Sound, Ears, Brains and the World

Josh McDermott
Dept. of Brain and Cognitive Sciences, MIT
9.523

Consider some examples of typical auditory input:

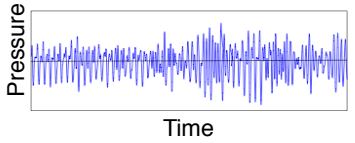
Scene from cafe:

Scene from sports bar:

Radio excerpt:

Barry White:

The ear receives a pressure waveform.



AUDITION

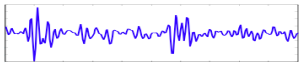
When objects in the world vibrate, they transmit acoustic energy through surrounding medium in the form of a wave.

The ears measure this sound energy and transmit it to the brain.

The task of the brain is to interpret this signal, and use it to figure out what is out there in the world.

The listener is interested in what happened in the world to cause the sound:

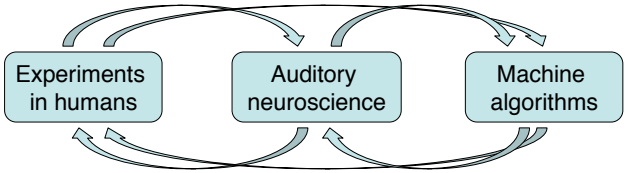
- Most properties of interest are not explicit in the waveform:



How do we derive information about the world from sound?

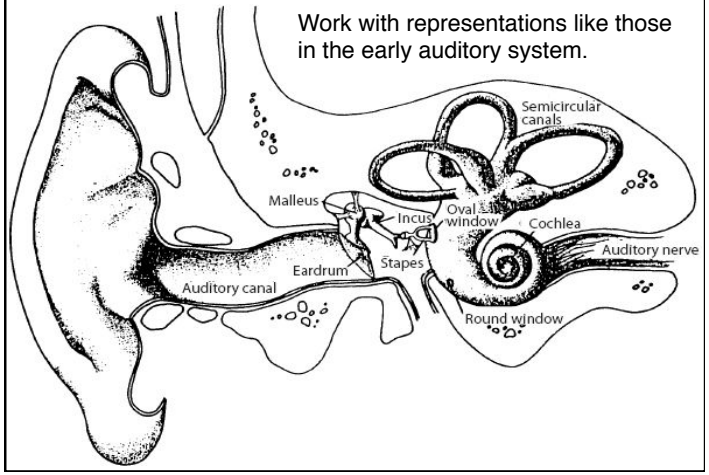
My research group: Laboratory for Computational Audition

Psychology **Neuroscience** **Engineering**



General approach: start with what the brain has to work with.

Work with representations like those in the early auditory system.

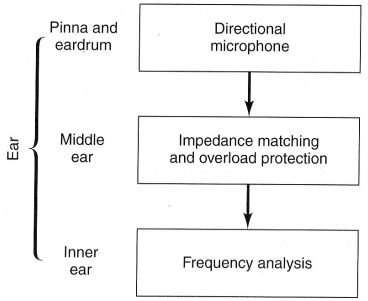


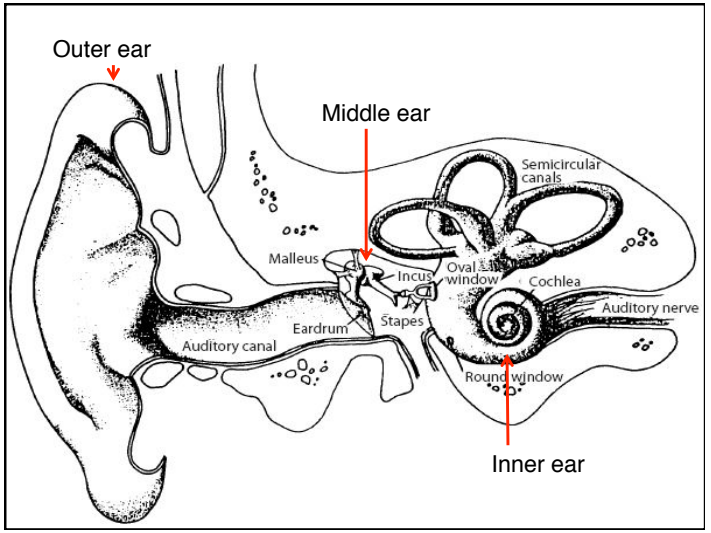
Plan for today:

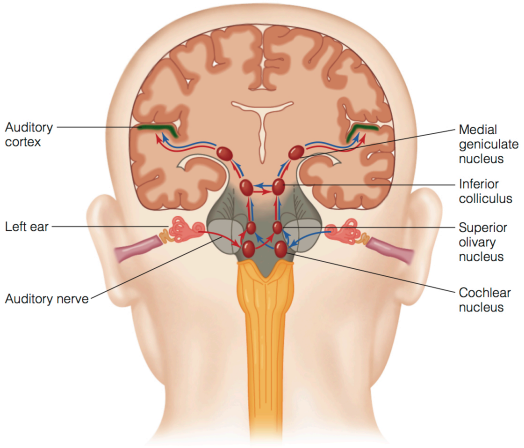
1. Overview of Auditory System
2. Sound Texture Perception

Part 1: Overview of Auditory System

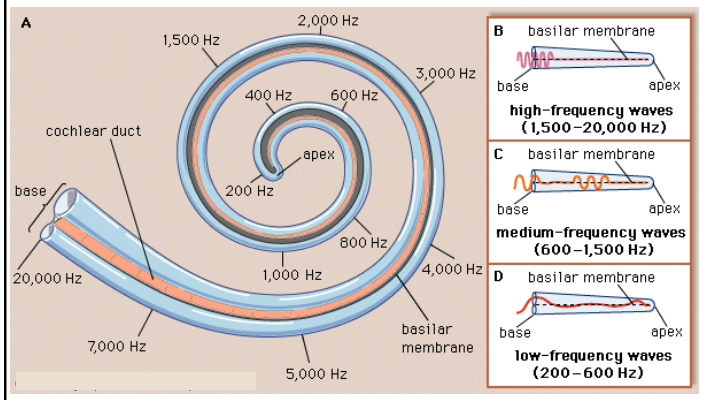
Functional schematic of the ear:



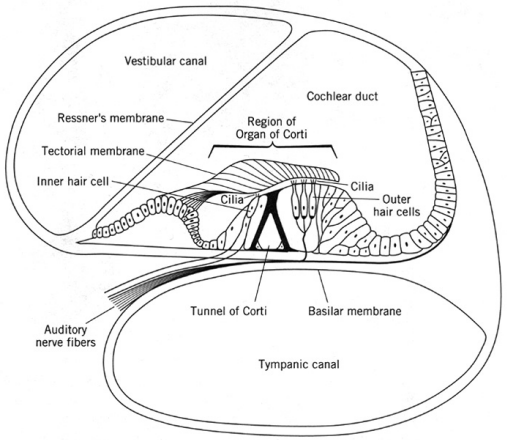




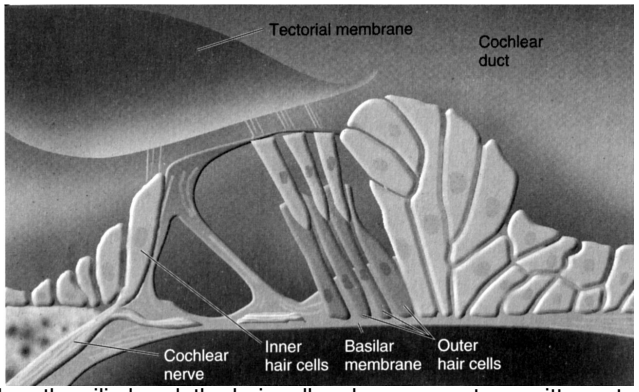
Cochlear transduction is frequency tuned:



Cross section of cochlea



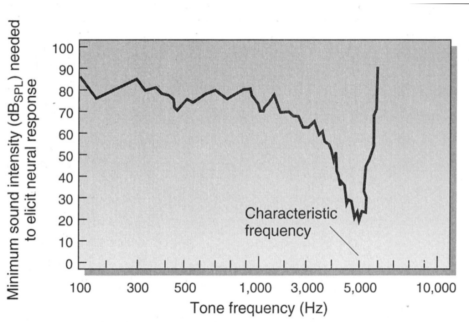
Movement of the basilar membrane causes the hair cells to move against the tectorial membrane, which causes the cilia to bend.



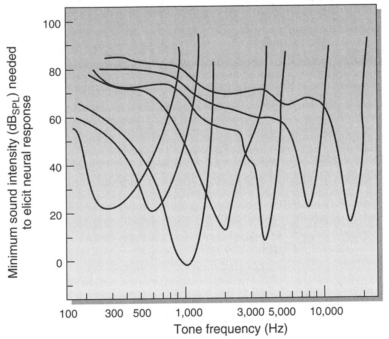
When the cilia bend, the hair cells release neurotransmitter onto synapses with auditory nerve fibers that send signals to the brain.

But because only part of the basilar membrane moves for a given frequency of sound, each hair cell and auditory nerve fiber signal only particular frequencies of sound.

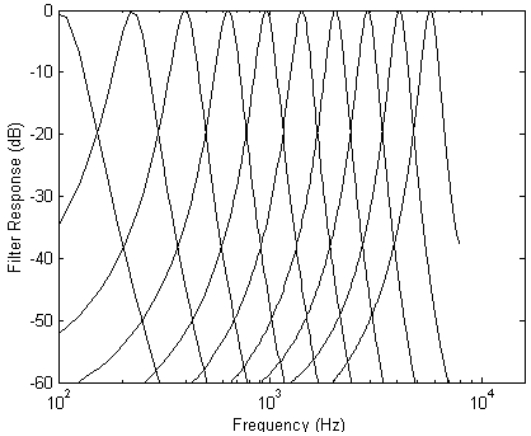
One example:



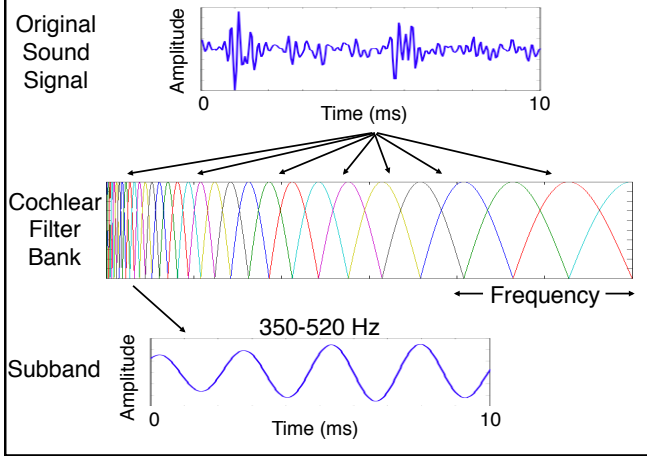
Different nerve fibers (synapsing at different points along the basilar membrane) are tuned to different frequencies:



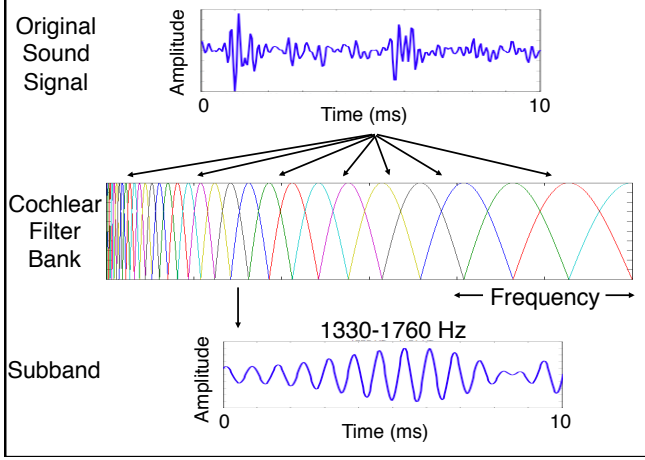
Auditory nerve fibers usually approximated as bandpass filters:



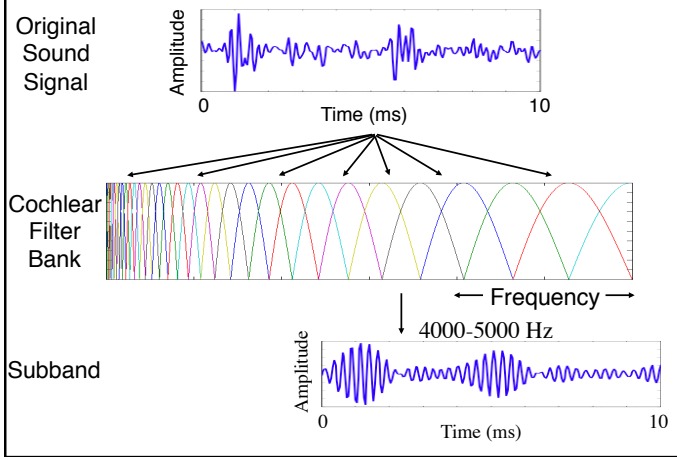
1. Sound signal represented with “subbands”, like auditory nerve:



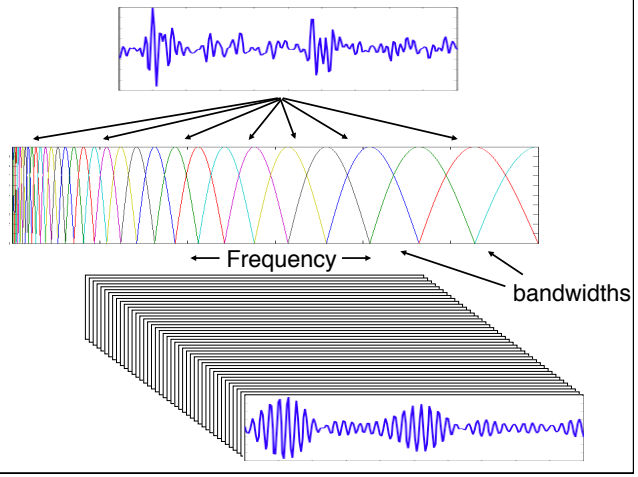
1. Sound signal represented with “subbands”, like auditory nerve:



1. Sound signal represented with “subbands”, like auditory nerve:

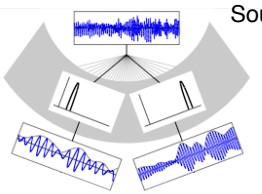


1. Sound signal represented with “subbands”, like auditory nerve:

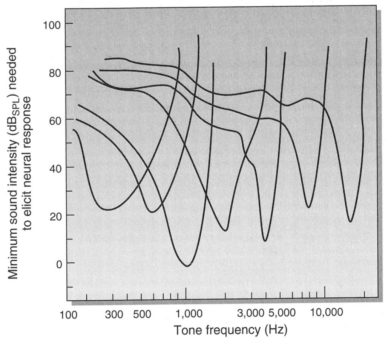


AUDITORY MODEL

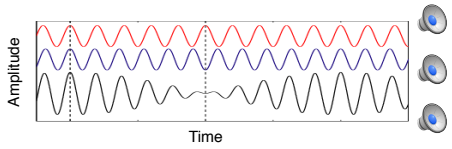
1. Cochlear filters



Frequency selectivity has a host of perceptual consequences.



Perception of beating constrained by freq. selectivity



Superposition of two pure tones waxes and wanes in amplitude.

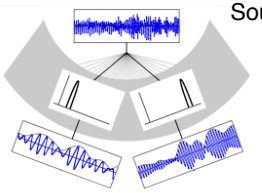
The perceptual correlate of rapid beating is known as roughness.

•Perception of beating is constrained by the cochlea:
-Beats are only heard if two frequency components fall within the filter bandwidth of the cochlea:

1 semitone frequency difference:
3 semitones:
8 semitones:

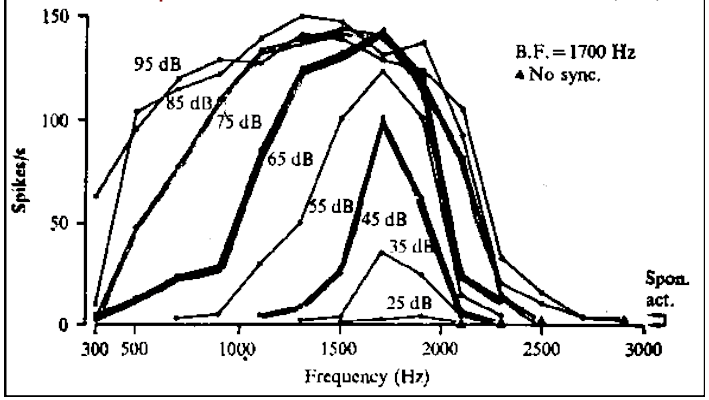
AUDITORY MODEL

1. Cochlear filters



But… linear filtering provides only an approximate description of cochlear tuning. At high levels, frequency tuning broadens:

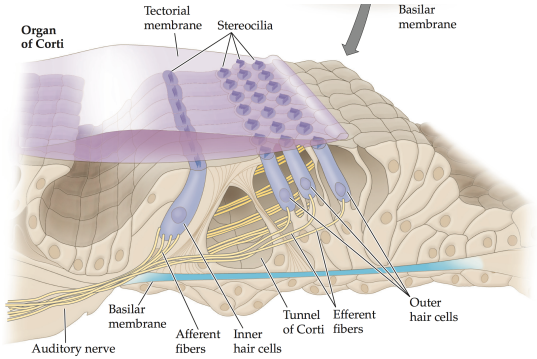
Auditory nerve fiber response to pure tones at different frequencies and levels: Rose (1971)



A related complication - there are two kinds of hair cells:

1. Inner hair cells (transduce sound energy)

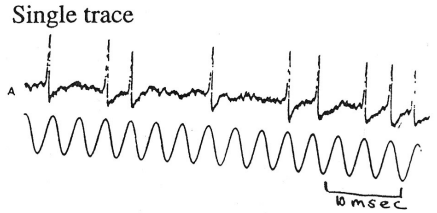
2. Outer hair cells (part of a feedback system that amplifies the motion of the basilar membrane and sharpens its tuning).



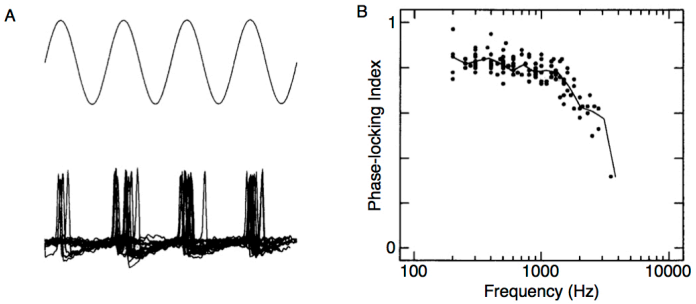
Neural Coding of Sound

Another important response property of the cochlea:

For low frequencies, auditory nerve spikes are phase-locked to the stimulus:



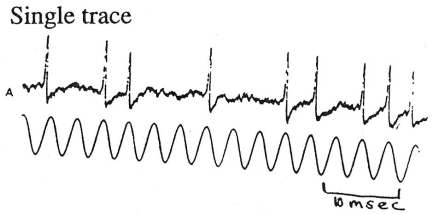
Phase locking occurs for frequencies under ~4kHz
(in nonhuman animals – no available data in humans).



Most nerve fibers don't fire with every stimulus cycle, especially for higher frequencies.

A stimulus is encoded by many nerve fibers at once.

This form of population coding may be one reason why there are many more auditory nerve fibers (30,000 per ear) than there are inner hair cells (3500 per ear).



Some interesting numbers:

Per ear:

3500 inner hair cells
12,000 outer hair cells
30,000 auditory nerve fibers

Per hemisphere:

60 million neurons in primary auditory cortex?

Per eye:

5 million cones
100 million rods
1.5 million optic nerve fibers

140 million neurons in primary visual cortex

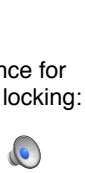
There are thus two cues to the frequencies in a sound:

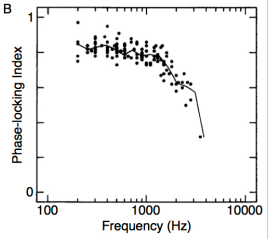
1) The place of excitation in the cochlea (which nerve fibers fire the most), i.e. a rate code.

2) The timing of spikes that are fired.

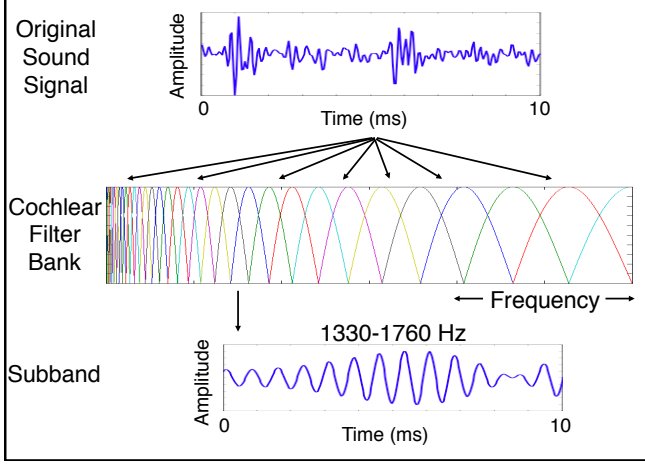
Debate on the importance of these two cues has raged for over a century, and is still unresolved.

Circumstantial evidence for importance of phase locking:

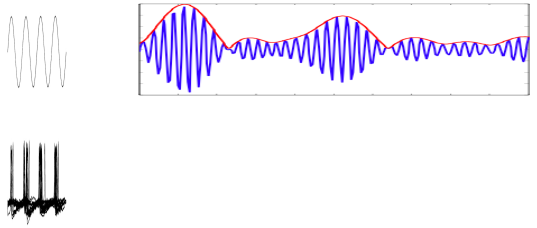




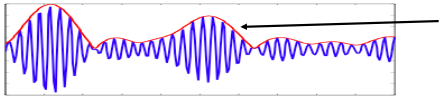
How does a subband relate to what we see in the auditory nerve?



Subbands can be characterized by instantaneous amplitude and phase, loosely mapping onto rate and spike timing in auditory nerve:



Much of the information in sound is carried by the way that frequencies are modulated over time, measured by the instantaneous amplitude in a subband:

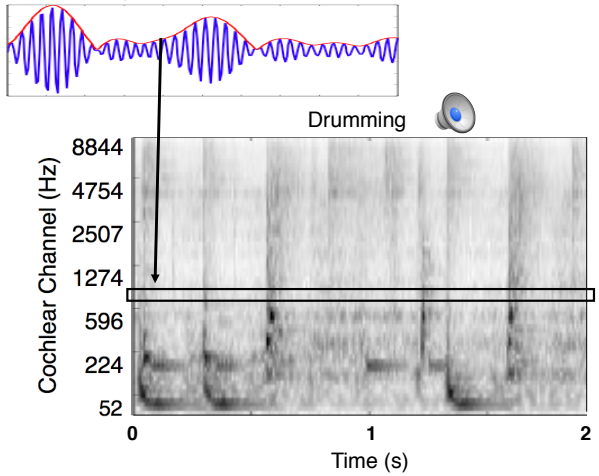


Amplitude modulations are captured by the **envelope**

-envelope is easy to extract from auditory nerve responses (firing rate over local time windows)

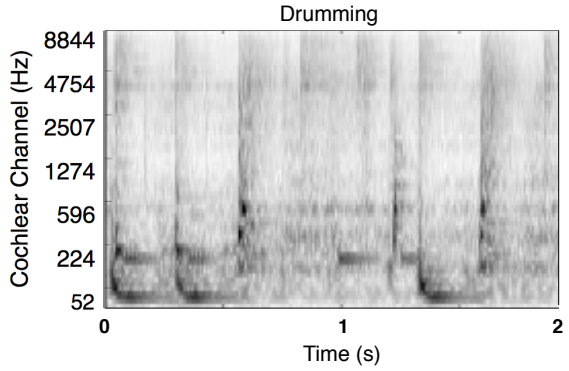
-we often extract it with Hilbert transform
(magnitude of analytic signal)

A spectrogram contains the envelope of each subband:

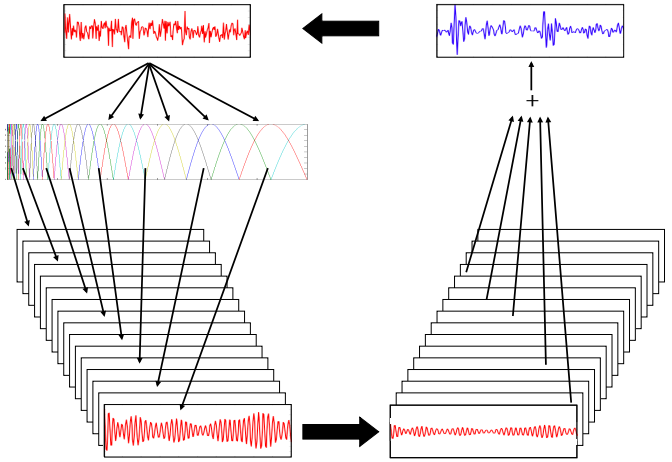


Envelopes often capture all the information that matters perceptually.

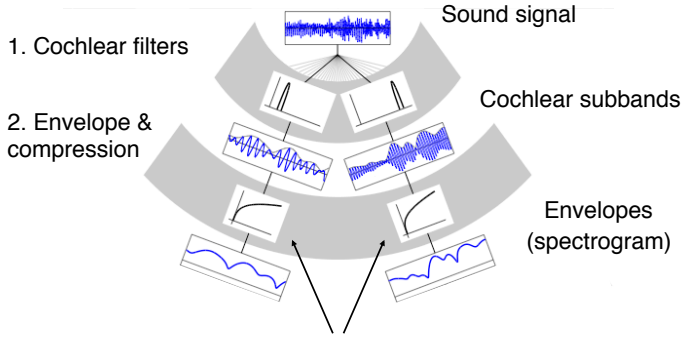
Sounds can be reconstructed just from the envelopes.



Start with noise, replace with envelopes, resynthesize, iterate:

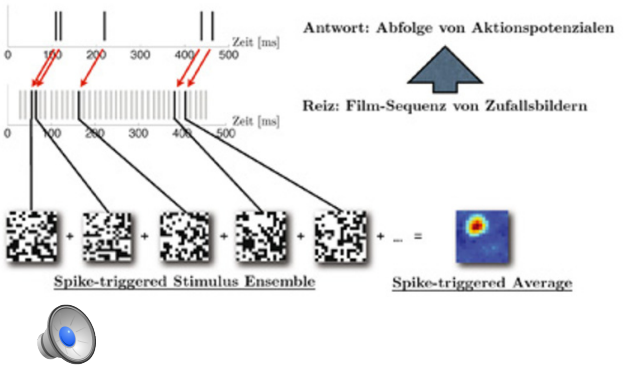


AUDITORY MODEL



Amplitude compression, simulating that of cochlea

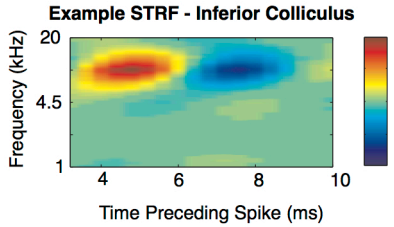
Spike-triggered average: a method to characterize a neuron's receptive field.

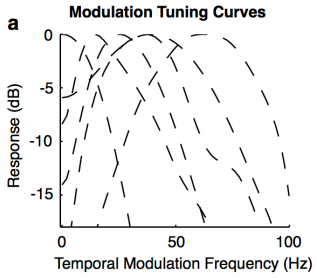


STRF = spectro-temporal receptive field

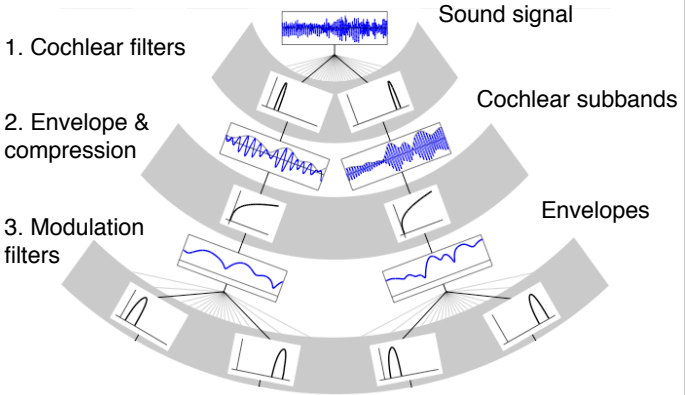
Derived from methods like the STA applied to stimulus spectrogram.

As early as the midbrain, auditory neurons are tuned to particular modulation rates.



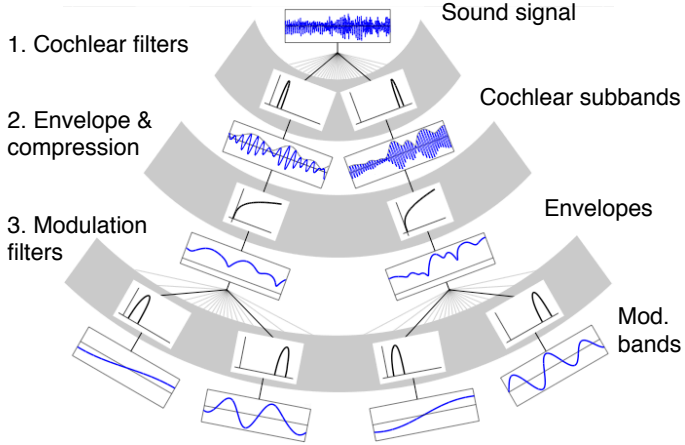


Envelope structure measured with second filter bank:

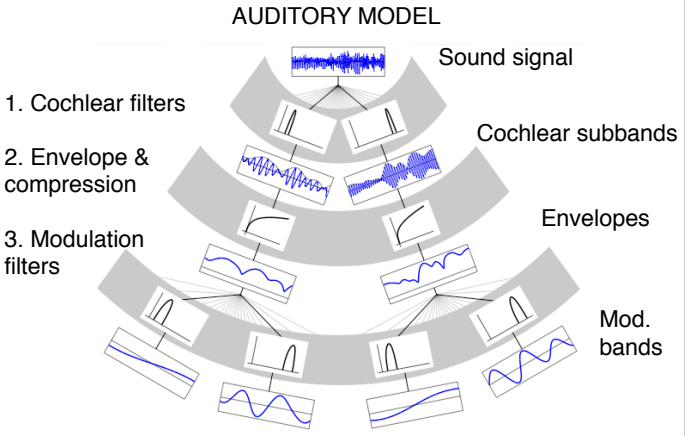


Old idea (Dau, Viemeister etc.), with fair bit of empirical support.

Envelope structure measured with second filter bank:

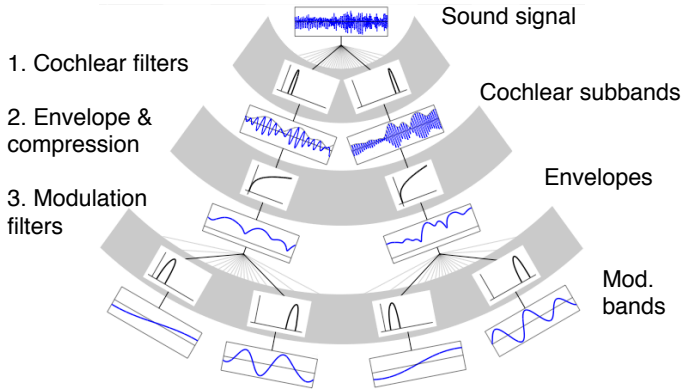


Model of auditory signal processing from cochlea to midbrain/ thalamus:



Given these representations, how do we recognize sounds and their properties?

AUDITORY MODEL

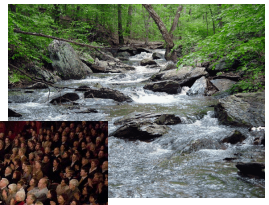


Part 2: Sound Texture

SOUND TEXTURE

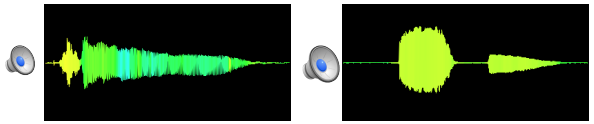
Textures result from large numbers of acoustic events.

- rain
- wind
- birds in a forest
- running water
- insects at night
- crowd noise
- applause
- fire

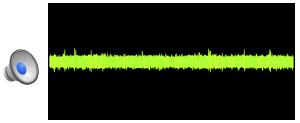



Common in the world, though historically neglected.

Much of hearing research is concerned with the sounds of individual events:

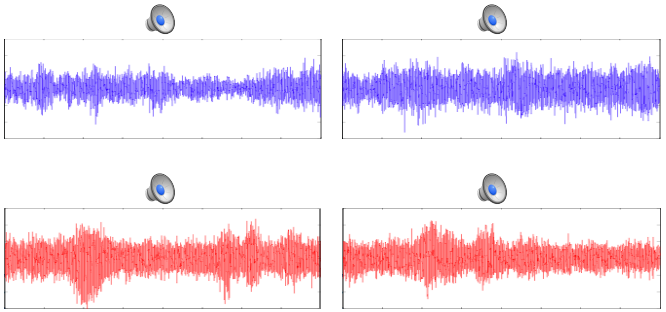


Unlike event sounds, textures are stationary - essential properties do not change over time.



- Stationarity makes textures a good starting point for understanding auditory representation.

How do people represent, recognize sound textures?



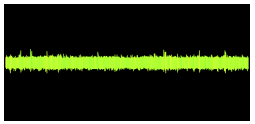
What do you extract and store about these waveforms to recognize that they are the same kind of thing?

Key Theoretical Proposal:

•Because they are stationary, textures can be captured by statistics that are *time-averages* of acoustic measurements.

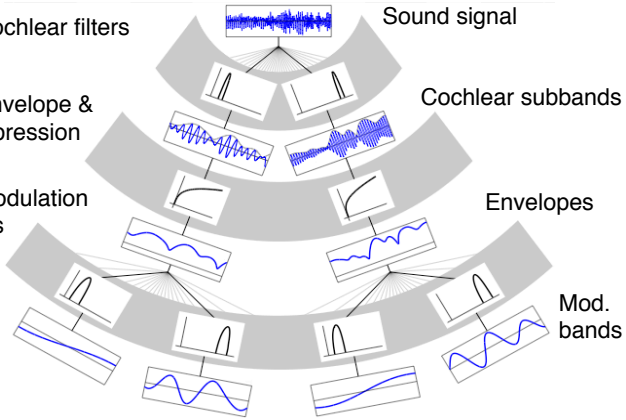
•When you recognize the sound of fire or the sound of rain, you may be recognizing these statistics.

What kinds of statistics might we be measuring?



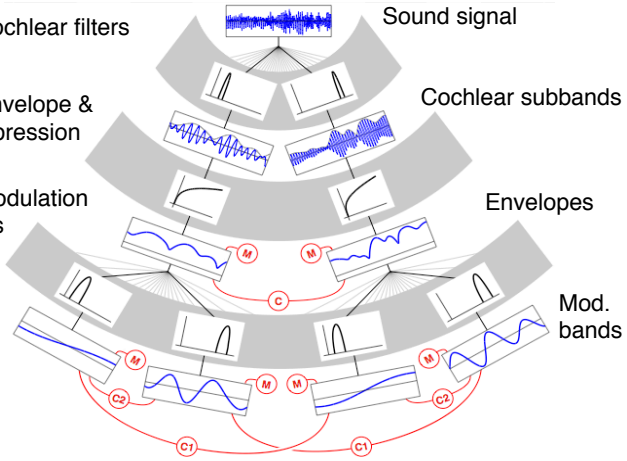
Whatever statistics the auditory system measures are presumably derived from these representations:

1. Cochlear filters
2. Envelope & compression
3. Modulation filters

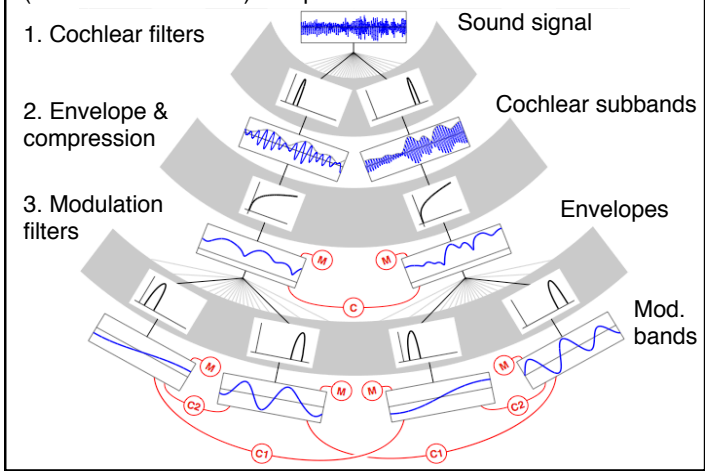


How far can we get with generic statistics of standard auditory representations?

1. Cochlear filters
2. Envelope & compression
3. Modulation filters

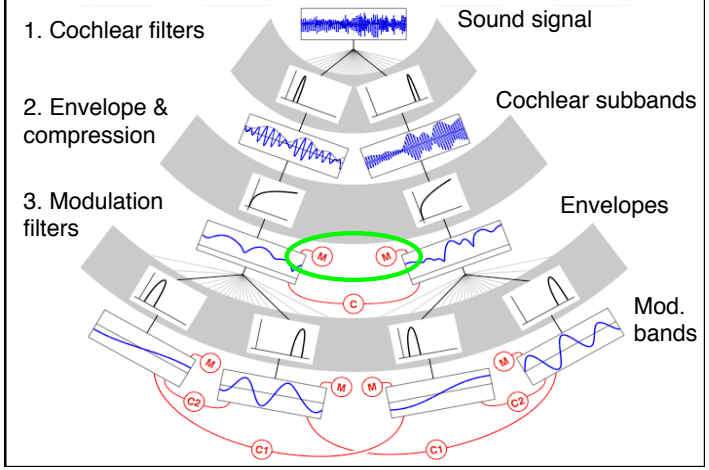


How far can we get with generic statistics – marginal moments (mean/variance/skew) and pairwise correlations?

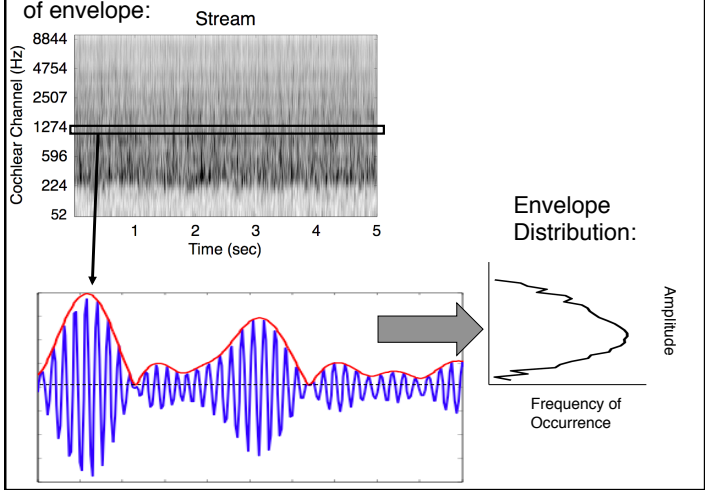


- Statistics are not specifically tailored to natural sounds
- Ultimately, would be nice to learn statistics from data (stay tuned…)
- But moments and correlations are interpretable, and might give insight.

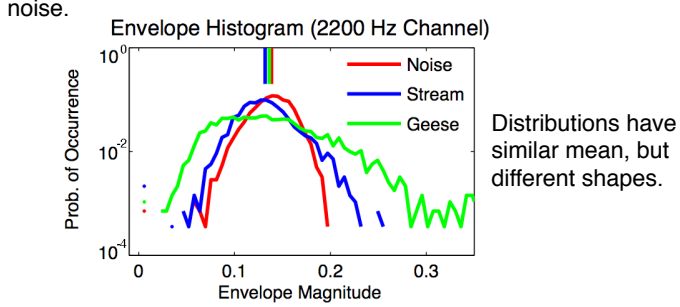
To be useful for recognition, statistics need to give different values for different sounds…



Marginal moments (mean, variance, skew) describe distribution of envelope:



Envelope distributions for natural signals differ from those for noise.

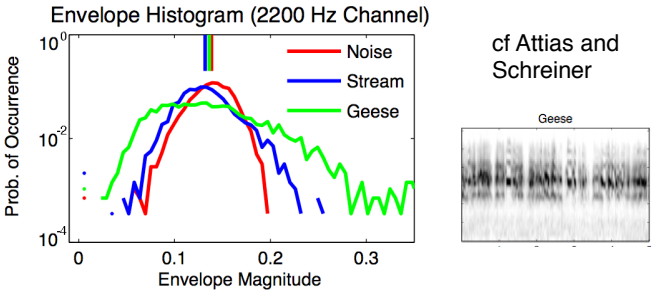


Distributions have similar mean, but different shapes.

Natural signals are **sparser** than noise.

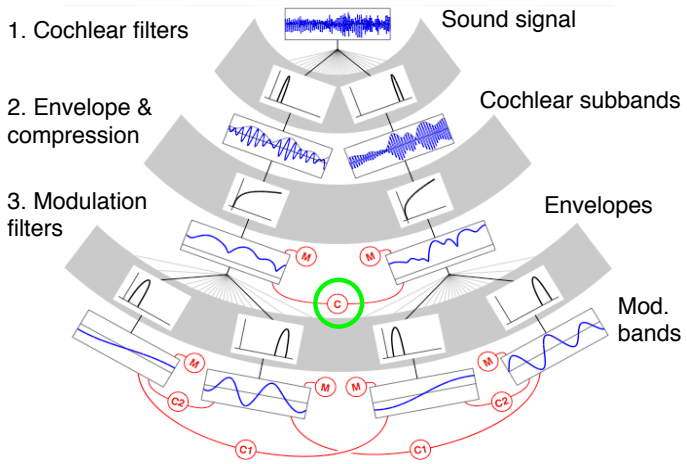
Intuition: natural sounds contain events (raindrops, geese calls)

These events are infrequent,
but when they occur, they produce large amplitudes.

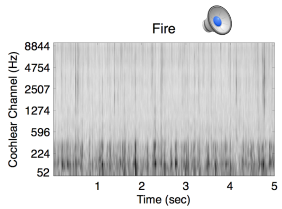


cf Attias and Schreiner

Sparsity reflected in envelope moments (cf Strickland, Lorenzi).

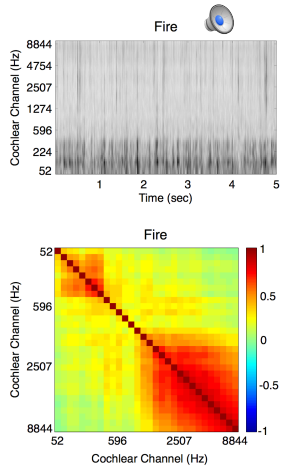


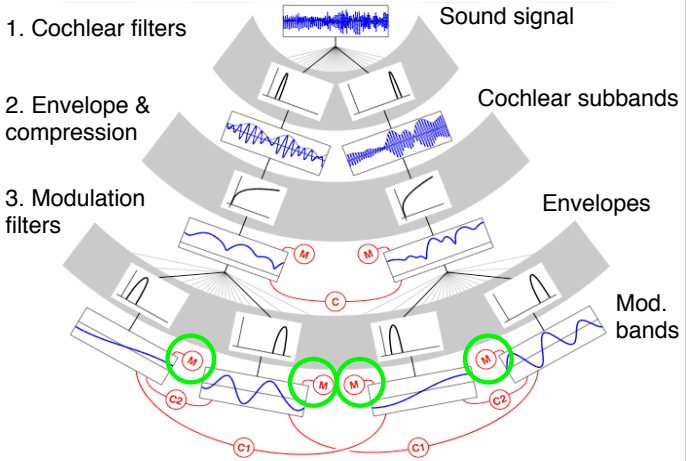
Correlations between envelopes vary across sounds.



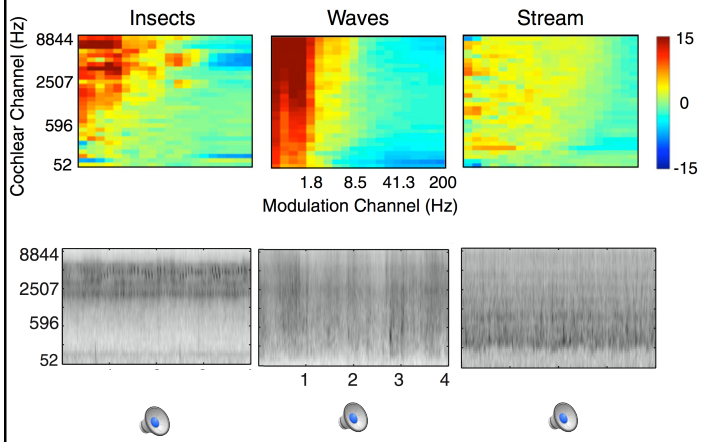
Broadband events induce dependencies between channels.

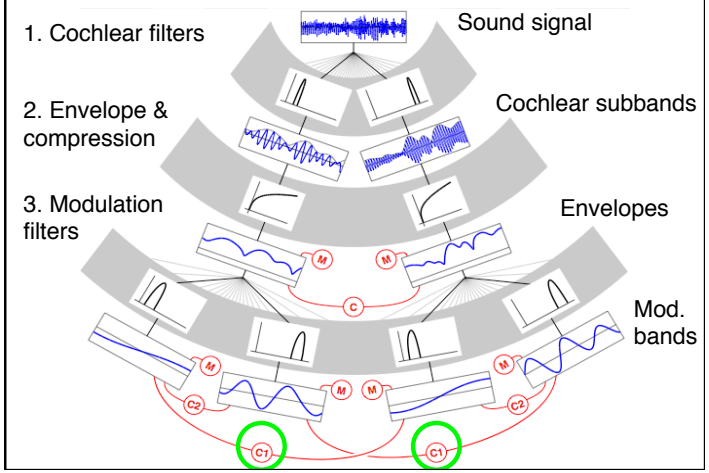
Correlations reflect broadband events (crackles, claps):

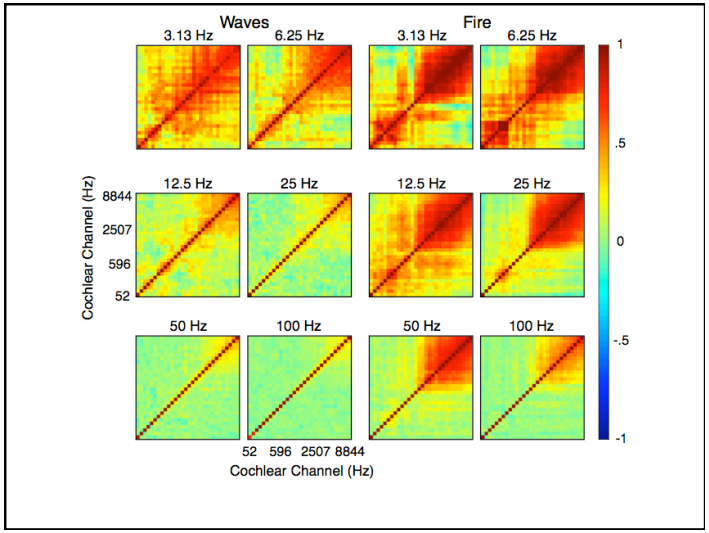


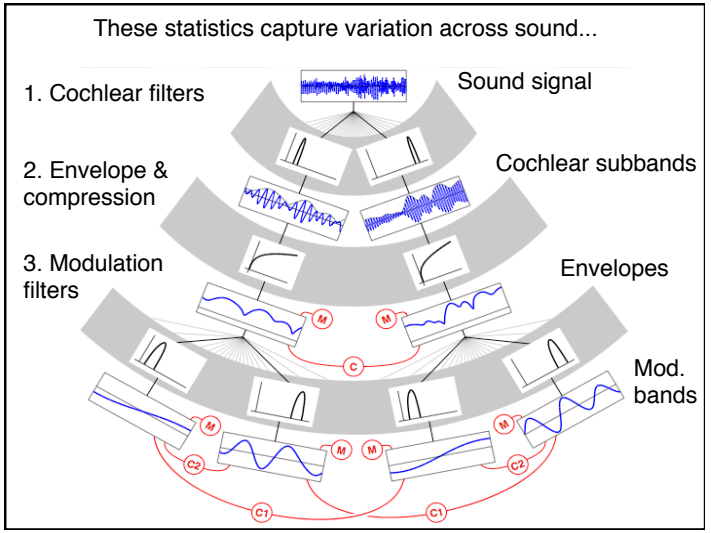


Textures vary in distribution of modulation power:









Can they account for perception of real-world textures?

Key Methodological Proposal:

•Synthesis is a powerful way to test a perceptual theory.

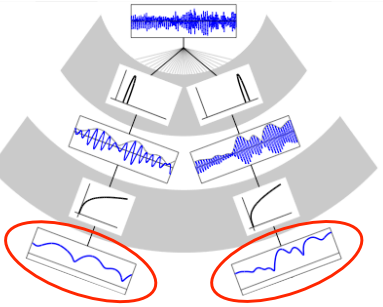
•If your brain represents sounds with a set of measurements, then:

Signals with the same values of those measurements should sound the same.

•Sounds synthesized to have the same measurements as a real-world recording should sound like it…

IF the measurements are what the brain is using to represent sound.

Simple example: test the role of the mean of each cochlear envelope (power spectrum)

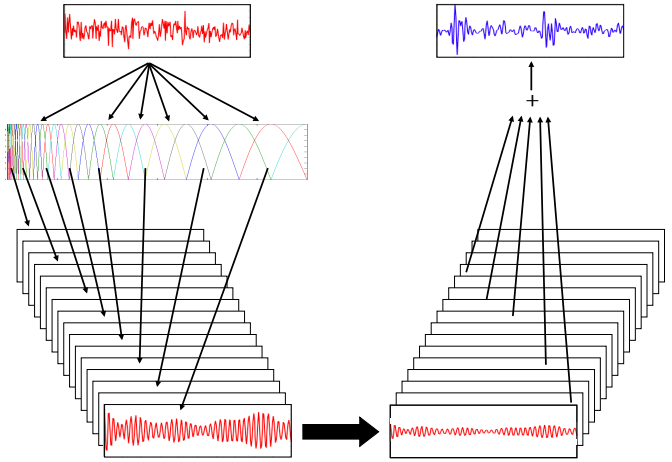


Envelopes

•Measure average value of each envelope in real-world texture

•Then synthesize random signal with same envelope means.

Start with noise, rescale noise subbands, resynthesize:



What do they sound like?

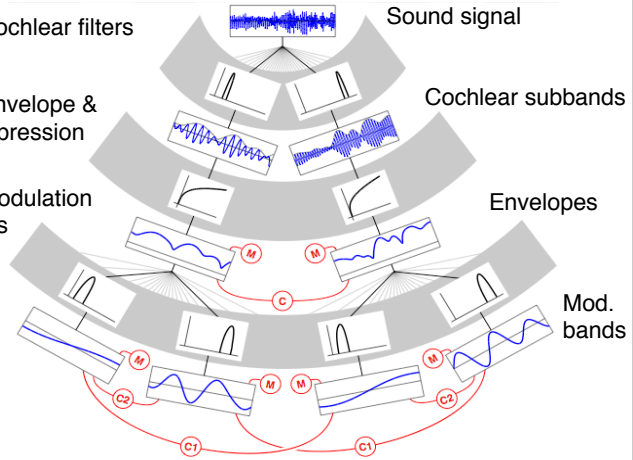
	Original	Power
Rain		
Stream		
Bubbles		
Fire		
Applause		

•Synthesis is not realistic (everything sounds like noise):

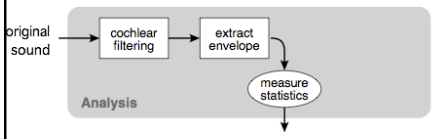
•We aren't simply registering the spectrum (mean values of envelopes) when we recognize textures.

Will additional simple statistics do any better?

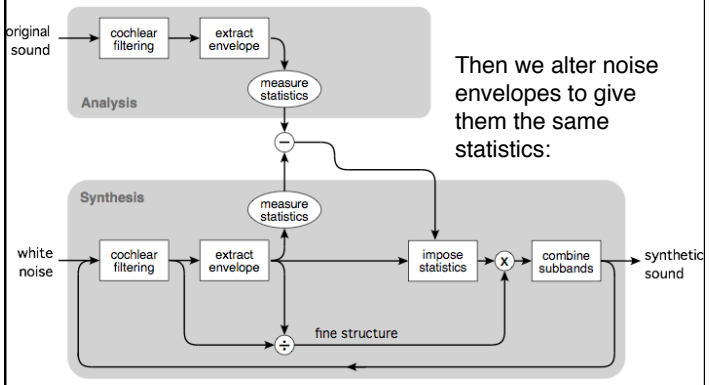
1. Cochlear filters
2. Envelope & compression
3. Modulation filters



First, we measure the statistics of a real-world sound texture:



First, we measure the statistics of a real-world sound texture:



Then we alter noise envelopes to give them the same statistics:

*synthesis code now available

McDermott & Simoncelli, Neuron, 2011

The result: a signal that shares the statistics of a real-world sound.

How do they sound?

If statistics account for texture perception, synthetic signals should sound like new examples of the real thing…

With marginal moments and pairwise correlations, synthesis is often compelling:

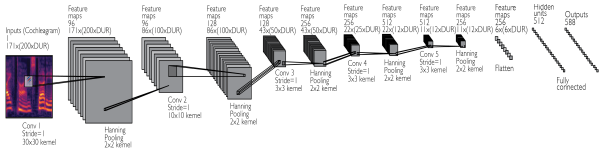
	Original	Power		All Stats
Rain				
Stream				
Bubbles				
Fire				
Applause				
Wind				
Insects				
Birds				
Crowd				

Also works for many “unnatural” sounds:

	Original	Power	All Stats
Rustling Paper			
Train			
Helicopter			
Jackhammer			

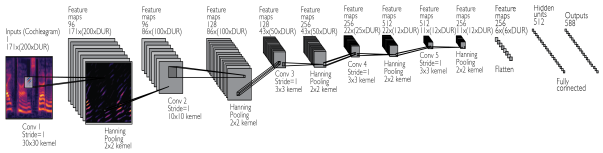
Success of synthesis suggests these statistics could underlie representation and recognition of textures.

Can we learn good representations of texture?



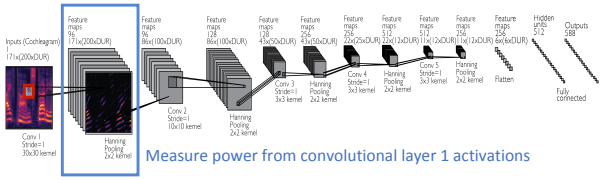
Jenelle Feather

Can we learn good representations of texture?



Jenelle Feather

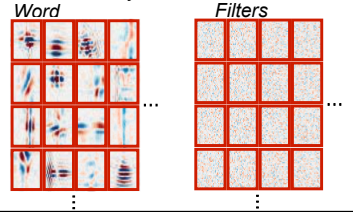
Can we learn good representations of texture?



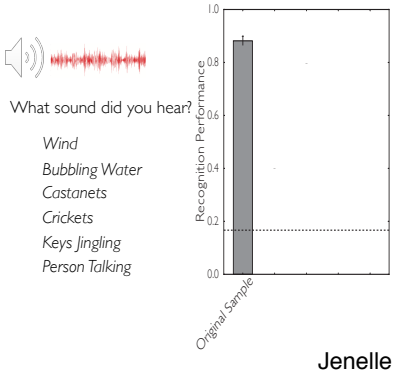
Measure power from convolutional layer 1 activations

Task : Classify Word

Random Filters

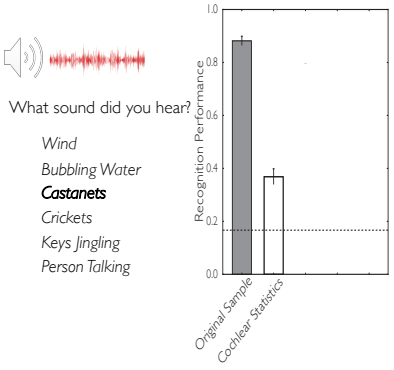


Recognition of synthetic textures:

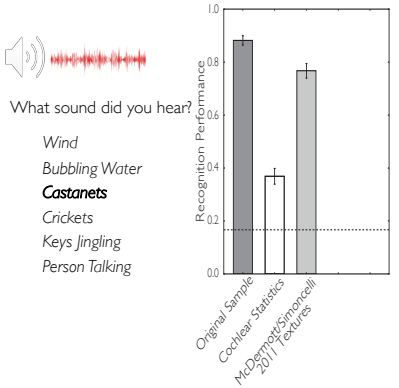


Jenelle Feather

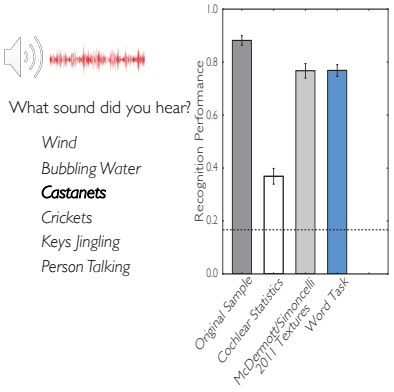
Recognition of synthetic textures:



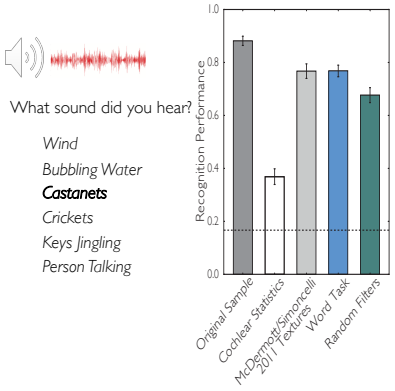
Recognition of synthetic textures:



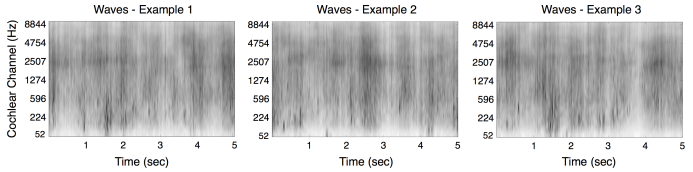
First layer of convolutional filters produce recognizable textures:



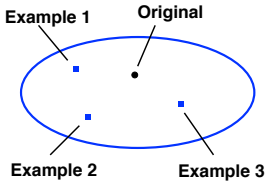
Learned filters are important:



Many sound signals have the same statistics:

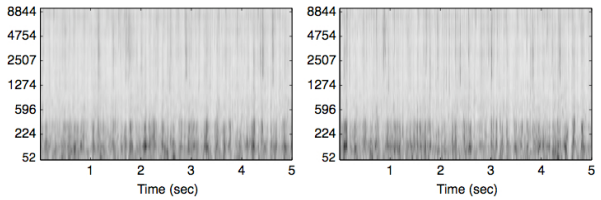


•Time-averages provide invariance.

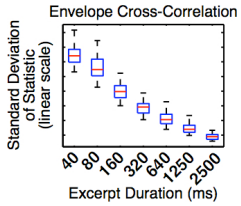


Interesting possibility:

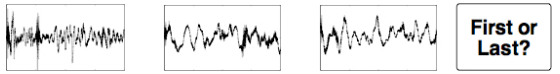
If the brain just represents time-averaged statistics, different exemplars of a texture should be difficult to discriminate



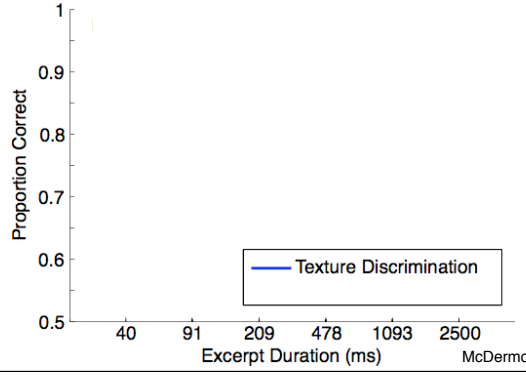
But only for long excerpts:



First, discrimination of different textures (diff. long-term statistics):

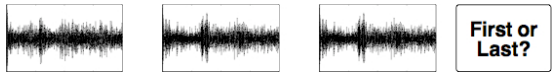


Which sound was produced by a different source?

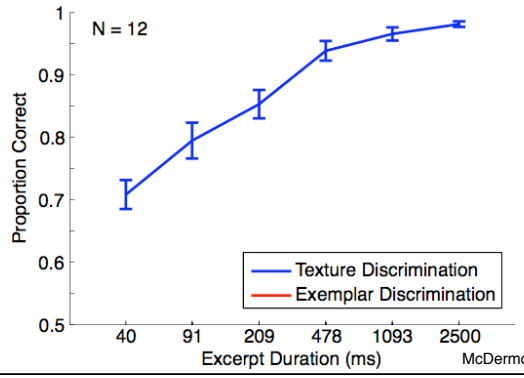


McDermott et al. 2013

What about excerpts of same texture (same long-term statistics)?

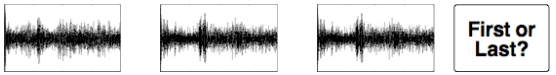


Which sound was different from the other two?

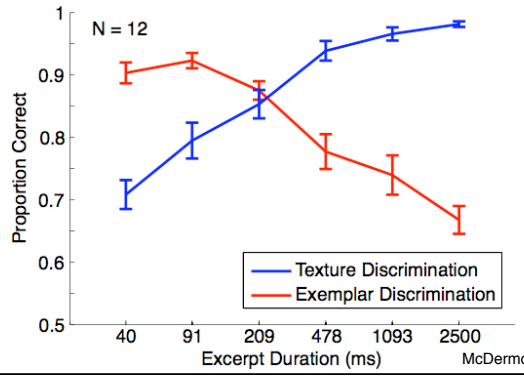


McDermott et al. 2013

Although information increases w. duration, performance declines!



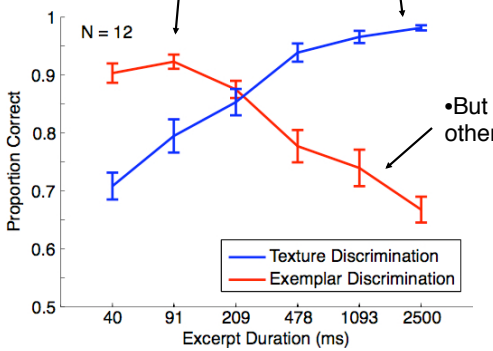
Which sound was different from the other two?



McDermott et al. 2013

•Textures contain discriminable detail…

•Details are apparently accrued into statistics…

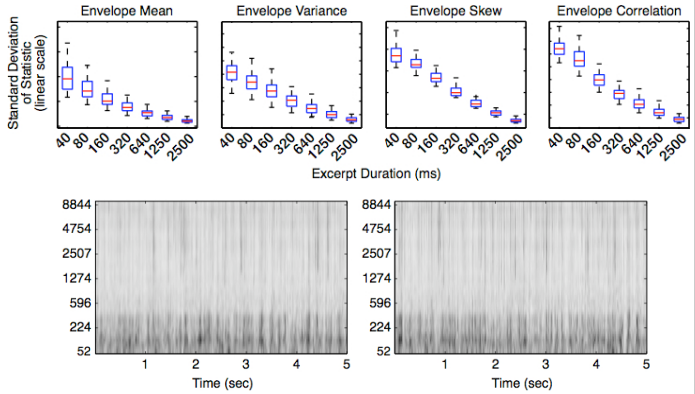


•But details are not otherwise retained.

McDermott et al. 2013

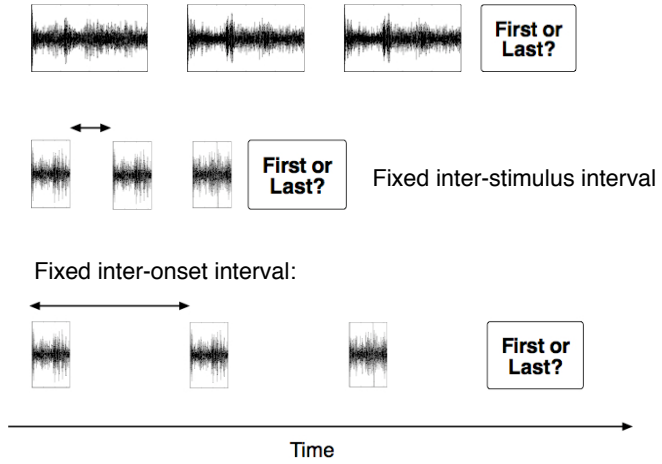
Results suggest representation of time-averaged statistics.

(because statistics converge as duration increases)

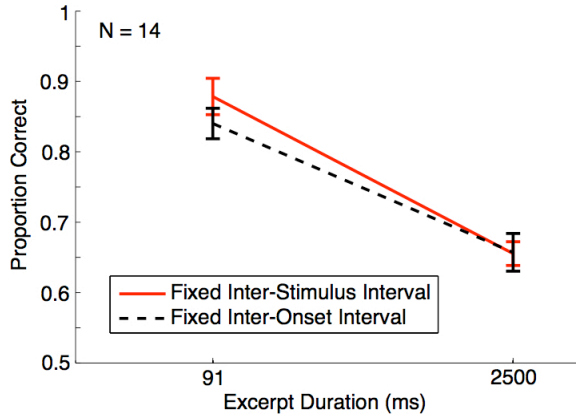


McDermott et al. 2013

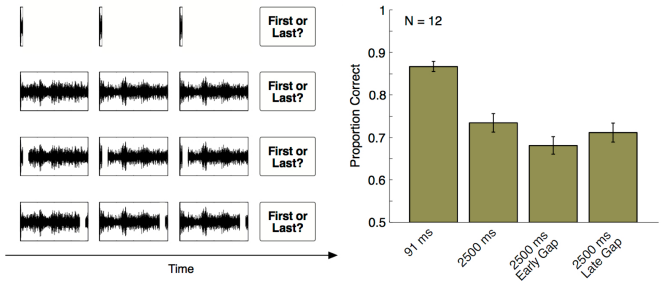
Could difference be due to decay of memory for detail with time?



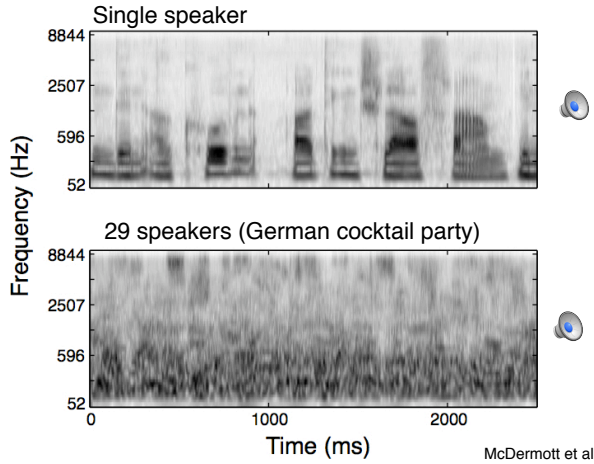
Time delay per se has little effect:



Cueing subjects to beginning or end has little effect:

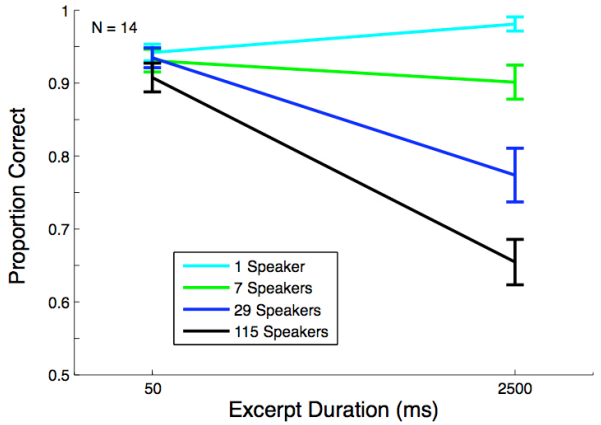


Textures are normally generated from superposition of sources…

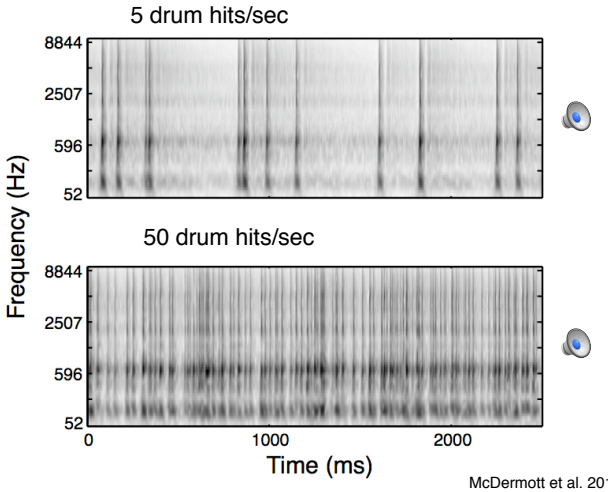


McDermott et al. 2013

Impairment at long durations is specific to textures, not present for single sources:

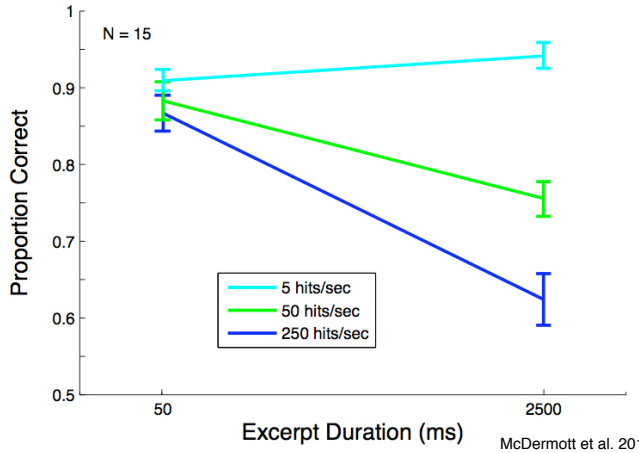


McDermott et al. 2013



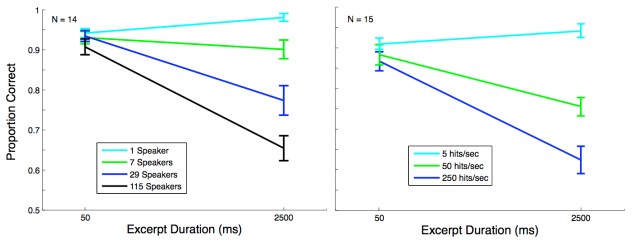
McDermott et al. 2013

Impairment at long durations again specific to dense textures.



McDermott et al. 2013

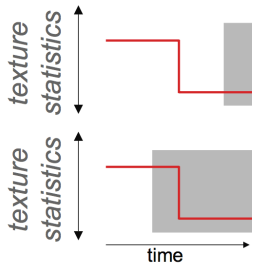
- High performance with short excerpts indicates that all stimuli have discriminable variation.
- But temporal detail is not retained when signals are texture-like.
- Suggests different representations of “stuff” and “things”.



McDermott et al. 2013

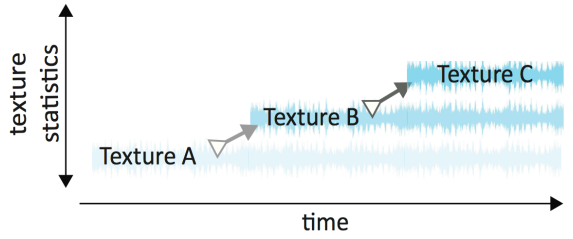
Over what time window are statistics averaged?

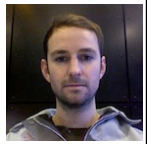
Idea: present signals whose statistics undergo a change:



Judgments of end state of texture should be biased depending on length of averaging window.

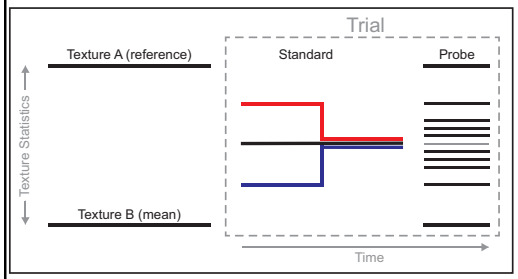
Changes in statistics enabled by morphing:





Richard McWalter

Experimental Paradigm

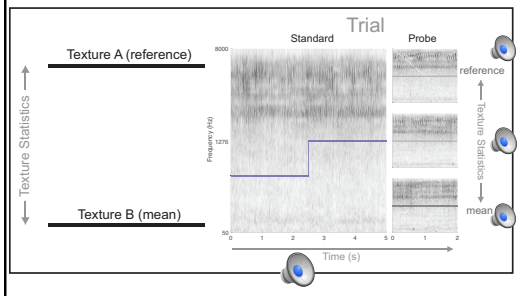


Which signal was closer to reference texture?

(Base judgments on end of standard)

McWalter & McDermott, 2018

Experimental Paradigm

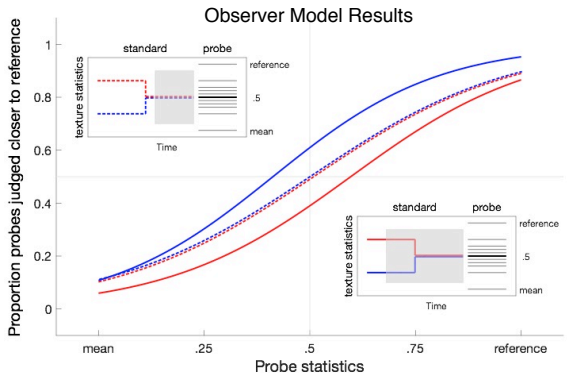


Which signal was closer to reference texture?

(Base judgments on end of standard)

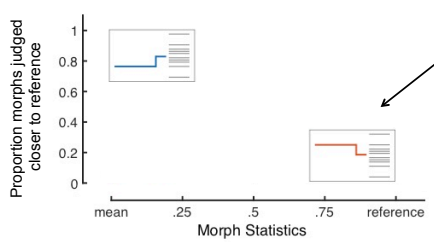
McWalter & McDermott, 2018

Bias in judgments should depend on averaging window extent:



McWalter & McDermott, 2018

Bias in judgments indicates integration of stimulus history:

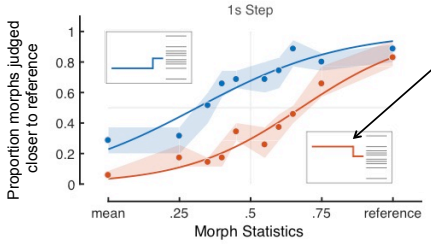


Step 1s before end:

Which signal was closer to reference texture?

McWalter & McDermott, 2018

Bias in judgments indicates integration of stimulus history:

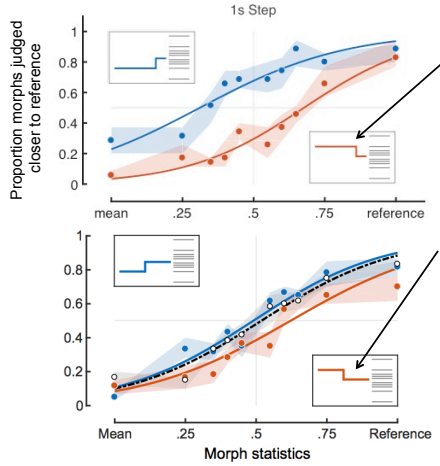


Step 1s before end:

Which signal was closer to reference texture?

McWalter & McDermott, 2018

Bias in judgments indicates integration of stimulus history:



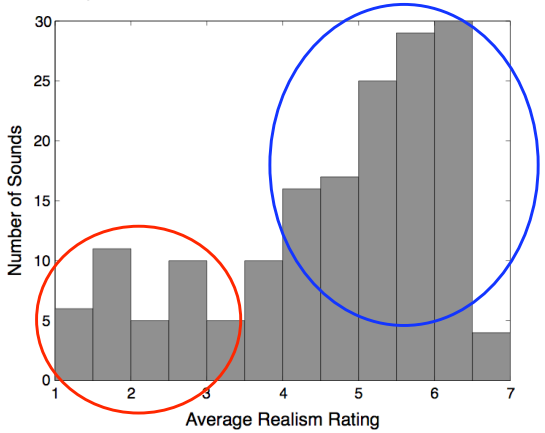
Step 1s before end:

Which signal was closer to reference texture?

Step 2.5s before end mostly does not affect judgments.

Suggestive of averaging window of several seconds.

Listen to original, then synthetic; rate realism. (170 sounds)



McDermott and Simoncelli, Neuron, 2011

Lowest rated sounds are among most interesting, as they imply brain is measuring something model is not:

Pitch	1.93	Railroad crossing
Rhythm	1.90	Tapping rhythm - quarter note pairs
Pitch	1.77	Wind chimes
Reverb	1.77	Running up stairs
Rhythm	1.70	Tapping rhythm - quarter note triplets
Reverb	1.67	Snare drum beats
	1.63	Walking on gravel
Reverb	1.60	Snare drum rimshot sequence
Rhythm	1.60	Music - drum break
Pitch	1.50	Music - mambo
Rhythm	1.50	Bongo drum loop
Reverb	1.47	Firecracker explosions
Pitch	1.40	Person speaking French
Pitch	1.37	Church bells
Pitch	1.20	Person speaking English

TAKE-HOME MESSAGES

•Sound synthesis can help us test/explore theories of audition.
- -variables that produce compelling synthesis could underlie perception.
- -synthesis failures point the way to new variables that might be important for the perceptual system.

•Textures are a nice point of entry into real-world hearing

•Many natural sounds may be recognized with relatively simple statistics of early auditory representations
- -simplest statistics (spectrum) are not that informative
- -slightly more complex statistics are quite powerful
- -for textures of moderate length, statistics may be all we retain

OPEN QUESTIONS

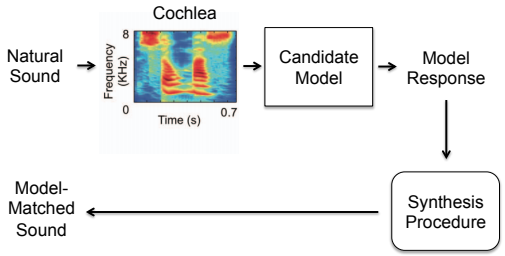
•Locus of time-averaging?
- •Presumptive integration windows of several seconds are long relative to typical timescales in auditory system…

•Relation to scene analysis?
- •What happens when foreground sounds are superimposed on a texture?

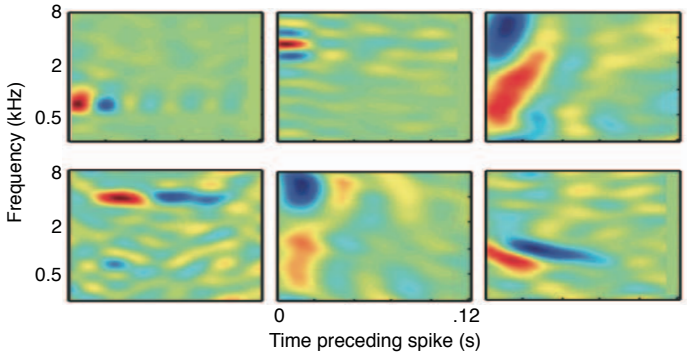
•What statistics are needed to account for synthesis failures?

Another application of model-based synthesis:

- Idea: present a natural sound, and a synthetic signal that produces same response in a model
- If model is good description of neural response, responses to natural and model-matched sounds should be similar.

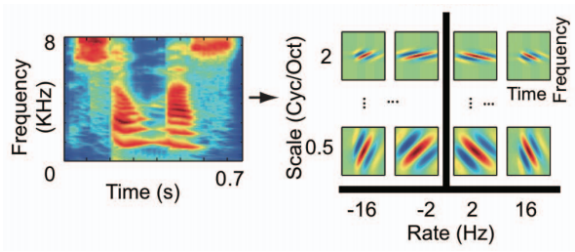


Linear filter characterizations in auditory cortex of animals (w.r.t. spectrogram):



Mesgarani et al. 2007

Common model of auditory cortex: linear spectrotemporal filtering

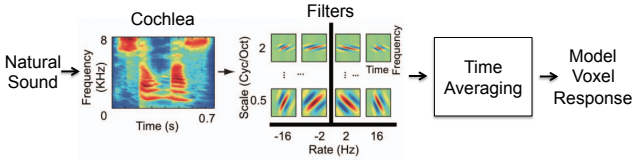


Very useful (and widely used), but clearly limited.

Shamma and colleagues

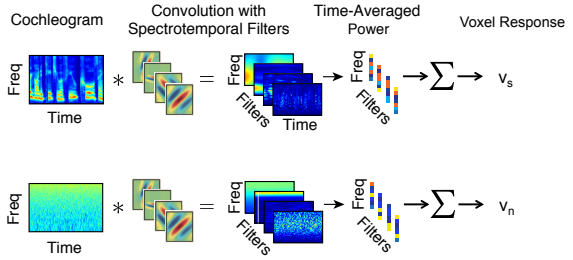
Another application of model-based synthesis:

- Idea: present a natural sound, and a synthetic signal that produces same response in a model
- If model is good description of neural response, responses to natural and model-matched sounds should be similar.
- To test models with fMRI, we model the BOLD signal from a voxel as a sum of time-averaged responses of each unit (filter) in model.



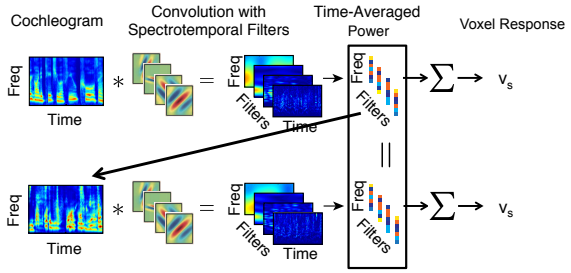
– Work by Sam Norman-Haignere

Model-Matched Synthetic Stimuli



Measure time-averaged statistics from a natural sound
Measure same statistics for a noise sound

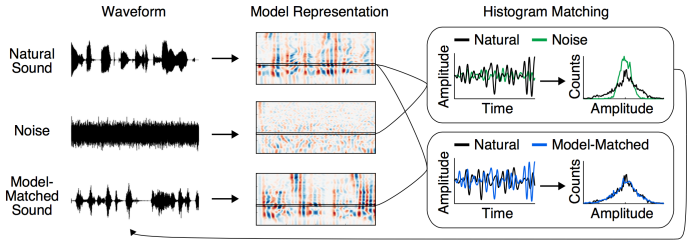
Model-Matched Synthetic Stimuli



Measure time-averaged statistics from a natural sound
Measure same statistics for a noise sound
Coerce a noise signal to have the matched statistics
⇒ via texture-synthesis procedure (e.g. McDermott & Simoncelli, 2011)

Synthesis of model-matched sounds

Synthesis algorithm used relies on histogram matching (Heeger and Bergen, 1995)



Matches time-average of any instantaneous function of the filter response.
(matches power, but also other moments)

$$\int g(u_k(t, s_{mm}))dt = \int g(u_k(t, s_{nat}))dt \quad \forall k$$

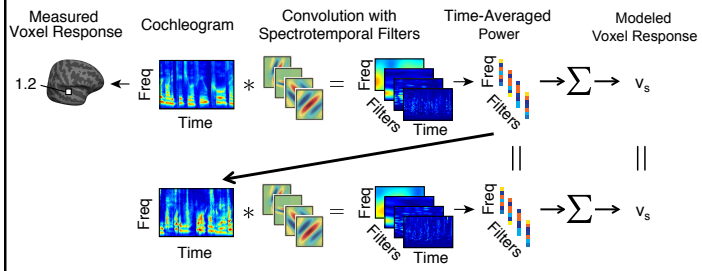
Synthesis of model-matched sounds

Synthesis algorithm used relies on histogram matching (Heeger and Bergen, 1995)

Matches time-average of any instantaneous function of the filter response.
(matches power, but also other moments)

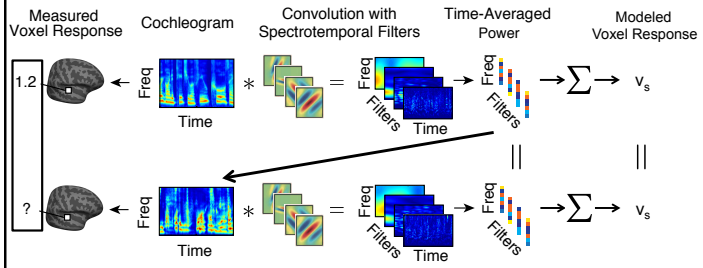
$$\int g(u_k(t, s_{mm}))dt = \int g(u_k(t, s_{nat}))dt \quad \forall k$$

Model-Matched Synthetic Stimuli



Natural and synthetic sound by definition give the same model response.

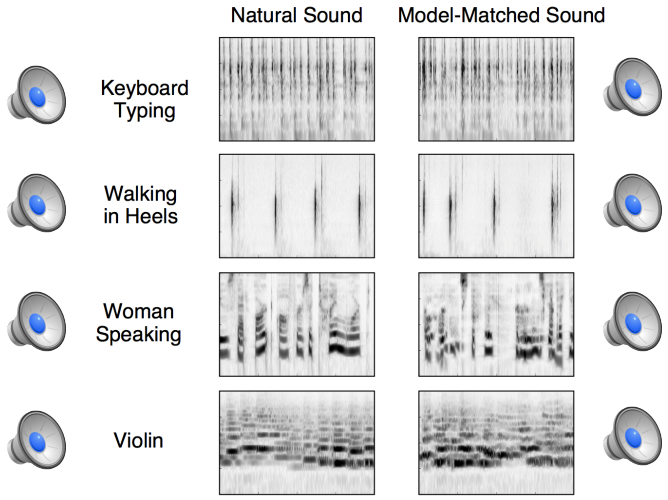
Model-Matched Synthetic Stimuli



Natural and synthetic sound by definition give the same model response

Similar neural response?

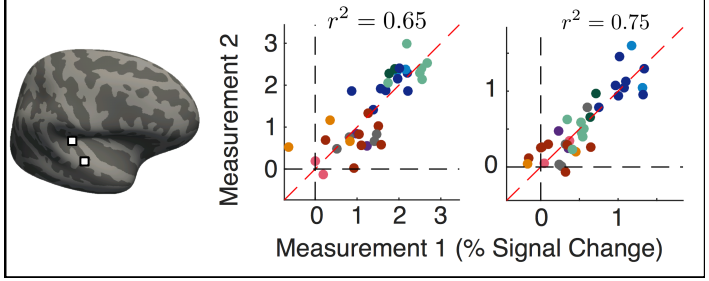
⇒ If model is accurate, neural response should also be matched



Approach: Measure and model tuning of individual voxels

- We measure fMRI response in each voxel to each sound
- Sounds are short (2s), so response is one number

- Individual voxels exhibit reliable responses to sounds:



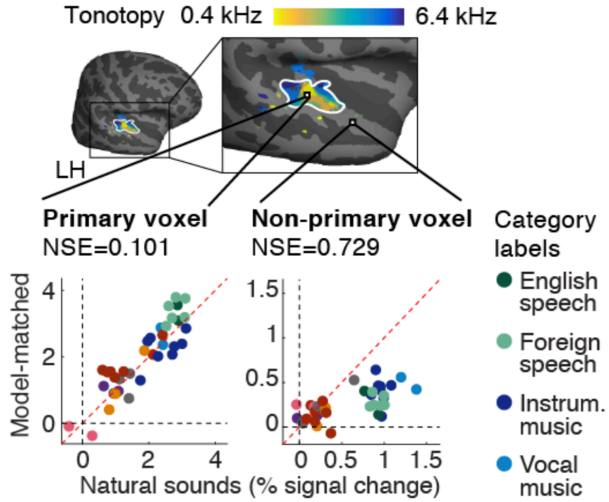
Participants heard real and model-matched versions of these sounds:

1. Woman speaking
2. Man speaking
3. Spanish
4. French
5. Italian
6. German
7. Hindi
8. Russian
9. Big band music
10. Bluegrass
11. Cello
12. Orchestra
13. Piano
14. Saxophone
15. Violin
16. Latin music
17. Country song
18. R&B song
19. Biting & chewing
20. Finger tapping
21. Walking on leaves
22. Scratching
23. Walking in heels
24. Writing on paper
25. Heart beat
26. Cicadas
27. Crickets
28. Baby Crying
29. Breathing
30. Clock ticking
31. Siren
32. Keyboard Typing
33. Chimes
34. Chopping food
35. Crumpling paper
36. Keys jingling

Category Labels
- English Speech
- Foreign Speech
- Instr. Music
- Vocal Music
- Human NonVocal
- Animal NonVocal
- NonSpeech Voc
- Mechanical
- Env. Sounds

- We measure average response of each voxel in auditory cortex to each natural and model-matched sound.
- Yielding two 36-dimensional vectors per voxel
- We measure squared error between natural and model-matched response vector in each voxel.

- Voxels for whom model is a good description should yield low error.



Dissimilarity between voxel response to real and synth sounds:

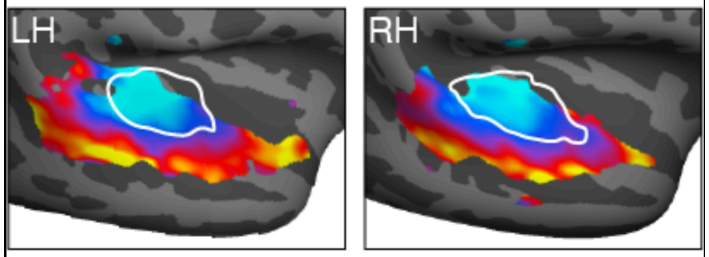
0  1

- Normalized squared error (NSE)
- Plotted for all voxels that respond reliably to sound.

Dissimilarity between voxel response to real and synth sounds:

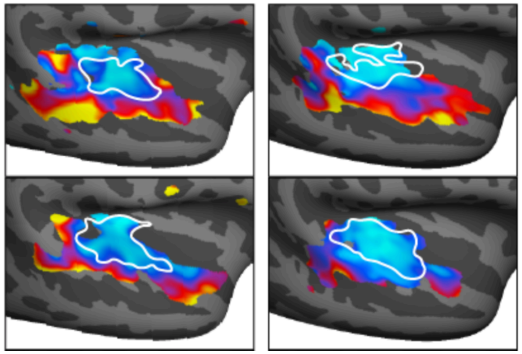
0  1

Left Hemisphere **Right Hemisphere**



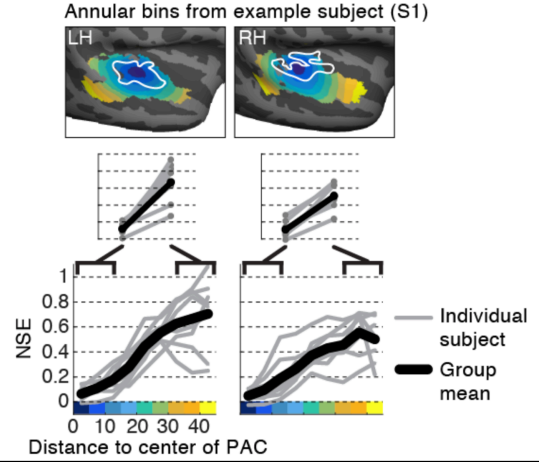
- Standard linear model accounts for much of the voxel tuning near primary auditory cortex, very little in non-primary regions.

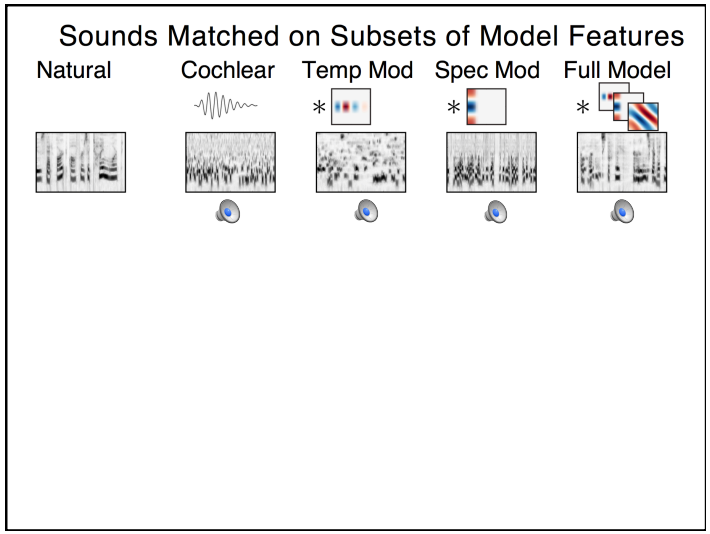
Maps from example individual subjects:

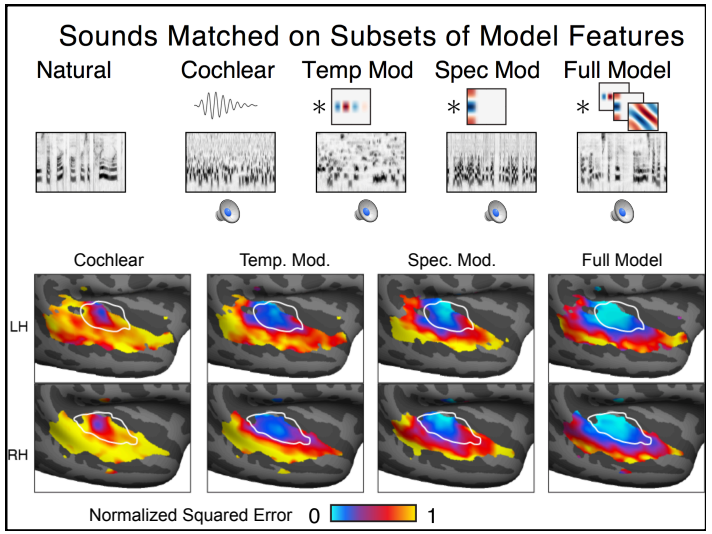


Voxels are excluded that are below a threshold of test-retest reliability (hence holes in maps).

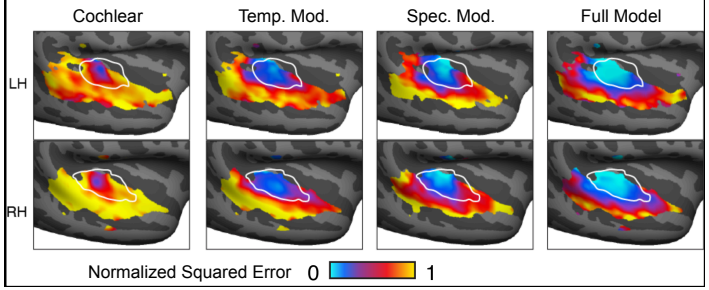
Dissimilarity vs. distance to primary auditory cortex:

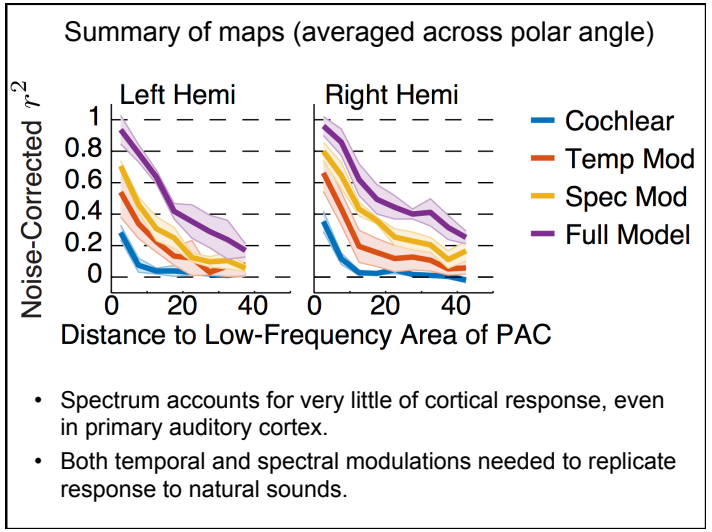






- Spectrum accounts for very little of cortical response, even in primary auditory cortex.
- Both temporal and spectral modulations needed to replicate response to natural sounds.



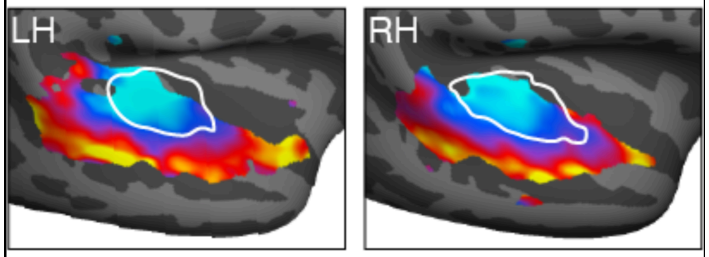


- Spectrum accounts for very little of cortical response, even in primary auditory cortex.
- Both temporal and spectral modulations needed to replicate response to natural sounds.

Dissimilarity between voxel response to real and synth sounds:

0 1

Left Hemisphere **Right Hemisphere**



- Standard linear model accounts for much of the voxel tuning near primary auditory cortex, very little in non-primary regions.

ACKNOWLEDGMENTS

Eero Simoncelli  Michael Schemitsch

Richard McWalter  Jenelle Feather 

Sam Norman-Haignere 

Howard Hughes Medical Institute
James S McDonnell Foundation
National Science Foundation