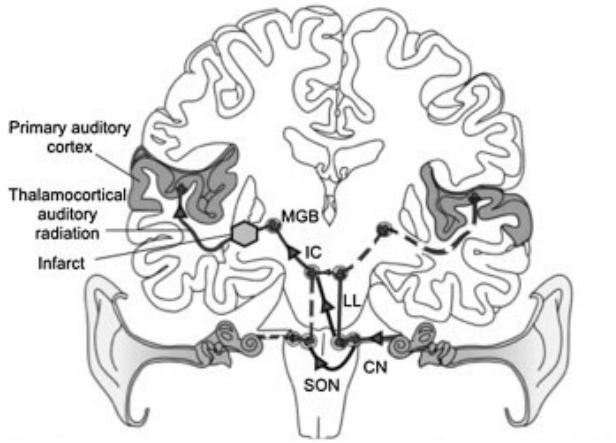


Figure 1.3: Central auditory pathway. Medial geniculate body (MGB), cochlear nucleus (CN), superior olivary nucleus (SON) and inferior colliculus (IC). [Figure taken from doi:10.1186/1752-1947-8-400]

1.1.3 Auditory Cortex

The anatomical landmarks which define the auditory cortex include Heschl's gyrus and planum temporale. These have been identified by either their tonotopic response to acoustic input frequency and the connections to the thalamus (Schnupp et al., 2011). When considered in more general terms (across mammals), researchers have identified primary (or core) regions and secondary (or belt) areas of auditory cortex, which both appear to interact with higher-order cognitive structures (Schnupp et al., 2011; Snell, 2010). The difficulty in identifying specific landmarks and their processing function is that human auditory cortex differs from other mammals, which is not necessarily the case for the subcortical processing stages (Schnupp et al., 2011).

From a signal processing standpoint, there are several acoustic features which seem to be represented at the cortical level of the auditory system (Depireux et al., 2001). Frequency tonotopy is the main functional localizer for primary auditory cortex and has a high-low-high frequency structure (Humphries et al., 2010). The auditory cortex appears to also be sensitive to frequency and amplitude modulations (Theunissen et al., 2000; Theunissen et al., 2001). They have been measured as two-dimensional receptive fields in several mammals and appear to capture many important cues relevant to auditory perception . However, much of our understanding of cortical function can be better defined by perceptually based signal properties (such as pitch) or specific stimulus sound categories (e.g. speech or music) (Belin et al., 2000; Norman-Haignere et al., 2015a; Zatorre et al., 2002).

Natural sounds are also represented in the auditory system, with some aspects extending beyond

the acoustic (spectro-temporal tuning) or fundamental perception (e.g. pitch) (Norman-Haignere et al., 2013). Speech is the most notable natural sound, as it is primarily relegated to humans and it is the focal point of most hearing research. The temporal structure of speech appears to be important and researchers have identified speech selective regions beyond primary auditory cortex (Belin et al., 2000; Overath et al., 2015a). In addition, music appears to cover a broad anatomical region of the auditory cortex and is spatially collocated with other important auditory signal cues.

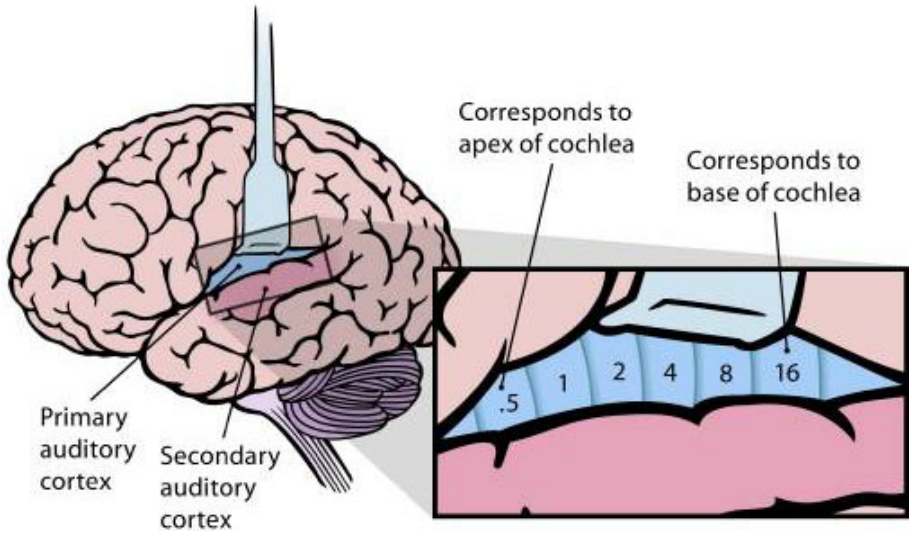


Figure 1.4: The primary auditory cortex responds topographically to the cochlear frequency spectrum (shown in kHz). [Figure taken from doi:10.1371/journal.pbio.0030137]

1.2 Auditory Modeling

Many anatomical features of the mammalian auditory system have been well defined, but how might one approach modelling their function? Several models have been proposed that capture the tuning properties of the auditory systems. Some models take a theoretical approach while others take a more phenomenological approach. There are models constructed around neural data, perceptual data, stimulus statistics, or a combination thereof.

The cochlea is arguably the most well defined component of the auditory system. Cochlear modelling has classically taken on two forms: a frequency-selective bandpass filter(bank) model or an inner-ear mechanical model. The former is a more functional approach while the latter more accurately simulates the biophysical properties of the cochlea, particularly the basilar membrane. Both types of models can take the frequency tonotopy of the inner ear into account, as well as the non-

linear behavior of the cochlea. Perhaps the simplest model available is a linear filterbank model (e.g. in the form of bank of symmetric filters or Gammatone filters with modest asymmetries (Hohmann, 2002; Patterson et al., 1987; Slaney, 1993). The gammatone filter appears to be able to account for both physiological and psychophysical data, and has been used extensively to model aspects of auditory perception. The individual bandpass filter center frequencies have also been identified as an equivalent rectangular band (ERB)-rate, which amounts to approximately 30 logarithmically spaced filters that cover the frequency hearing range of a human listener with normal hearing. Although frequency selectivity depends nonlinearly on sound pressure level and acoustical context, this simple functional model has been used successfully for over 25 years.

Modelling the transduction of the mechanical vibrations of the basilar membrane to the neural patterns in the auditory nerve is a critical aspect of all neural models of the auditory system. Several successful models have used auditory-nerve recordings in mammals and reproduced the averaged response bahavior . These models have been successful in capturing more detail about the auditory periphery, including compression, temporal codding/phase locking and adaptation (Carney, 1993; Heinz et al., 2001; Zhang et al., 2001; Zilany and Bruce, 2006). Meddis and colleagues also have well developed compartment models spanning the auditory nerve to the cochlea nucleus (Meddis, 1986). A more pragmatic approach to model hair-cell responses is to extract the envelope (e.g. Hilbert envelope or half-wave rectification and lowpass filter) of the acoustic signal. This method can account for many aspects of higher level representations and perception (Chi et al., 2005; Jørgensen and Dau, 2011).

The latter stages of the auditory system are less well defined; however, there have been several noble attempts to model their function. The difficulty in modeling higher auditory processing stages, such as primary auditory cortex, is that it appears that responses are context dependent and model predictions are typically valid only within the same context (Eggermont, 2010). This insinuates that cortical neurons have adaptive receptive fields. Spectro-temporal receptive fields appear to be strong candidates as a foundation of cortical processing. Spectro-temporal filters have been used with some success to describe cortical responses to artificial stimuli (moving ripples) and natural sounds (speech, environmental sounds) (Chi et al., 2005; Theunissen et al., 2000).

Summary models combine several components of the auditory system, which decompose the incoming signal into a form that may facilitate some task relevant function. These models primarily use a frequency-selective front end, modeling core aspects of the cochlear processing, followed by some comprehensive model of higher-level auditory processing. Two successful models include

the spectro-temporal receptive field model and the cascaded linear filtering modulation filterbank model, both of which have had broad reaching applications, from speech intelligibility predictions to natural sound perception (Chi et al., 2005; Dau et al., 1997; Jørgensen and Dau, 2011; Theunissen et al., 2000).

The modulation filterbank model shown in Figure 3 will be used throughout this thesis. The model captures the tuning properties of the early auditory system, namely frequency selectivity (ERB-rate spaced Gammatone filterbank) and amplitude modulation frequency selectivity (octave spaced constant-Q filterbank). Although the exact neural locus of the model has yet to be identified, particularly the modulation filterbank element, the role of amplitude modulation frequency selectivity in all manners of auditory perception is undeniable.

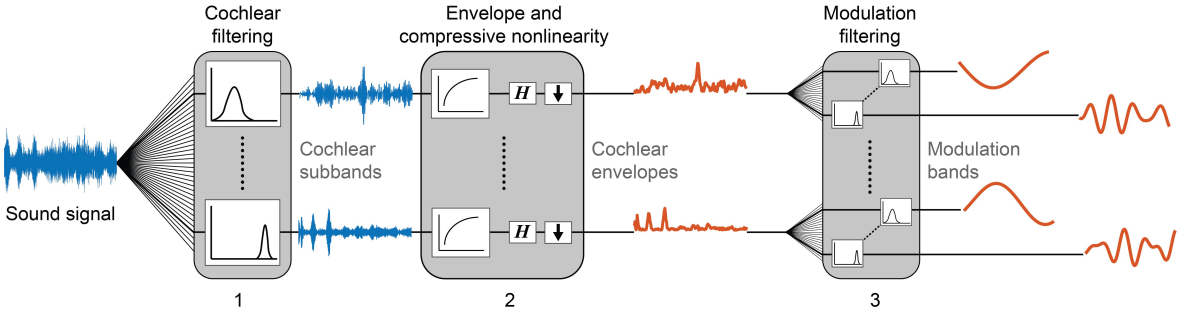


Figure 1.5: Auditory model (adapted from/inspired by Dau et al. (1997)). The functional auditory model captures the tuning properties of the early (peripheral and subcortical) auditory system: (1) auditory filterbank modeled on the resonance frequencies of the cochlear, (2) nonlinearity captures the compression of the cochlear followed by computation of the Hilbert envelope, functionally modelling the transduction from the mechanical vibrations of the cochlear to the auditory nerve and down-sampling, and (3) modulation filterbank capturing the selectivity of the auditory system to different envelope fluctuation rates.

1.3 Our Approach

In order to simulate the processes of the auditory system, we need to address the structure of the natural world the system represents. There is mounting evidence that the auditory system, and sensory systems more generally, are optimized to process natural stimuli (Field, 1987; Freeman et al., 2013; McDermott et al., 2013; Moerel et al., 2012; Olshausen and Field, 1996; Portilla and Simoncelli, 2000; Santoro et al., 2014; Simoncelli and Olshausen, 2001; Theunissen and Elie, 2014; Turner and Sahani, 2008; Ziemba et al., 2016). The inherent structure found in natural sounds may offer insight

into the processing mechanisms the auditory system has adopted to navigate through different environments.

Natural stimuli contain statistical regularities that sensory systems may use for recognition (Freeman et al., 2013). The structure may make part of the signal redundant, for example when two acoustic properties are known to covary in the natural world (Schwartz and Simoncelli, 2001). This notion has been leveraged in the visual domain and appears to also play a role in audition (Simoncelli and Olshausen, 2001; Theunissen and Elie, 2014). The most obvious regularities arise from signal periodicity. At fast rates, this may be related to pitch perception, and at slow rates this may be related to rhythm (Plack et al., 2006; Poeppel, 2003). The latter also seems to play a role in extracting features from more complex sound sources, such as speech or music (Ding et al., 2016; Overath et al., 2015a). Acoustic signals can also be temporally redundant, in which case a time-averaged summary statistic may be an optimal approach, balancing recognition/computation load (McDermott et al., 2013).

The auditory scene often represents combination of sound sources, and these sources vary greatly in their temporal complexity. How does the auditory system cope with such scenes? For temporally complex sounds, such as speech, it has been suggested that the auditory system uses multiple analysis windows and operates in an abstract high-dimensional feature space (Ding et al., 2016). For more temporally homogeneous sounds, such as a stream, fire or insect swarm, the auditory system may operate in a relatively low-dimensional feature space (McDermott and Simoncelli, 2011). The reason may be embedded in the amount of information that is carried in the signal (Saint-Arnaud and Popat, 1995). Depending on the stimulus, the auditory system may opt for the most utilitarian representation.

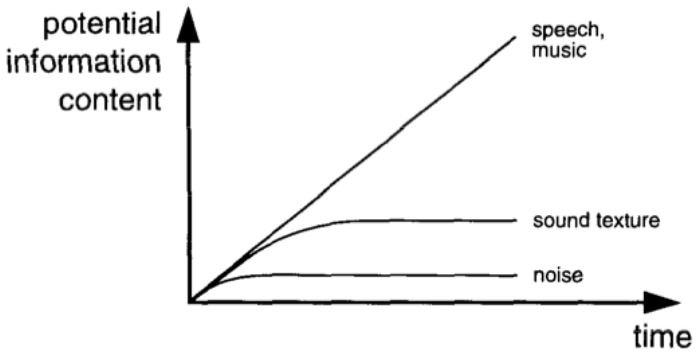


Figure 1.6: Sound textures and noise show constant long-term characteristics (taken from (Saint-Arnaud and Popat, 1995)).

Temporally homogeneous sounds appear to be well represented with a set of time-averaged

summary statistics. Inspired by models of early visual representations which appear to average information over space (especially in the visual periphery (Balas et al., 2009; Freeman and Simoncelli, 2011; Portilla and Simoncelli, 2000), temporally homogeneous natural sounds appear to be well characterized by statistics measured from early auditory representations, including marginal moments and pair-wise correlations (McDermott and Simoncelli, 2011). This time-averaging, although in contrast to many temporally acute traits of the auditory system, would enable a compressed representation of background sounds in the auditory scene (McDermott et al., 2013). From an efficient coding standpoint, it would make sense for a sensory system to neglect redundant information.

A standard auditory model inspired by psychophysical and physiological data appears to capture the pertinent tuning properties of the early auditory system. The two main dimensions the model operates in are frequency selectivity and amplitude modulation frequency selectivity. The first stage of the model includes frequency-selective 'cochlear' filters. The following stage includes cochlear compression and envelope extraction. Lastly, the model applies a second layer of linear filtering, separating the signal into rate-selective amplitude modulation channels. This model has been attributed to subcortical processes. However, it should be noted that the frequency tonotopy of the cochlear is preserved up to the auditory cortex, and the neural basis for amplitude modulation selectivity has been relatively abstruse to the research community.

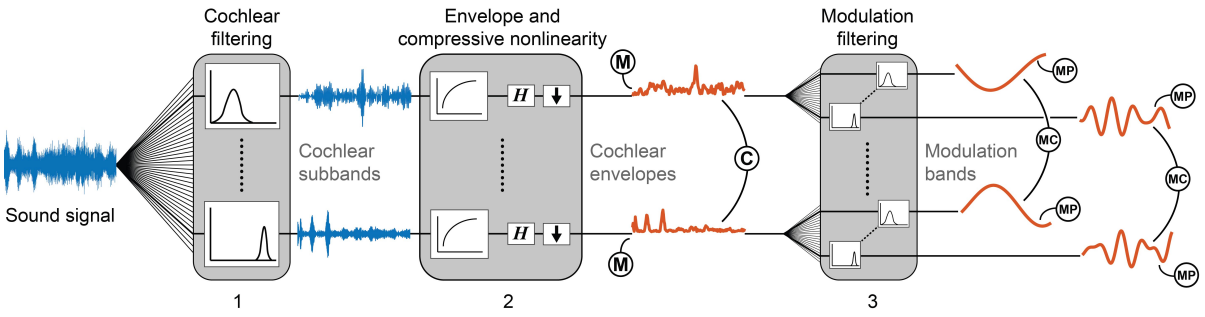


Figure 1.7: Texture analysis model (inspired by McDermott and Simoncelli (2011)). The functional auditory model captures the tuning properties of the early (peripheral and subcortical) auditory system: (1) cochlear filterbank, (2) cochlear compression followed by computation Hilbert envelope and down-sampling, and (3) modulation filterbank. Texture statistics include marginal moments (M) of cochlear envelopes (C), modulation power (MP), pair-wise correlations between cochlear envelopes (MC1), and modulation-rate pairwise correlations between cochlear envelopes (MC2).

The summary statistics can be coupled with the early auditory system model to capture the perceptually relevant features of temporally homogeneous sound textures. Textures can be characterized as the sounds generated from the superposition of many similar acoustic events. Examples of

textures include a galloping horse, sparrows chirping or an applauding crowd. They are ubiquitous in the natural auditory scene and often described as background noise. The summary statistics identified by McDermott and Simoncelli (2011) varied across stimuli, which may lend to the notion of texture recognition being based on summary statistics. Figure **??** shows statistics for two textures (ocean waves and swamp insects).

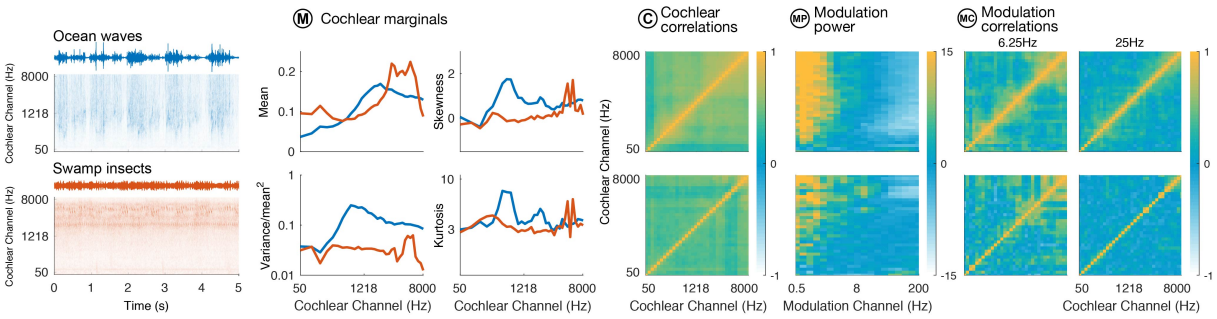


Figure 1.8: Example texture statistics for two real-world texture recordings: ocean waves and swamp insects. Left panels show waveform (top) of an example sound texture. Spectrograms (bottom left) show 2s excerpts at 2 different time positions within the waveform. Right panels show texture statistics measured for each 2s excerpt. Cochlear marginal moments for each excerpt are plotted on same graphs. For brevity, modulation correlations, are shown for only two subbands. For these example textures, statistical measurements vary greatly between the two sounds.

Texture perception appears to also rely on a similar time-averaged summary statistics representation. McDermott et al. (2013) revealed an inability of the auditory system to access the temporal detail of textures. This was demonstrated in a texture exemplar discrimination task, where listeners performance decreased with increasing stimulus duration. Conversely, discriminability increased with duration when the task used sounds from different sources (with different time-averaged statistics). This finding was explained by assuming the auditory system is optimized for a time-averaged representation, pooling acoustics information over time. A similar task was conducted with speech, ranging from single-talker to multi-talker babble, as well as drum hits, ranging from several repetitions per second to hundreds.

The perceptual characteristics of texture have been explored in an analysis-via-synthesis approach (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). The auditory texture model was embedded in a synthesis system that shaped a Gaussian noise seed to have time-averaged statistics that matched those measured from a real-world texture recording. This approach had several benefits; namely the ability to generated many examples of unique textures with the same time averaged statistics in addition to generating naturalistic stimuli with a reasonable degree of

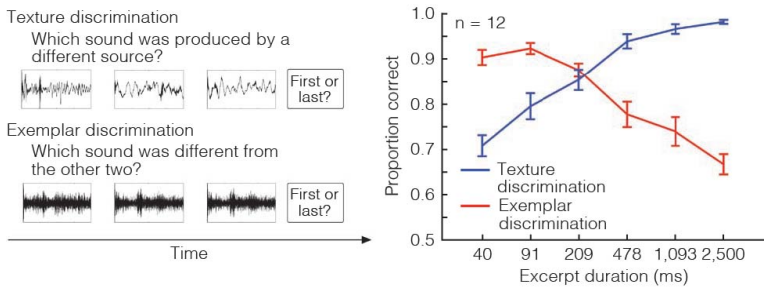


Figure 1.9: (Figure and text adapted from McDermott et al. (2013)) Schematic of trial structure for texture discrimination and exemplar discrimination (left panel). Three sounds were played in succession, separated by a fixed interval. In the texture-discrimination task, two of the sounds were distinct excerpts of the same texture and the third (presented first or last) was an excerpt of a different texture. In the exemplar-discrimination task, two of the sounds were physically identical excerpts of a texture and the third was a different excerpt from the same texture. The results of texture and exemplar discrimination (right panel).

control over their statistical properties. The method enabled the exploration of texture perception from exemplar discriminability to identification, and the necessity of a biologically plausible auditory model.

The sound texture synthesis system of McDermott and Simoncelli (2011) produced compelling exemplars for temporally homogeneous textures but failed for stimuli which depended on more abstract or complex representations. This could be a particular acoustic feature, such as rhythmic patterns, a perceptual attribute, such as pitch, or simply a complex temporal signal, such as speech or music. For the latter case, it may be a combination of several more basic features which make up the representation. The evaluation of synthesis performance was conducted in a realism rating experiment. The failures could be attributed to either the processing structure, the temporal analysis, or the texture statistics.

The results suggested that the higher-order statistics that characterize sound textures provide a potentially effective foundation for investigating fundamental auditory perception and its neural basis. The signal computable texture space incorporated within the biologically plausible auditory model is a strong foundation to investigate such questions as the time-averaging of sound statistics, the neural locus of texture recognition and the incorporation of relevant perceptual cues to the auditory texture model. This dissertation attempted to follow these paths in four successive chapters.

1.4 Navigating the Thesis

The thesis is separated into four main chapters, all of which attempt to combine aspects of natural sound statistics, auditory perception, and neural responses. The projects aim to expand on research output from Josh McDermott's computational audition lab and Eero Simoncelli's computational vision lab. Specifically, the research focused on the perception of texture and its neural basis. All of the projects used synthetic sound textures as the primary experimental stimuli (in-house developed system inspired by McDermott and Simoncelli).

The project described in chapter two delved into the nature of the time-averaging process that underlies sound texture perception. McDermott et al. (2013) revealed a time-averaging of summary statistics in the perception of texture. This opened an interesting new research avenue as to what the characteristics of such a time-averaging process might be. The project probed the averaging process for a broad range of textures, how the averaging process responded to continuity, whether the integration could be cognitively controlled, and how such a process might operate in an auditory scene.

The third chapter incorporated neuroimaging in an attempt to uncover the neural locus of texture perception in the auditory system. Using function Magnetic Resonance Imaging (fMRI) and sound texture synthesis, we attempted to identify what region of the system might be selective to texture, in particular the higher-order statistics that define texture. The measured responses were then compared to behavioral data to identify a neural correlate of texture perception.

The fourth chapter attempted to expand the perceptual space of texture by advancing the model of McDermott and Simoncelli (2011) to incorporate second-order amplitude modulations. While the perception of artificial sounds comprised of second-order amplitude modulation have been explored, their role in natural sound perception is unexplored. We present an updated auditory texture model that can capture second-order amplitude modulation and explore the perceptual consequences for the time-averaging of texture.

Lastly, the fifth chapter investigated the perception of textures synthesized from models of hearing impairment. Because the auditory texture model identifies a set of signal computable statistics, the idea was to investigate how these statistics changed when the model was altered. Specifically, we applied classic impairments, such as loss of cochlear compression and broadening of auditory frequency selective filters, and monitored the output statistics. In addition, we performed a series of behavioral experiments to identify the perceptual effects of altering the auditory texture model.

Finally, we offered a possible optimization that could recoup the texture statistics from a modified auditory model.

2

Adaptive Time-Averaging of Auditory Scenes[a]

Abstract

To overcome variability, estimate scene characteristics, and compress sensory input, perceptual systems pool data into statistical summaries. One example occurs in auditory scenes, where background texture appears to be represented with time-averaged sound statistics. How are averages computed? We probed the averaging mechanism using 'texture steps' - subtle shifts in stimulus statistics. Although generally imperceptible, steps occurring in the previous several seconds biased texture judgments, indicative of a multi-second integration window. Listeners seemed unable to willfully extend or restrict this window, but showed signatures of longer integration times for temporally variable textures. Moreover, step-induced biases were reduced by salient intervening texture discontinuities. Biases were also absent for steps produced by texture segments that perceptually segregated from the background. The results suggest an integration process that adapts to stimulus characteristics, excluding stimulus components likely to have distinct causes in the world, and extending integration when it benefits statistical estimation of variable signals.

2.1 Introduction

Sensory receptors measure signals over short time scales, but perception often entails combining these measurements over much longer durations. In some cases this is because the structures that we must recognize are revealed gradually over time. In other cases the presence of noise or variability means that samples must be gathered over some period of time in order for quantities of interest to be robustly estimated. Although much is known about the integration of information over space in the visual system, where dendritic trees instantiate receptive fields of progressively larger extent and complexity (Dumoulin and Wandell, 2008), little is known about analogous processes in time.

[a] This chapter is based on McWalter and McDermott (in Review - submitted March 2017).

Integration plays a fundamental role in the domain of textures - signals generated by the superposition of many similar events or objects. In vision and touch, texture often indicates surface material (bark, grass, gravel etc.) (Brodatz, 1966). In audition, textures provide signatures of the surrounding environment, as when produced by rain, swarms of insects, or galloping horses (McDermott and Simoncelli, 2011; Saint-Arnaud and Popat, 1995; Schwarz, 2011). Textures are believed to be represented in the brain by statistics that summarize signal properties over space or time (Landy, 2013; McDermott et al., 2013; Portilla and Simoncelli, 2000; Ziemba et al., 2016).

Statistical representations are believed to play a role in many aspects of perception (Alvarez and Oliva, 2009; Ariely, 2001; Balas et al., 2009; Brady et al., 2017; Brunton et al., 2013; Freeman and Simoncelli, 2011; Greenwood et al., 2009; Haberman and Whitney, 2009; Lorenzi et al., 1999; Nelken and De Cheveigné, 2013; Parkes et al., 2001; Piazza et al., 2013; Strickland and Viemeister, 1996), but texture is a particularly appealing domain in which to study them. Textures are rich and varied, ubiquitous in the world, relevant to multiple sensory systems, and arguably the only type of signal that is well described (at present) by signal-computable, biologically plausible models (Figure 2.1A). Moreover, the core integration operation for texture is plausibly simple: the statistics underlying texture perception can all be written as averages of sensory measurements of various sorts (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). In sound this averaging occurs over time, providing a canonical example of temporal integration.

The goal of this paper was to characterize the extent and nature of the averaging mechanism. What is the temporal extent over which averaging occurs? Is information averaged blindly within such windows, or is integration subject to principles of perceptual grouping? And does integration adapt to the intrinsic time scales of the signal being integrated?

It was not obvious a priori what sort of averaging process(es) to expect. Longer-term averages are more robust to variability and noise, but run the risk of pooling together signals with distinct statistical properties. Moreover, textures vary in the time scale at which they are stationary, reflected in the variability of their statistics (Figure 2.1B-C). Some textures, such as dense rain, have statistics that are stable over even short time scales, and that could be reliably measured with a short-term time-average. But others, such as ocean waves, are less homogeneous on local time scales despite being generated by a process with fixed long-term parameters. In the latter case a longer averaging window might be necessary to measure the long-term statistical properties accurately. It thus seemed plausible that an optimal integration process would vary in the time scale of integration depending on the nature of the acoustic input. An additional complexity is that textures in real-world scenes

typically co-occur with other sounds (Bronkhorst, 2000; Lee et al., 2017; McDermott et al., 2011; Moore et al., 2013), raising the question of whether texture statistics are averaged blindly over all the sound occurring within some window.

We developed a methodology to probe the averaging process underlying texture perception using synthetic textures whose statistics underwent a subtle shift at some point during their duration ('texture steps'; Figure 2.2A). The rationale was that judgments of texture should be biased by the stimulus history if the stimulus beyond the change point is included in the average. The results suggest an integration process that pools information over several seconds, and that cannot be willfully extended or shortened, but that adapts to stimulus homogeneity. Integration also appears to be partially reset by salient stimulus discontinuities attributed to the texture and to exclude stimulus segments heard as belonging to a distinct source. The methodology developed here could be readily applied to characterize averaging mechanisms in other sensory modalities.

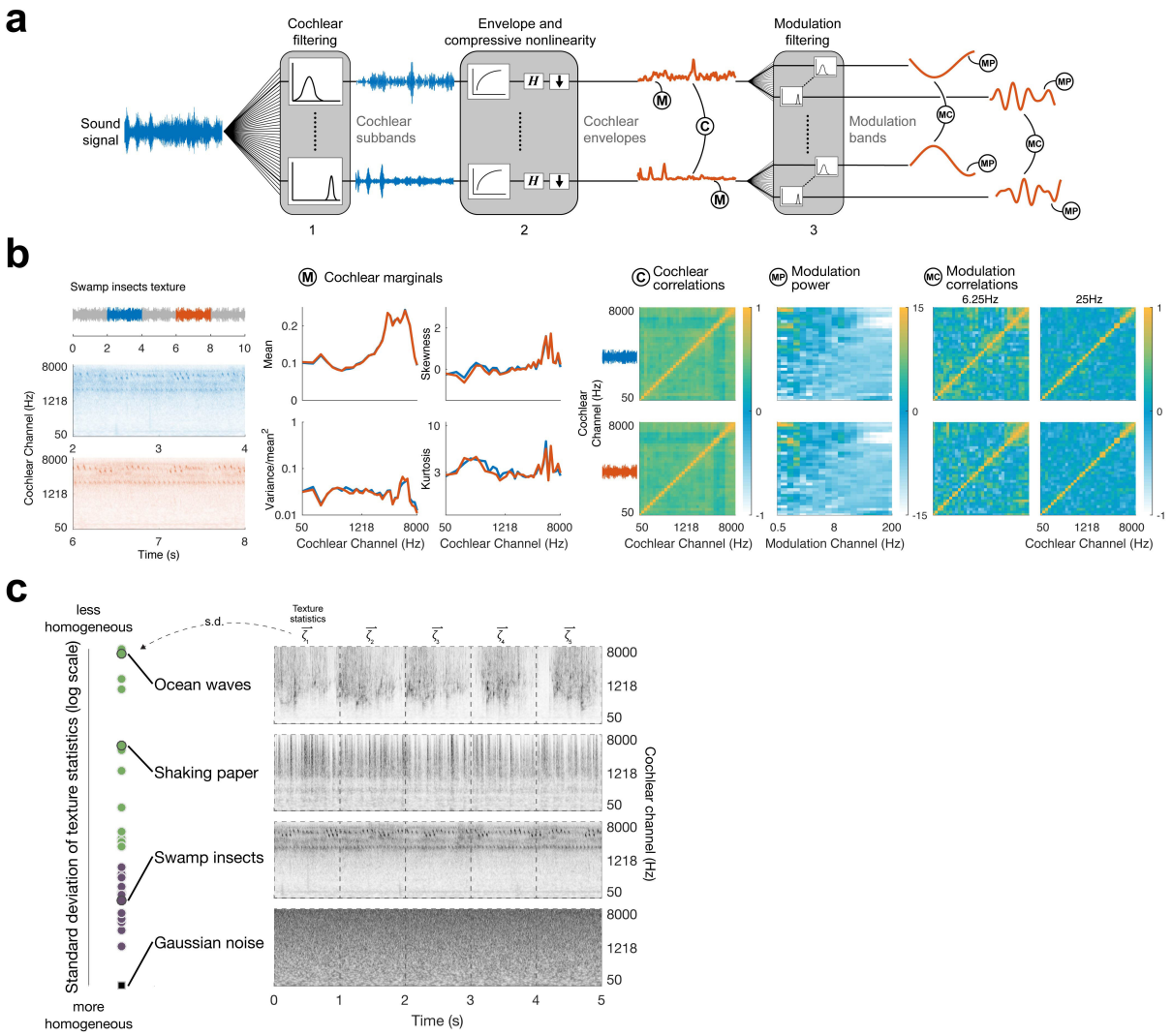


Figure 2.1: Variability in real-world texture statistics. (a) Auditory texture model (McDermott and Simoncelli, 2011). Statistics are measured from an auditory model capturing the tuning properties of the peripheral and subcortical auditory system in three stages: (1) a simulated cochlear filterbank, (2) nonlinear amplitude compression followed by amplitude envelope extraction, and (3) a modulation filterbank capturing selectivity to different envelope fluctuation rates. The statistics measured from this model include marginal moments of cochlear envelopes, modulation power, pair-wise correlations between cochlear envelopes, and pairwise correlations between modulation subbands. (b) Example of texture stationarity. Left panels show waveform (top) of an example sound texture (swamp insects). Spectrograms (bottom left) show 2s excerpts at 2 different time positions within the waveform. Right panels show texture statistics measured for each 2s excerpt. Cochlear marginal moments for each excerpt are plotted on same graphs. For brevity, modulation correlations, are shown for only two subbands. For this example texture, statistical measurements are fairly stable at the sampled time scale. (c) Variability of statistics in real-world textures. Left panel shows the standard deviation of texture statistics measured from multiple 1s excerpts of each of a set of 27 textures used in the subsequent experiments. Right panel shows spectrograms of 5 s excerpts of example textures (ocean waves, shaking paper, and swamp insects, as well as Gaussian noise). Some real-world textures have statistics that are quite stable at a time scale of 1s, while others exhibit variability (and would only produce stable estimates at longer time scales).

2.2 Results

Our main experimental paradigm involved asking listeners to compare a texture whose statistics were constant to a texture whose statistics subtly shifted mid-way through its duration (Figure 2.2A). The design thus necessitated generating stimuli at arbitrary points in a texture space ('morphs') and stimuli whose statistics underwent a change mid-way through their duration ('steps') from one such point to another (Figure 2.2B). Morphs were generated from statistics that fell on a line between the statistics of the reference texture and the statistics of a 'mean' texture (created by averaging the statistics of 50 real-world textures). Steps were generated by first synthesizing a morph for the starting values of the step statistics, and then running the synthesis procedure again on the latter part of the signal to shift the statistics to the end values of the step (Figure 2.2B).

Textures were synthesized using an extended version of the McDermott and Simoncelli (2011) sound synthesis procedure in which Gaussian noise was shaped to have particular values for a set of statistics. Statistics were measured from a model of the peripheral auditory system that simulated cochlear and modulation filtering (Dau et al., 1997). The statistics employed included marginal moments and pair-wise correlations measured from both sets of filters.

Figure 2.2C displays an example of a texture step and several morphs generated for the reference texture of swamp insects (see Figures S1 and S2 for visualizations of the statistic trajectories in example step stimuli). In practice we used several different reference textures per experiment, and the experiments were divided into blocks of trials based on a particular reference, so that listeners could be familiarized with the reference (Figure 2.2A).

The task required listeners to judge which of the two texture stimuli was more similar to a reference texture. To ensure that the stimuli would have the desired effect on this judgment were it in fact based on time-averaged statistics, we implemented an observer model (Figure 2.2D). The model measured statistics from the experimental stimuli using an averaging window of some duration, and chose the stimulus whose statistics were closest to those of the reference. Figure 2.2E shows the result the model performing the task with two different averaging windows. When the averaging window includes the step, the psychometric functions are shifted depending on the direction of the step, as intended. We tested whether human listeners would exhibit similar biases, and if so, over what time scales.

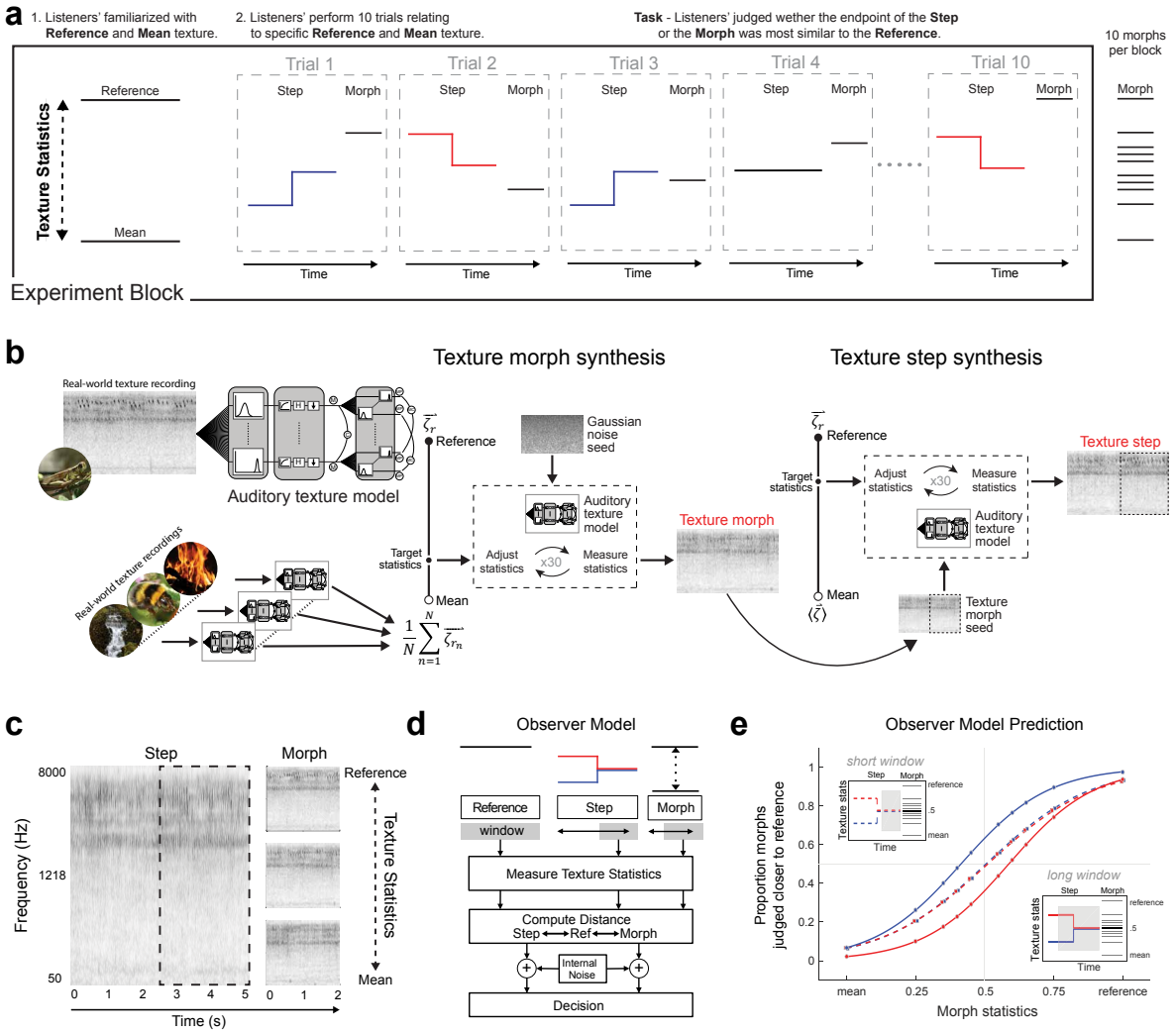


Figure 2.2: Texture step experiment methodology. (a) Schematic of trial structure for Experiments 1, 3, 4, and 5. Listeners were asked to judge whether the step or the morph was most similar to a reference texture that varied from block to block. Listeners were informed that the first stimulus on a trial (the step) could undergo a change and to base their judgments on the end of the stimulus. (b) Synthesis of texture morphs and texture steps. Two textures (reference and mean) were passed to the auditory model, which measured their texture statistics and generated target texture statistics at intermediate points along a line in the space of statistics. Synthesis began with Gaussian noise and adjusted the statistics to those of the desired intermediate point between the reference and mean texture. Texture steps were created by further adjusting a portion (dashed region) of a texture morph to match the statistics of another point on the line between the reference and mean texture. Here and elsewhere, we use a Euclidean metric in the space of texture statistics. (c) Example spectrograms for step experiment stimuli for the swamp insects reference texture. The example step stimulus contains a step in statistics 2.5s from the endpoint. The three example morph stimuli have statistics from the reference, midpoint and mean. (d) Observer model. The model averages statistics within a rectangular window extending from the endpoint of each stimulus and compares them to the statistics of the reference texture. (e) Performance of the observer model on a texture step experiment using two different window sizes. Plot shows the proportion of morphs judged closer to reference as a function of the morph statistics (points on a line between mean and reference statistics). When the integration window extends beyond the step (solid lines, bottom right inlay) the observer model exhibits a difference in the point of subjective equality between the two step conditions, but not otherwise (dashed lines, top-right inlay).

2.2.1 Experiment 1: Effect of stimulus history on texture judgments

We first examined the extent to which the stimulus history was included in texture judgments by positioning the step at two different points in time (either 1s or 2.5s from the endpoint). The step moved either towards or away from the reference, but in all cases was typically not salient to listeners (or to the authors). The texture morph (second stimulus in a trial) varied from trial to trial at points between the reference and mean texture. The experiment also included a baseline standard condition with constant statistics at the midpoint between the reference and mean texture.

Figure 2.3A plots the proportion of trials on which the morph was judged closer to the reference, as a function of the morph statistics. The step condition with constant statistics yielded the expected psychometric function: the proportion of trials on which listeners chose the morph increased as the morph statistics approached the reference, with the point of subjective equality at the midpoint on the statistic continuum. However, the psychometric functions for the step conditions were offset in either direction, indicating that the stimulus history influenced listeners' judgments despite the instructions to base their judgments on the endpoint of the step. We quantified this bias as the difference in the point of subjective equality for the two step directions (Figure 2.3B). The bias was statistically significant for both the 1s steps ($p<0.0001$; obtained by bootstrap) and 2.5s steps ($p<0.0001$; obtained by bootstrap), but was significantly reduced in the latter condition ($p=0.0034$; obtained by bootstrap). The results are suggestive of an averaging window on the order of a few seconds.

2.2.2 Experiment 2: Can listeners extend their averaging window?

The results of Experiment 1 suggest that listeners integrate over several seconds to estimate texture statistics even when they are warned that the signals are changing and that they should attend to their endings. These findings are consistent with the idea that listeners have difficulty restricting texture integration to a shorter interval. To investigate whether listeners could instead average over a longer temporal extent when it would benefit performance, we designed an alternative task. Listeners were presented with two texture excerpts whose statistics were fixed throughout their duration: a 'standard' excerpt of a set duration, and a 'morph' whose duration varied from 0.2 to 7.5 seconds and whose statistics could either be closer or further from the reference with equal probability (Figure 2.4A). Listeners again judged which of the two sounds was more similar to a reference texture, and the sounds were again generated from statistics drawn from the line between the mean texture and the reference texture statistics.

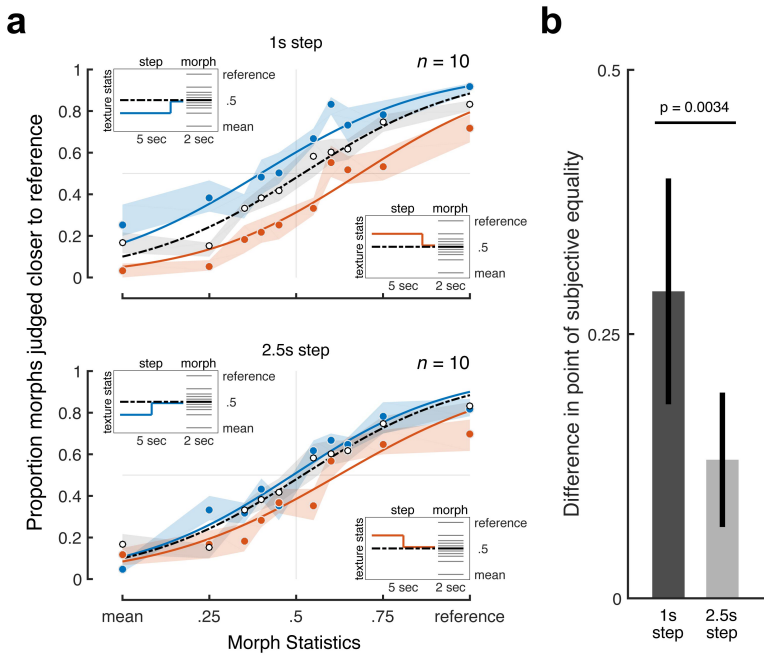


Figure 2.3: Results of Experiment 1 (Effect of texture step). (a) Texture discrimination for 1-s (upper panel) and 2.5-s (lower panel) step conditions. Insets show schematic stimuli (standard intervals with and without steps, and morph stimuli along continuum from mean to reference texture). Here and elsewhere, shaded regions show bootstrapped 68% confidence intervals on individual data points. Solid lines show corresponding logistic function fits. (b) Difference between points of subjective equality for the upward and downward step conditions (computed separately for 1-s and 2.5-s step locations). Here and elsewhere, error bars show bootstrapped 95% confidence intervals on the difference.

We reasoned that statistical estimation should benefit from averaging over the entire probe duration, such that an ideal observer would improve continuously as duration increases (Figure 2.4B, gray curve). But if listeners' have an integration window of fixed duration and retain only its most recent output for a given stimulus, performance might be expected to plateau after the probe duration exceeds it (Figure 2.4B, blue curve). We used two different standard durations (1 and 2 seconds) to verify that integration over the morph was not somehow constrained by the duration of the sound it was being compared to. Results were similar for the two standard durations and are averaged together here.

As shown in Figure 2.4C, performance increased with the duration of the morph interval up to approximately 2 seconds, but then plateaued, with no improvement evident for the longest two durations used. We assessed the location of the plateau by fitting an elbow function (Overath et al., 2015a) to the data, bootstrapping to obtain confidence intervals on the elbow point (Figure 2.4C&D). The best-fitting elbow point was 1.96 seconds. This finding is consistent with the idea that listeners cannot average over more than a few seconds, at least for the textures we used here, even when it

would benefit their performance.

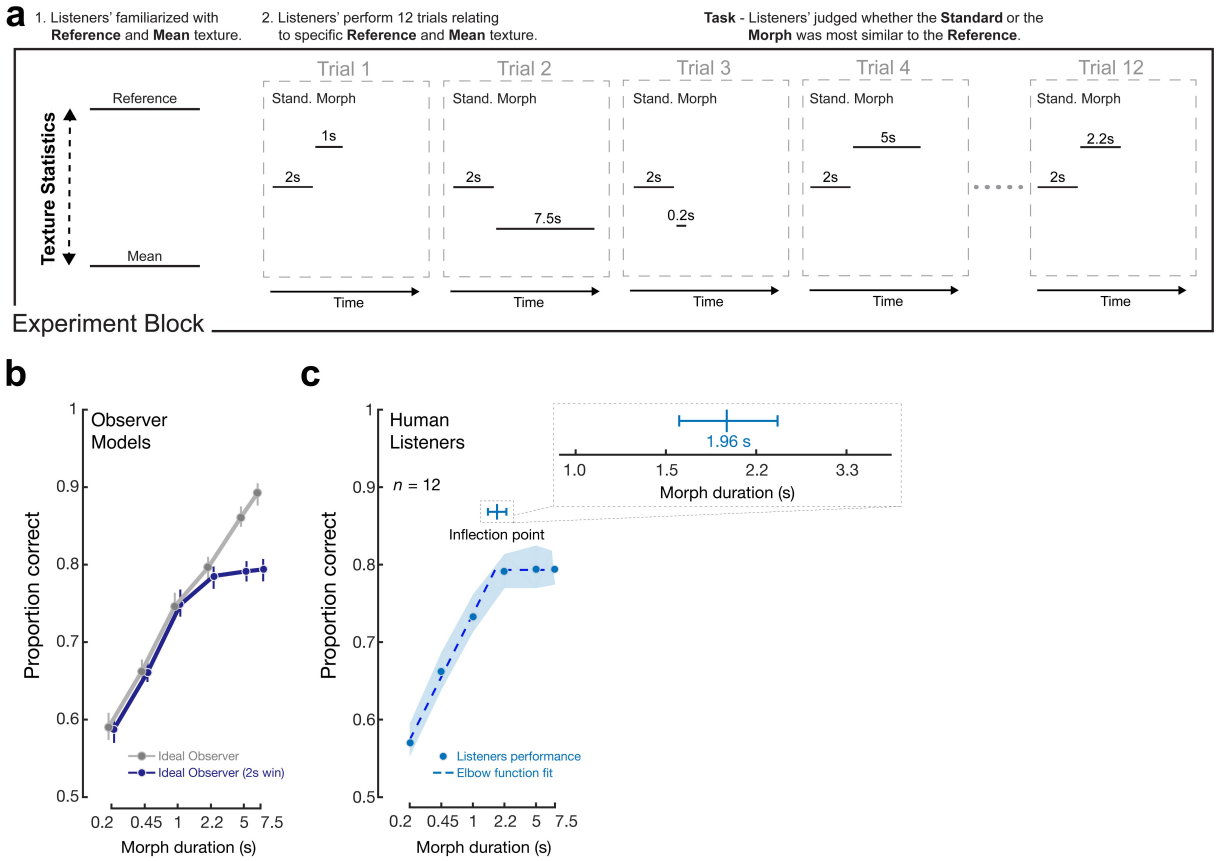


Figure 2.4: Design and results of Experiment 2 (Effect of texture duration). (a) Schematic of trial structure. Listeners were asked to judge which of two stimuli (here termed the Standard and the Morph) was most similar to a reference texture that varied from block to block. The Standard was fixed in duration and statistics; the Morph varied in duration and statistics across trials. Listeners were informed that the second stimulus (the Morph) would vary in duration from trial to trial. (b) Discrimination performance of observer models vs. probe duration (run on experimental stimuli). Performance of ideal observer (grey) increases with probe duration because statistical estimates become more accurate as more samples are available for averaging. An otherwise identical model with a 2s analysis window placed at the end of each stimulus (blue) shows a performance plateau once the probe duration exceeds the analysis window. Error bars show 95% confidence intervals obtained using bootstrap (over stimuli). (c) Discrimination performance of human listeners vs. duration. Shaded region indicates 68% confidence interval on individual data points, obtained via bootstrap. Dashed line shows piecewise linear elbow function fit. Solid blue line shows the median elbow point and 95% confidence intervals on the elbow point obtained using bootstrapping. Inset shows expanded view of elbow function inflection point.

2.2.3 Experiment 3: Effect of texture homogeneity on temporal integration

If the goal of texture integration is to estimate the statistical properties of a texture, optimal estimation should involve a tradeoff between the variability of the estimator (which decreases as the integration window lengthens) and the likelihood that the estimator pools across portions of the signal with

distinct statistics (which increases as the integration window lengthens). Because textures vary in their homogeneity, the window length at which this tradeoff is optimized should also vary. For highly homogeneous textures, such as dense rain, a short window might suffice for stable estimates, whereas for less homogeneous textures, such as the sound of ocean waves, a longer window could be better. We thus explored whether the auditory system adjusts the time scale of integration to the homogeneity of the signal it receives.

We first evaluated the homogeneity of a large set of textures by measuring the variability in statistics measured in 1s analysis windows (Figure 2.5A). We selected the six textures with maximum variability (less homogeneous) and the six with minimum variability (more homogeneous) for use as reference textures in an experiment. We repeated the paradigm of Experiment 1 with a step at 2.5s. All other parameters were identical to those of Experiment 1 except for the reference textures used and the morph duration, which was set to 5s given that more variable textures were used.

Figure 2.5B&C shows the average results for each of the two sets of textures (results were averaged across the reference textures within each set to produce stable psychometric functions given the amount of data collected). Both sets produced biases from the step at 2.5s, but the bias was substantially larger for the textures that were less homogeneous. This result suggests that more of the stimulus history is incorporated into statistic estimates for the less homogeneous textures, and is consistent with an averaging process whose temporal extent is linked to the temporal complexity of the texture. It also appears that the degree of bias for the more homogeneous textures is greater than that observed in Experiment 1 for the 2.5s condition. Although the participants are different between experiments, this difference could indicate that adjustments in the extent of averaging are relatively sluggish, such that the presence of less homogeneous textures within the experiment could affect the extent of averaging in the blocks with more homogeneous textures.

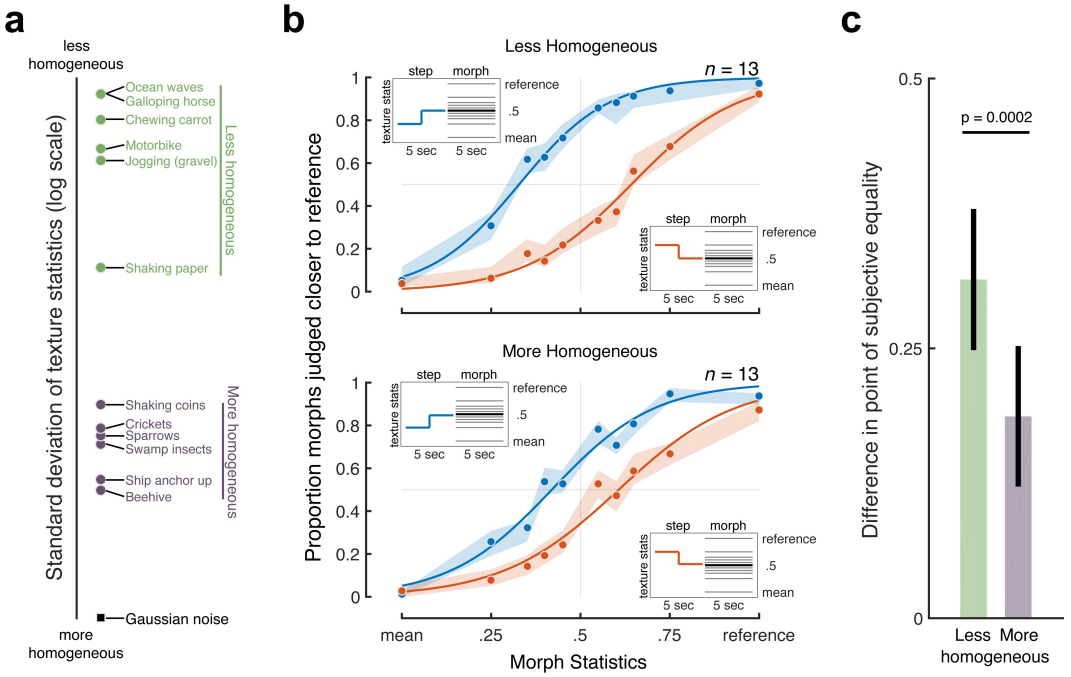


Figure 2.5: Results of Experiment 3 (Effect of texture homogeneity). (a) Variability in texture statistics measured across 1-s windows for the 12 reference textures (selected to include 6 more homogeneous and 6 less homogeneous textures). Variability of Gaussian noise statistics is provided for reference. (b) Texture discrimination judgments for a step at 2.5s for less homogeneous (upper panel) and more homogeneous (lower panel) textures. (c) Difference between points of subjective equality for the upward and downward step conditions for more and less homogenous textures.

As a further test of the influence of texture homogeneity, we reanalyzed the effect of probe duration from Experiment 2, dividing the 20 reference textures according to the variability of their statistics (Figure 2.6A). As shown in Figure 2.6B, the dependence of performance on the probe duration was different for the two groups of textures, with the less homogeneous textures appearing to plateau at a longer probe duration. To quantify this effect, we again fit piecewise linear 'elbow' functions to the data (Figure 2.6B&C). The elbow points of the best fitting functions were 1.3s and 2.9s for the more and less homogenous textures, respectively, and were significantly different (p=0.0002, bootstrapped). Although asymptotic performance was somewhat lower for the more-homogenous textures, it is not obvious how this in itself would produce an earlier plateau point. Overall, the results provide further evidence for an averaging process that pools information over a temporal extent that varies with the signal homogeneity.

2.2.4 Experiment 4: Effect of stimulus continuity on texture integration

The apparent presence of a multi-second averaging window that adjusts itself to stimulus statistics but is not under willful control raises the question of whether the integration process is 'blind', or

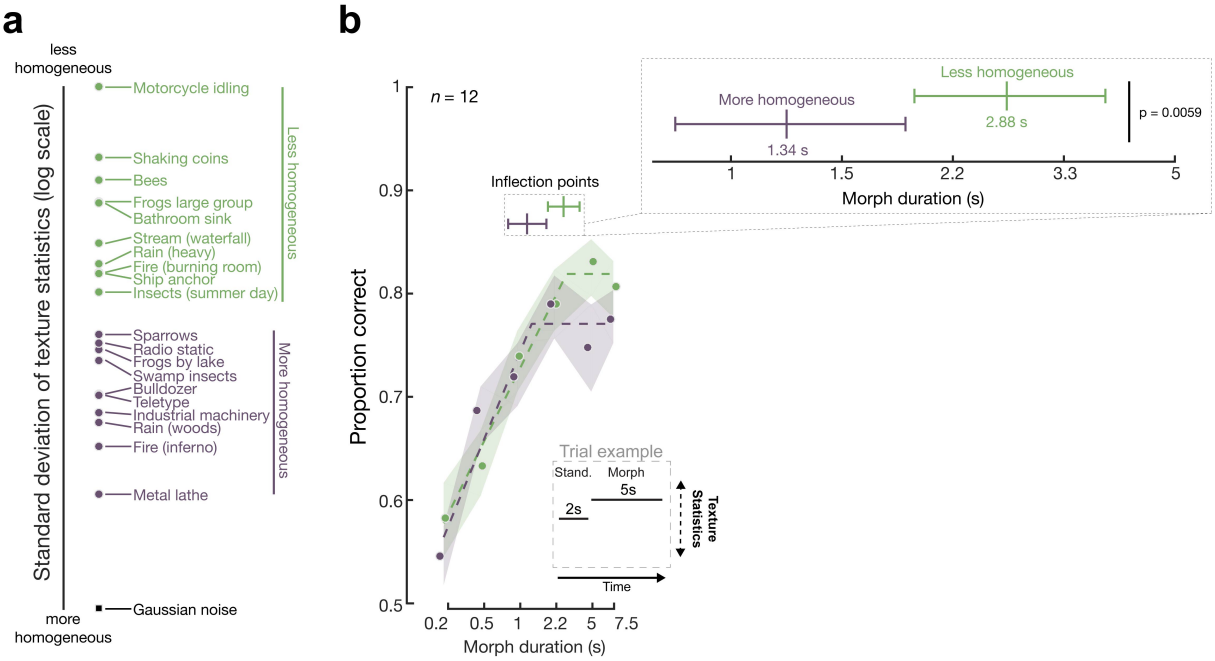


Figure 2.6: Influence of texture homogeneity on effect of duration (Experiment 2). (a) Variability in texture statistics measured across 1-s windows for the 20 reference textures used in Experiment 2. (b) Discrimination performance vs. duration for more homogeneous and less homogeneous textures. Bottom inlay shows schematic of trial structure. Top inset shows expanded view of inflection points of piecewise linear functions fit to data for the more and less homogeneous texture groups (with 95% confidence intervals obtained by bootstrap).

whether integration might be restricted to segments of the sound signal that are likely to be part of the same texture. We explored this issue by introducing audible discontinuities in the step texture immediately following the step, with the notion that the discontinuity might cause the history prior to it to be excluded from or down-weighted in the averaging process. We created three variants with a step positioned 1s from the endpoint (Figure 2.7A). The first condition preserved the continuity of the step (as in Experiment 1), while the second replaced the segment of the texture immediately following the step with a silent gap (200ms in duration). A third condition tested the effect of perceptual rather than physical discontinuity by filling the gap with a spectrally matched noise burst, causing perceptual continuity of the underlying texture (Carlyon et al., 2004; Warren, 1970). Listeners were again instructed to base their judgments using the endpoint of the step stimulus.

As shown in Figure 2.7B, the gap (middle panel) substantially reduced the bias produced by the step compared to the continuous (top panel) or noise burst (bottom panel) conditions. This reduction was statistically significant in both cases (gap vs. continuous: p=0.0003; gap vs. noise burst: p=0.0035; Figure 2.7C). The biases measured for the continuous step and noise burst conditions were not significantly different ($p = 0.46$). These results suggest that the integration process underlying

texture judgments is partially reset by discontinuities attributed to the texture but not by those attributed to interfering sources (noise burst). The results are consistent with the idea that integration occur preferentially over parts of the sound signal that are likely to have been generated by the same process in the world.

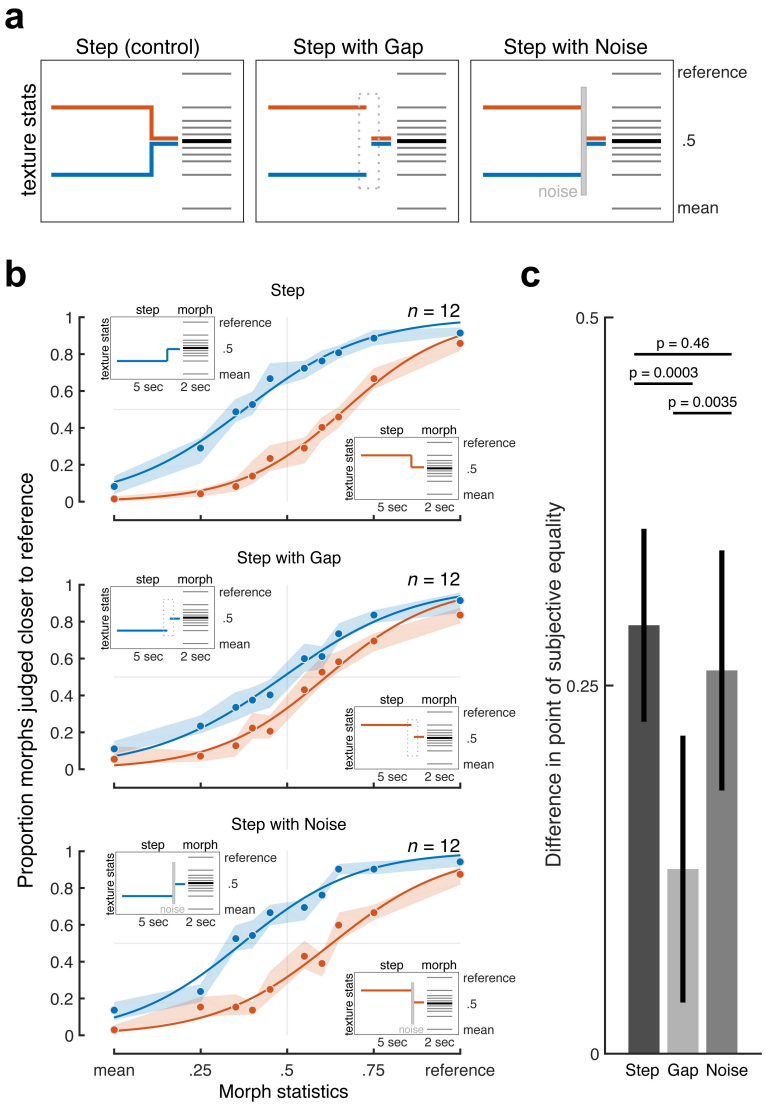


Figure 2.7: Results of Experiment 4 (Effect of texture continuity). (a) Schematic of stimuli. The gap condition included a 200ms silent gap positioned 1s from the endpoint of the step interval (i.e., immediately following the step). The noise burst condition included a 200ms spectrally matched noise burst in place of the gap. The intensity of the noise burst was set to produce perceptual continuity between the texture before and after it. (b) Texture discrimination for the three conditions. (c) Difference between points of subjective equality for the upward and downward steps for each condition.

2.2.5 Experiment 5: Effect of foreground/background on texture grouping

The finding that integration appears to occur across an intervening noise burst (Experiment 4) raises the question of whether such extraneous sounds are included in the texture integration process. Textures in auditory scenes are often superimposed with other sound sources, as when a bird calls next to a stream, or a person speaks during a rainstorm. Texture statistic estimates would be erroneously biased by these sounds if they were included in the integration process.

To test whether foreground sounds embedded in a texture are excluded from integration, we extended our texture step paradigm. Stimuli were generated with three segments (producing two steps, at 2s and 1s from the endpoint). In one condition ("Background") the segments were all the same intensity and appeared to all be part of the same continuous background texture (Figure 2.8A, left panel). In the other condition ("Foreground") the level of the second segment was 12dB higher than the other segments (Figure 2.8A, right panel). The level increment caused the middle segment to perceptually segregate from the other two segments, which were heard as continuing 'behind' it (Carlyon et al., 2004; Warren, 1970). The segment statistics were chosen such that integration over several seconds would yield biases in opposite directions depending on whether the middle segment was included in the integration. Listeners were told that the step interval might undergo a change in loudness, and that they should base their judgments on the endpoint of the step stimulus.

Without the level increment, listeners' judgments exhibited a bias toward the statistics of the middle segment (Figure 2.8B, top), consistent with its inclusion in the averaging process used to estimate texture statistics. However, the bias was reversed when the middle segment was higher in level than its neighbors (Figure 2.8B, bottom; the biases in the two conditions were significantly different: $p<0.0001$, Figure 2.8C). This bias is consistent with what might be expected if the middle segment were excluded from the integration (and if integration extended across the middle segment to include part of the first segment, as might be expected given the results of Experiment 4). These results provide further evidence that texture integration is restricted to sounds that are likely to have come from a similar source. Overall, the results of Experiments 4 and 5 indicate that texture perception functions as part of auditory scene analysis, concurrent with the grouping or streaming of sound sources.

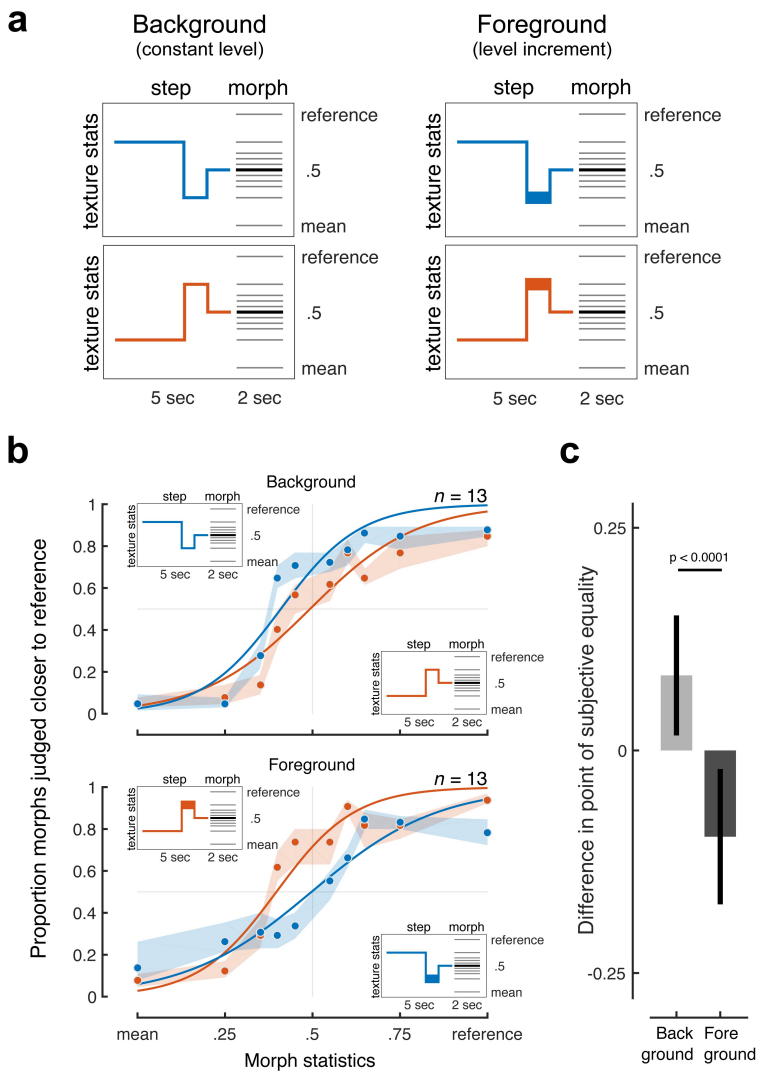


Figure 2.8: Results from Experiment 5 (Effect of foreground sounds on texture integration). (a) The background condition was composed of 3 segments with different statistics, creating steps 2s and 1s from the endpoint of the step interval. The grouping of the second segment with the other two was manipulated by increasing the level of the second segment by 12dB (indicated by thicker line in experiment schematic). The level increment caused the second segment to be heard as a distinct foreground sound, "behind" which the other segments perceptually completed. (b) Texture discrimination for foreground and background step conditions. (c) Difference between points of subjective equality for the two step directions for the continuous and level increment step conditions.

2.3 Discussion

Textures, be they visual, auditory, or tactile, are believed to be represented with statistics that are averages over time and/or space of sensory measurements. We conducted a set of experiments on sound textures to probe the nature of the averaging process. We found that the judgments of human listeners were biased by subtle changes in statistics that occurred in the previous several seconds, implicating an integration process over this extent (Experiment 1). This effect was present

even though participants were instructed to attend to the end of the stimulus and were warned that the stimulus would undergo changes. When given the opportunity to average over more than a few seconds (Experiment 2), listeners did not appear able to do so. However, the biasing effects of the stimulus history were more pronounced when textures were more variable, suggesting that the integration process extends itself in the presence of higher stimulus variability, perhaps as needed to achieve stable statistic estimates (Experiment 3). We also found that biasing effects were diminished when there was a salient change (silent gap) in the texture, suggesting that the integration process partially resets itself in such conditions (Experiment 4). Lastly, texture integration appears to occur across foreground sounds that interrupt a texture (Experiment 4), but to exclude such foreground sounds from the calculation of the texture's statistics (Experiment 5). The results indicate an adaptive integration process extending over several seconds that adjusts to the temporal complexity of the auditory input, and that appears to be inseparable from auditory scene analysis.

2.3.1 Adaptive time-averaging

Our results appear to provide an example of a sensory process adjusting its computation to the statistics of the stimulus. The integration process underlying listeners' judgments shows evidence of changing in temporal extent depending on the stimulus despite listeners being unable to willfully extend or shorten it when listening to a particular texture. The results suggest that the auditory system senses the variability of the stimulus and adjusts its integration accordingly. Stimulus variability could be sensed indirectly, perhaps by monitoring the stability of internal statistical estimates. It remains unclear over what time scale this variability adaptation occurs, but the fact that the bias for less homogeneous textures in Experiment 3 was somewhat greater than that in Experiment 1 suggests that the more homogeneous textures influenced integration on other blocks of trials. This time scale adaptation suggests a potential general principle for time-averaging that will be interesting to explore in other domains where statistical representations are used (Alvarez and Oliva, 2009; Ariely, 2001; Balas et al., 2009; Brady et al., 2017; Brunton et al., 2013; Greenwood et al., 2009; Haberman and Whitney, 2009).

2.3.2 Temporal integration in the auditory system

Although to our knowledge our experiments represent the first attempt to measure an integration window for texture, there has been longstanding interest in temporal integration windows for other aspects of hearing. Loudness integration is perhaps most closely analogous to the averaging that

seems to occur for texture, and is believed to reflect a nonlinear function of stimulus intensity averaged over a window of a few hundred milliseconds (Buus et al., 1997; Glasberg and Moore, 2002; Scharf, 1978; Zwislocki, 1969). Integration processes also occur in binaural hearing, on a similar time scale (Buell and Hafter, 1988). The integration time scale we have characterized for sound texture seems noteworthy both in that it seems to adapt to signal characteristics and in that it is quite long relative to typical time scales in the auditory system.

The long time scale raises the question of the neural mechanism computing the average. Receptive fields in the auditory cortex are rarely longer than a few hundred milliseconds (Atiani et al., 2014; Hullett et al., 2016; Miller et al., 2002b). To our knowledge the only phenomenon in the auditory system that extends over multiple seconds is stimulus-specific adaptation (Condon and Weinberger, 1991; Dean et al., 2005a; Kohn, 2007; Kvale and Schreiner, 2004b; Natan et al., 2015; Ulanovsky et al., 2004; Ulanovsky et al., 2003). Adaptation is widespread in sensory systems, and may also scale with stimulus statistics (Fairhall et al., 2001) but its computational role remains unclear. Whether adaptation could serve to compute quantities like texture statistics is an intriguing direction for future research.

Although we have suggested an averaging 'window' of several seconds, we still know little about the shape of any such window. Systematic exploration of the effects of steps at different time lags could help to determine whether different parts of the signal are weighted differently. However, we also note that there need not be a single window - texture is determined by many different statistics, and some might pool information over longer windows than others. The optimal integration time might vary across statistics depending on their intrinsic variability, much as it seemed that the averaging of stimulus history depended on the texture variability. Our results suggest a rough overall integration time constant of several seconds, but this could represent the net effect of different integration windows for different statistics. This issue could be investigated by generating steps in particular statistics. We also note that the apparent sophistication of the averaging (excluding distinct foreground sounds, for instance) renders the metaphor of a 'window' limited as a description of the integration process.

2.3.3 Role of texture integration in perception

One consequence of a multi-second averaging process is that sensitivity to changes in statistics should be limited - an averaging window will blur together the statistics on either side of a change. In this respect the step schematics in our figures represent only the generative parameters of the

stimuli, not something that could be directly measured from the stimulus (Figure S2). Consistent with this idea, the steps in our stimuli were usually inaudible. It will be interesting to explore the perception of stimuli longer than the texture averaging window. Do listeners retain any sense of drift in statistical properties when they slowly change over time, or when they undergo an abrupt change larger than those used in the present study (Boubenec et al., 2017)? The presumptive advantage of a limited averaging window is to retain some sensitivity to such changes. We know that listeners can retain estimates of texture statistics for multiple discrete excerpts, because they must in order to perform the discrimination tasks used here and elsewhere (McDermott et al., 2013). But it also appears that listeners do not retain distinct estimates from different time points within the same statistical process (otherwise they would have shown some improvement with duration beyond a few seconds in Experiment 2) (Viemeister and Wakefield, 1991). It thus remains to be seen how the temporal evolution of a changing statistical process is represented; here we simply used the change to characterize the statistic estimate at the end of a stimulus.

2.3.4 Texture perception and scene analysis

Our results indicate that texture integration is intertwined with the process of segregating an auditory scene into distinct sound sources. We found that a silent gap was sufficient to substantially reduce the influence of the stimulus history, suggestive of an integration process that selectively averages those stimulus elements likely to be part of the same texture. We also found evidence that concurrent foreground sounds are not included in the estimate of texture statistics when there is strong evidence that they should be segregated from the texture. These effects may be critical for accurate perception of textures in real-world conditions with multiple sound sources. One open question is whether the similarity of foreground sounds sound to the background texture affects the extent to which they are integrated.

2.3.5 Relation to statistical representations in other sensory modalities

The results presented here seem likely to have analogues in statistical representations in other modalities (Alvarez and Oliva, 2009; Balas et al., 2009; Brady et al., 2017; Greenwood et al., 2009; Haberman and Whitney, 2009; Parkes et al., 2001). Visual texture representation is usually conceived as resulting from averages over spatial position (Landy, 2013; Ziemba et al., 2016), but similar computational principles could be present. For instance, the spatial extent of averaging could potentially be influenced by the homogeneity of the texture. Visual textures also sometimes occur

over time, as when we look at leaves rustling in the wind (“Estimating the material properties of fabric from video”), and the time-averaging evident with sound textures could also occur in such cases. Tactile textures, where spatial detail is typically registered by sweeping a finger over a surface, also involve temporal integration (Hollins and Risner, 2000; Weber et al., 2013), the basis of which remains uncharacterized. In all these cases, the experimental methodology developed here could be used to study the underlying mechanisms.

2.4 Methods

2.4.1 Auditory Texture Model

To simulate cochlear frequency analysis, sounds were filtered into subbands by convolving the input with a bank of bandpass filters with different center frequencies and bandwidths. We used fourth-order gammatone filters as they closely approximate the tuning properties of auditory filters and, as a filterbank, can be designed to be paraunitary (allowing perfect signal reconstruction via a paraconjugate filterbank). The filterbank consisted of 34 bandpass filters with center frequencies defined by the equivalent rectangular bandwidth (ERBN) scale (50Hz to 8097Hz) (Glasberg and Moore, 1990). The output of the filterbank represents the first processing stage from our model (Figure 2.1A).

The resulting "cochlear" subbands were subsequently processed with a power-law compression (0.3) which models the non-linear behavior of the cochlea (Ruggero, 1992a). Subband envelopes were then computed from the analytic signal (Hilbert transform), intended to approximate the transduction from the mechanical vibrations of the basilar membrane to the auditory nerve response. Lastly, the subband envelopes were downsampled to 400Hz prior to the second processing stage.

The final processing stage filtered each cochlear envelope into amplitude modulation rate subbands by convolving each envelope with a second bank of bandpass filters. The modulation filterbank consisted of 18 half-octave spaced bandpass filters (0.5 to 200Hz) with constant Q = 2. The modulation filterbank models' the selectivity of the human auditory system and is hypothesized to be a result of thalamic processing (Dau et al., 1997; Jepsen et al., 2008; Miller et al., 2002b). The modulation bands represent the output of the final processing stage of our auditory texture model.

The model input was a discrete time-domain waveform, $x(t)$, usually several seconds in duration ($\sim$ 5s). The texture statistics were computed on the cochlear envelope subbands, $x_k(t)$, and the modulation subbands, $b_{k,n}(t)$, where k indexes the cochlear channel and n indexes the modulation channel. The windowing function, $w(t)$, obeyed the constraint that $\sum_t w(t) = 1$.

The envelope statistics include the mean, coefficient of variance, skewness and kurtosis, and represent the first four marginal moments. The marginal moments capture the sparsity of the time-averaged subband envelopes. The moments were defined as (in ascending ordering)

$$\mu_k = \sum_t w(t) x_k(t)$$

$$\frac{\sigma_k^2}{\mu_k^2} = \frac{\sum_t w(t)\big(x_k(t)-\mu_k\big)^2}{\mu_k^2},$$

$$\eta_k = \frac{\sum_t w(t)\big(x_k(t)-\mu_k\big)^3}{\sigma_k^3},$$

$$\kappa_k = \frac{\sum_t w(t)\big(x_k(t)-\mu_k\big)^2}{\sigma_k^4}, k \in [1...34] \quad \text{in each case.}$$

Pair-wise correlations were computed between neighboring cochlear bands. The correlation captures broadband events that would activate cochlear bands simultaneously (McDermott and Simoncelli, 2011; McDermott et al., 2009). The measure can be computed as a square of sums or in the more condensed form can be written as

$$c_{jk} = \frac{\sum_t w(t)\big(x_j(t)-\mu_j\big)\big(x_k(t)-\mu_k\big)}{\sigma_k^3}, j,k, \in [1...34],$$

such that $\quad k-j = [1,2,3,4,5,6,7,8].$

To capture the envelope power at different modulation rates, the modulation subband variance normalized by the corresponding total cochlear envelope variance was measured. The modulation power measure takes the following form

$$\sigma_{k,n}^2 = \frac{\sum_t w(t)\big(b_{k,n}(t)-\mu_k\big)^2}{\mu_k^2}, k \in [1...34], n \in [1...18].$$

Lastly, the texture representation included correlations between modulation subbands of distinct cochlear channels. Some sounds feature correlations across many modulation subbands (e.g. fire), whereas others have correlations only between a subset of modulation subbands (ocean waves and wind, for instance, exhibit correlated modulation at slow but not high rates (McDermott and Simoncelli, 2011). These correlations are given by

$$c_{jk} = \frac{\sum_t w(t)\big(b_{j,n}(t)-\mu_j\big)\big(b_{k,n}(t)-\mu_k\big)}{\sigma_{j,n}\sigma_{k,n}},$$

$$j \in [1...34], (k-j) = [1,2], n \in [3,5,7,9,11,13].$$

The texture statistics identified here resulted in a parameter vector, ζ , which was used to generated the synthetic textures.

2.4.2 Synthesis

Synthetic texture stimuli were generated using a variant of the McDermott and Simoncelli (2011) synthesis system. The original system was modified to facilitate the generation of 'texture steps' (sound textures that underwent a change in their statistics at some point during their duration), and 'texture morphs' (sound textures generated from statistics sampled at points along a line between two textures) (Figure 2.2B). The synthesis process also allowed for the generation of distinct exemplars that possessed similar texture statistics by seeding the synthesis system with different samples of random noise.

First, sound texture statistics were measured from 7-s excerpts of real-world texture recordings. The measured statistics comprised the mean, coefficient of variation, skewness, and kurtosis of each cochlear channel, pair-wise correlations across cochlear channels, the power from a set of modulation filters, and pair-wise correlations across modulation bands (McDermott and Simoncelli, 2011). Statistics were measured from 50 real-world texture recordings (Supplementary Table 1). Statistics from individual real-world recordings, ζ_{ref}, were used to synthesize 'reference' textures. The statistics of all 50 recordings were averaged to yield the statistics, ζ_{mean} , from which the 'mean' texture was synthesized. The measured statistics were imposed on a random noise seed whose duration depended on the stimulus type.

Texture steps were generated such that their statistical properties changed at some point in time by stepping from one set of texture statistics to another. The process began by synthesizing a texture from one set of statistics, $\zeta_{\Delta 1}$, usually set to a point along a line in the space of statistics between the mean and the reference $\big(\zeta_{\Delta 1} = \Delta 1 \zeta_{ref} + (1-\Delta 1)\zeta_{mean}\big)$, using Gaussian noise as the seed. We then passed the synthesis system a second set of statistics, typically a different point along the same line in the space of statistics, $\zeta_{\Delta 2}$, and the previous synthetic texture as the seed. The seed was modified to cause the portion after the step point to have the desired statistics; the portion prior to the step point

was not altered. This synthesis process minimized artifacts that might otherwise occur by simply concatenating texture segments generated from distinct statistics - the seeding procedure helped ensure that the degree and location of amplitude modulations at the border between segments with distinct statistics were compatible. We were thus able to generate textures whose statistics varied over time, yet had no obvious local indication of the change in statistics. In practice the changes in the statistics were difficult to notice, even to the authors.

2.4.3 Human Subjects

Participants were recruited from a university specific job posting site and had self-reported normal hearing. Participants completed the required consent form (overseen by the Danish Science-Ethics Committee or the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology) and were compensated with an hourly wage for their time.

2.4.4 Experiment 1: Effect of stimulus history on texture judgments

Stimuli The experiment used 6 reference textures: sounds generated by bees, sparrows, shaking coins, swamp insects, rain, and a campfire (Figure S3). These textures were selected because they were perceptually and statistically unique and produced relatively realistic synthetic exemplars. Each reference texture was paired with a mean texture whose statistics were the average of those of 50 real-world texture recordings, except for the time-averaged spectrum (i.e. cochlear envelope mean statistics), which was matched to the reference texture. The first stimulus (the 'step') presented within a trial was 5-s in duration. It consisted either of a texture with constant statistics set to the midpoint between mean and reference, or a step texture that started either 25% or 75% of the distance between mean and reference, stepping to the midpoint between reference and mean (Figure 1C). The step occurred either 1-s or 2.5-s from the endpoint. There were a total of 5 conditions for each texture family (constant statistics, step up at 2.5s, step down at 2.5s, step up at 1s, and step down at 1s). Five exemplars were synthesized for each step condition for each texture. The second stimulus on a trial was a 2-s morph, which was generated with statistics drawn from one of 10 positions on the line between the mean and the reference (indicated as the relative position along the line: 0 - mean, 0.25, 0.35, 0.4, 0.45 0.55, 0.6, 0.65, 0.75, and 1 - reference). The two second morph was extracted from a single 5-s synthetic texture exemplar with the desired statistics. The two stimuli on each trial were separated by an inter-stimulus-interval of 250 ms.

Procedure At the beginning of each block, the participant was required to listen to the reference and

mean textures at least once, after which they could begin a block of 10 trials. Each block presented one trial for each of the 10 morph positions paired with a randomly selected step stimulus. The step stimulus was randomly chosen from the set of 5 exemplars. The 2-s morph was randomly excerpted from the 5-s exemplar. Participants were required to select the interval that was most similar to the reference texture. They were informed that the step interval could change over time and were instructed to use the endpoint when comparing the two intervals. Participants had the option of listening to the reference or mean textures during the block, in order to refresh their memory as needed. Feedback was not provided. The experiment consisted of 300 trials in total, separated into 30 blocks (5 for each of the six reference textures). The order of the blocks was random subject to the constraint that two blocks with the same reference texture never occurred consecutively. Participants completed the experiment in sections of 5 blocks (60 trials). Prior to beginning the experiment, participants completed a practice session consisting of 6 blocks of 6 trials with feedback. The format of the practice trials differed from those of the main experiment: the step had a duration of 2-s and always had statistics at the midpoint between the reference and mean, and morphs were drawn from six positions (0, 0.25, 0.4, 0.6, 0.75, 1 on the continuum from the mean to the reference).

Participants Sixteen participants completed Experiment 1 (7 female, mean age = 26.0, SD = 5.4). We excluded the data from poorly performing participants with an inclusion criterion that was neutral with respect to the hypotheses: we required participants to perform at least 85% correct when the morph was set to its most extreme values (with the statistics of the reference or mean) for at least one of the 5 step conditions were included. Ten participants met this inclusion criterion.

2.4.5 Experiment 2: Effect of probe duration

Stimuli The experiment was separated into blocks formed by 20 reference textures (Supplementary Table 2) again selected from the larger set of 50 textures based on their statistical uniqueness and the realism of the resulting synthetic exemplars. Each trial consisted of a 'standard' stimulus, with statistics set to the midpoint between the mean and reference, followed by a 'probe' stimulus which varied in duration (0.2, 0.45, 1, 2.2, 5, 7.5-s) and had statistics set to either the 25% or 75% point between the mean and reference. The standard was either 1s (N=6) or 2s (N=10) in duration depending on the participant. We experimented with both durations of the standard interval to ensure that the standard duration would not have a large effect on the apparent integration window. The inter-stimulus-interval was 250 ms.

Procedure The experiment protocol was similar to that of the step experiment, with a reference tex-

ture synthesized from the statistics of a real-world texture recording and a mean texture synthesized from statistics averaged across 50 real-world texture recordings. At the beginning of each trial, the participant was required to listen to the reference and mean textures at least once, after which they could begin the first block of trials. Each block consisted of 12 trials: 6 probe durations at each of the two probe statistic values. The standard and probe were randomly selected from within a 10-s synthetic exemplar with constant statistics. The task was to select the interval most similar to the reference texture. Participants again had the option to listen to the reference or mean at any point during the block, to refresh their memory. Participants were informed that the probe stimulus would vary in duration from trial to trial and were instructed to use as much of the stimulus as possible when making their judgments. The experiment consisted of 240 trials separated into 20 blocks of 12 trials, each based around a particular reference texture. Trials were randomly ordered within blocks, and the block order was random subject to the constraint that two blocks from the same reference texture never happened consecutively. Participants completed the experiment in sections of 96, 96, and 48 trials.

Participants Sixteen participants completed Experiment 2 (12 female, mean age = 23.3, SD = 2.1). We analyzed only those participants that achieved at least 65% correct across all trials. Twelve participants met the inclusion criterion.

2.4.6 Experiment 3: Effect of texture homogeneity

Stimuli The experiment again consisted of blocks formed by particular reference textures. 12 reference textures were used: 6 more homogenous (derived from the sounds of birds, insects, beehive, shaking coins, ship anchor up, crickets) and 6 less homogenous textures (derived from the sounds of frogs, a motorbike, chewing carrot, galloping horses, shaking paper and ocean waves). These textures were selected by measuring the standard deviation of the texture statistics across 1-s windows in a large set of textures, and then choosing the textures that had highest and lowest variability. The 'step' stimuli contained a step either towards or away from the reference at 2.5-s (starting at either .25 or .75 on the mean-reference continuum and ending at the midpoint). The morph duration was extended to 5-seconds to facilitate judgments of the less homogenous textures.

Procedure The experiment contained 240 trials (two step conditions with each of 10 morph positions for each of the 12 reference textures, with the trials for a particular reference texture divided randomly into two nonconsecutive blocks of 10 trials), which were completed in sections of 24 blocks. The procedure was otherwise the same as that of Experiment 1.

Participants Sixteen participants completed Experiment 3 (9 female, mean age = 24.5, SD = 3.54). Thirteen participants met the inclusion criterion, by exceeding 85% correct in at least one endpoint morph conditions in one of the step conditions.

2.4.7 Experiment 4: Effect of noise burst and silent gap

Stimuli The 6 reference textures were the same as Experiment 1. All of the conditions had a step 1s from the endpoint (towards or away from the reference). One third of the trials had a 200ms gap immediately after the step (replacing the texture waveform with silence), starting 1s from the endpoint of the step stimulus. Another third of the trials replaced 200ms of the texture with a spectrally matched noise burst, again starting 1s from the endpoint of the step. The rms level of the noise was set to the peak amplitude of the step for each trial. The remaining third of the step trials were identical to the 1s step conditions of Experiment 1. There were thus six conditions, each with a different type of step stimulus (steps up and down, step up and down with a gap, and steps up and down with a noise-filled gap). The steps started at the .25 or .75 point and ended midway between the reference and mean texture, as in Experiment 1.

Procedure As in Experiment 1, the experiment was divided into blocks of 10 trials featuring a particular reference texture, and the participant was required to listen to the reference and mean textures at least once at the start of each block. Each block presented one trial for each of the 10 morph positions paired with a randomly selected step stimulus (corresponding to one of the six conditions). This randomization had the consequence that participants did not have prior knowledge of which trials contained a burst or a gap. Participants were informed that the step stimulus would change over time and that some trials would contain a noise burst or a silent gap. In other respects the experimental procedure was the same as in Experiment 1. In particular, participants were again instructed to attend to the endpoint of the step interval when comparing the two intervals.

Participants Sixteen participants completed Experiment 4 (8 female, mean age = 25.1, SD = 2.7). The experiment contained 360 trials, which were completed in sections of 120 trials (3 blocks). We used the same inclusion criterion as in Experiment 1, analyzing only those participants who exceeded 85% correct in at least one of the endpoint morph conditions in at least one of the step conditions. Twelve participants met the inclusion criterion.

2.4.8 Experiment 5: Effect of foreground/background on texture grouping

Stimuli Five reference textures were included in the experiment (sparrows, applause, swamp insects, rain, campfire). Both conditions featured a step stimulus composed of three segments (3s, 1s, and 1s in duration, respectively) with different statistics, such that there were steps 2s and 1s from the endpoint (Figure 8A). The first segment (3 seconds) began at either the 25% or 75% point between the mean and reference, depending on the step direction. The second segment (1 second) stepped from 25% to 75% or 75% to 25%. The third segment (1 second) fell midway between the reference and mean texture. The segment durations and statistic positions were chosen such that judgments would be biased in opposite directions depending on whether the second segment was included in the integration process. The level increment conditions were identical except that the level of the second segment was increased by 12dB (20ms tapered cosine window at onset and offset of segment 2), an amount sufficient to cause the second segment to perceptually segregate from the other segments and for the first and third segments to be heard as continuing 'behind' the second segment. There were thus four conditions, each with a different type of step stimulus.

Procedure As in Experiment 4, the experiment was dividing into blocks of 10 trials featuring a particular reference texture, and the participant was required to listen to the reference and mean textures at least once at the start of each block. Participants were informed the step interval would change over time and that some trials would contain an intermediate increase in loudness at some point. In all other respects the experimental procedure was the same as in Experiment 4. In particular, participants were instructed to attend to the endpoint of the standard interval when comparing the two intervals.

Participants Fourteen participants completed Experiment 5 (6 female, mean age = 25.8, SD = 3). The experiment contained 200 trials, which were completed in sections of 100 trials (2 blocks). We used the same inclusion criterion as in Experiment 1, analyzing only those participants who exceeded 85% correct in at least one of the endpoint morph conditions in at least one of the standard conditions. Thirteen participants met the inclusion criterion.

2.4.9 Statistics

Step Experiments (1, 3, 4, 5) To estimate psychometric functions for each standard condition in the step experiments, we fit a logistic function to the mean data for each morph position. Confidence intervals on the data points and on the difference between the points of subjective equality for two step conditions were derived from bootstrap (10,000 samples). For every bootstrap iteration, a

random set of participants (N = 10, 13, 12, 13 respectively for Experiments 1, 3, 4, and 5) was selected with replacement and the points of subjective equality for each step condition were obtained from the psychometric functions fit to the data sample. Significance values were estimated from the bootstrap distributions by fitting a Gaussian and then computing the p-value from the Gaussian (this allows estimation of the significance of values in the tail of the bootstrap distribution where there are not enough samples to reliably estimate the p-value from the histogram alone (Norman-Haignere et al., 2015b)).

Duration Experiment (2) To evaluate the reliability of the inflection point and response plateau, we fit a piecewise linear function ('elbow function') to the data by bootstrapping (10,000 samples). On each iteration of the bootstrap, we estimated the inflection point for the elbow function from a random set of the 12 participants with replacement. The 95% confidence intervals were estimated from the resulting distributions. P-values were estimated as described above.

2.4.10 Observer Model

We created an observer model to quantify the effect of an integration window on texture discrimination. The model (Figure 2.2D) measured the texture statistics of the two stimuli in each trial of the experiment using an averaging window extending backwards in time for some duration, and compared them to the statistics of the reference texture. The step and morph stimuli were generated from statistics on the line between the mean and reference texture statistics, but because the observer model computed statistics over finite windows, the measured statistics were always displaced from the line to some extent due to measurement error from the finite sample. To determine which stimulus was closer to the reference, the statistics of the step ($\overrightarrow{\zeta}_{\text{step}}$) and morph ($\overrightarrow{\zeta}_{\text{morph}}$) were thus projected onto the line between the statistics of the mean ($\overrightarrow{\zeta}_{\text{mean}}$) and reference ($\overrightarrow{\zeta}_{\text{ref}}$) texture:

$$\overrightarrow{\zeta}'_{\text{ref},i} = \overrightarrow{\zeta}_{\text{ref},i} - \overrightarrow{\zeta}_{\text{mean},i},$$
$$\overrightarrow{\zeta}'_{\text{step},i} = \overrightarrow{\zeta}_{\text{step},i} - \overrightarrow{\zeta}_{\text{mean},i},$$
$$\overrightarrow{\zeta}''_{\text{step},i} = \frac{\left(\overrightarrow{\zeta}'_{\text{step},i}\right)^T \overrightarrow{\zeta}'_{\text{ref},i}}{\left(\overrightarrow{\zeta}'_{\text{ref},i}\right)^T \overrightarrow{\zeta}'_{\text{ref},i}} \overrightarrow{\zeta}'_{\text{ref},i},$$
$$\overrightarrow{\zeta}'_{\text{morph},i} = \overrightarrow{\zeta}_{\text{morph},i} - \overrightarrow{\zeta}_{\text{mean},i},$$
$$\overrightarrow{\zeta}''_{\text{morph},i} = \frac{\left(\overrightarrow{\zeta}'_{\text{morph},i}\right)^T \overrightarrow{\zeta}'_{\text{ref},i}}{\left(\overrightarrow{\zeta}'_{\text{ref},i}\right)^T \overrightarrow{\zeta}'_{\text{ref},i}} \overrightarrow{\zeta}'_{\text{ref},i},$$

where i denotes the class of texture statistics (mean, coefficient of variation, skew, kurtosis, cochlear correlation, modulation power, modulation correlation). The stimulus whose projected statistics ($\overrightarrow{\zeta}''_{\text{step},i}$ or $\overrightarrow{\zeta}''_{\text{morph},i}$) had the smallest Minkowski distance P (summed across statistic classes) to the reference was chosen by the model. The distance was computed for each statistical class separately and normalized by the distance of the reference to the mean ($\overrightarrow{\zeta}'_{\text{ref},i}$) for that statistic group (because different classes of statistics have different units and cover different ranges). The distance was computed as follows:

$$\mathrm{P} = \min\left\{\sum_i \left(\mathrm{d}_{\text{step},i}\right), \sum_i \left(\mathrm{d}_{\text{morph},i}\right)\right\},$$
$$\mathrm{d}_{\text{step},i} = \frac{\left(\sum_k \left|\overrightarrow{\zeta}'_{\text{ref},i,k} - \overrightarrow{\zeta}'_{\text{step},i,k}\right|^p\right)^{\frac{1}{p}}}{\left(\sum_k \left|\overrightarrow{\zeta}_{\text{ref},i,k} - \overrightarrow{\zeta}_{\text{mean},i,k}\right|^p\right)^{\frac{1}{p}}},$$
$$\mathrm{d}_{\text{morph},i} = \frac{\left(\sum_k \left|\overrightarrow{\zeta}'_{\text{ref},i,k} - \overrightarrow{\zeta}'_{\text{morph},i,k}\right|^p\right)^{\frac{1}{p}}}{\left(\sum_k \left|\overrightarrow{\zeta}_{\text{ref},i,k} - \overrightarrow{\zeta}_{\text{mean},i,k}\right|^p\right)^{\frac{1}{p}}}.$$

where d is the distance to the reference texture statistics ($\overrightarrow{\zeta}'_{\text{ref},i}$), k indexes over statistics, and p is the order of the Minkowski distance (the observer results used $p = 2$). To adjust the observer model's overall performance to that of human listeners, Gaussian noise was added to the statistic distances prior to the decision.

To evaluate optimal performance characteristics for the duration experiment (Figure 4B), we created an ideal observer. The structure of the ideal observer was similar to the observer model, with the exception that the ideal observer operated on the entire length of the experimental stimulus, and that the statistics of the reference and mean texture were measured from 7.5s excerpts, equal to the longest morph duration.

Supplemental Figures

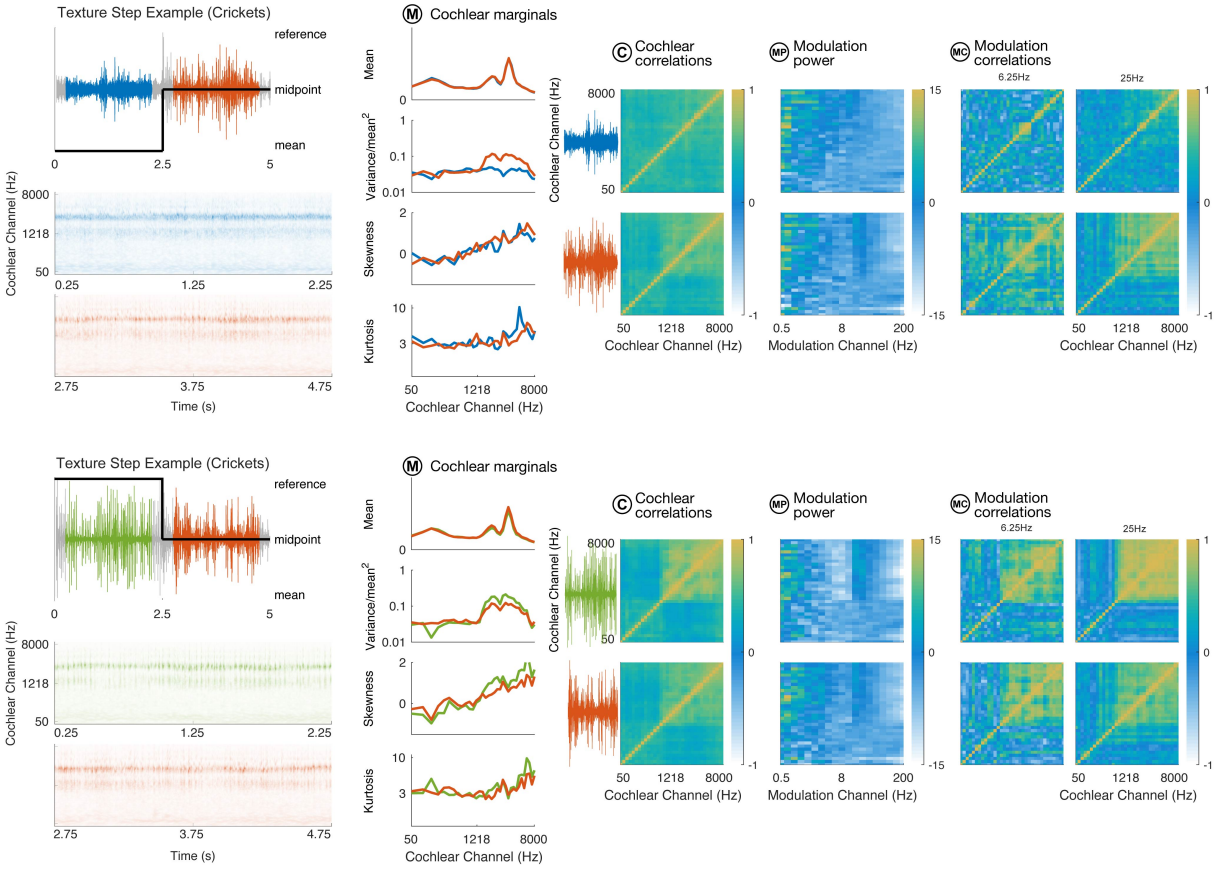


Figure 2.9: Statistical variation within synthetic texture steps. Two 2s excerpts are taken from each of two example stimuli in which an upward or downward step in texture statistics occurs at 2.5s. The statistics of each excerpt are plotted to the right (same statistics and format as Figure 1B). The time-averaged spectrum of the reference texture was imposed for the entire stimulus duration, hence the similarity across excerpts in the cochlear envelope marginal mean statistics (top panel within the column of marginal statistic plots).

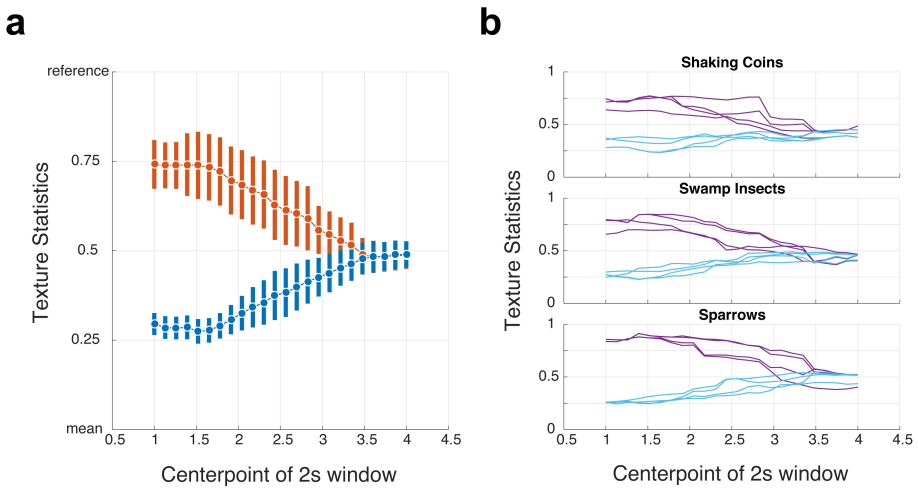


Figure 2.10: Texture statistics measured for 2.5s step condition in Experiment 1. (a) Average measured statistic trajectories for steps moving towards (blue) or away (red) from the reference texture with a shared endpoint midway between the reference and mean. Statistics were measured with a 2s analysis window centered at different positions within the stimulus and then projected on the line between the reference and the mean texture statistics. Because the analysis window is long, the measured statistics gradually change over time even though the statistics from which the signal was generated change discretely. Error bars show s.d. (b) Example statistic trajectories measured from individual stimuli.

3

Sensitivity to Sound Texture Statistics in Auditory Cortex[a]

Abstract

A primary objective of sensory neuroscience is to understand how the brain responds to the natural world. However much of our understanding of the auditory system has been gathered using carefully crafted artificial stimuli. In this study we employed sound textures, such as sparrows chirping or a stream flowing, to investigate the functional organization of the human auditory system using fMRI. Behavioral evidence suggests that such sound textures are identified using time-averaged summary statistics, but possible neural mechanism have yet to be explored. Synthetic textures were generated from ?texture statistics?, including marginal moments and pair-wise correlations, measured from a model that captures the frequency selectivity and amplitude modulation selectivity of the auditory system. In the first experiment, we generated textures by cumulatively including five subgroups of statistics for seven different texture sounds. An increase in BOLD-signal strength was observed in auditory cortex as the statistical subgroups were included. However, the BOLD-signal strength in inferior colliculus was similar for all statistical subgroups. In a corresponding behavioral experiment, we found that listeners? ability to identify textures generated with different statistical subgroups correlated with the parametric increase in BOLD-signal strength. Collectively, these results point towards a cortical mechanism that could identify auditory textures via time-averaged statistics.

3.1 Introduction

The perception of natural sound entails a neural representation that is transformed through a cascade of auditory processing layers. The early stages of the auditory system isolate the incoming acoustic

[a] This chapter is based on McWalter et al. (in Prep.).

waveform into a frequency tonotopy that is preserved to cortical layers, transmitting basic features of the sound. Later processing stages along the ascending auditory pathway extract more complex features to form abstract representations that support sound recognition and categorization. The exact processing hierarchy and the relationship to specific stimuli continue to be debated.

One reason the processing in cortical regions of the auditory system has been difficult to identify is the variability in response across natural stimuli. Natural sounds vary from temporally homogenous (such as rain) to highly variable and complex (such as speech or music). The auditory cortex preserves the basic frequency tonotopy found in the early auditory system as well as sensitive to frequency and amplitude modulations (Depireux et al., 2001; Humphries et al., 2010; Moerel et al., 2012; Theunissen et al., 2000). The sensitivity to natural sounds extends beyond the basic spectro-temporal features, where speech and music have been shown to elicit a unique response in primary regions of auditory cortex (Norman-Haignere et al., 2015a; Overath et al., 2015b). What other properties of natural sounds might be represented by cortical regions of the auditory system?

We were interested in how the auditory system processes natural sounds with inherent statistical regularity, referred to as texture. Sound textures ? such as rain, insect swarms or campfire ? are thought to be represented with time-average statistics measured from early auditory representations (McDermott et al., 2013; McDermott and Simoncelli, 2011). However, the possible mechanisms and processing organization of the auditory system to texture has yet to be explored. The recent use of texture to characterize visual perception and neural representations is an appealing approach to investigate processing hierarchy in sensory systems (Freeman et al., 2013; Ziemba et al., 2016). Responses to the higher-order statistical dependencies found in natural visual textures have recently been shown to differentiate the second visual cortical area V2 from primary visual cortex. It is still unclear whether stages of the auditory system may be similarly functionally characterized by their sensitivity to the statistical structure of natural sounds.

To investigate this, we employed sound texture synthesis to parametrically vary the statistical properties of texture sounds and measured the BOLD-signal response in human listeners. We began by measuring texture statistics from real-world recordings and generated families of textures from subsets of those statistics (McDermott and Simoncelli, 2011). Whereas the measured response in the auditory mid-brain (Inferior Colliculus) was nominal for textures generated from higher-order statistics, we found that primary and secondary auditory cortices responded selectively to texture statistics. We found that the response in auditory cortex mirrored that of perception in both identification and discrimination tasks. Collectively, our results indicate a cortical mechanism that

may facilitate the perception of sound texture.

3.2 Results

3.2.1 Synthesis of naturalistic texture

Our experiments used synthetic sound texture stimuli with statistics matching those of natural sounds. We began by measuring statistics from real-world audio recordings of sound textures processed through a biologically inspired model of the auditory periphery (Figure 1A). The model includes cochlear frequency selective filters, cochlear compression and auditory nerve transduction via envelope extraction. The cochlear envelopes are subsequently processed by modulation selective filters capturing modulation processing in the auditory midbrain. The texture analysis system measures marginal moments (mean, variance, skewness, and kurtosis) and pair-wise correlations of cochlear envelopes as well as modulation band power and across-cochlear envelope modulation rate correlations. The texture statistics are passed to a synthesis system that transforms a Gaussian noise seed to match the statistics of an original texture recording (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000) (Figure 1B). The iterative process results in a synthetic texture exemplar with similar perceptual qualities to the original (McDermott and Simoncelli, 2011).

Synthetic textures provide a means for using naturalistic stimuli while maintaining control over their time-averaged properties. The experiments leverage this by generating textures that cumulatively add statistics to spectrally matched noise stimuli to investigate where senstivity to the higher-order statistics may emerge along the auditory pathway.

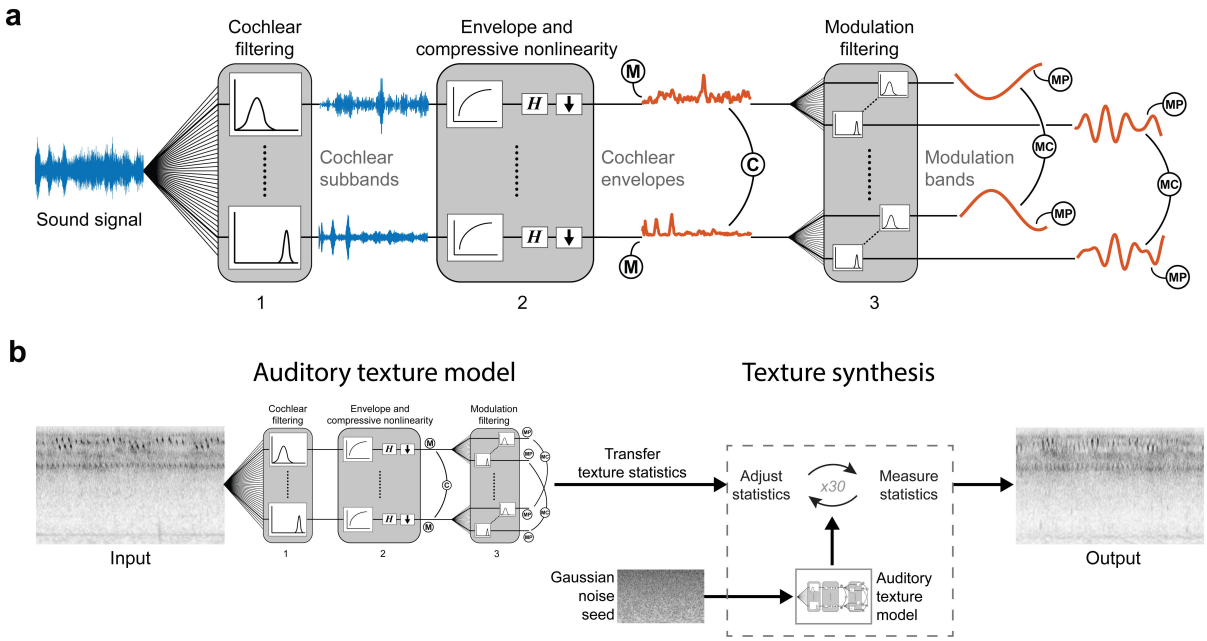


Figure 3.1: Stimulus generation via a texture analysis and synthesis system. (a) Texture analysis model (McDermott and Simoncelli, 2011). The functional auditory model captures the tuning properties of the peripheral and subcortical auditory system: (1) auditory filterbank modeled on the resonance frequencies of the cochlear, (2) nonlinearity captures the compression of the cochlear followed by computation of the Hilbert envelope, functionally modelling the transduction from the mechanical vibrations of the cochlea to neural signals in the auditory nerve, and (3) modulation filterbank capturing auditory midbrain selectivity to different envelope fluctuation rates. The statistics computed at various processing stages include marginal moments of cochlear envelopes (M), modulation power (MP), pair-wise correlations between cochlear envelopes (C), and pairwise correlations between modulation subbands (MC). (b) Texture synthesis system. Synthetic textures are generated by adjusting the time-averaged statistics of a Gaussian noise seed to match those measured from a real-world sound texture recording. The spectrogram of an original texture sound used in the study (swamp insects) is shown at the input of the synthesis system. The corresponding spectrogram of the synthetic texture is shown as the output of the synthesis system.

Figure 3.2: Texture statistics groups used in Experiment 1. (a) Spectrogram of 3 textures (sparrows ? left, insect swarm ? center, campfire ? right) used in the experiment (2s excerpts). (b) Five texture subgroups used to generate synthetic stimuli. The texture statistics groups are added cumulatively when generating the synthetic textures. 1- The cochlear channel mean statistic (first marginal moment, M1) for the three textures. Colors correspond to the texture spectrograms in Figure 2 (a). 2- Cochlear marginal moments (M2,3,4), including coefficient of variance, skewness and kurtosis. 3- Cochlear correlations (C) for the sparrows (left), insect swarm (center), and campfire (right). The same ordering is used in the subsequent statistics. 4- Modulation power (MP) for the textures. The subplots show the modulation power normalized to Gaussian noise. 5- Modulation correlation (MC) measured for a specific modulation band across cochlear channels for the textures.

3.2.2 Experiment 1a: BOLD fMRI responses to higher-order texture statistics

In the first experiment, we were interested in how synthetic stimuli generated from higher-order texture statistics might elicit differential responses in different parts of the auditory system. We presented human listeners with synthetic textures generated from subgroups of texture statistics and measured the BOLD-signal (Blood Oxygen Level Dependent). The texture statistics were separated into five subgroups and synthetic textures were generated by cumulatively including the statistics subgroups in those five steps. The first step matches the averaged response from the cochlear filters. This corresponds to spectrally matched noise, i.e. sounds that have the same overall spectrum by lack higher-order statistical dependencies. In step two we imposed the marginal moments (variance, skewness, and kurtosis) measured at the output of the cochlear filters. Step three we imposed cochlear pair-wise correlations. Step four we imposed modulation power. Step five we imposed modulation correlation.

The texture statistics from 3 example texture families (sparrows, insect swarm, and campfire) presented to all subjects are shown in Figure 2. Additional texture families were presented to a subset of subjects to investigate how well responses generalize across different texture sounds.

First we identified sound selective voxels in the inferior colliculus and auditory cortex as voxels that responded reliably to sound stimulation across the different stimuli (Figure 3B&D). Within these sound selective voxels, we found that only auditory cortical regions increased in response magnitude with the inclusion of higher-order texture statistic subgroups (Figure 3C). In particular, the mean response of all sound selective voxels increased with the inclusion of the cochlear marginal statistics and the modulation power (paired t-test, $p < 0.01$). No significant difference was observed in the inferior colliculus (paired t-test, $p > 0.01$). To identify ?texture sensitive? voxels responding linearly to the increase in statistics, we modelled the BOLD-signal response with a parametric modulation general linear model (GLM) with 5 steps for each statistics class (highlighted cluster Figure 3B, $p < 0.05$ FDR). Sensitivity to texture statistics was found to increase in a posterior direction from the lateral Heschl?s gyrus into the planum temporale region. In contrast, the response in inferior colliculus was nominal across all five subgroups of texture statistics and did not yield a significant cluster (Figure 3D, $p < 0.05$ FDR). This pattern of BOLD-signal increase in auditory cortex with the inclusion of higher-order texture statistics was observed for the individual participants (Figure 3F) and for the individual texture sounds (Figure 3G).

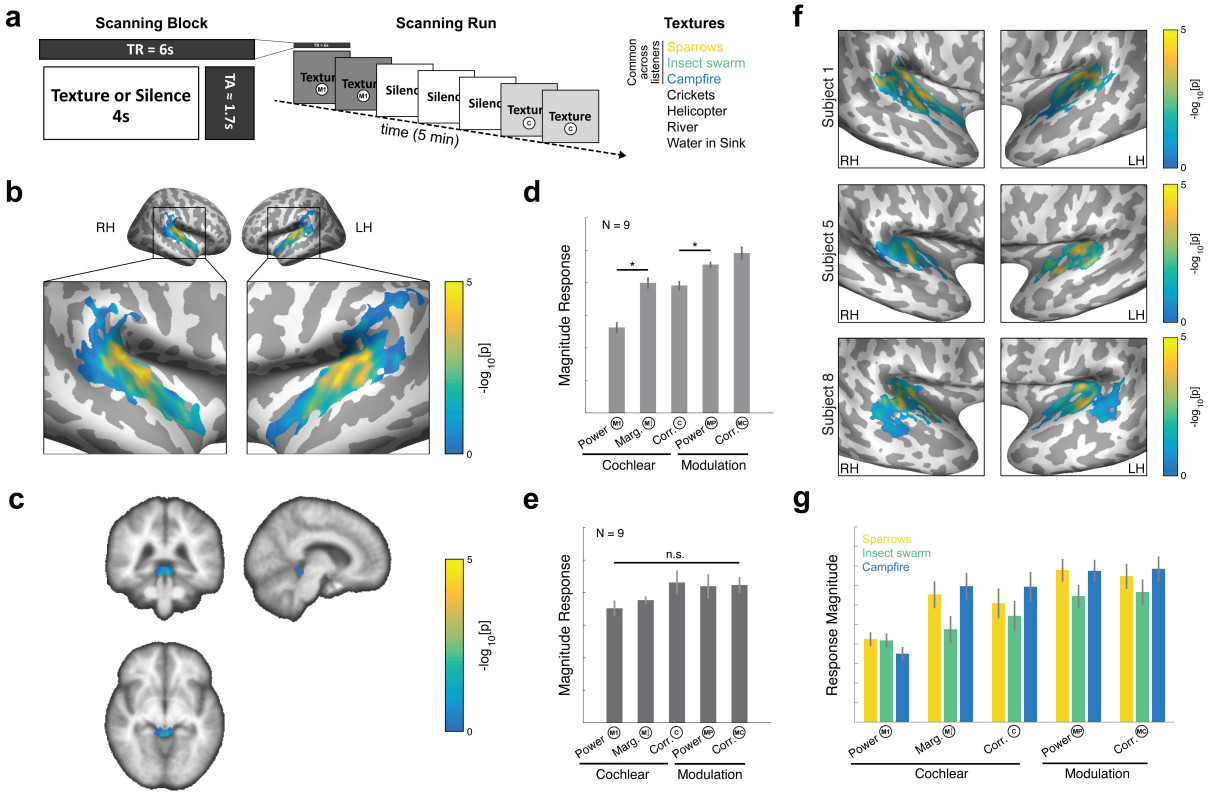


Figure 3.3: fMRI responses to parametrically varied texture statistics. (a) The scanning block was comprised of a 4-s texture stimulus presented during the quite period of a 6-s sparse scanning sequence (acquisition time TA = 1.7s). The scanning run consisted of two instances of the same texture family and statistics step presented in sequential blocks followed by three silent blocks. In total seven textures were presented. Three common textures were presented to all 9 listeners. (b) Sound selective cortical voxels. The colors indicate the degree to which the neural responses increase linearly with cumulative inclusion of texture statistics. (c) The response for each of the individual texture statistic subgroups averaged across sound selective voxels. The response increases with the inclusion of texture statistics. Error bars denote the standard error of the mean (s.e.m.). Asterisks show significant change in response magnitude (paired t-test, $p < 0.05$). (d) Sound selective subcortical voxels, composed exclusively of inferior colliculus. (e) Responses in the inferior colliculus did not differ for the different texture statistics subgroups (paired t-test, $p > 0.05$). (f) Texture sensitivity in sound selective voxels for 3 individual listeners. (g) The response magnitude to 3 texture families (sparrows, insect swarm and campfire) for stimuli generated with graduated statistics.

3.2.3 Experiment 1b: Behavioral texture identification

The inclusion of higher-order statistics has been shown to allow recognition and categorization of texture sounds (McDermott and Simoncelli 2011) and that different statistical subgroups are relevant for different texture families. We wondered whether the variation in the cortical response profiles for the individual texture families seen in Experiment 1 (Figure 3G) might be associated with their relevance in behavioral identification. To examine this, we conducted a psychophysical

identification task with human listeners using the same stimuli generated for each of the 5 texture statistics subgroups (similar to that conducted by McDonalds 2011).The listeners were presented with a 4 second texture and required to identify the sound from a list of 5 label descriptors (Figure 4A).

Replicating previous results, identification performance increased when the stimuli included more of the texture subgroups (Figure 4B, $F[6,49] = 118.11$, $p < 0.0001$). Identification performance was poor when only the cochlear mean statistics were used to the generate the stimuli, but gradually increased with the inclusion of higher-order statistics. The identification performance varied considerably across textures, with some texture identified by only the cochlear marginal (e.g. small stream), while others required inclusion of the modulation statistics (e.g. helicopter). Recognition performance for the three textures used in the fMRI experiment are shown in Figure 4C.

We observed that responses in texture sensitive regions in the lateral Hechl?s Gyrus and along the posterior superior temporal plane varied parametrically with identification performance (Figure 4D). The same pattern of response was seen when examining each the texture families individually (Figure 4E).

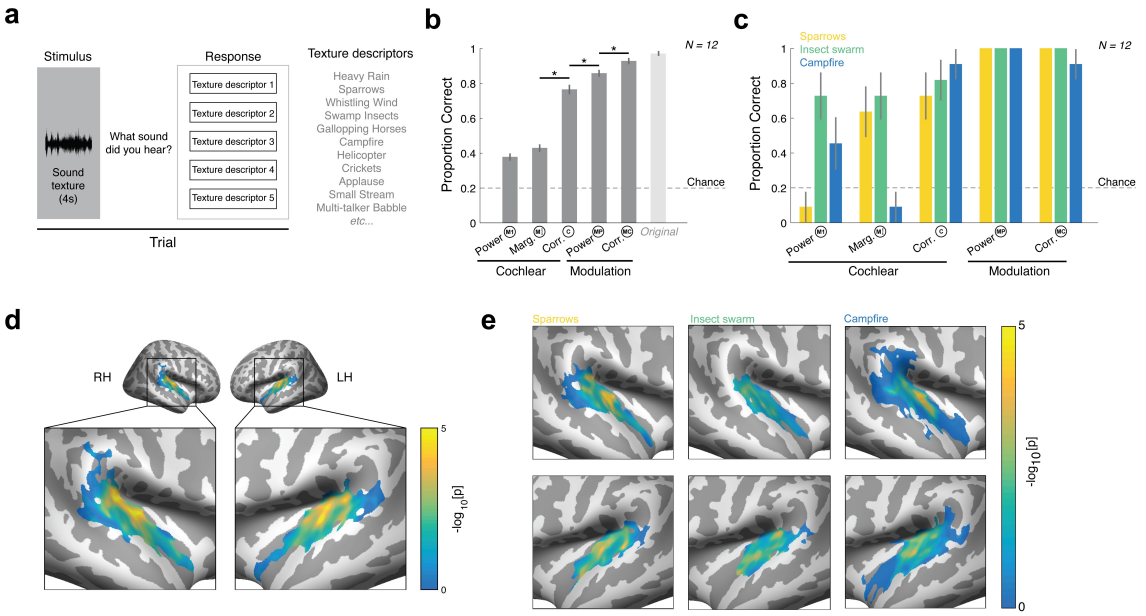


Figure 3.4: Texture identification performance correlation to BOLD-signal response. (a) Schematic of behavioral experiment paradigm. Listeners identified a texture from a list of five label descriptors. (b) Identification of texture improved with the inclusion of higher-order texture statistics. Asterisk denotes significance between paired conditions ($p < 0.01$, paired t-test). Error bars denote standard error of the mean (s.e.m.) (c) Identification performance for 3 textures common across all fMRI listeners. (d) BOLD-signal response in listeners presented with synthetic texture cumulatively including statistics. Color indicates significance, which captures the extent to which the fMRI voxel responses increase as a function of the behavioral performance (proportion correct from plot b). (e) BOLD-signal response in listeners for select texture families. Color indicates the statistical significance of the parametric increase in BOLD signal strength with the proportion of correctly identified textures as statistical subgroups were added (plot c).

3.2.4 Experiment 2: Response to texture morphs

In Experiment 1 we identified cortical regions sensitive to texture statistics by cumulatively adding groups of statistics as defined in our model of the early auditory system. With sound signals arising in the natural world, however, different statistical properties co-vary (Swartz & Simoncelli, 2001). In Experiment 2, we examined texture sounds where all statistical properties varied together between the natural texture sound and spectrally matched noise. We built ?texture morphs? by gradually interpolating all statistical parameters between the noise and the texture in 7 steps on a logarithmic scale (0.10, 0.16, 0.25, 0.40, 0.63, 1.00 - texture). Gradual titration of higher-order statistics allowed us to examine perception and neural sensitivity to varying degrees of ?naturalness? of the texture (Freeman et al., 2013). In a separate psychophysical and fMRI experiment, we presented the textures to human listener?s and measured the fMRI responses for 6 morph steps relative to spectrally matched noise (Figure 5A). Listeners in the psychophysical test performed a 3-interval, 2-alternative forced choice (odd-one-out) task to discriminate each step from the noise reference. Similarly, in the fMRI experiment listeners were presented with the each of the steps alternating with the spectrally matched noise.

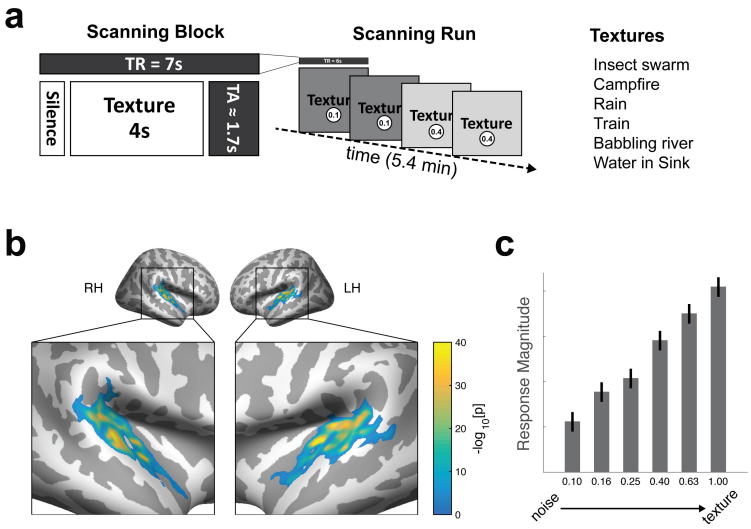


Figure 3.5: BOLD-signal response to the naturalness of the texture. (a) Listeners? were presented with 2 texture morphs followed by 2 spectrally matched noise sparse scanning sequence for a given texture family. The two texture morphs were from the same morph step (0.1, 0.16, 0.25, 0.4, 0.63 or 1.00), but presented in a random order in the scanning run. (b) Voxels that respond to a change in the texture naturalness. Response is presented in primary regions of auditory cortex. (c) The right panel shows the response to each of the 6 texture morphs relative to spectrally matched noise. The bars show standard error of the mean (s.e.m.).

As the textures morphed away from the noise, we again observed a pattern of increasing activity in

primary and secondary auditory cortices (Figure 5B&C). Contrasting each of the 6 texture morph steps to spectrally matched noise within sound selective voxels, responses increased monotonically as the ensemble of higher-order statistics was added. We did not observe any response in inferior colliculus, which may be attributed to the higher baseline BOLD-signal response to spectrally matched noise as opposed to silence (Experiment 1). In the companion perceptual task using the same texture morph stimuli we found that discriminability increased with the naturalness of textures (Figure 6A&B). Correlating with the fMRI responses, we found that responses across auditory cortex predicted their behavioral discriminability (Figure 6C).

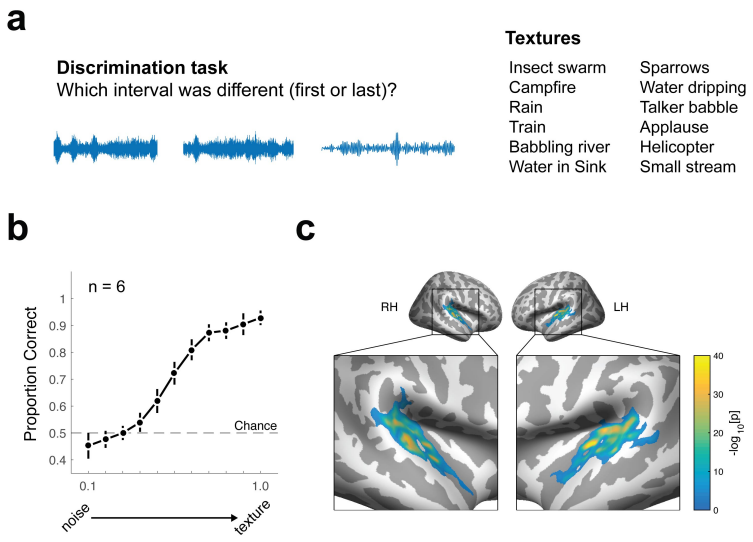


Figure 3.6: Discriminability (behavioral) of texture naturalness (a) Three-interval two-alternative forced choice experiment. The listeners were presented with 3 textures. Two were synthesized from the noise (0.1) while the other varied in its naturalness between 0.1 and 1.0. (b) Behavioral response to texture discrimination task. Performance increased with the naturalness of the texture. Error bars show s.e.m. (c) Parametric modulation response based on behavioral results for Subject 1.

3.3 Discussion

In this study, we investigated the neural response of the human auditory system to sound texture. By manipulating the statistical properties of synthetic sound textures, we were able to identify auditory cortical regions that respond selectively to stimuli generated from higher-order texture statistics. In the first study, we presented synthetic textures with varying degrees of statistical complexity ? from cochlear mean (first-order) to the modulation correlations (higher-order). Our results suggest that the sensitivity of the auditory system to higher-order texture statistics is reflected in a cortical response, whereas the auditory midbrain (inferior colliculus) appears to respond to sound regardless

of their statistics. These results were further examined by analyzing the BOLD-signal with behavioral texture identification data, which again pointed towards the auditory cortex as a neural locus for sound texture perception. Finally, we examined the auditory systems BOLD-signal response and behavioral discriminability to the naturalness of texture. This secondary experiment also suggested that texture perception may be mediated by cortical processes.

3.3.1 Caveats

Our results show that auditory cortex responds to synthetic texture stimuli generated from higher-order statistics. However, it remains unclear if this demonstrates a sensitivity to sound statistics or whether more fundamental acoustic properties of our stimuli are driving the response. Auditory cortex is known to be selective to audio frequency, and often spectro-temporal modulations also drive cortical responses (Chi et al., 2005; David et al., 2009; Depireux et al., 2001; Mesgarani et al., 2006; Pasley et al., 2012; Theunissen et al., 2000; Theunissen et al., 2001). Both the cochlear marginal (coeff. of var., skew, kurtosis) and modulation power affect the amplitude modulation spectrum. Therefore, it may be possible that the observed changes in the BOLD-signal response are simply due to changes in the acoustic properties of the stimuli. It is interesting to note that there are many voxels in primary regions of auditory cortex that seem to be modulated by the recognition performance.

Perceptually, many textures appear to rely on correlation statistics for identification (McDermott and Simoncelli, 2011). However, our results show no BOLD-signal increase with the inclusion of cochlear correlation of modulation correlation statistics. It may be the case that representation of across cochlear channel information is not represented in a specific region, but in the correlation of activity across cortex. This could be investigated by investigation the correlation of texture with the activation in across regions of cortex.

Our paradigm yielded a response to sound stimuli in inferior colliculus, but this did not appear to be modulated by the inclusion of higher-order statistics (Experiment 1). Inferior colliculus demonstrates a frequency tonotopy, similar to cortex, and has been identified as a potential processing node for envelope amplitude modulation selectivity (Dau et al., 1997; De Martino et al., 2013; Joris et al., 2004b). There is also some evidence that midbrain regions, such as inferior colliculus and thalamus, are sensitive to sound statistics (Ayala et al., 2013). A priori, it was not obvious that textures may be represented in auditory cortex as the model used to synthesize textures is primarily attributed to peripheral or subcortical processes (Dau et al., 1997; Jepsen et al., 2008). Although our results point towards a cortical mechanism underlying the perception of texture, other more temporally acute

electrophysiological techniques, may provide additional insight into this.

3.3.2 Neural Sensitivity to Natural Sound Statistics

There is mounting evidence that sensory systems have evolved to handle natural sounds (Theunissen and Elie, 2014). Our experiments adopted the approach of using naturalistic stimuli to probe sensory systems. Natural stimulus statistics have been successful in delineating cortical areas (Freeman et al., 2013), and we show that responses in auditory cortex may be driven in part by time-averaged summary statistics. Although the acoustic properties of the stimuli in experiment 1 and 2 are fundamentally different at each interval, we reliably observed an increase in the neural response with the inclusion of higher-order statistics. Additionally, the relationship between sound recognition in a behaviorally relevant context, and BOLD-signal response is also an intriguing hypothesis that our findings support.

3.3.3 Hierarchy of Processing in the Auditory System

We investigated auditory responses by parametrically including higher-order texture statistics which yielded a hierarchy of processing between the midbrain and auditory cortex. There is some evidence to suggest that there is a transformation in the receptive fields of neurons between inferior colliculus and auditory cortex (Atencio et al., 2012). Our findings provide a compelling example where a distinct response increases is recorded at later stages that are not observed at earlier stages. Other work involving speech, music and environmental sounds also point towards region selective processing beyond basic acoustic features. These findings are relevant for to the scope to which the auditory system uses texture for particularly tasks. For example, are texture statistics only relevant for temporally homogenous natural sounds or do they play a role in the perception of more complex stimuli, such as speech.

3.3.4 Sensory Perception of Texture

Sensory systems navigate the natural world, and an efficient approach may be to share neural coding strategies across modalities. Texture perception, in particular, appears to be relevant for different sensory modalities (Balas et al., 2009; Freeman and Simoncelli, 2011; McDermott et al., 2013). Relevant perceptual qualities of both image and sound textures appear to be captured by space or time summary statistics (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). It is yet unexplored if combing behavioral tasks, such as the sound produced by touching a surface,

elicits a shared response. Experiments in this domain could provide clues as to what sensory systems require and expect from events in the natural world.

3.4 Methods

Model. The auditory model captures the tuning properties of the early auditory system (Figure 1a). The model follows a linear-nonlinear-linear (LNL) processing structure. The first stage is a linear Gammatone (34 channel) filterbank, modeling the frequency selectivity of the cochlear. Gammatone filter center frequencies were set using the equivalent rectangular band scale, which is roughly logarithmic, from 50Hz to 8kHz (Glasberg and Moore, 1990). The second stage is a power-law nonlinearity, modeling cochlear compression (Ruggero, 1992a), with an exponent of 0.3. Next, the Hilbert envelope is extracted which models the transduction of the cochlear (basilar membrane) vibration to the auditory nerve (Joris et al., 2004b). The signal is subsequently down-sampled to 400Hz. The final stage is a linear modulation (19 channel) filterbank, which is primarily derived from behavioral experiments, but is thought to be a subcortical (thalamus) processing stage (Dau et al., 1997). The spacing of the modulation filters is on a half-octave scale from 0.5Hz to 200Hz (down-sampled Nyquist frequency), with the lower and upper filters of type low-pass and high-pass respectively.

We measured time-averaged texture statistics from real-world recordings. The texture statistics include cochlear envelope marginal moments (mean, variance, skew, and kurtosis) and pair-wise correlations, modulation power and pair-wise modulation rate correlation across cochlear channels. The texture statistics were similar to those proposed by McDermott and Simoncelli (2011).

Synthesis. The synthetic textures were generated by adjusting a Gaussian noise seed to match the statistics of a particular input texture. The noise seed is processed by the auditory texture model to the modulation bands. The modulation power and correlation statistics are imposed by minimizing the error between the input texture and noise statistics using gradient descent (McDermott and Simoncelli 2011). The modulation bands are reconstructed, and the cochlea envelope marginal and correlation statics are imposed. The cochlear bands are then reconstructed back to the single channel. Reconstruction is achieved using synthesis filterbanks generated from the Gammatone and modulation filterbanks, both of which are paraunitary. A convergence criteria, based on the SNR of the texture statistics (Portilla and Simoncelli, 2000), was monitored and usually conformed well within the maximum allowable iterations (SNR = 30). To generate novel synthetic stimuli with the

same time-averaged statistics, the synthesis process was run with different Gaussian noise seeds with the same target statistics.

3.4.1 Experiment 1 - Graduated texture statistics

fMRI. Subjects. Data were acquired from 9 self-reported normal hearing subjects (2 female, mean age = 28.9, std = 8.2). Two subjects were authors. The subjects gave their consent to participate in the experiments, which was approved by the Danish Science-Ethics Committee (Den Nationale Videnskabsetiske Komités). The subjects participated in a 3.5 hour visit, which included 2 blocks of 1.5 hours and 1 hour of scanning with a break between the blocks. The first block contained a data collection for experiment 1. The second block was used to obtain a high-resolution anatomical scan and auditory frequency tonotopic maps.

Stimuli. We generated textures with graduated statistics divided into five groups: (1) cochlear mean, (2) cochlear marginal, (3) cochlear correlations, (4) modulation power, and (5) modulation correlation. The stimuli were generated by imposing a subset of the texture statistics during the synthesis process. The texture groups were added cumulatively. The first step included only the cochlear mean (power), which resulted in a texture which was spectrally matched to the target. The last step included all texture statistics. From seven real world texture recordings (texture families), five 5s synthetic exemplars were generated.

Protocol. Each scanning run related to a given texture family. The synthetic texture exemplars were presented in 4-s segments during the quite period of a 6-s sparse scanning block. Two sequential blocks presented textures from the same statistic group, followed by three blocks of silence. The statistic group order was randomized, but controlled to be presented within 120 seconds of each other during the run. Each run consisted of 52 blocks, including 2 initial dummy scans to minimize B0 effects. Subjects performed 3 runs of each texture family. The textures were played from a PC running a custom python script that was triggered by the MR-scanner via an RME-Babyface USB sound card. The textures were present to the subjects in the scanner with a set of MR-Confon circumaural passive attenuating headphones. Auditory frequency tonotopic maps were also measured during the scanning session for 6 subjects. Four runs were conducted with a variable tone frequency sweep - 2-up and 2-down sweeps - (Humphries et al., 2010). The sweep stimuli were presented with the same setup as above, but presented using Sensimetric S14 insert earphones.

MRI Acquisition. MRI data were acquired using a Philips Achieva (3.2.3) 3T scanner using a 32 channel transmit/receive head coil (Philips SENSE Head). Functional scans were acquired using a

sparse spin-echo sequence to measure the blood-oxygen level dependent changes (3mm isotropic voxels, 32 slices (64x64 voxel grid), TR = 6s, TA 1.7s, TE = 30ms). The center of the field of view was aligned with the lateral sulcus, through Heschl?s gyrus, capturing significant portions of subcortex.

Preprocessing. Cortical surface meshes were created from the high-resolution anatomical scans using Freesurfer software (Dale et al., 1999). Functional data preprocessing was accomplished using a combination of Freesurfer, FSL and custom Matlab scripts. The data was motion corrected (per run) using mcFLIRT (FSL). Low frequency voxel response drift was compensated for using a 2nd order Savitzky-Golay filter with a 114 second window (per voxel, per run). The mean response for each voxel was subtracted and scaled to have unit variance. Registration of functional data to the high-resolution anatomical scan was achieved using boundary based registration (BBRegister). The functional data was spatially smoothed with a kernel of 3mm-FWHM and registered to the volumetric, left and right cortical surface meshes using Freesurfer.

Analysis. We first identified the sound-selective voxels using a standard general linear model (GLM) in FS-FAST ($p < 0.05$, FDR). A second repressor modelled each texture statistic class from 1 (cochlear mean) to 5 (modulation correlations). The significance levels (p-value) therefore represent voxels that increase with a parametric modulation analysis for steps [1,2,3,4,5]. The results are shown as a group average for auditory cortex (Figure 3B) and inferior colliculus (Figure 3C) and in native space for auditory cortex (Figure 3F). The texture statistics classes were also modelled as five individual regressors in the listeners? native space. The BOLD-signal response magnitude was computed for the individual subjects and then averaged across listeners (Figure 3D&E). The magnitude response for the 3 common textures (sparrows, insect warm, campfire) were computed in a similar manner.

3.4.2 Psychophysics (Identification Task)

Subjects. Twelve self-reported normal hearing subjects (7 female, mean age = 24.3, std = 3.5) performed the listening experiment. The subjects gave their consent to participate in the experiments, which was approved by the Danish Science-Ethics Committee (Den Nationale Videnskabsetiske Komité)?.

Procedure. The listeners performed an identification task by listening to synthetic textures over headphones and making a selection from a list of five label descriptors. The experiment included 6 variations of 39 textures: textures from the five statistics groups used in the fMRI task along with the original real-world texture recording used as the input to the auditory texture model to generate the synthetic textures. The listeners were presented with a single 4-s texture from the 6 variants and

provided with a list of 5 text label descriptors. The sounds were divided into 5 categories (animal, environment, mechanical, people, and water). The correct texture text label along with 4 alternate choices from different sound texture categories were presented. There were 234 trials in total (39 textures and 6 variants) which were performed in a single session.

The listeners performed the experiment in a single-walled sound treated booth. The sounds were presented using Sennheiser HD-650 headphones via an RME-UFX sound card using a custom experiment designed using PsychToolbox/Matlab (PC). The sound textures were presented at a level of 70dB SPL.

Analysis. A F-test and paired t-test was performed on the proportion correct.

3.4.3 Experiment 2 - Texture morphs

fMRI. Subjects. Data were acquired from 2 self-reported normal hearing subjects (2 male, mean age = 25.0, std = 3.0). The subjects gave their consent to participate in the experiments, which was approved by the Danish National Research Ethics Committee (Den Nationale Videnskabsetiske Komités). The subjects participated in a 3 hour visit, which included 2 blocks of 1 hour scanning, with a break between the blocks. Both blocks were used for functional data collection of texture morph stimuli.

Stimuli. We generated textures that morphed from spectrally matched noise $\overrightarrow{s}_{\text{noise}}$ (cochlear mean) to synthetic textures generated from a parameter vector $\overrightarrow{s}_{\text{nat}}$ (texture statistics) measured from real-world texture recordings. Synthetic textures were then generated by introducing the higher-order texture statistics $\overrightarrow{s}_{\text{interp}} = \Delta\overrightarrow{s}_{\text{nat}} + (1-\Delta)\overrightarrow{s}_{\text{noise}}$, which consisted of 7 logarithmically spaced steps. We generated exemplars for 6 texture families for each interpolated parameter vector (84 stimuli).

Protocol. Each scanning run related to a 1 of 6 texture families. The synthetic texture exemplars were presented in 4-s segments during the quite period of a 7-s sparse scanning block. Two sequential blocks presented textures from the same interpolated parameter vector (different exemplars), followed by two blocks of spectrally matched noise. Each run consisted of 50 blocks, including 2 initial dummy scans to minimize B0 effects. Subjects performed 3 runs of each texture family, where the order was randomized across runs. Presentation of the textures was the same as the first experiment.

MRI Acquisition & Preprocessing. The same scanner and pipeline from the first experiment was used.

Analysis. The same analysis was performed as was described in Experiment 1.

3.4.4 Psychophysics (Discrimination Task)

Subjects. Six self-reported normal hearing subjects (2 female, mean age = 25.2, std = 3.3) performed the listening experiment. The subjects gave their consent to participate in the experiments, which was approved by the Danish Science-Ethics Committee (Den Nationale Videnskabsetiske Komité)?.

Procedure. The listeners performed a three interval, two alternative forced choice discrimination task. The listeners? were required to identify the texture that was different, positioned in either the first or last interval. The experiment included 12 texture families and 6 texture morphs per family as used in the fMRI experiment. The listeners were presented with a three intervals of 2-s textures, where either the first or last interval was a texture morph , and the other two textures were spectrally matched noise . There were 198 trials in total (9 textures, 11 morphs, 2 repetitions) which were performed in a single session. The experiments occurred in the same location with the same equipment as the psychophysics of experiment 1.

Analysis. TBD.

4

Amplitude Modulation Sensitivity in Sound Texture Perception [a]

Abstract

Sound textures, such as crackling fire or chirping crickets, represent a broad class of sounds defined by their homogeneous temporal structure. It has been suggested that the perception of texture is mediated by time-averaged summary statistics measured from early auditory representations. In this study, we investigated the perception of sound textures that contain rhythmic structure, specifically second-order amplitude modulations that arise from "beating" in the envelope-frequency domain. We developed an auditory texture model that utilizes a cascade of modulation filterbanks that capture the structure of simple rhythmic patterns. The model was examined in a series of psychophysical listening experiments using synthetic sound textures - stimuli generated using time-averaged statistics measured from real-world textures. In a texture identification task, our results indicated that second-order amplitude modulation sensitivity enhanced recognition. Next, we examined the contribution of the second-order modulation analysis in a preference task, where the proposed auditory texture model was preferred over a range of model deviants that lacked second-order modulation rate sensitivity. Lastly, the discriminability of textures that included second-order amplitude modulations appeared to be perceived using a time-averaging process. Overall, our results demonstrate that second-order modulation statistics vary across textures and that the inclusion of a second-order modulation analysis generates improvements in the perceived quality of synthetic textures compared to the first-order modulation analysis considered in previous approaches.

[a] This chapter is based on McWalter (in Review).

4.1 Introduction

Many of the sounds encountered in everyday life contain unique temporal sequences. The uniqueness of an acoustic signal and its subsequent neural representation allow a listener to identify and differentiate sounds. The representation of such an acoustic signal at higher stages of the auditory system inevitably depends on peripheral auditory processes, particularly the spectral and temporal tuning properties of the early auditory system. The fundamental understanding of these tuning properties has led to several models which decompose the incoming sound into structures that may facilitate their recognition (Chi et al., 2005; Dau et al., 1997; Irino and Patterson, 2006; Patterson et al., 1987). The structure of the decomposed sound at different stages of the model can inform as to which features may be relevant for recognition.

A particular class of sounds, referred to as sound textures, has been used to examine representations in the auditory system (McDermott and Simoncelli, 2011). Sound textures can be characterized by their temporal homogeneity and it has been suggested that their perception relies on a relatively compact set of time-averaged summary statistics measured from early auditory representations (McDermott et al., 2013). Sound textures are ubiquitous in the natural world and although the summary statistic representation can be expressed in a relatively compact form, textures span a broad perceptual range (e.g. rain, fire, ocean waves, insect swarms etc...).

One aspect of sound textures that makes them unique is that their statistical properties remain relatively constant over time. This temporal homogeneity nevertheless depends on the duration of the analysis window (McDermott et al., 2013). As the analysis window increases the statistical properties of the texture stabilize. In addition, it has been shown that the ability of listeners to discriminate texture exemplars decreases with stimulus duration. This is in contrast to results in a corresponding discrimination task obtained with auditory stimuli that varied over time (e.g. single-talker speech), where performance increases with stimulus duration. Even though sound textures are temporally homogeneous and do not contain complex temporal features or unique acoustic events, they are readily identifiable and differentiable (McDermott and Simoncelli, 2011).

The texture synthesis system of McDermott and Simoncelli (2011) described spectral and temporal tuning properties of the early auditory system that are crucial for texture perception. Synthetic textures were generated by measuring time-averaged texture statistics at the output of several processing stages of a biologically plausible auditory model, which were subsequently used to shape a Gaussian noise seed to have matching statistics. The auditory texture model included frequency-

selective auditory filters and amplitude-modulation selective filters derived from both psychophysical and physiological data (Dau et al., 1997). The authors demonstrated that when the auditory model deviated in its biological plausibility, such as applying linearly spaced auditory filters, the perceptual quality of the texture exemplars was reduced. In addition, McDermott and Simoncelli (2011) identified which texture statistics were necessary for correct identification, revealing subsets of statistics that were requisite for different sound textures. Collectively, the results suggested that textures synthesized with the complete set of texture statistics and a biologically plausible auditory model were preferred over all other identified synthesis system configurations.

The sound synthesis system proposed by McDermott and Simoncelli (2011) generated compelling exemplars for a broad range of sounds, but there were also sounds for which the auditory texture model failed to capture some of the perceptually significant features. The failures were identified by means of a realism rating performed by human listeners', who compared synthetic textures to corresponding original real-world texture recordings. The shortcomings were attributed to either the processing structure or the statistics measured from the auditory texture model. One such texture group were sounds that contained rhythmic structure (McDermott and Simoncelli, 2011).

In the present study, the auditory texture model of McDermott and Simoncelli (2011) was extended to include sensitivity to second-order amplitude modulations. Second-order amplitude modulations arise from beating in the envelope-frequency domain and, at slow modulation rates, have the perceptual quality of simple rhythms (Lorenzi et al., 2001a). This type of amplitude modulation has been shown to be salient in numerous behavioral experiments (Ewert et al., 2002; Füllgrabe et al., 2005; Lorenzi et al., 2001a; Lorenzi et al., 2001b; Verhey et al., 2003). The perception of second-order amplitude modulation has also been modelled by applying non-linear processing and modulation-selective filtering to a signal's envelope (Ewert et al., 2002). While the role of second-order amplitude modulation in sound perception has been investigated using artificial stimuli, their significance in natural sound perception has yet to be examined.

We undertook an analysis-via-synthesis approach to examine the role of second-order amplitude modulations in sound texture perception (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). This entailed generating synthetic sounds from time-averaged statistics measured at different stages of our auditory texture model. The synthetic sounds were controlled by two main factors: the structure of the auditory texture model and the statistics passed to the texture synthesis system. We first ensured that the sound texture synthesis system was able to capture the temporal structure of a second-order amplitude modulated signal. Subsequently, we examined the significance of the

auditory texture model in a series of behavioral texture identification and preference tasks. Lastly, we attempted to quantify the role of time-averaging in the perception of second-order amplitude modulation stimuli.

4.2 Methods

4.2.1 Auditory Texture Model

The auditory texture model is based on a cascaded filterbank structure that separates the signal into frequency subbands (Figure 4.1). The first stage of the model uses 34 gammatone filters, equally spaced on the equivalent rectangular bandwidth (ERB)N scale from 50Hz to 8kHz (Glasberg and Moore, 1990):

$$g(t) = c\,t^3 e^{2\pi i f_c} e^{-2\pi\beta t},$$

where f_c is the gammatone center frequency, β is a bandwidth tuning parameter and c is a scale coefficient. Although gammatone filters only capture the basic frequency selectivity of the auditory system, more advanced and dynamic filterbank architectures, such as gammachirp filters (Irino and Patterson, 2006), did not yield any improvement in texture synthesis as observed in pilot experiments. To allow for the reconstruction of the subbands, a paraconjugate filter, $\tilde{G}(z)$, was created for each gammatone filter, $G(z)$ (Bolcskei et al., 1998):

$$\tilde{G}z = \frac{1}{G(z)}\left(G(z)G(z)^T + G^*(z)G^*(z)^T\right),$$

where $G(z)$ is the Fourier transform of $g(t)$, and $G^*(t)$ is the complex conjugate of $G(t)$. Perfect reconstruction is achieved as long as:

$$\tilde{G}zG(z) = 1.$$

To model fundamental properties of the peripheral auditory system, we applied compression and envelope extraction to the subband signals. The compression was used to model the non-linear behavior of the cochlea (e.g., Ruggero (1992a)) and was implemented as a power-law compression with an exponent value of 0.3. As all textures were presented at a sound pressure level (SPL) of 70 dB, it was deemed not necessary to include level-dependent compression. To functionally model the transduction from the cochlear to the auditory nerve, the envelopes of the compressed subbands

were extracted using the Hilbert transform and down-sampled to 400 Hz (McDermott and Simoncelli, 2011). The compressed, down-sampled envelopes roughly estimate the transduction from basilar-membrane vibrations to inner hair-cell receptor potentials.

The model then processed each cochlear channel signal by a modulation filterbank, accounting for the first-order modulation sensitivity and selectivity of the auditory system. The filterbank applied to each cochlear channel comprised of 19 filters, half-octave spaced from 0.5 to 200 Hz. This type of functional modeling is consistent with previous perceptual models of modulation sensitivity (Dau et al., 1997) and shares similarities with neurophysiological findings (Joris et al., 2004a; Malone et al., 2015; Miller et al., 2002b). The broadly tuned modulation filters have a constant $Q = 2$ and a shape defined by a Kaiser-Bessel window. Reconstruction of the modulation filterbank was achieved with the same method as the frequency selective gammatone filterbank.

The output of each modulation filter was subsequently processed by a second modulation filterbank, accounting for the sensitivity of the auditory system to second-order amplitude modulations. Each second-order modulation filterbank contained 17, half-octave spaced bands in the range from 0.25 to 64 Hz. The model was inspired by behavioral experiments and simulations revealing an auditory sensitivity to second-order modulations that is similar in nature to the sensitivity to first-order amplitude modulations (Ewert et al., 2002; Füllgrabe et al., 2005; Lorenzi et al., 2001a; Lorenzi et al., 2001b). The model processing layer proposed here has some shared attributes as the model presented in Ewert et al. (2002), but has the added benefit of being easily invertible. The second-order modulation filters have a constant and a Kaiser-Bessel window.

4.2.2 Texture Statistics

The goal of statistics selection is to find a description of sound textures that is consistent with human sensory perception (Portilla and Simoncelli, 2000). The selected statistics should be based on relatively simple operations that could plausibly occur in the neural domain. The values of the measured should also vary across textures, facilitating the recognition of sound textures by the difference in the statistical representation. Lastly, there should be a perceptual salience to the textures, such that the use of their statistics contributes to the realism of the corresponding synthetic texture.

The statistics measured from the auditory model include marginal moments and pair-wise correlations (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). The included texture statistics are similar to those described in McDermott and Simoncelli (2011). They were computed

from the envelope of the cochlea channels, including the first- and second-order modulation filters, and were measured over texture excerpts of several seconds. Examples of the statistics for three textures (insect swarm, campfire, and small stream) measured from the auditory texture model (Figure 4.1) are shown in Figure 4.2.

The envelope statistics include the mean, coefficient of variance, skewness and kurtosis, and represent the first four marginal moments, defined as:

$$\mu_n = \overline{\overrightarrow{x}_n}, \quad \frac{\sigma_n^2}{\mu_n^2} = \frac{\overline{(\overrightarrow{x}_n - \mu_n)^2}}{\mu_n^2}, \quad \eta_n = \frac{\overline{(\overrightarrow{x}_n - \mu_n)^3}}{\sigma_n^3}, \quad \kappa_n = \frac{\overline{(\overrightarrow{x}_n - \mu_n)^4}}{\sigma_n^4},$$

where n is the signal channel. Pair-wise correlations were computed as a cross-covariance with the form:

$$c_{nm} = \frac{\overline{(\overrightarrow{x}_m - \mu_m)(\overrightarrow{x}_n - \mu_n)}}{\sigma_m \sigma_n},$$

where n and m are the signal channel pairs. The final statistic captures envelope phase:

$$c_{nm} = \frac{\overline{\overrightarrow{d}^*_m \overrightarrow{a}_n}}{\sigma_m \sigma_n}, \quad d_m = \frac{\overrightarrow{a}_m}{||\overrightarrow{a}_m||}, \quad \overrightarrow{a}_m = \overrightarrow{b}_m + H(\overrightarrow{b}_m),$$

where H is the analytic signal, and d^* is the complex conjugate of d.

The marginal moments (M) describe the distribution of the individual subbands (Figure 4.2a) and capture the overall level as well as the sparsity of the signal (Field, 1987). The correlation statistics (C) capture how neighboring signals co-vary. The correlation statistics are measured between the eight neighboring cochlear channels (Figure 4.2b). There are 372 statistics measured at the cochlear stage of the auditory model (M = 128, and C = 236).

The statistics measured at the first-order modulation stage include the coefficient of variance (M1P, Figure 4.2c), the correlation measured across cochlear channels and first-order modulation channels (MC1, Figure 4.2d), and the correlation measured across modulation channels for the first-order modulations (MC2, Figure 4.2e). Because the outputs of the modulation filters have zero mean, the variance effectively reflects a measure of the modulation channel power. The variance was measured for cochlear channels that have a center frequency at least four times that of the modulation frequency (Dau et al., 1997). The modulation correlations measured across cochlear channels (MC1) reflect a cross-covariance measure. The correlation was measured for two neighboring cochlear channels. The modulation correlation measured across modulation rates (MC2) included phase

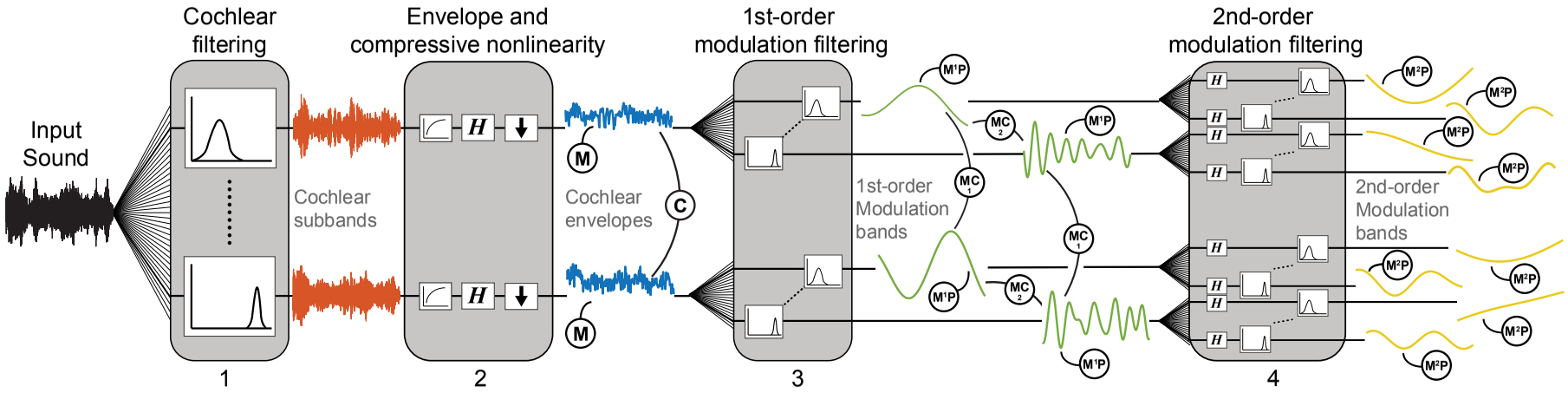


Figure 4.1: Texture analysis model. The functional auditory model captures the tuning properties of the peripheral and subcortical auditory system: (1) An auditory filterbank simulated the resonance frequencies of the cochlear, (2) a nonlinearity captures the compression of the cochlear followed by computation of the Hilbert envelope, functionally modelling the transduction from the mechanical vibrations on the basilar membrane to the receptor potentials in the hair cells, (3) a first-order modulation filterbank captures the selectivity of the auditory system to different envelope fluctuation rates, and (4) a second-order modulation filterbank captures the sensitivity of the auditory system to beating in the envelope frequency domain. Texture statistics include marginal moments of cochlear envelopes (M), 1st-order modulation power (M1P), pair-wise correlations between cochlear envelopes (C), pairwise correlations between modulation subbands (MC1), phase correlation between octave spaced modulation bands (MC2), and 2nd-order modulation power (M2P).

information and was computed for octave-spaced modulation frequencies. The number of statistics considered in the modulation domain was 1258 (M1P = 646, MC1 = 408, and MC2 = 204).

The last analysis stage is the second-order modulation envelope bands, where the modulation power was measured for each band (M2P, Figure 4.2f). The power was measured for first-order modulation rates that are at least twice that of the second-order modulation rate. The 2nd-order modulation power required the largest overall number of statistics (M2P = 3400).

4.2.3 Synthesis System

The synthesis of sound textures was accomplished by modifying a Gaussian noise seed to have statistics that match those measured from a real-world texture recording (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). The original texture recording was decomposed using our biologically motivated auditory model where the texture statistics were measured. The statistics were then passed to the synthesis algorithm which imposed the measured statistics on the decomposed Gaussian noise signal. The modified signals were reconstructed back to a single-channel waveform. A schematic of the synthesis system can be seen in Figure 4.3a.

The imposition of texture statistics on the noise input was achieved using the LF-BFGS variant of gradient descent (limited-memory Broyden-Fletcher-Goldfarb-Shanno). The noise signal was decomposed to the second-order modulation bands, where the power statistics were imposed. The bands were then reconstructed to the first-order modulation bands, and the modulation power and correlation statistics were imposed. The modulation bands were then reconstructed to the cochlear envelopes, where the marginal moments and pair-wise correlations statistics were imposed. Lastly, the cochlear envelopes were combined with the fine-structure of the noise seed and the cochlear channels were resynthesized to the single channel waveform.

The synthesis process requires many iterations in order to attain convergence for each of the texture statistics due to the reconstruction of the subbands and tiered imposition of statistics. The reconstruction of the filterbanks modified the statistics of each subband due to the overlap in frequency of neighboring filters. The reconstruction from the cochlear envelopes to the cochlear channels was also affected by the combination of the envelope and fine structure. In addition, the texture statistics were modified at 3 layers (cochlear envelopes, 1st-order modulations, and 2nd-order modulations) of the auditory model, and the modification at each level had an impact on the other two. Due to these two factors, an iterative process for imposing texture statistics was required.

The synthesis was deemed successful if the synthetic texture possessed statistics approached

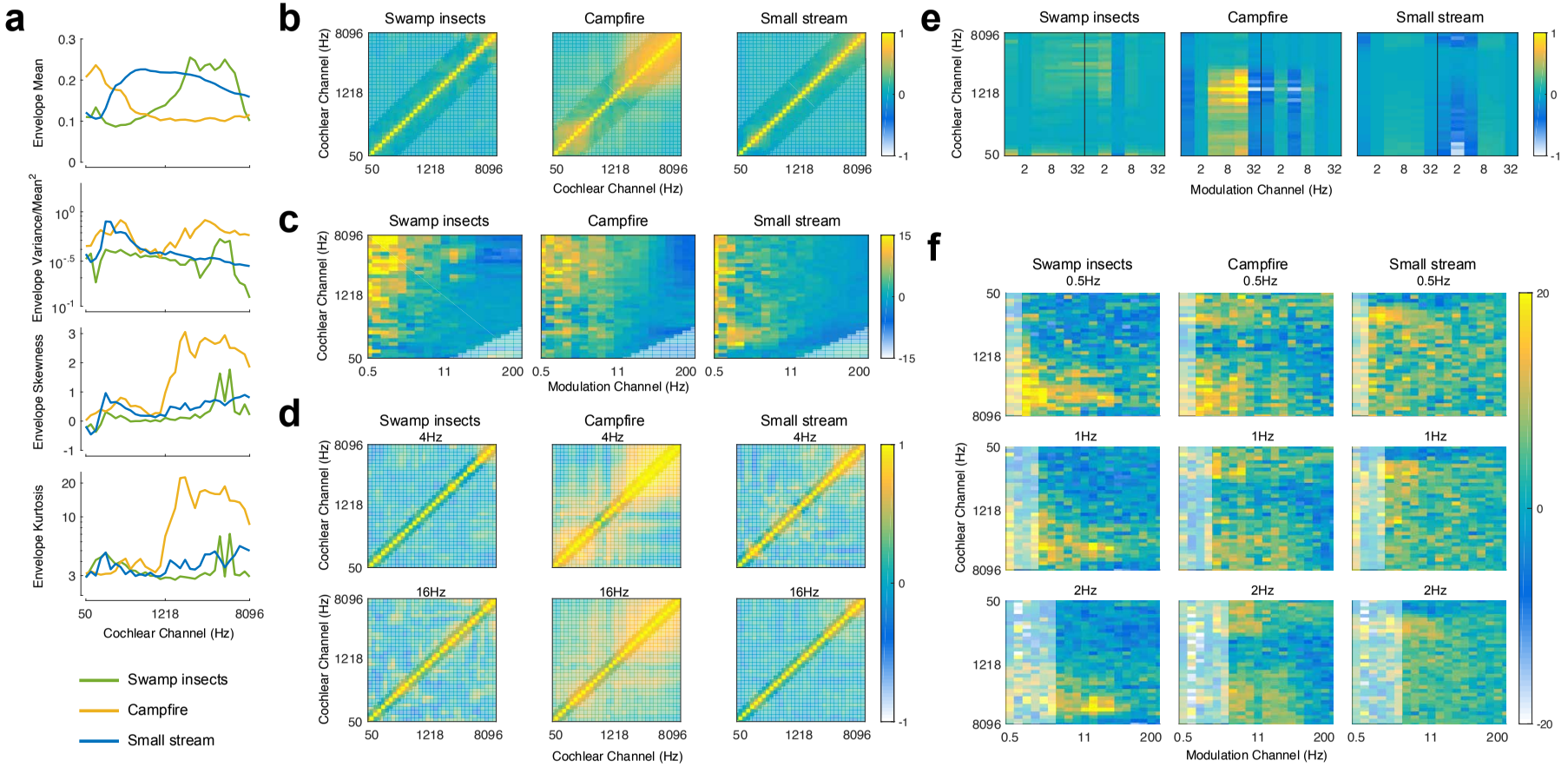


Figure 4.2: Texture Statistics. (a) Cochlear envelope marginal moments (mean, coefficient of variance, skewness, kurtosis) measured from three real-world texture recordings (Swamp insects, Campfire, Small stream). (b) Cochlear envelope pair-wise correlations measured between different cochlear channels. The label of the texture analyzed is located above the subfigure (and for all subsequent subfigures). Lightened regions here and elsewhere denote texture statistics that are not imposed during the synthesis process. (c) Modulation band power (variance). The figure is normalized by the modulation power of Gaussian noise and plotted on a log (dB) scale. (d) Modulation correlation measured for a particular rate across cochlear channels. The modulation rate is indicated above the subfigure. (e) Modulation phase correlation measured between octave-spaced modulation bands. (f) Second-order modulation band power (variance). The second-order modulation frequency is indicated above the individual subfigures for a selection of rates (0.5, 1, and 2Hz). The statistics are plotted relative to Gaussian noise on a log (dB) scale.

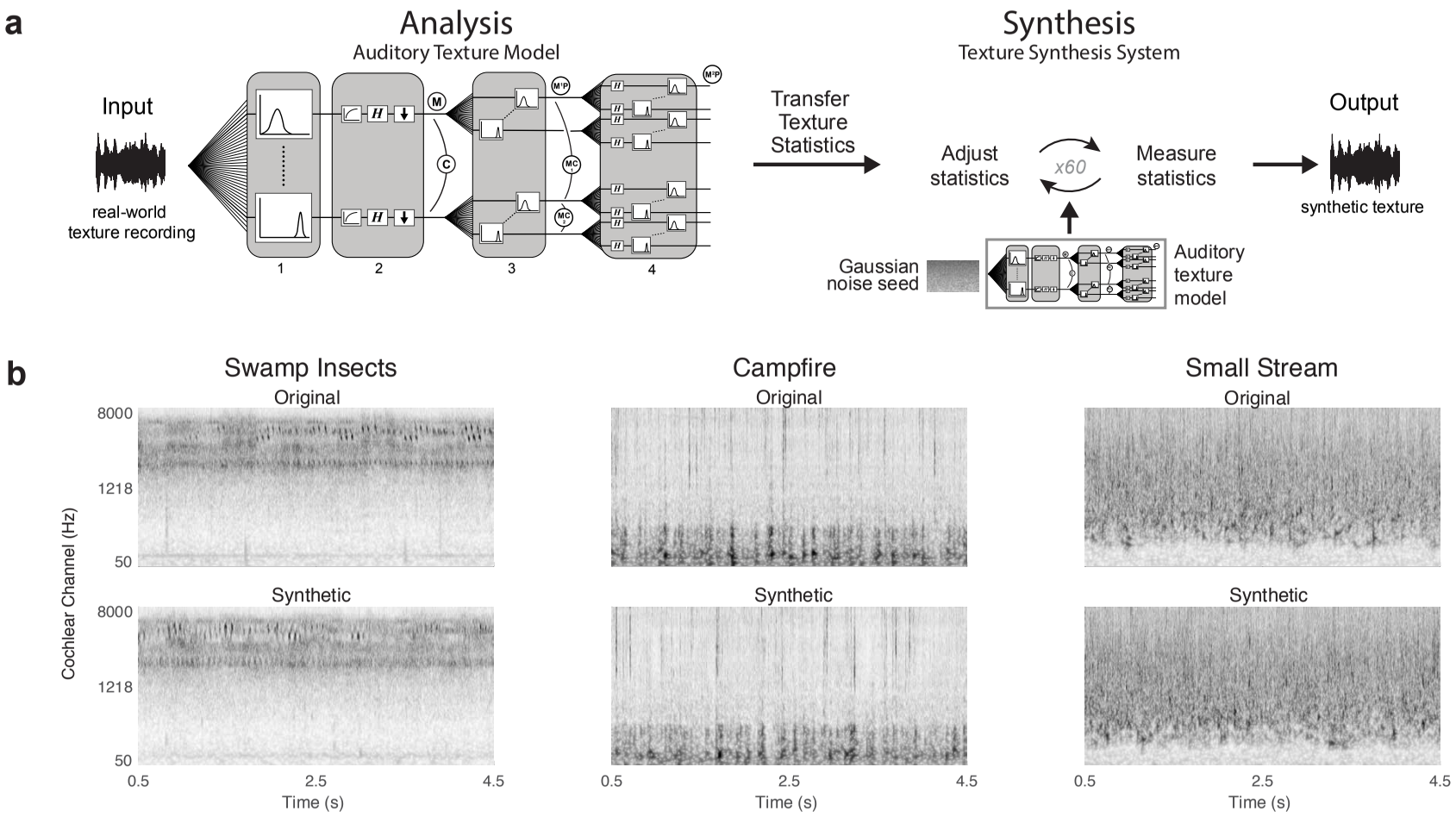


Figure 4.3: Texture synthesis system and synthetic examples. (a) Texture synthesis is accomplished by measuring statistics from a real-world texture recording at different stages of the auditory texture model. The statistics are then passed to the synthesis system that adjusts the statistics of a Gaussian noise seed to match the input statistics. The iterative process outputs a synthetic texture with the same time-averaged statistics as the real-world texture recording. (b) Original real-world texture recordings and their synthetic counterparts. The synthetic textures were generated with a complete set of texture statistics.

those measured from the original real-world texture recoding. The convergence was evaluated based on the signal-to-noise ratio (SNR) between the synthetic and original texture statistics (Portilla and Simoncelli, 2000). When the synthesis process reached an SNR of 30 dB or higher across the texture statistics, the process ended, generating a synthetic texture. The system also had a maximum synthesis iteration limit of 60. However, the convergence criterion was often met within 60 iterations. The cochleograms of the original and synthetic textures are shown in Figure 4.3b.

Texture Synthesis System Validation

The proposed auditory texture model and adjoining synthesis system were validated with a second-order amplitude modulation signal identified by (McDermott and Simoncelli, 2011)). The stimulus has a repetition frequency of 2Hz and modulators at 8Hz and 6Hz. Perceptually, the stimulus contained a long noise burst ($t = 0.1875$ or $\frac{3}{16}s$), followed by two short noise bursts ($t = 0.0625$ or $\frac{1}{16}s$) that were repeated every 500ms.

4.2.4 Psychophysical Experiments

The listeners were recruited from a university specific job posting site. The listeners completed the required consent form and were compensated with an hourly wage for their time. All experiments were approved by the Science Ethics Committee for the Capital Region of Denmark.

The listeners performed the experiment in a single-walled IAC sound isolating booth. The sounds were presented at 70 dB SPL via Sennheiser HD 650 headphones. The playback system included an RME Fireface UCX soundcard and the experiments were all created using Mathworks MATLAB and the PsychToolBox (psychtoolbox.org) software.

The synthetic textures used in experiments 1 and 2 were generated in 5-s long samples. Multiple exemplars were generated for each texture. Each exemplar was created using a different Gaussian noise seed such that no sample was identical in terms of the waveform, but had the same time-averaged texture statistics. 4-s long excerpts were taken from the middle portion of the texture samples with a tapered cosine (Tukey) window with 20-ms ramps at the onset and offset.

Experiment 1 - Texture Identification

Each trial consisted of a 4-s texture synthesized from subsets of texture statistics that were cumulatively included from the cochlear envelope mean to the 2nd-order modulation power. The listeners were required to identify the sound from a list of 5 label descriptors. The experiment consisted of 59

sound textures. The textures were divided into 5 texture groups, defined by the authors: animals, environment, mechanical, human, and water sounds. The list of 4 incorrect labels for each texture was selected from different texture groups. There were 7 conditions per texture (6 synthetic and 1 original) and 413 trials per experiment. Eleven self-reported normal-hearing listeners participated in the experiment (6 female, 23.3 mean age).

Experiment 2 - Modulation Processing Model Comparison

Each trial consisted of three intervals; the original real-world texture recording, a synthetic texture generated from the above-mentioned texture synthesis system (reference), and a synthetic texture generated from a modified version of the auditory model. The real-world texture was presented first. Textures generated from the reference system and a modified auditory model were then presented in intervals 2 and 3, where by the order of presentation was randomized. Each interval was 4s long with an inter-stimulus-interval of 400 ms. The listeners were asked to select the interval that was most similar to the real-world texture recording. The same 59 textures were used in the experiment, presented in 236 trails. Eleven self-reported-normal hearing listeners participated in the experiment (7 female, 24.2 mean age).

Synthetic textures generated from a reference auditory model and four alternate auditory models were included in the experiment. The reference model is described in Figure 4.1, including texture statistics measured from the cochlear envelope, 1st and 2nd order modulation bands. The first alternate model removed the 2nd-order modulation bands, and was in principle similar to that of McDermott and Simoncelli (2011). The second alternate model removed the 2nd-order modulation bands and replaced the half-octave spaced 1st-order modulation filterbank by an octave-spaced variant. Octave-spaced modulation selectivity has been suggested in several models of auditory perception (Dau et al., 1997; Jørgensen and Dau, 2011). The third alternate model removed the 2nd-order modulation bands and substituted the half-octave spaced modulation filterbank with a low-pass filter of 150Hz. The low-pass characteristic of amplitude modulation perception has been proposed, and here we used a model that preserves the sensitivity to modulation rates but lacks the selectivity of the filterbank model (Joris et al., 2004a; Kohlrausch et al., 2000). The fourth alternate model also removed the 2nd-order modulation bands and substituted the half-octave spaced modulation filterbank with a low-pass filter with a cutoff frequency of . The sluggishness of the auditory system to amplitude modulation perception is reflected in the heightened sensitivity to slow modulation rates (Dau et al., 1996; Viemeister, 1979).

Experiment 3 - Second-order Modulation Discrimination

Each trial consisted of three 2-s intervals. The listeners performed an odd-one-out experiment, where they were required to identify the interval (first or last) that was different from the other two. The stimulus sets described below were evaluated in separate experiment blocks. Twelve self-reported-normal hearing listeners participated in the experiment (3 female, 23.0 mean age).

The first stimulus set was generated from second-order amplitude modulated white noise using the following equation:

$$s(t) = \big(1 + \big(0.5 + \sin(2\pi f_{m1} t + \phi)\big)\sin(2\pi f_{m2} t + \phi)\big)\, n(t)$$

where f_{m1} is the first modulator, t is time, ϕ is the phase of the first modulator, f_{m2} is the second modulator, and n is the noise. f_{m1} had a modulation frequency of 2, 4, 8, 16, 32 or 64. f_{m2} had a modulation rate of $f_{m2} = [0.1, 0.13, 0.17, 0.22, 0.28, 0.36, 0.46, 0.60, 0.77, \text{or} 1.00]$. f_{m2} was randomized for each trial. The exemplars were 5 seconds in duration. Two intervals were sampled from the first 2 seconds, and the "odd" interval was sampled from the last 2 seconds. Each condition was repeated 4 times, for a total of 240 trials.

The next stimulus set used second-order amplitude modulated white noise generated from a combination of f_{m1} and f_{m2} pairs, creating a complex amplitude modulated signal. Each stimulus was created using the six f_{m1} frequencies, each paired with a corresponding f_{m2} frequency that was randomly selected from the list of 10, modulating the same white noise seed. The six second-order modulated signals were then summed to created one stimulus. The exemplars were 5 s in duration. Two intervals were sampled from the first 2 seconds, and the "odd" interval was sampled from the last 2 seconds. There were 48 stimuli presented, one per trial.

The final stimulus set was composed of sound textures generated with the complete set of texture statistics, including second-order amplitude modulations power. The 59 textures used in experiments 1 and 2 were used in this experiment. The exemplars were 5 s in duration. Two intervals were sampled from the first 2 seconds, and the "odd" interval was sampled from the last 2 seconds. There were 59 trials in total.

4.3 Results

The auditory model proposed in the present study includes frequency-selective filtering (in the audio-frequency domain) as well as a cascade of amplitude modulation filterbanks to capture time-

averaged amplitude modulations and simple rhythmic structure. The model was combined with a sound synthesis system to generate synthetic textures that were then examined in several behavioral listening experiments. The results show three main findings: (1) the model captures simple rhythmic structure by way of second-order amplitude modulation analysis, (2) the inclusion of second-order amplitude modulation analysis contributes to the recognition of the synthetic textures, and (3) second-order amplitude modulations in textures may be perceived using time-average statistics measured from early auditory representations.

4.3.1 Synthesis Verification for 2nd-order Modulations

Although the second-order texture statistics varied across textures, it was unclear how the synthesis process would perform in creating new sound examples. To test this, we used a second-order amplitude modulation signal identified by McDermott and Simoncelli (2011) that has a salient rhythmic structure. Figure 4.4a shows the original sound (top), a synthetic version with second-order modulation analysis (middle) and a synthetic version without second-order analysis (bottom). The synthetic sound generated from texture statistics that include second-order amplitude modulation analysis captures the rhythmic pattern of the original sound, whereas the version without second-order analysis fails to capture the rhythmic structure even though the duration of the noise bursts is comparable to that in the original sound. The successful synthesis of the rhythmic sound suggests that the cascaded modulation filterbank analysis can capture rhythmic structure.

The second-order amplitude modulation statistics for the example rhythmic sound are shown in Figure 4.4b. The majority of the modulation power can be found in the 2Hz second-order modulation channel (bottom left panel) across several first-order modulation rates. For a relatively simple rhythmic sound, there is considerable modulation power across frequencies. This is primarily due to amplitude modulation interactions between the modulation frequencies and the broadband (Gaussian) noise carrier. If a second-order amplitude modulated tone was used instead of the noise with its intrinsic modulations, the modulation power would be relegated entirely to the 2-Hz band.

4.3.2 Texture Perception: Identification and Preference

Our first behavioral experiment investigated the ability of listeners to identify sound textures generated from subsets of statistics. Listeners were presented with a 4s texture and asked to identify the sound from a list of 5 text label descriptors. The textures synthesized with the cochlear channel power resulted in low performance, but the performance increased with the inclusion of higher-order

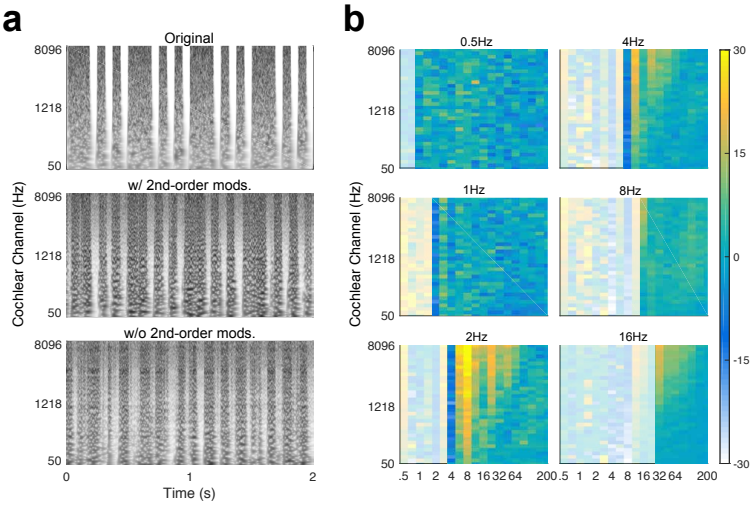


Figure 4.4: Verification of second-order texture synthesis. (a) Spectrogram of example rhythmic (second-order modulated) noise bursts with 500ms repetition pattern. The upper panel shows the original sound, the middle panel shows the synthetic version with second-order modulation texture statistics (w/ 2nd-order mods.) and the bottom panel shows the synthetic version without second-order modulation texture statistics (w/o 2nd-order mods.). (b) Second-order modulation power statistics. The 500ms period is reflected in the majority of power held within the 2Hz 2nd-order modulation band (lower-left panel).

texture statistics and approached that of the original real-world texture recording when second-order amplitude modulation statistics were used (Figure 4.5a; $F[6,49] = 123.51$, $p < 0.0001$). The results suggest that listeners benefited from the addition of second-order amplitude modulation analysis to the auditory texture model.

Next, we were interested in how synthetic textures generated with alternate amplitude modulation processing models compared to our auditory texture models. To investigate this, we generated textures from four models that included only the first-order amplitude modulation analysis (Figure 4.5b). The results show that our auditory texture model, with second-order amplitude modulation analysis, was preferred over all other model variants (Figure 4.5c; $p < 0.01$ relative to chance). Notably, the inclusion of second-order modulation analysis yielded a modest yet significant improvement over the half-octave spaced first-order modulation, which is comparable to that developed by McDermott and Simoncelli (2011).

The results from the preference experiment revealed which textures benefited most from second-order amplitude modulation analysis. Figure 4.6a shows a list of the top 8 preferred textures measured between the half-octave spaced filterbank and our auditory texture model. The list includes a broad range of sounds, from mechanical/machine noises to animal/insect sounds. The least preferred textures are also shown, which yield sounds which may not depend greatly on amplitude modulation

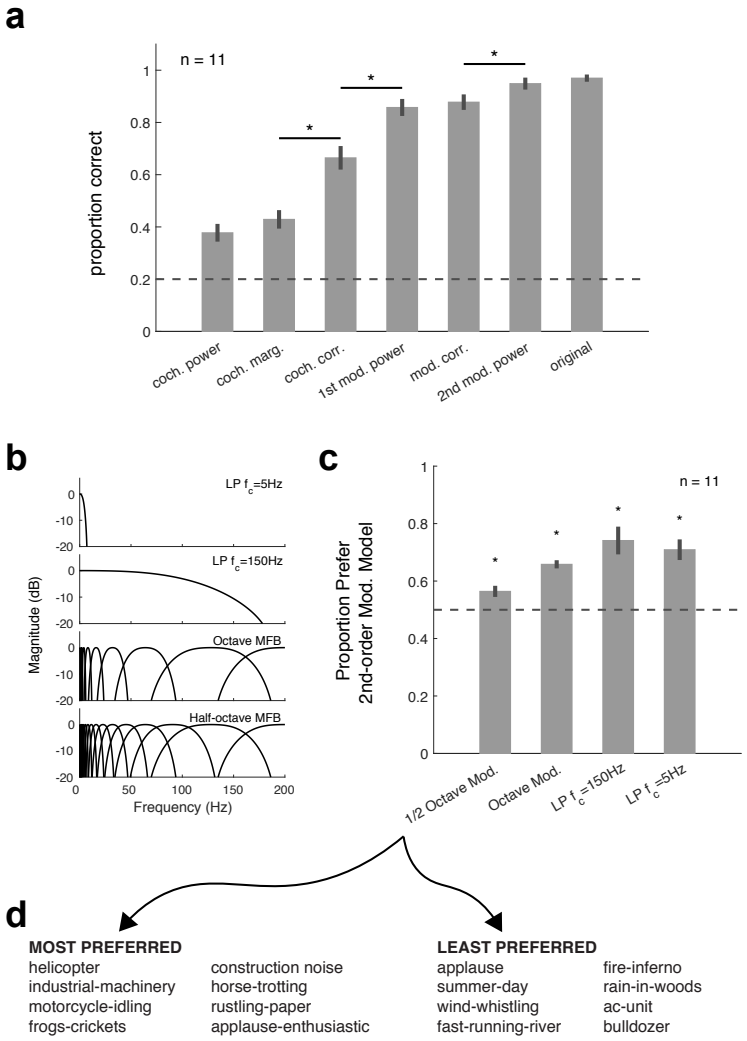


Figure 4.5: Synthetic texture identification and preference tasks. (a) Identification of sound textures improves with the inclusion of more statistics. Asterisks denote significant differences between conditions, $p < 0.01$ (paired t tests, corrected for multiple comparisons). Error bars here and elsewhere show standard error. Dashed lines here and elsewhere show chance performance. (b) Modulation filter(bank) structure used in the listening experiments. For low-pass (LP) conditions, only the statistics of the signal in the passband were modified. (c) Sounds synthesized with the 2nd-order modulation statistics were preferred over all other auditory texture models. Asterisk denotes significance from chance ($p < 0.01$). (d) Eight most preferred (left) and least preferred (right) textures from experiment 2 (relative to 1/2 octave modulation filterbank model).

texture statistics (i.e. cochlear envelope marginal moments and pair-wise correlations). Two example textures, helicopter and frogs-crickets, are shown in Figure 4.6b. For each texture, the left panel shows the 2nd-order modulation texture statistics for selected bands and the right panel shows the original (top) and synthetic (bottom) texture cochleogram. Notably, the second-order amplitude modulation power differs between the two textures, suggesting that the additional analysis contributes to sound texture recognition.

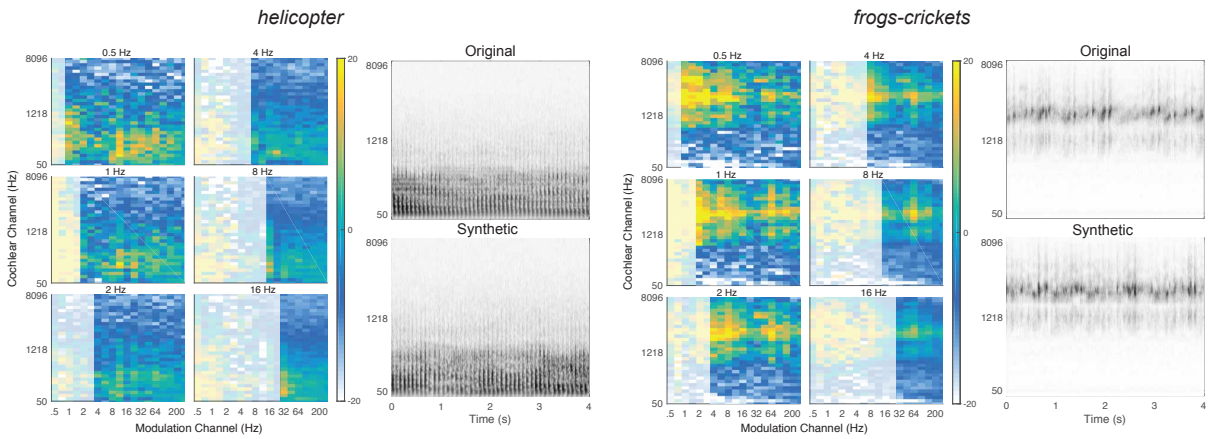


Figure 4.6: Textures that benefit from second-order modulation statistics. Two example texture from the preferred list: Helicopter (left) and frogs-crickets (right). The left panel shows the second-order modulation statistics for six selected bands. The right panel shows the spectrogram of the original texture (top) and the synthetic texture (bottom).

4.3.3 Second-order modulation discrimination

To examine if second-order amplitude modulations are processed by the auditory system similarly to textures, i.e., integrated over modest time windows of a few seconds, or if the auditory system has the temporal acuity to identify and discriminate second-order modulations with higher precision, a set of discrimination experiments was performed where synthetic sound textures were compared to artificial control stimuli generated from amplitude modulated Gaussian noise. The experiments covered three stimulus groups: rate-specific second-order amplitude modulations, complex second-order amplitude modulation noise from a set of modulation rates, and synthetic sound textures generated using second-order amplitude modulation statistics.

The first experiment included second-order amplitude modulations of increasing rate from 2Hz to 64Hz. The results showed that, at low rates, the listeners have the ability to discriminate modulated noise exemplars (Figure 4.7 - left panel). The performance decreased with increasing modulation rate and approached chance level for modulation rates above 16 Hz. For these control stimuli, the results suggest that the auditory system may have access the modulation phase for rates 16 Hz and below.

The discriminability of the complex modulated Gaussian noise and/from the synthetic texture was poor (Figure 4.7 - right panel) compared to the low modulation rates considered in the previous experiment. This suggests that, for texture sounds, access to the modulation phase is limited in the auditory system. Isolating the top eight most preferred textures from Experiment 2 revealed comparable performance to the complete set of textures. The performance observed for sound

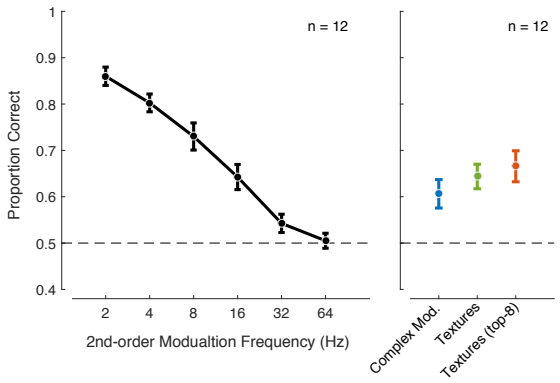


Figure 4.7: Second-order amplitude modulation and texture exemplar discrimination. The black symbols show the response to second-order amplitude modulated Gaussian noise exemplar discrimination as a function of modulation rate. Error bars indicate the standard error. The blue symbol indicates exemplar discrimination performance for complex second-order amplitude modulated Gaussian noise. The green symbol indicates exemplar discrimination performance for synthetic sound textures that include all indicated texture statistics (including second-order amplitude modulation statistics). The red symbol indicates exemplar discrimination performance for top-8 synthetic (Experiment 2) sound textures that include all indicated texture statistics.

textures in a similar odd-one-out discrimination task was comparable to that reported in McDermott et al. (2013) for an interval duration of about 2 seconds. Collectively, the results suggest that textures, including those that benefit from second-order modulation analysis, may be perceived using time-average statistics, whereas the auditory system appears to retain more temporal detail for our second-order modulation control stimuli for rates below 16Hz.

4.4 Discussion

The perception of sound texture can be characterized by a set of time-averaged statistics measured from early auditory representations. We extended the auditory texture model of McDermott and Simoncelli (2011) to account for simple rhythmic structures in sound textures via a cascade of amplitude modulation filterbanks. The auditory texture model was coupled with a sound synthesis system to generate texture exemplars from the statistics measured at different stages of the model. The synthetic stimuli were first used in a texture identification experiment, where the listeners' ability to recognize a texture improved with the inclusion of the subgroups of statistics. We found that the performance obtained using the second-order amplitude modulation analysis approached that of the original real-world texture recordings and was higher than the performance obtained using only a first-order amplitude modulation analysis (Exp. 1). We also generated synthetic textures from alternate auditory models of amplitude modulation sensitivity. The synthetic textures were used

in a preference task, where listeners' preferred sounds synthesized using second-order amplitude modulation over all other model variants (Exp. 2). Lastly, we performed an experiment focusing on second-order amplitude modulation perception in a discrimination task. The listeners? ability to discriminate second-order modulation sound exemplars decreased with increasing modulation rate, and complex second-order modulated Gaussian noise and synthetic textures appear to be perceived using a time-averaging mechanism (Exp. 3).

4.4.1 Amplitude modulations in texture perception

The auditory texture model described by McDermott and Simoncelli (2011) included a biologically plausible first-order modulation filterbank operating on individual cochlear channel envelope. The textures synthesized with this model produced many compelling textures, including sounds generated from machinery (e.g., helicopter, printing press) with relatively uniform short-time repetitions as well as environmental sounds (e.g. wind, ocean waves) with variable slow modulations. Our texture model built upon this work and provided further evidence for the importance of modulation selectivity in sound texture perception. For first-order modulation analysis, the results from the preference task (Experiment 2) demonstrated that using half-octave spaced modulation filterbank yields the best performance out of the model variants. The model has a slightly higher selectivity than has that reported in earlier models (Dau et al., 1997). One reason may be that the selectivity of the auditory system for natural sounds, such as textures, may be slightly different than that for artificial stimuli used to identify the auditory systems' modulation tuning curves and selectivity. Another possible explanation is that natural sounds do not conform to octave spaced modulation frequencies, and if the modulation power in a natural sound has a maximum between two modulation bands with fixed center frequencies, the synthetic sounds vary to a greater degree from the original real-world recording.

The results from the preference experiment also identified which textures were most improved (preferred) by the inclusion of the second-order modulation analysis. In some cases, these were sounds that also had strong first-order amplitude modulations. As texture statistics are measured from time-averages spanning several seconds, the measured statistics may not be entirely reflective of some signal variation. In addition, the synthesis of textures operates on the same time-averages; therefore, the signal might adhere to the measured statistic but how it varies over time may differ from exemplar to exemplar. One plausible explanation is that a second-order modulation analysis captures the variability of first-order modulations over time. For example, mechanical sounds

(helicopter or industrial machinery) with relatively constant modulation rate and amplitude result in low second-order modulation power. Therefore, second-order modulation sensitivity appears to be beneficial beyond capturing simple rhythmic structure, to mediate first-order modulation amplitude across time.

4.4.2 Model architecture and statistics

There might be several auditory model architectures that can successfully capture rhythmic structure in sound textures. Our proposed model, using a cascade of modulation filterbanks, seems to provide a compelling approach, as it is relatively intuitive and straight forward to implement in the already established texture analysis-synthesis framework. Another option, however, would be the "venelope" model proposed by Ewert et al. (2002) which used a side-chain analysis to measure the second-order amplitude modulations. In this model, the second-order modulations are extracted from the cochlear envelope and analyzed using a single modulation filterbank. The "venelope" model is more efficient than our cascaded model and there is some evidence to suggest that second-order modulation are processed in the auditory system using the same mechanism as the first-order modulation (Verhey et al., 2003). However, the cascaded modulation filterbank model considered in this study can capture simple rhythmic structure and provided an easier means to reconstruct the filters and thus synthesize textures.

Our approach to modelling of the auditory system, based on audio-frequency and amplitude-modulation-frequency selective filtering, is consisted with biological evidence from the mammalian auditory system (Joris et al., 2004a; Rodríguez et al., 2010; Ruggero, 1992a). This is found in the auditory-inspired filter structure for both cochlear channels and modulation-selective channels, which culminated in a cascade of filterbanks with intermediate envelope extraction using the Hilbert transform. A similar hierarchical processing architecture has also been well defined by Mallat and colleagues as scattering moments (Bruna and Mallat, 2013; Mallat, 2012). The scattering moments have been shown to capture a wide range of structure in natural stimuli (Andén and Mallat, 2014; Andén and Mallat, 2011; Andén and Mallat, 2012), in addition to being used for sound texture synthesis (Bruna and Mallat, 2013).

A consequence of the cascaded filterbank model proposed here is that the number of statistics required to capture the auditory features increases with each layer. This is particularly the case for the second-order modulation analysis, where we measure 3400 parameters, which increases the number of texture statistics by a factor of 3 as compared to the model of McDermott and Simoncelli

(2011). It may be possible to optimize the number of parameters identifying which modulation rates are most significant for texture perception. Alternatively, using a different model, such as the "venelope" model of Ewert et al. (2002), could reduce the number of parameters needed to capture the second-order amplitude modulation.

4.4.3 Temporal regularity in texture perception

Sounds textures have been defined as the superposition of many similar acoustic events, therefore it was not obvious a priori that sounds with temporal regularities would be perceived in the same way - as time-averages of sensory measurements. Temporal patterns are important for sound perception, and their contribution has been investigated in terms of auditory streaming (Andreou et al., 2011; Bendixen et al., 2010). In addition, sensitivity to temporal regularities in the auditory system has also been shown in complex listening environments (Barascud et al., 2016). Our results show that second-order modulation statistics vary across textures, and the inclusion of this second modulation analysis generated modest improvements in the perceived quality of the synthetic textures. Textures generated with second-order amplitude modulation analysis seemed to result in similar discriminability, suggesting that the features captured by the cascaded modulation filterbank model may be perceived via a similar time-averaging mechanism that has been proposed for more noise-like textures.

4.4.4 Relationship to visual texture perception

One of the interesting ideas about texture perception is that of a unified representation across sensory modalities. Textures have been investigated in the visual system (Freeman and Simoncelli, 2011; Julesz, 1962; Portilla and Simoncelli, 2000), the somatosensory system (Connor and Johnson, 1992) and the auditory system (McDermott and Simoncelli, 2011; Saint-Arnaud and Popat, 1995). Of particular relevance to our work is how the sound texture synthesis system proposed by McDermott and Simoncelli (2011) is comparable in processing structure and analysis to that presented by Portilla and Simoncelli (2000) for visual textures. In both models, the input signal is processed by layers of linear filter and envelope extraction, while the texture analysis statistics, which are primarily composed of marginal moments and pair-wise correlations, are also similar between the two models. Our model of cascaded filterbanks also overlaps with other models of the image texture perception (Wang et al., 2012). It therefore seems valuable to look across sensory modalities for shared perceptual spaces (Zaidi et al., 2013).

Our investigation of second-order modulation analysis in sound texture perception may also be relatable to spatial texture patterns, or maximally regular textures, in the visual system. P. J. Kohler, et al. (Kohler et al., 2016) showed a neural sensitivity to image texture patterns that repeat in space. Our work is also indicative of sound texture pattern sensitivity in time. Previous work in both sound and image texture perception has also made the comparison of perceptual pooling over time and space, respectively (Balas et al., 2009; Freeman and Simoncelli, 2011; McDermott et al., 2013). Conceptually, the apparent texture time-averaging in audition draws compelling parallels to the spatial averaging observed in visual texture perception.

4.4.5 Perspectives and Implications

In this study, we investigated the significance of second-order amplitude modulations in natural sound texture perception. The generation of synthetic sound textures using a cascade of modulation filterbanks appears to contribute positively to the perception of texture. We also observed that the auditory system is sensitive to specific rates of second-order modulations, showing heightened acuity to isolated modulations for rates below 16 Hz. Future experiments would be useful to understand the role of temporal regularity in texture at different modulations rates and spectral frequencies. In addition, such stimuli could be useful to understand the perception of texture in complex auditory scenes, such as the perceptual segregation of speech in the presence of different types of background textures.

5

Statistical Representation of Sound Textures in the Impaired Auditory System[a]

Abstract

Many challenges exist when it comes to understanding and compensating for hearing impairment. Traditional methods, such as pure tone audiometry and speech intelligibility tests, offer insight into the deficiencies of a hearing-impaired listener, but can only partially reveal the mechanisms that underlie the hearing loss. An alternative approach is to investigate the statistical representation of sounds for hearing-impaired listeners along the auditory pathway. Using models of the auditory periphery and sound synthesis, we aimed to probe hearing impaired perception for sound textures - temporally homogeneous sounds such as rain, birds, or fire. It has been suggested that sound texture perception is mediated by time-averaged statistics measured from early auditory representations (McDermott et al., 2013). Changes to early auditory processing, such as broader "peripheral" filters or reduced compression, alter the statistical representation of sound textures. We show that these changes in the statistical representation are reflected in perception, where listeners can discriminate between synthetic textures generated from normal and impaired models of the auditory periphery. Further, a simple compensation strategy was investigated to recover the perceptual qualities of a synthetic sound texture generated from an impaired model.

5.1 Introduction

The healthy auditory system is capable of processing many sounds with varying spectral and temporal features. These sounds range from the simplest artificial stimuli, such as a tone, to the most complex auditory scene, composed of such elements as the "cocktail party", music, or environmental sounds. A sensorineural hearing-impaired system, on the other hand, demonstrates weakness in processing

[a] This chapter is based on McWalter and Dau (ISAAR 2015).

almost all sounds as compared to the normal, healthy ear. The simple artificial tones are no longer audible for particular levels and frequencies. The auditory scene becomes overwhelming as the attention-driven source separation is no longer able to track the target sound. These changes are mostly attributed to the degradation of early auditory processing, such as broadening of "peripheral" filters and loss of compression, which in turn modifies the representation of sounds at higher stages of the auditory system.

Although environmental sounds have been used in speech-in-noise experiments, their processing and perception remains relatively unstudied in the impaired auditory system. Investigating the perception of environmental sounds in the impaired auditory system could prove beneficial for understanding the difficulties such listeners have in complex listening environments. One possible avenue is to explore the representation of sound textures - temporally homogeneous sounds such as rain, birds chirping or fire - that are composed of the superposition of many similar acoustic events. It has been shown that the perceptual qualities of sound textures can be captured using a standard model of the auditory system and a set of *texture* statistics (McDermott and Simoncelli, 2011).

In this study, we investigated the auditory systems' sensitivity to synthetic sound textures generated with various impaired models of the auditory periphery. Using normal-hearing listeners we probed the response to two major factors in sensorineural hearing loss; broader peripheral filters and loss of compression. In addition, we quantified the effects of the impaired synthetic textures by parametrically varying the synthesis system statistics. Lastly, we developed a compensation strategy to optimize the texture statistics in an attempt to regain the perceptual qualities of sounds generated from impaired models towards that of an original texture.

5.2 Sound Texture Analysis and Synthesis

The generation of sound textures can be accomplished by *shaping* Gaussian noise with original sound texture statistics measured from a standard model of the auditory system (McDermott and Simoncelli, 2011). The model accounts for fundamental spectral and temporal processing by using a set of cascaded filter banks. The *texture* statistics are measured on the envelope of a filtered original sound texture, which capture the time-averaged envelope distributions as well as the covariance between pairs of neighboring filterbank channels. A companion synthesis component accepts the statistics and modifies a Gaussian noise signal, such that the statistics of the original sound texture are imposed on the synthetic sound. The synthesis process facilitates the exploration of the model

structure and the statistical parameters to investigate the change in texture representation and their consequences on perception.

The auditory model is composed of three main components; peripheral frequency filtering, compression and envelope extraction, and modulation filtering as shown in Figure **??**: Analysis System. The peripheral filtering is accomplished by means of a gammatone filterbank, where the normal-hearing system uses equivalent rectangular bandwidth (ERB) spaced filters (Glasberg and Moore, 1990). A power-law compression is applied to the output of each peripheral filter signal followed by computing the absolute value of the discrete time analytic signal, resulting in the subband envelope (Harte et al., 2005). The final stage is a modulation filterbank, which is composed of octave-spaced bandpass filters (Dau et al., 1997).

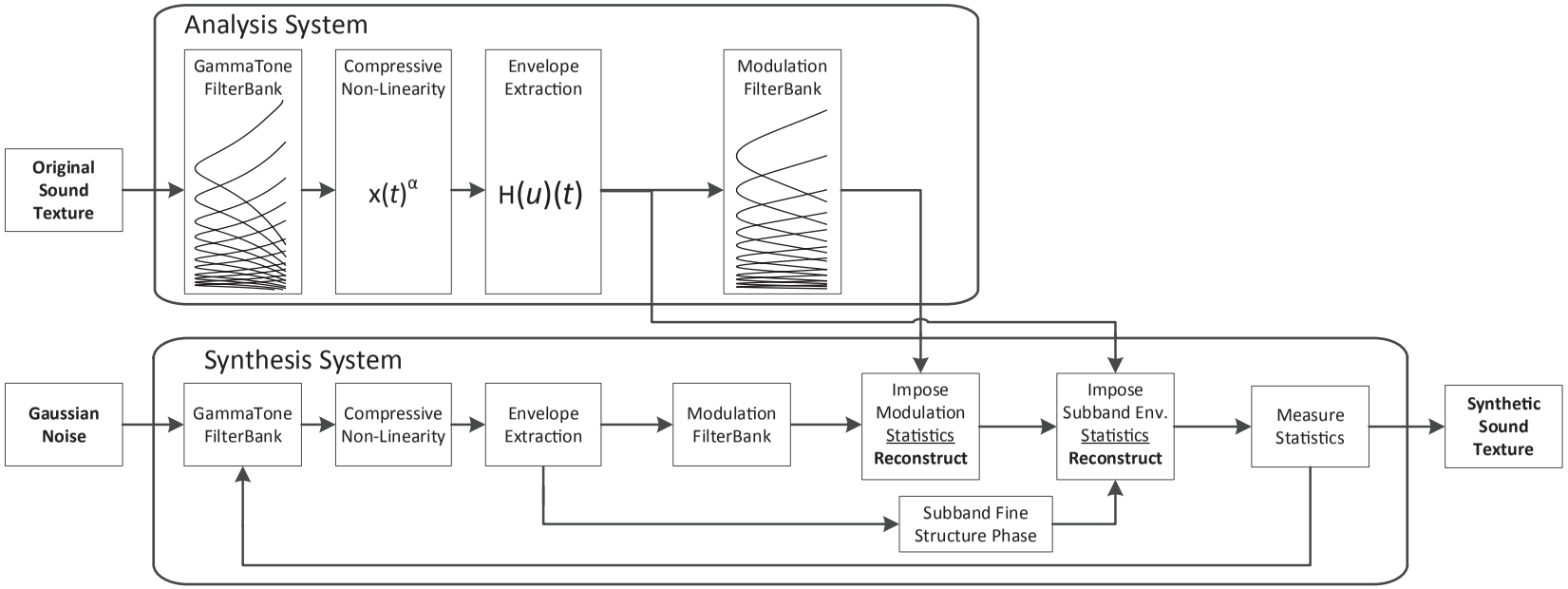


Figure 5.1: Implementation of the texture synthesis system (McDermott and Simoncelli, 2011). System is comprised of an auditory inspired analysis component, which measures marginal moments and pair-wise correlations. The statistics are passed to the synthesis component, which imposes the *texture* statistics on a noise input.

Statistics that capture many perceptually significant features of sound textures have been identified by McDermott and Simoncelli (2011). These include marginal moments and pair-wise correlations, measured on the envelope signals of the peripheral filters and modulation filters. The envelope signals are down-sampled to 400 Hz at the output of the peripheral filter, as shown in Figure **??**: Synthesis System. The statistics can be grouped into two main categories; the subband envelope statistics and the modulation statistics. The subband envelope statistics include marginal moments (mean, coefficient of variance, skewness, and kurtosis) and pair-wise correlations measured across the eight neighboring subbands. The modulation statistics include the modulation power measured at the output of each modulation filter, as well as pair-wise correlations measured for a specific modulation filter center frequency across the neighboring peripheral subbands.

The synthesis of sound textures is accomplished by imposing the statistics measured from the auditory model (Analysis System) to a Gaussian noise input. The synthesis system operates in two domains; the subband envelope and modulation domain. The synthesis system begins by deconstructing the noise signal to the modulation domain and applying both the modulation power statistics and modulation correlation statistics. The modulation filtered signals are then reconstructed to the subband envelope form, where the marginal moments and pair-wise correlation statistics are imposed. The subband envelope signals are then recombined with the subband fine structure phase signal and reconstructed to the time-domain signal.

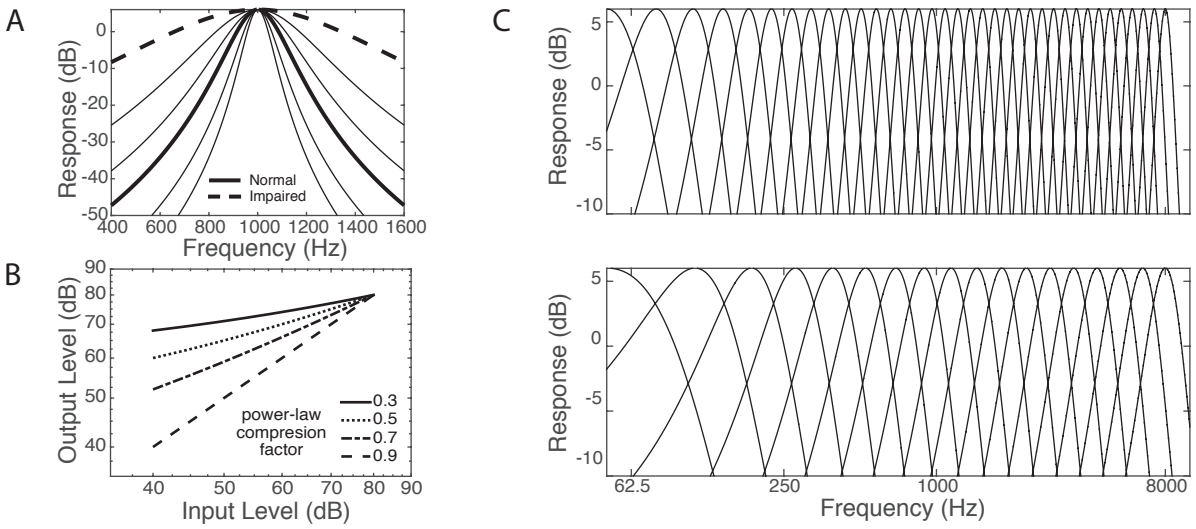


Figure 5.2: Comparison of normal and imaipred model configurations. (A) Simulated peripheral filter bandwidth for normal and impaired (4x) listeners. (B) Power-law compression ratio input-output level between normal ($\alpha = 0.3$) and impaired ($\alpha = 0.9$). (C) Filterbank model of frequency selectivity for normal (upper) and impaired (lower) hearing.

Synthetic textures were generated to functionally account for changes to the auditory system caused by sensorineural hearing loss. The limited frequency selectivity is modeled by broadening the peripheral gammatone filters and the loss of compression is modeled as an increase in the power-law compression (Moore, 2007; Rosengard et al., 2005). Figures **??**A and **??**B show the filter bandwidth and compression ratio used to generate the synthetic textures. The cross-over level for neighboring filters was preserved in all models, which resulted in fewer peripheral filters being used for the impaired auditory model. In turn, this reduced number of peripheral filters reduces the number of parameters measured for each textures. A comparison of the peripheral filterbank structure is shown in Figure **??**C.

Textures synthesized with impaired models of the auditory periphery alter the statistical representation of the sound textures, as shown in Figure **??**. In order to characterize this change, we generated 45 different textures with a normal and an impaired model with four times broader filters. The textures, including birds chirping, babble, river flowing, and jackhammer sounds, were selected to span the space of statistics, and therefore also covered a broad range of perception. The synthetic sounds were then analyzed using a reference normal auditory model. To make the normal and impaired synthetic textures more comparable, each parameter was transformed such that they varied linearly. The coefficient of variance was computed on the individual statistics. As can be seen in Figure **??**, the variation is not consistent for all textures suggesting that some parameter groups are more affected by changes in the early auditory processing than others.

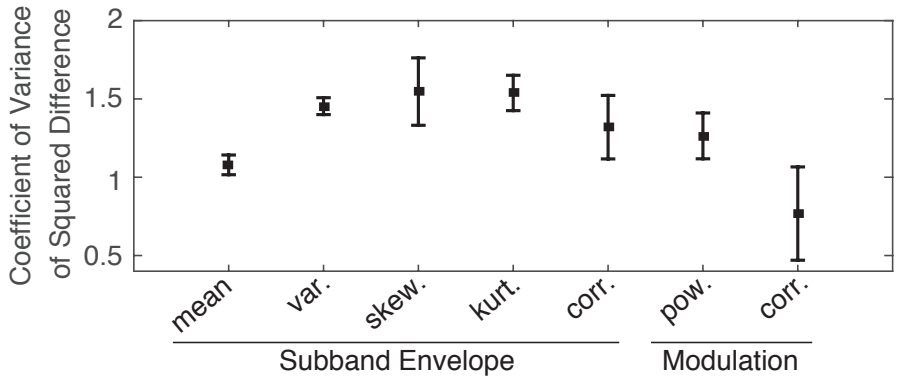


Figure 5.3: Normalized coefficient of variance comparing the normal and impaired synthetic texture statistics.

Although it is valuable to compare the average variation in texture statistics between normal and impaired auditory models, it is perhaps more intuitive to examine the individual statistics for a given texture. Figure **??** shows this comparison for the sound texture *birds chirping*. The marginal statistics

vary (Figure **??**A), particularly for the high frequency channels and higher-order marginal moments. However, for this texture, the time-averaged frequency spectrum is well preserved, as shown by the similarity between the normal and impaired mean statistics (Figure **??**A, top-right). The correlation statistics (Figure **??**B) vary as well, showing a noticeable increase in the co-variance of neighboring peripheral channels. This was expected for the hearing-impaired filters, as there is considerably more overlap between neighboring filters (see Figure **??**C). Lastly, the modulation power (Figure **??**C) reveals a difference between the two synthetic textures, particularly in the frequency region around 1.5 kHz for slow modulations.

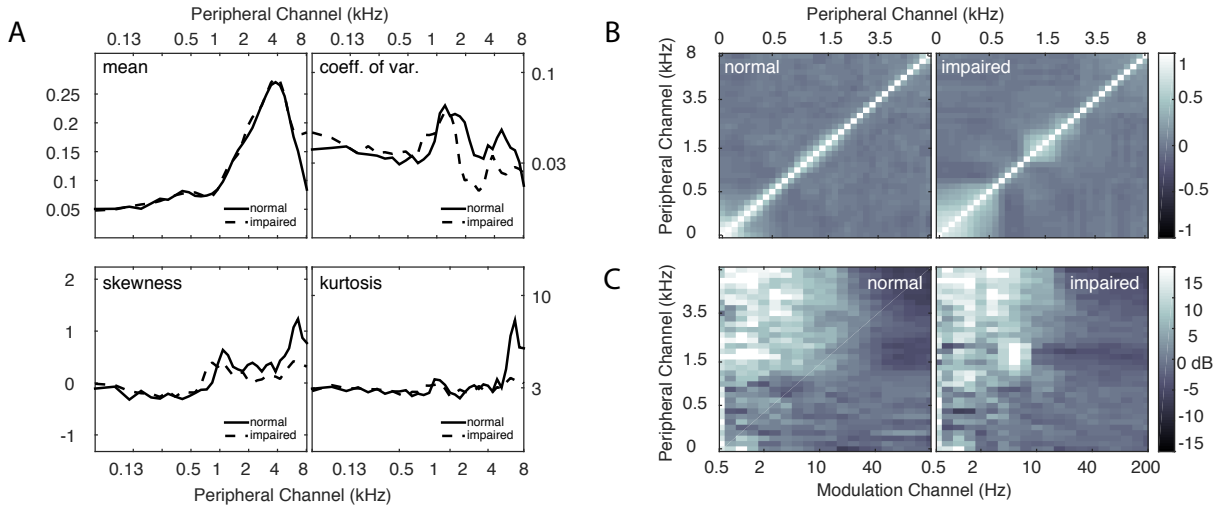


Figure 5.4: Comparison of normal and imparied texture statistics for *birds chipring*. (A) Marginal moments (mean, coeff. of variance, skewness, kurtosis), (B) pair-wise correlations for subband envelope, (C) modulation power. Note the modulation pair-wise correlation statistics are not shown.

5.3 Experiments

In order to investigate the significance of frequency selectivity and compression in sound texture perception, we asked listeners to discriminate between synthetic textures generated with normal and modified auditory models. The listeners were presented with three intervals, each 2 seconds in duration, and required to find the *odd* or modified interval, where two intervals were generated with a normal hearing model and the *odd* interval was generated with a modified hearing model. The stimuli were presented via open-ear headphones at a sound pressure level (SPL) of 65 dB. The modified texture could either be the first interval or the last interval. The two intervals generated from a normal hearing model were from the same texture family, but different sound instances,

ensuring that listeners could not use unique acoustic features in their judgments.

Figure **??**A shows the results for textures generated with broader as well as narrower peripheral filters, where the textures generated using ERB spaced filters are the reference. Fifteen self-reported normal-hearing listeners participated in the experiment. The results show an increase in discrimination performance as the model deviated from the reference. This is particularly the case when the synthetic textures were generated with broader filters. However, it can also be seen that performance increases with narrower filters, suggesting that the higher number of filters may capture some additional frequency cues. Figure **??**B shows the results for textures generated with reduced compression. Eight self-reported normal-hearing listeners participated in the experiment. The results show an increase in discrimination performance as the auditory model parameters deviated from normal hearing. The listeners reported audible artifacts in some of the intervals, and indeed, the change in compression seemed to offer cues when listening to modified compression settings. In addition, the synthesis process applies the compression during the analysis and removes the compression during the synthesis process, essential by reversing the effects of the compression. Therefore, the synthesis process seems to negate the possibility of exploring the perceptual consequences of compression with texture synthesis.

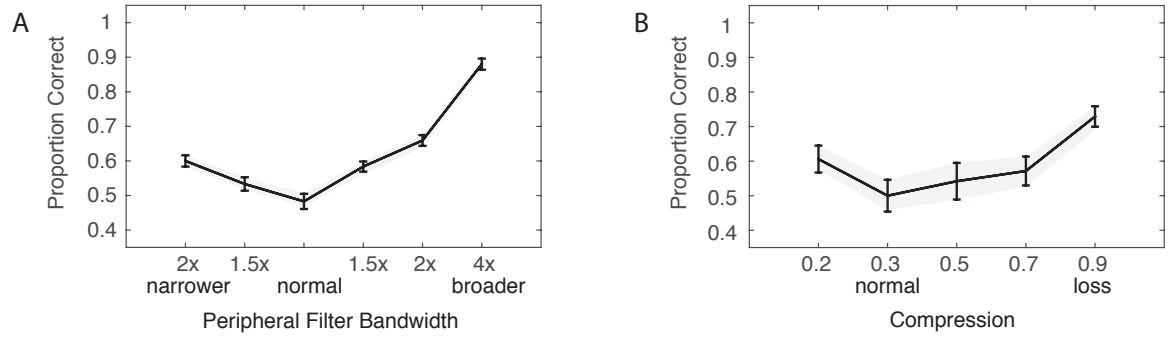


Figure 5.5: Discrimination results for synthetic textures generated with impaired models of the auditory periphery shown as a proportion correct for (A) broader/narrower peripheral filter and (B) loss/change in compression. Error bars show standard error.

To better quantify the contribution of the texture statistics to the perception of normal and impaired synthetic textures, we designed a preference experiment with stimuli that impaired particular statistical groups; marginal moments, pair-wise correlations, or modulation power. The listeners' were presented an original sound texture which was compared to two synthetic textures generated from a normal and parametrically impaired auditory model. The three intervals were each 4 seconds

in duration. The presentation of the synthetic intervals was randomized. The stimuli were presented via headphones at a level of 65dB SPL.

The results from the parametrically impaired auditory model with 4x broader filters are shown in Figure **??**A. 12 self-reported normal hearing listeners participated. The figure shows the pair-wise correlation parameter group was the most sensitive to impairment, as 72% of synthetic textures generated from a normal-hearing model were preferred over a pair-wise correlation-impaired model. The impaired marginal moments parameter group also showed an effect on the perception followed by the modulation power. It should be highlighted, that a common modulation selective filterbank structure was used for all synthetic textures. These results highlight the impact of the individual impaired parameter groups on hearing impairment.

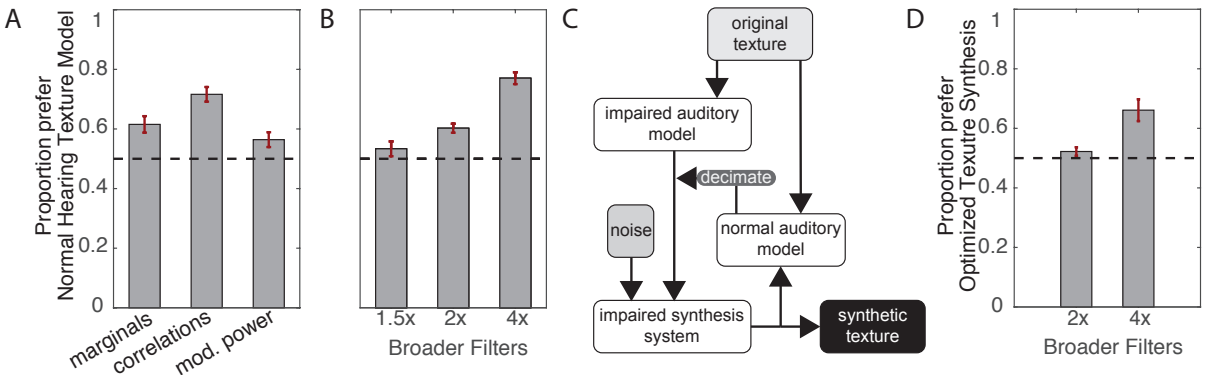


Figure 5.6: Results of listeners who prefer textures synthesized with normal hearing model for (A) impairing individual parameter groups and (B) varying severity of the model impairment. (C) Compensation stategy for impaired sound texture synthesis. (D) Optimized impaired texture statistics results for 2x and 4x broader peripheral filters. Error bars show standard error.

As a control, we also asked listeners to perform a preference task with the wholly impaired auditory system with 3 configurations of peripheral filter broadening - 1.5x, 2x and 4x - shown in Figure **??**B. The results are consists with the results shown in Figure **??**A as well as the parametrically varied impaired auditory model results. The results show that the perceptual quality declined as the auditory model deviated from that of a normal system.

5.4 Compensation Strategy

Given that the representation and perception of synthetic sound textures change with the impairment of the peripheral auditory model, the question is whether it possible to modify the statistical

representation to regain the perceptual quality towards the original texture. The results from experiments 1 and 2 revealed that a broadening of peripheral filters is salient for synthetic sound textures and most affected by the changes in the representation of pair-wise correlation statistics. A possible optimization strategy for an impaired auditory system could be a decimated version of the normal hearing statistics. However, the textures synthesized with an impaired model and decimated normal hearing statistics yielded poor synthetic versions, and often the synthesis failed. A different structure was implemented that used parallel normal and impaired model analysis systems, which is shown in Figure **??**C. The coupled analysis adjusts the impaired statistics such that the synthetic output is optimized to yield a synthetic texture similar to the original texture as measured by a normal auditory model. This can be thought of as *nudging* the impaired model representation to output a texture with similar perceptual qualities to the original texture.

Listeners performed a preference task to reveal the significance of the impaired auditory model optimization system. In each trial, listeners were presented with an original texture recording followed by two randomly presented synthetic textures; one synthesized with an impaired auditory model and another synthesized with the impaired auditory model optimization system. The stimuli were presented via headphones at a level of 65dB SPL and each interval was 4 seconds in duration. The results from the impaired auditory model texture optimization system in Figure **??**D show a modest improvement in subjective performance for the 4x broader peripheral filter case. In the case of the 2x broader filters, no improvements were found. Although the performance of the optimization system did not yield comparable results to the original, there is modest benefit and the method does warrant further investigation.

5.5 Summary

Sound textures offer a novel avenue for investigating the changes in representation due to hearing impairments, as well as the perceptual consequences of those changes. The differences in sound textures synthesized with auditory models that deviated from the normal hearing system were identifiable by normal-hearing listeners. The model impairments introduced changes to the statistical representation of sound textures, which related to perception to varying degrees. The results showed that pair-wise correlation statistics offer a primary auditory cue that affects the quality of the texture synthesis. Understanding how such *noise* signals are represented in the normal and impaired auditory system may offer some insight into the processing involved in “cocktail party” scenarios.

6

General Discussion

The aim of this thesis was to explore the perception and neural representation of sound texture in the auditory system. The central theme in all of the projects was the use of naturalistic stimuli to characterize the auditory system and generalize aspects of auditory perception beyond artificial stimuli. The results from chapter 2 exposed some interesting properties of texture perception and how texture might be perceived in the greater context of an auditory scene. The results from chapter 3 offered some evidence that textures may be processed at cortical levels of the auditory system. We also delved into expanding the perceptual space of texture to include simple rhythmic patterns (second-order amplitude modulations). Lastly, we offered some preliminary perspective on how the synthesis of naturalistic stimuli might be used to characterize and compensate for hearing impairment. This work offers some modest advances in our understanding of the auditory system. But more importantly, yields interesting experimental results that may not have been attainable without the use of naturalistic stimuli.

6.1 Summary and discussion of main results

This section will gloss over the main findings of each project and discusses some broader perspectives and implications in terms of auditory texture processing and perception.

6.1.1 Texture time-averaging

The general aim of this project was to characterize the nature of the time-averaging process which seems to underlie texture perception (McDermott et al., 2013). We devised a novel experimental listening paradigm to gauge the time window of integration based on the amount of bias induced by the stimulus history. This was achieved by presenting the listeners with naturalistic texture stimuli that changed (stepped in) statistics at some point during their duration. The listeners were then presented with a morph (probe) texture and asked to judge whether the step or the morph was most similar to a reference texture. With modest training, listeners were able to complete the experiment,

varying degrees of bias was observed in their judgments from texture to texture.

The notion of a multi-second time-averaging mechanism is quite long for the auditory system. Sensory systems have varying degrees of temporal acuity, and it is arguably audition that possesses the most fine-grained detail of the natural environment (Nuetzel and Hafter, 1976). The evidence presented by McDermott et al. (2013) suggested a time-averaged summary statistic may be involved in texture perception. In this case, we isolated the higher-order texture statistics (beyond time-averaged spectrum - spectrally matched noise). The finding that the time-averaging of texture is on the order of seconds begs the question of how changes in the auditory scene might be captured by the system. The translation of this basic paradigm to a change detection task seems a likely candidate for future experiments (Boubenec et al., 2017; Sohoglu and Chait, 2016).

The adaptive nature of the time-averaging process was another interesting finding. In hindsight, it seems like a reasonable approach for a sensory system to adapt to the characteristics of the input signal (Schwartz and Simoncelli, 2001). We showed that for texture, the auditory system appears to vary the perceptual window of integration. The averaging window becomes longer for more variable textures (e.g. ocean waves) and shorter for less variable textures (e.g. rain). This offers some perspective as to what type of strategy the auditory system may undertake for natural sound perception such as texture. It may be the case that the window must be sufficiently long to accumulate a good estimate of the signal statistics while also minimizing the window in order to detect changes in the auditory scene.

The time-averaging of texture responds to perceptual continuity and therefore does not operating blindly. We exposed this auditory strategy by introducing silent gaps and noise bursts during the stimulus presentation. Our findings suggest that the auditory system begins to accumulate evidence at the onset of a stimulus. However, if the texture is masked, the listener appears to assume that the texture is continuous in the background, regardless of its actual presence in the stimulus. This finding began to shed light on the role of texture in the auditory scene.

The auditory scene is often comprised of many different sound sources. There is some recent evidence to suggest that multiple sound textures may be grouped independently (Andreou et al., 2011). We designed an experiment to simulate foreground and background sounds in order to isolate the grouping of textures over time. Our results suggest that texture perception in an auditory scene is likely to group sound sources that are similar. This was observed by a reversal in judgment bias when the foreground was included in the results. This may have some interesting implications for how listeners segregate foreground sounds and background sounds. Even though they may represented

as texture in nature, they are likely not all grouped together and the auditory system segregates perceptual streams of different sound sources. A logical extension of this project would be to include speech as the foreground sound with various background textures.

6.1.2 Sound texture and fMRI

Our results suggesting that texture perception may be mediated by summary statistics operating on a multi-second adaptive time-averaging window (building on McDermott et al. (2013)) raised the question of where such a process could occur in the auditory system. This project attempted to address that question by analyzing the BOLD-signal (fMRI) response to textures defined by higher-order statistics. Previous work in vision by Freeman et al. (2013) found a well-defined area of visual cortex (V2) that responded selectively to higher-order image texture statistics. The translation to sound seemed like an interesting avenue to pursue, given that the perceptual space of image and sound textures can be captured by similar signal computable statistics (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000).

The first part of the study used synthetic sound stimuli generated with subsets of texture statistics. The statistics were cumulatively included (in 5 steps) from spectrally matched noise to textures synthesized from the defined set of texture statistics. The synthetic texture stimuli were presented to listeners in an MRI scanner and we measured the BOLD-signal response. With our scanning paradigm, we found sound selective voxels in primary auditory cortex as well as inferior colliculus (midbrain). However, it appeared that only voxels in auditory cortex responded to the inclusion of the higher-order texture statistics, beyond the spectrally matched noise. This result suggested that the response to texture in the auditory system may occur at cortical stages.

The BOLD-signal responses in auditory cortex were then related to behavioral texture identification performance. Listeners were presented with the same synthetic textures generated by cumulatively including statistics and asked to identify the sound from a list of 5 text label descriptors (McDermott and Simoncelli, 2011; Portilla and Simoncelli, 2000). The proportion correct from the behavioral experiment was then used to model the BOLD-signal response. The results suggeste that voxels that varied with the result from the behavioral experiment were located in primary auditory cortex.

As a sanity check, a second experiment was run with synthetic stimuli that morphed between the spectrally matched noise and the texture. This was more akin to the approach taken by Freeman and colleagues (2013). With a slight variation of the paradigm, we observed similar results, where

the BOLD-signal increased in auditory cortex with the titration of higher-order texture statistics. The BOLD-signal responses were then modeled using behavioral discrimination data obtained using the same stimuli. Again, the results pointed towards a cortical mechanism mediating the perception of sound texture.

The study was quite exploratory in nature and was intended to be the foundation for a neuroimaging project that explored BOLD-signal decoding models (Kay et al., 2008; Naselaris et al., 2011; Nishimoto et al., 2011). A next step would be to construct models to retrieve the texture statistics from the BOLD-signal response. This would offer further insight into how the texture summary statistics are computed in the auditory system. Are there statistics selective regions? Are they overlapping in regions of cortex? How might they vary across textures? It remains to be seen whether this is indeed a realistic avenue to pursue.

6.1.3 Texture and rhythmic structure

One main utility of sound textures and using naturalistic stimuli in behavioral experiments is that fundamental aspects of auditory perception learned from artificial sound can be generalized to a greater degree. Essentially, can we generalize from responses to specific artificial stimuli to a broader class of natural sounds? This project attempted to expand on the sound texture synthesis system of McDermott and Simoncelli (2011) to include second-order amplitude modulations. Second-order amplitude modulations arise from beating in the envelope-frequency domain, and, at slow rates, have a rhythmic perceptual quality. We developed an auditory texture model that used a cascade of modulation filterbanks to capture second-order amplitude modulations. The model was coupled with a texture synthesis system to generated novel texture stimuli for use in behavioral experiments.

The first part of the project looked at whether the model could capture simple second-order amplitude modulations and how the inclusion of such a system would expand on earlier auditory texture models. We found that the cascaded modulation filterbank structure was able to account for second-order amplitude modulation. In addition, the inclusion of the second modulation filterbank and accompanying power statistic improved the recognition of synthetic textures. The improved recognition was attributed to the model being able to account for simple rhythms in texture as well as mediate the time-varying modulation depth of first order modulations.

Two follow up experiments were conducted. The first examined the role of amplitude modulation selectivity in sound texture perception. We found that synthetic textures generated with the inclusion of modulation selective filters were preferred over those generated from low-pass models. Also,

the inclusion of a cascaded modulation filterbank offered modest improvement over a first-order modulation filterbank analysis. This performance increase was at the expense of expanding the model parameters by a factor of 3. The second experiment looked at the basic perception of second-order amplitude modulations and whether they may be perceived using a time-averaged summary mechanism. The results suggested that the time-averaging mechanism may be present for textures and stimuli that contain several second-order amplitude modulators. However, the auditory system appears to retain higher temporal acuity for artificial stimuli generated from second-order amplitude modulations at slow rates, below 16 Hz.

This project aimed at expanding the texture analysis system of McDermott and Simoncelli (2011) to include second-order amplitude modulation sensitivity. It seems likely that other perceptually relevant sound features, such as pitch, might be an interesting avenue to pursue within the realm of sound texture perception and synthesis (McDermott and Simoncelli, 2011; Plack et al., 2006). Again, the analysis-via-synthesis approach provided an interesting method to probe auditory perception with naturalistic stimuli. The model could be expanded in the perceptual space of texture. However, it may also be interesting to extend the synthesis approach to more temporally dynamic environmental sounds.

6.1.4 Impaired texture models

The characterization of human hearing performance is often performed with pure-tone audiometry. In some cases, this is expanded upon with speech intelligibility testing and other cognitive tasks (digit experiment). In other cases, objective experiments are performed, such as middle-ear impedance or otoacoustic emissions measurements. Although there is a relatively rich battery of tests to characterize the auditory perception performance of a human listener, there are sometimes missing components. With hearing-impaired listeners, the goal is to return audibility and listening performance to a normal hearing level. This is often difficult to attain for many reasons. One reason could be that the characterization of the listeners hearing impairment is not complete.

In this study, we began to explore the representation of texture in the impaired auditory system. We approached the problem from a modeling standpoint, where we have signal computable statistics measured from a standard auditory model. Simple modifications were applied to the model to account for changes in the auditory system related to hearing impairment, such as broader peripheral filters or loss of compression (Moore, 2007). The model was then modified for different classes of statistics, and their perceptual consequences were measured in a series of preference listening

experiments. The results suggested that the perceptual qualities of the texture defined by across-cochlear channel statistics were affected the most by the changes to the auditory model structure. These preliminary findings could be used to weight a compensation strategy for a particular statistic class over another.

The second part of the study investigated how to compensate in the domain of texture statistics due to changes in the model. This was equated to a compensation strategy based on sound statistics. Although the ability to regain the statistics was quite poor and rudimentary, this approach is slightly different than the classic approach of hearing aid compensation, aimed at regaining audibility at the level of the cochlea.

The project opens many possible applied and more fundamental research directions. The first would be to run the same experiments with hearing impaired listeners, and investigate if the synthetic textures generated from the modified auditory texture model (and, in turn, the modified statistical representation) are audible. It would also be interesting to try to implement an offline compensation strategy for recovering the quality of texture for specific hearing-impaired listeners. It's unclear, however, whether this would have any benefit for applied technologies, such as assistive listening devices, hearing aids, or cochlear implants.

6.2 Implication for Perception in Auditory Scenes

At the root of each study in this thesis is the understanding how the auditory system navigates natural acoustic environments. As has been highlighted several times throughout this thesis, mounting evidence suggests that sensory systems have evolved to handle naturalistic stimuli. Texture, be they tactile, visual or auditory, seems like a compelling stimulus base to investigate sensory perception. For audition, auditory scenes are often comprised of many sources, some of which may be texture-like in composition. The collective results presented across the chapters displayed some interesting findings related to how the auditory system perceives and represents environmental texture stimuli.

The auditory system appears to group stimuli that are likely to originate from a similar source. Although the findings support the notion of multiple streams, it is still not clear how the auditory system groups simultaneous natural sound sources based on their higher-order statistical regularities (textures). As there is variability in the estimate of statistics in textures over modest time windows (seconds), it is likely that texture streams are also robust to some variability of estimate. But how much distance is required for two texture to be streamed independently? What if they are not collocated

(in the middle of the head... as in our experiments), but are spatially positioned around the auditory scene? How would the streaming behave with a more significant foreground sound such as speech? What we can say is the fundamental mechanisms underlying sound perception are only coarsely defined.

If textures are a fundamental component of auditory streams, and neural responses are located in the primary regions of auditory cortex, it can be assumed that more complex stimuli extend to surrounding secondary regions. This would be in line with recent work on speech and music, which seems to identify secondary regions of auditory cortex and superior temporal sulcus as neural processing loci (Norman-Haignere et al., 2015a; Overath et al., 2015a). Although there are perceptual attributes that evolve in the ascending auditory pathway, it would be interesting to investigate whether any of the texture like statistics are computed in the auditory midbrain, cortex, or at all?

Bibliography

Alvarez, G. A. and A. Oliva (2009). "Spatial ensemble statistics are efficient codes that can be represented with reduced attention". In: *Proc. Natl. Acad. Sci. USA* 106.18, pp. 7345–7350.

Andén, J. and S. Mallat (2014). "Deep Scattering Spectrum". In: *IEEE Transactions on Signal Processing* 62.16, pp. 4114–4128.

Andén, J. and S. Mallat (2011). "Multiscale Scattering for Audio Classification". In: pp. 657–662.

Andén, J. and S. Mallat (2012). "Scattering representation of modulated sounds". In: *15th Int. Conf. Digital Audio Effects* 9.

Andreou, L. V., M. Kashino, and M. Chait (2011). "The role of temporal regularity in auditory segregation". In: *Hear. Res.* 280.1-2, pp. 228–235.

Ariely, D. (2001). "Seeing sets: Representation by statistical properties". In: *Psychol. Sci.* 12.2, pp. 157–162.

Atencio, C. A., T. O. Sharpee, and C. E. Schreiner (2012). "Receptive field dimensionality increases from the auditory midbrain to cortex". In: *Journal of neurophysiology* 107.10, pp. 2594–2603.

Atiani, S. et al. (2014). "Emergent selectivity for task-relevant stimuli in higher-order auditory cortex". In: *Neuron* 82.2, pp. 486–499.

Ayala, Y. A., D. Pérez-González, D. Duque, I. Nelken, and M. S. Malmierca (2013). "Frequency discrimination and stimulus deviance in the inferior colliculus and cochlear nucleus". In:

Balas, B., L. Nakano, and R. Rosenholtz (2009). "A summary-statistic representation in peripheral vision explains visual crowding". In: *J. Vis.* 9.12.

Barascud, N., M. T. Pearce, T. D. Griffiths, K. J. Friston, and M. Chait (2016). "Brain responses in humans reveal ideal observer-like sensitivity to complex acoustic patterns". In: *Proc. Natl. Acad. Sci. USA* 113.5, E616–E625.

Belin, P., R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike (2000). "Voice-selective areas in human auditory cortex". In: *Nature* 403.6767, pp. 309–312.

Bendixen, A., S. L. Denham, K. Gyimesi, and I. Winkler (2010). "Regular patterns stabilize auditory streams". In: *J. Acoust. Soc. Am.* 128.6, pp. 3658–3666.

Bolcskei, H., F. Hlawatsch, and H. G. Feichtinger (1998). “Frame-theoretic analysis of oversampled filter banks”. In: *IEEE Transactions on Signal Processing* 46.12, pp. 3256–3268.

Boubenec, Y., J. Lawlor, U. Górska, S. Shamma, and B. Englitz (2017). “Detecting changes in dynamic and complex acoustic environments”. In: *eLife* 6, e24910.

Bouman, K. L., B. Xiao, P. Battaglia, and W. T. Freeman. “Estimating the material properties of fabric from video”. In: pp. 1984–1991.

Brady, T., A Shafer-Skelton, and G. Alvarez (2017). “Global ensemble texture representations are critical to rapid scene perception”. In: *Journal of experimental psychology. Human perception and performance*.

Brodatz, P. (1966). *Textures: a photographic album for artists and designers*. Dover Pubns.

Bronkhorst, A. W. (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.

Bruna, J. and S. Mallat (2013). “Invariant scattering convolution networks”. In: *IEEE Trans Pattern Anal Mach Intell* 35.8, pp. 1872–1886.

Brunton, B. W., M. M. Botvinick, and C. D. Brody (2013). “Rats and humans can optimally accumulate evidence for decision-making”. In: *Science* 340.6128, pp. 95–98.

Buell, T. N. and E. R. Hafter (1988). “Discrimination of interaural differences of time in the envelopes of high?frequency signals: Integration times”. In: *J. Acoust. Soc. Am.* 84.6, pp. 2063–2066.

Buus, S., M. Florentine, and T. Poulsen (1997). “Temporal integration of loudness, loudness discrimination, and the form of the loudness function”. In: *J. Acoust. Soc. Am.* 101.2, pp. 669–680.

Carlyon, R. P., C. Micheyl, J. M. Deeks, and B. C. Moore (2004). “Auditory processing of real and illusory changes in frequency modulation (FM) phase”. In: *J. Acoust. Soc. Am.* 116.6, pp. 3629–3639.

Carney, L. H. (1993). “A model for the responses of low?frequency auditory?nerve fibers in cat”. In: *The Journal of the Acoustical Society of America* 93.1, pp. 401–417.

Chi, T., P. Ru, and S. A. Shamma (2005). “Multiresolution spectrotemporal analysis of complex sounds”. In: *J. Acoust. Soc. Am.* 118.2, pp. 887–906.

Condon, C. D. and N. M. Weinberger (1991). “Habituation produces frequency-specific plasticity of receptive fields in the auditory cortex”. In: *Behav. Neurosci.* 105.3, pp. 416–430.

Connor, C. E. and K. O. Johnson (1992). “Neural coding of tactile texture: comparison of spatial and temporal mechanisms for roughness perception”. In: *J. Neurosci.* 12.9, pp. 3414–3426.

Dale, A. M., B. Fischl, and M. I. Sereno (1999). “Cortical surface-based analysis. I. Segmentation and surface reconstruction”. In: *Neuroimage* 9.2, pp. 179–194.

Dau, T., B. Kollmeier, and A. Kohlrausch (1997). “Modeling auditory processing of amplitude modulation 1. Detection and masking with narrow-band carriers”. In: *J. Acoust. Soc. Am.* 102.5, pp. 2892–2905.

Dau, T., D. Püschel, and A. Kohlrausch (1996). “A quantitative model of the “effective” signal processing in the auditory system. I. Model structure”. In: *J. Acoust. Soc. Am.* 99.6, pp. 3615–3622.

David, S. V., N. Mesgarani, J. B. Fritz, and S. A. Shamma (2009). “Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli”. In: *Journal of Neuroscience* 29.11, pp. 3374–3386.

De Martino, F. et al. (2013). “Spatial organization of frequency preference and selectivity in the human inferior colliculus”. In: *Nature communications* 4, p. 1386.

Dean, I., N. S. Harper, and D. McAlpine (2005a). “Neural population coding of sound level adapts to stimulus statistics”. In: *Nat. Neurosci.* 8.12, pp. 1684–1689.

Dean, I., N. S. Harper, and D. McAlpine (2005b). “Neural population coding of sound level adapts to stimulus statistics”. In: *Nature neuroscience* 8.12, pp. 1684–1689.

Dean, I., B. L. Robinson, N. S. Harper, and D. McAlpine (2008). “Rapid neural adaptation to sound level statistics”. In: *Journal of Neuroscience* 28.25, pp. 6430–6438.

Depireux, D. A., J. Z. Simon, D. J. Klein, and S. A. Shamma (2001). “Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex”. In: *J. Neurophysiol.* 85.3, pp. 1220–1234.

Ding, N., L. Melloni, H. Zhang, X. Tian, and D. Poeppel (2016). “Cortical tracking of hierarchical linguistic structures in connected speech”. In: *Nature neuroscience* 19.1, pp. 158–164.

Dumoulin, S. O. and B. A. Wandell (2008). “Population receptive field estimates in human visual cortex”. In: *Neuroimage* 39.2, pp. 647–660.

Eggermont, J. J. (2010). “The auditory cortex: the final frontier”. In: *Computational models of the auditory system*. Springer, pp. 97–127.

Ewert, S. D., J. L. Verhey, and T. Dau (2002). “Spectro-temporal processing in the envelope-frequency domain”. In: *J. Acoust. Soc. Am.* 112.6, pp. 2921–2931.

Fairhall, A. L., G. D. Lewen, W. Bialek, and R. R. de Ruyter Van Steveninck (2001). “Efficiency and ambiguity in an adaptive neural code”. In: *Nature* 412.6849, pp. 787–792.

Field, D. J. (1987). “Relations between the statistics of natural images and the response properties of cortical cells”. In: *J Opt Soc Am A* 4.12, pp. 2379–2394.

Freeman, J. and E. P. Simoncelli (2011). “Metamers of the ventral stream”. In: *Nat. Neurosci.* 14.9, pp. 1195–1201.

Freeman, J., C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon (2013). “A functional and perceptual signature of the second visual area in primates”. In: *Nat. Neurosci.* 16.7, pp. 974–981.

Füllgrabe, C., B. C. J. Moore, L. Demany, S. D. Ewert, S. Sheft, and C. Lorenzi (2005). “Modulation masking produced by second-order modulators”. In: *J. Acoust. Soc. Am.* 117.4, pp. 2158–2168.

Glasberg, B. R. and B. C. J. Moore (1990). “Derivation of auditory filter shapes from notched-noise data”. In: *Hear. Res.* 47.1-2, pp. 103–138.

Glasberg, B. R. and B. C. Moore (2002). “A model of loudness applicable to time-varying sounds”. In: *J. Audio Eng. Soc.* 50.5, pp. 331–342.

Greenwood, J. A., P. J. Bex, and S. C. Dakin (2009). “Positional averaging explains crowding with letter-like stimuli”. In: *Proc. Natl. Acad. Sci. USA* 106.31, pp. 13130–13135.

Haberman, J. and D. Whitney (2009). “Seeing the mean: Ensemble coding for sets of faces”. In: *Journal of Experimental Psychology: Human Perception and Performance* 35.3, p. 718.

Harte, J. M., S. J. Elliott, and H. J. Rice (2005). “A comparison of various nonlinear models of cochlear compression”. In: *The Journal of the Acoustical Society of America* 117.6, pp. 3777–3786.

Heinz, M. G., X. Zhang, I. C. Bruce, and L. H. Carney (2001). “Auditory nerve model for predicting performance limits of normal and impaired listeners”. In: *Acoustics Research Letters Online* 2.3, pp. 91–96.

Hohmann, V. (2002). “Frequency analysis and synthesis using a Gammatone filterbank”. In: *Acta Acustica united with Acustica* 88.3, pp. 433–442.

Hollins, M. and S. R. Risner (2000). “Evidence for the duplex theory of tactile texture perception”. In: *Attention, Perception, & Psychophysics* 62.4, pp. 695–705.

Hullett, P. W., L. S. Hamilton, N. Mesgarani, C. E. Schreiner, and E. F. Chang (2016). “Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli”. In: *J. Neurosci.* 36.6, pp. 2014–2026.

Humphries, C., E. Liebenthal, and J. R. Binder (2010). “Tonotopic organization of human auditory cortex”. In: *Neuroimage* 50.3, pp. 1202–1211.

Irino, T. and R. D. Patterson (2006). “Dynamic, compressive gammachirp auditory filterbank for perceptual signal processing”. In: *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, Vols 1-13*, pp. 4991–4994.

Jepsen, M. L., S. D. Ewert, and T. Dau (2008). “A computational model of human auditory signal processing and perception”. In: *J. Acoust. Soc. Am.* 124.1, pp. 422–438.

Jørgensen, S. and T. Dau (2011). “Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing”. In: *J. Acoust. Soc. Am.* 130.3, pp. 1475–1487.

Joris, P. X., C. E. Schreiner, and A. Rees (2004a). “Neural processing of amplitude-modulated sounds”. In: *Physiological Reviews* 84.2, pp. 541–577.

Joris, P. X., C. E. Schreiner, and A. Rees (2004b). “Neural processing of amplitude-modulated sounds”. In: *Physiol Rev* 84.2, pp. 541–577.

Julesz, B. (1962). “Visual pattern discrimination”. In: *IRE transactions on Information Theory* 8.2, pp. 84–92.

Kay, K. N., T. Naselaris, R. J. Prenger, and J. L. Gallant (2008). “Identifying natural images from human brain activity”. In: *Nature* 452.7185, pp. 352–355.

Kohler, P. J., A. Clarke, A. Yakovleva, Y. Liu, and A. M. Norcia (2016). “Representation of Maximally Regular Textures in Human Visual Cortex”. In: *J. Neurosci.* 36.3, pp. 714–729.

Kohlrausch, A., R. Fassel, and T. Dau (2000). “The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers”. In: *J. Acoust. Soc. Am.* 108.2, pp. 723–734.

Kohn, A. (2007). “Visual adaptation: physiology, mechanisms, and functional benefits”. In: *J. Neurophysiol.* 97.5, pp. 3155–3164.

Kvale, M. N. and C. E. Schreiner (2004a). “Short-term adaptation of auditory receptive fields to dynamic stimuli”. In: *Journal of Neurophysiology* 91.2, pp. 604–612.

Kvale, M. N. and C. E. Schreiner (2004b). “Short-term adaptation of auditory receptive fields to dynamic stimuli”. In: *J. Neurophysiol.* 91.2, pp. 604–612.

Landy, M. S. (2013). “Texture analysis and perception”. In: *The new visual neurosciences (ed. Werner JS, Chalupa LM)*, pp. 639–652.

Lee, N., J. L. Ward, A. Vélez, C. Micheyl, and M. A. Bee (2017). “Frogs exploit statistical regularities in noisy acoustic scenes to solve cocktail-party-like problems”. In: *Curr. Bio.* 27.5, pp. 743–750.

Lopez-Poveda, E. A. and R. Meddis (2001). “A human nonlinear cochlear filterbank”. In: *The Journal of the Acoustical Society of America* 110.6, pp. 3107–3118.

Lorenzi, C. et al. (2001a). “Second-order modulation detection thresholds for pure-tone and narrow-band noise carriers”. In: *J. Acoust. Soc. Am.* 110.5, pp. 2470–2478.

Lorenzi, C., C. Soares, and T. Vonner (2001b). “Second-order temporal modulation transfer functions”. In: *J. Acoust. Soc. Am.* 110.2, pp. 1030–1038.

Lorenzi, C., F. Berthommier, and L. Demany (1999). “Discrimination of amplitude-modulation phase spectrum”. In: *J. Acoust. Soc. Am.* 105.5, pp. 2987–2990.

Mallat, S. (2012). “Group Invariant Scattering”. In: *Communications on Pure and Applied Mathematics* 65.10, pp. 1331–1398.

Malone, B. J., R. E. Beitel, M. Vollmer, M. A. Heiser, and C. E. Schreiner (2015). “Modulation-frequency-specific adaptation in awake auditory cortex”. In: *J. Neurosci.* 35.15, pp. 5904–5916.

McDermott, J. H. and E. P. Simoncelli (2011). “Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis”. In: *Neuron* 71.5, pp. 926–940.

McDermott, J. H., M. Schemitsch, and E. P. Simoncelli (2013). “Summary statistics in auditory perception”. In: *Nat. Neurosci.* 16.4, pp. 493–498.

McDermott, J. H., A. J. Oxenham, and E. P. Simoncelli (2009). “Sound texture synthesis via filter statistics”. In: pp. 297–300.

McDermott, J. H., D. Wrobleski, and A. J. Oxenham (2011). “Recovering sound sources from embedded repetition”. In: *Proc. Natl. Acad. Sci. USA* 108.3, pp. 1188–1193.

Meddis, R. (1986). “Simulation of mechanical to neural transduction in the auditory receptor”. In: *The Journal of the Acoustical Society of America* 79.3, pp. 702–711.

Mesgarani, N., M. Slaney, and S. A. Shamma (2006). “Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.3, pp. 920–930.

Miller, L. M., M. A. Escabi, H. L. Read, and C. E. Schreiner (2002a). “Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex”. In: *J. Neurophysiol.* 87.1, pp. 516–527.

Miller, L. M., M. A. Escabi, H. L. Read, and C. E. Schreiner (2002b). “Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex”. In: *J. Neurophysiol.* 87.1, pp. 516–527.

Moerel, M., F. De Martino, and E. Formisano (2012). “Processing of natural sounds in human auditory cortex: tonotopy, spectral tuning, and relation to voice sensitivity”. In: *J. Neurosci.* 32.41, pp. 14205–14216.

Moore, B. C. (2007). *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons.

Moore, R. C., T. Lee, and F. E. Theunissen (2013). "Noise-invariant neurons in the avian auditory cortex: Hearing the song in noise". In: *PLoS Comput. Biol.* 9.3, e1002942.

Naselaris, T., K. N. Kay, S. Nishimoto, and J. L. Gallant (2011). "Encoding and decoding in fMRI". In: *Neuroimage* 56.2, pp. 400–410.

Natan, R. G. et al. (2015). "Complementary control of sensory adaptation by two types of cortical interneurons". In: *eLife* 4, e09868.

Nelken, I. and A. De Cheveigné (2013). "An ear for statistics". In: *Nat. Neurosci.* 16.4, pp. 381–382.

Nishimoto, S., A. T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J. L. Gallant (2011). "Reconstructing visual experiences from brain activity evoked by natural movies". In: *Current Biology* 21.19, pp. 1641–1646.

Norman-Haignere, S., N. G. Kanwisher, and J. H. McDermott (2013). "Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex". In: *Journal of Neuroscience* 33.50, pp. 19451–19469.

Norman-Haignere, S., N. G. Kanwisher, and J. H. McDermott (2015a). "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition". In: *Neuron* 88.6, pp. 1281–1296.

Norman-Haignere, S., N. G. Kanwisher, and J. H. McDermott (2015b). "Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition". In: *Neuron* 88.6, pp. 1281–1296.

Nuetzel, J. M. and E. R. Hafter (1976). "Lateralization of complex waveforms: effects of fine structure, amplitude, and duration". In: *J. Acoust. Soc. Am.* 60.6, pp. 1339–1346.

Ogawa, S., T.-M. Lee, A. R. Kay, and D. W. Tank (1990). "Brain magnetic resonance imaging with contrast dependent on blood oxygenation". In: *Proceedings of the National Academy of Sciences* 87.24, pp. 9868–9872.

Olshausen, B. A. and D. J. Field (1996). "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". In: *Nature* 381.6583, p. 607.

Overath, T., J. H. McDermott, J. M. Zarate, and D. Poeppel (2015a). "The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts". In: *Nat. Neurosci.* 18.6, pp. 903–911.

Overath, T., J. H. McDermott, J. M. Zarate, and D. Poeppel (2015b). “The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts”. In: *Nat. Neurosci.* 18.6, pp. 903–911.

Parkes, L., J. Lund, A. Angelucci, J. A. Solomon, and M. Morgan (2001). “Compulsory averaging of crowded orientation signals in human vision”. In: *Nat. Neurosci.* 4.7, pp. 739–744.

Pasley, B. N. et al. (2012). “Reconstructing speech from human auditory cortex”. In: *PLoS Biol* 10.1, e1001251.

Patterson, R., I. Nimmo-Smith, J. Holdsworth, and P. Rice (1987). “An efficient auditory filterbank based on the gammatone function”. In: 2.

Piazza, E. A., T. D. Sweeny, D. Wessel, M. A. Silver, and D. Whitney (2013). “Humans use summary statistics to perceive auditory sequences”. In: *Psychol. Sci.* 24.8, pp. 1389–1397.

Plack, C. J. (2005). *The sense of hearing*. Mahwah, N.J.: Lawrence Erlbaum Associates, xi, 267 p.

Plack, C. J., A. J. Oxenham, and R. R. Fay (2006). *Pitch: neural coding and perception*. Vol. 24. Springer Science & Business Media.

Poeppel, D. (2003). “The analysis of speech in different temporal integration windows: cerebral lateralization as ?asymmetric sampling in time?” In: *Speech communication* 41.1, pp. 245–255.

Portilla, J. and E. P. Simoncelli (2000). “A parametric texture model based on joint statistics of complex wavelet coefficients”. In: *Int. J. Comput. Vis.* 40.1, pp. 49–70.

Preuss, A and P Möller-Preuss (1990). “Processing of amplitude modulated sounds in the medial geniculate body of squirrel monkeys”. In: *Experimental brain research* 79.1, pp. 207–211.

Robles, L. and M. A. Ruggero (2001). “Mechanics of the mammalian cochlea”. In: *Physiological reviews* 81.3, pp. 1305–1352.

Rodríguez, F. A., C. Chen, H. L. Read, and M. A. Escabí (2010). “Neural modulation tuning characteristics scale to efficiently encode natural sound statistics”. In: *J. Neurosci.* 30.47, pp. 15969–15980.

Rosengard, P. S., A. J. Oxenham, and L. D. Braida (2005). “Comparing different estimates of cochlear compression in listeners with normal and impaired hearing”. In: *The Journal of the Acoustical Society of America* 117.5, pp. 3028–3041.

Rosowski, J. and E. Relkin (2001). “Introduction to the analysis of middle ear function”. In: *Physiology of the Ear. Singular, San Diego*, pp. 161–190.

Ruggero, M. A. (1992a). “Responses to sound of the basilar membrane of the mammalian cochlea”. In: *Curr. Opin. Neurobiol.* 2.4, pp. 449–456.

Ruggero, M. A. (1992b). “Responses to sound of the basilar membrane of the mammalian cochlea”. In: *Current opinion in neurobiology* 2.4, pp. 449–456.

Saint-Arnaud, N. and K. Popat (1995). “Analysis and synthesis of sound textures”. In: *Proc. AJCAI Workshop Comput. Auditory Scene Anal.* Pp. 293–308.

Sakmann, B. and E. Neher (1984). “Patch clamp techniques for studying ionic channels in excitable membranes”. In: *Annual review of physiology* 46.1, pp. 455–472.

Santoro, R. et al. (2014). “Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex”. In: *Plos Computational Biology* 10.1.

Scharf, B. (1978). “Loudness”. In: *Handbook of perception* 4, pp. 187–242.

Schnupp, J., I. Nelken, and A. King (2011). *Auditory neuroscience: Making sense of sound*. MIT press.

Schwartz, O. and E. P. Simoncelli (2001). “Natural signal statistics and sensory gain control”. In: *Nature neuroscience* 4.8, pp. 819–825.

Schwarz, D. (2011). “State of the art in sound texture synthesis”. In: pp. 221–231.

Sharpee, T. O., C. A. Atencio, and C. E. Schreiner (2011). “Hierarchical representations in the auditory cortex”. In: *Current opinion in neurobiology* 21.5, pp. 761–767.

Simoncelli, E. P. and B. A. Olshausen (2001). “Natural image statistics and neural representation”. In: *Annual review of neuroscience* 24.1, pp. 1193–1216.

Slaney, M. (1993). “An efficient implementation of the Patterson-Holdsworth auditory filter bank”. In: *Apple Computer, Perception Group, Tech. Rep* 35, p. 8.

Snell, R. S. (2010). *Clinical neuroanatomy*. Lippincott Williams and Wilkins.

Sohoglu, E. and M. Chait (2016). “Neural dynamics of change detection in crowded acoustic scenes”. In: *NeuroImage* 126, pp. 164–172.

Strickland, E. A. and N. F. Viemeister (1996). “Cues for discrimination of envelopes”. In: *J. Acoust. Soc. Am.* 99.6, pp. 3638–3646.

Theunissen, F. E., K. Sen, and A. J. Doupe (2000). “Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds”. In: *J. Neurosci.* 20.6, pp. 2315–2331.

Theunissen, F. E. and J. E. Elie (2014). “Neural processing of natural sounds”. In: *Nature Reviews Neuroscience* 15.6, pp. 355–366.

Theunissen, F. E., S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant (2001). “Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli”. In: *Network: Computation in Neural Systems* 12.3, pp. 289–316.

Turner, R. and M. Sahani (2008). “Modeling natural sounds with modulation cascade processes”. In: *Advances in neural information processing systems*, pp. 1545–1552.

Ulanovsky, N., L. Las, and I. Nelken (2003). “Processing of low-probability sounds by cortical neurons”. In: *Nat. Neurosci.* 6.4, pp. 391–398.

Ulanovsky, N., L. Las, D. Farkas, and I. Nelken (2004). “Multiple time scales of adaptation in auditory cortex neurons”. In: *J. Neurosci.* 24.46, pp. 10440–10453.

Verhey, J. L., S. D. Ewert, and T. Dau (2003). “Modulation masking produced by complex tone modulators”. In: *J. Acoust. Soc. Am.* 114.4 Pt 1, pp. 2135–2146.

Viemeister, N. F. and G. H. Wakefield (1991). “Temporal integration and multiple looks”. In: *J. Acoust. Soc. Am.* 90.2 Pt 1, pp. 858–865.

Viemeister, N. F. (1979). “Temporal modulation transfer functions based upon modulation thresholds”. In: *J. Acoust. Soc. Am.* 66.5, pp. 1364–1380.

Von Békésy, G. and E. G. Wever (1960). *Experiments in hearing*. Vol. 8. McGraw-Hill New York.

Wang, H. X., D. J. Heeger, and M. S. Landy (2012). “Responses to second-order texture modulations undergo surround suppression”. In: *Vision Res.* 62, pp. 192–200.

Warren, R. M. (1970). “Perceptual restoration of missing speech sounds”. In: *Science* 167.3917, pp. 392–393.

Weber, A. I. et al. (2013). “Spatial and temporal codes mediate the tactile perception of natural textures”. In: *Proc. Natl. Acad. Sci. USA* 110.42, pp. 17107–17112.

Zaidi, Q., J. Victor, J. McDermott, M. Geffen, S. Bensmaia, and T. A. Cleland (2013). “Perceptual spaces: mathematical structures to neural mechanisms”. In: *J. Neurosci.* 33.45, pp. 17597–17602.

Zatorre, R. J., P. Belin, and V. B. Penhune (2002). “Structure and function of auditory cortex: music and speech”. In: *Trends in cognitive sciences* 6.1, pp. 37–46.

Zhang, X., M. G. Heinz, I. C. Bruce, and L. H. Carney (2001). “A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression”. In: *The Journal of the Acoustical Society of America* 109.2, pp. 648–670.

Ziemba, C. M., J. Freeman, J. A. Movshon, and E. P. Simoncelli (2016). “Selectivity and tolerance for visual texture in macaque V2”. In: *Proc. Natl. Acad. Sci. USA* 113.22, E3140–E3149.

Zilany, M. S. and I. C. Bruce (2006). “Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery”. In: *The Journal of the Acoustical Society of America* 120.3, pp. 1446–1466.

Zwislocki, J. J. (1969). “Temporal summation of loudness - An analysis”. In: *J. Acoust. Soc. Am.* 46.2p2, pp. 431–441.