Sound, Ears, Brains and the World

Josh McDermott
Dept. of Brain and Cognitive Sciences, MIT
CBMM Summer School

Consider some examples of typical auditory input:

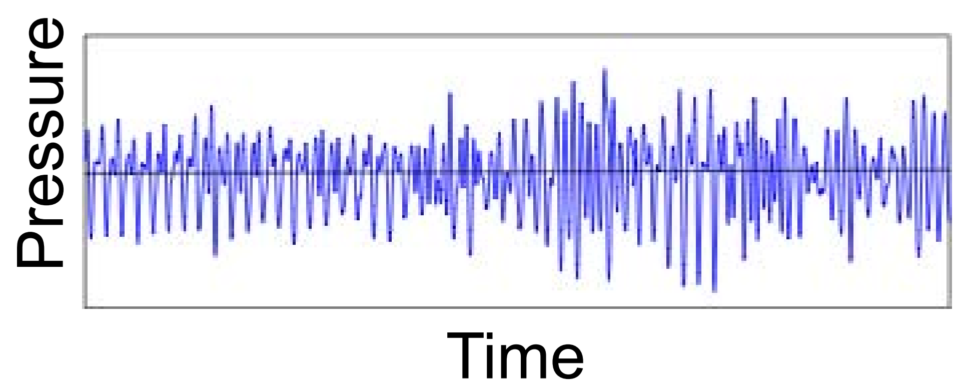
Scene from cafe:

Scene from sports bar:

Radio excerpt:

Barry White:

The ear receives a pressure waveform.



AUDITION

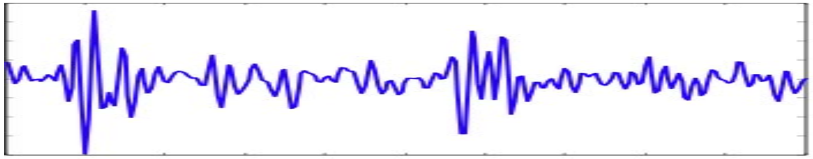
When objects in the world vibrate, they transmit acoustic energy through surrounding medium in the form of a wave.

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:

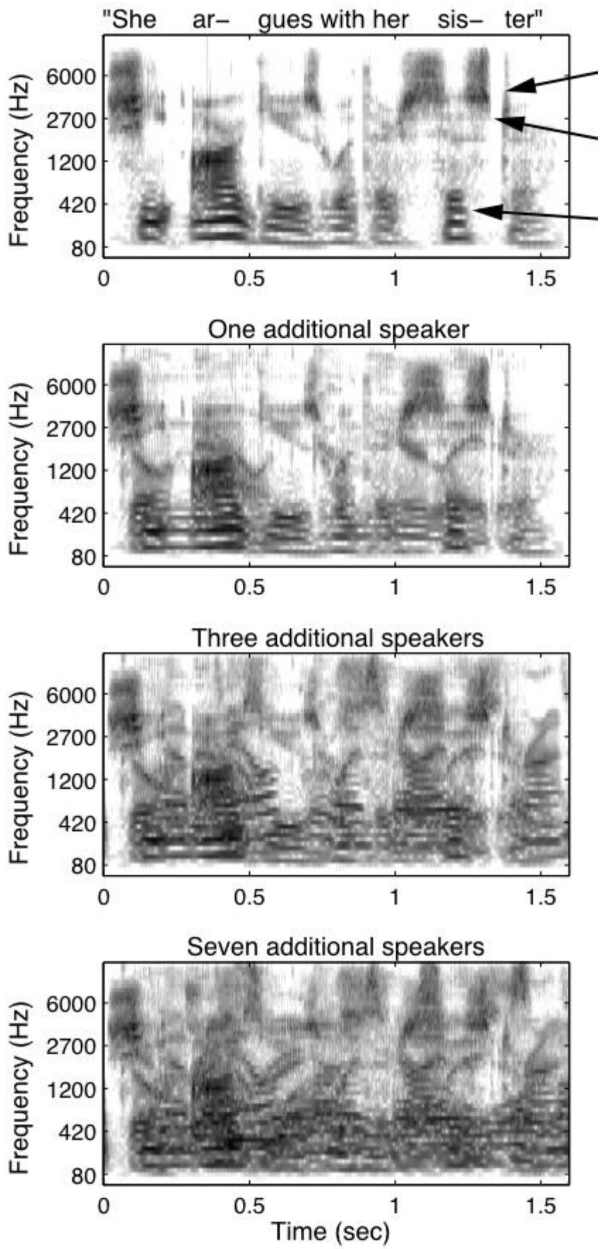
- Most properties of interest are not explicit in the waveform:



How do we derive information about the world from sound?

The Cocktail Party Problem

Real-world settings often involve concurrent sounds.



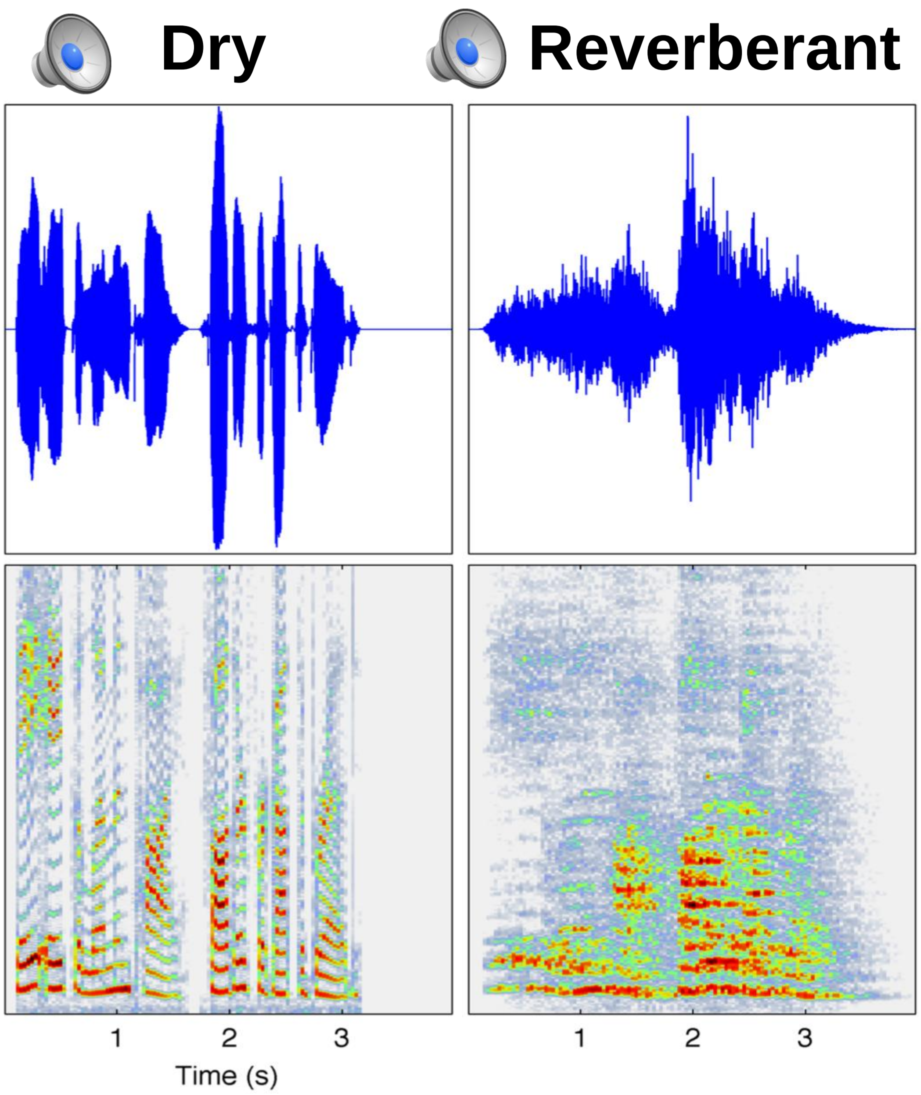
- Presence of other speakers obscures much structure of target utterance, but speech remains intelligible.

- Present-day speech recognition algorithms (e.g. in your iPhone) fall apart in such circumstances.

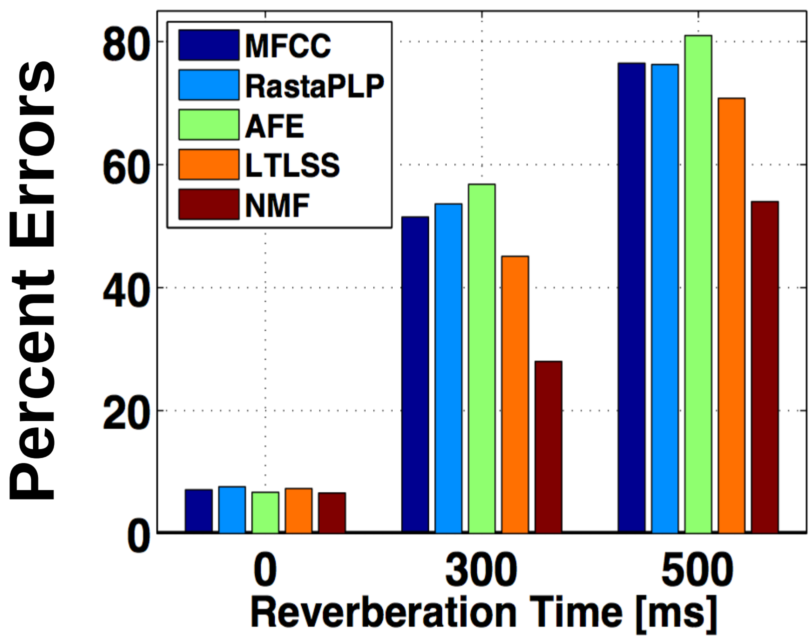
Source: McDermott, Josh H. "The cocktail party problem." Current Biology 19, no. 22 (2009): R1024-R1027.

Figure removed due to copyright restrictions. Please see the video.
Source: Chapter 8, Ochsner, Kevin, and Stephen M. Kosslyn. The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges. Vol. 2. Oxford University Press, 2013.

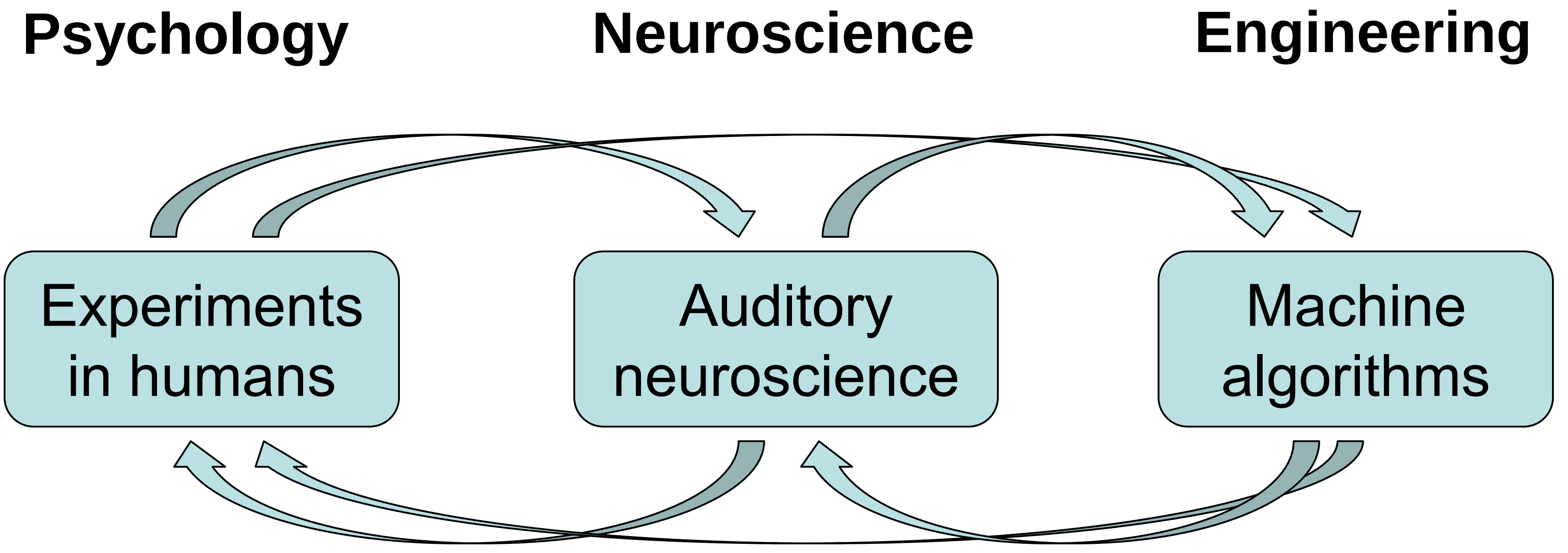
Human speech recognition is also remarkably invariant:



Machine speech recognition is not robust:

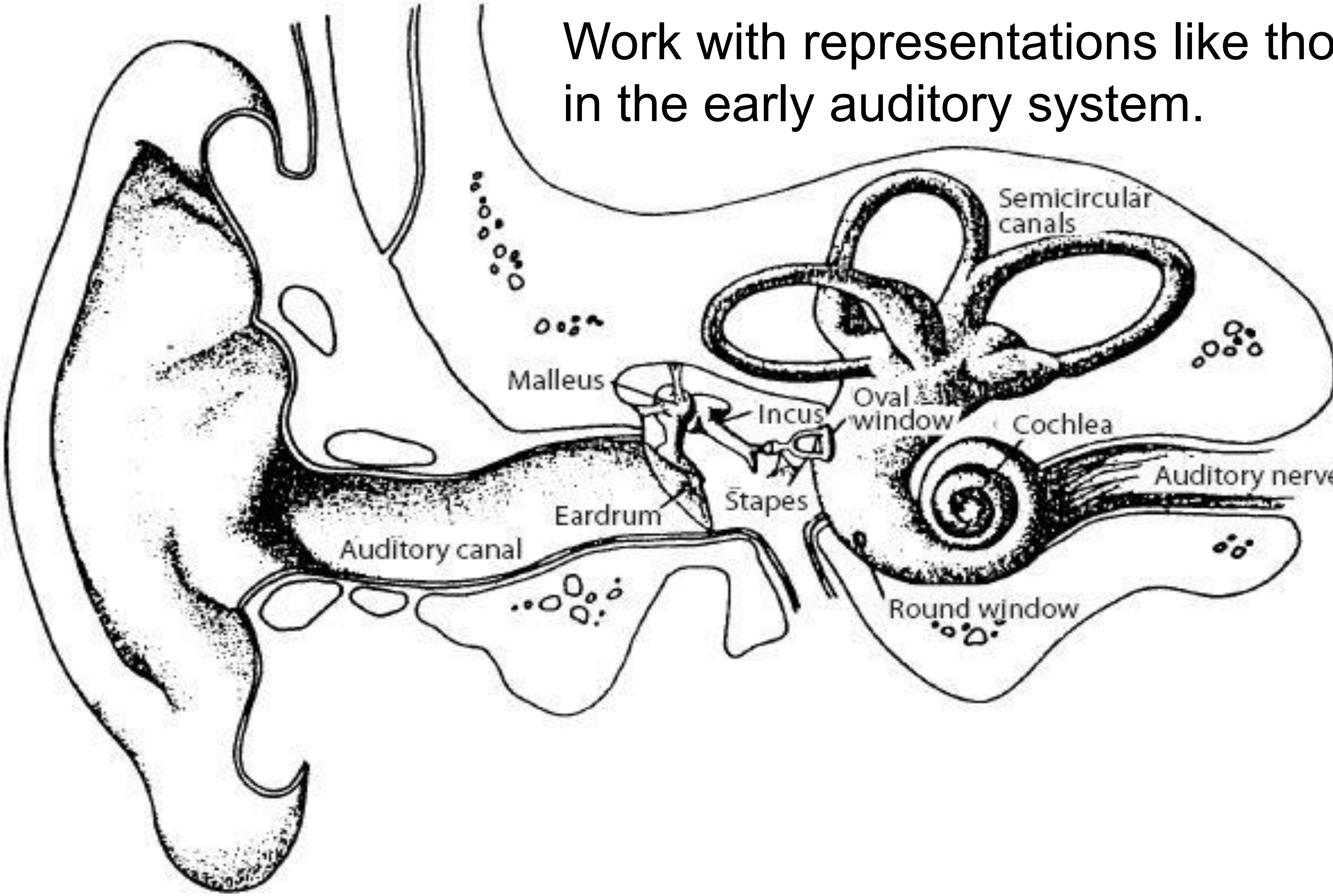


My research group: Laboratory for Computational Audition



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

Work with representations like those in the early auditory system.



Source: Lindsay, Peter H., and Donald A. Norman. Human information processing: An introduction to psychology. Academic press, 2013.

Plan for this morning:

1. Overview of Auditory System
2. Sound Texture Perception
3. Perception of Sound Sources
4. Auditory Scene Analysis

Part 1: Overview of Auditory System

Functional schematic of the ear:

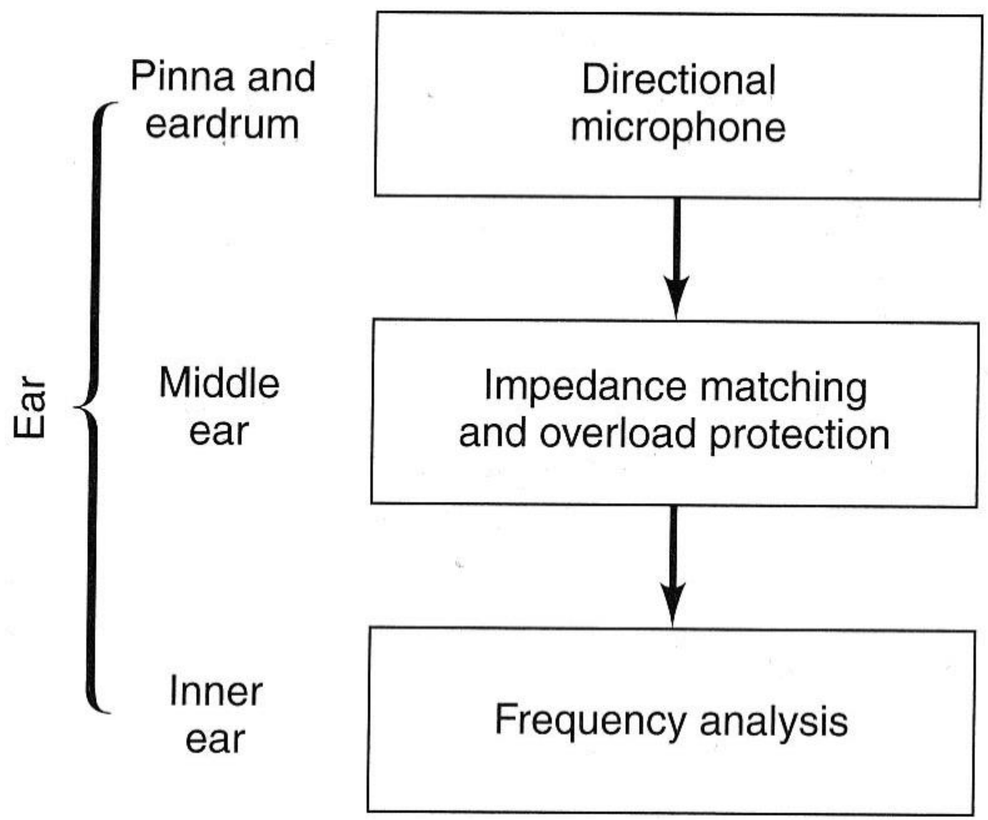
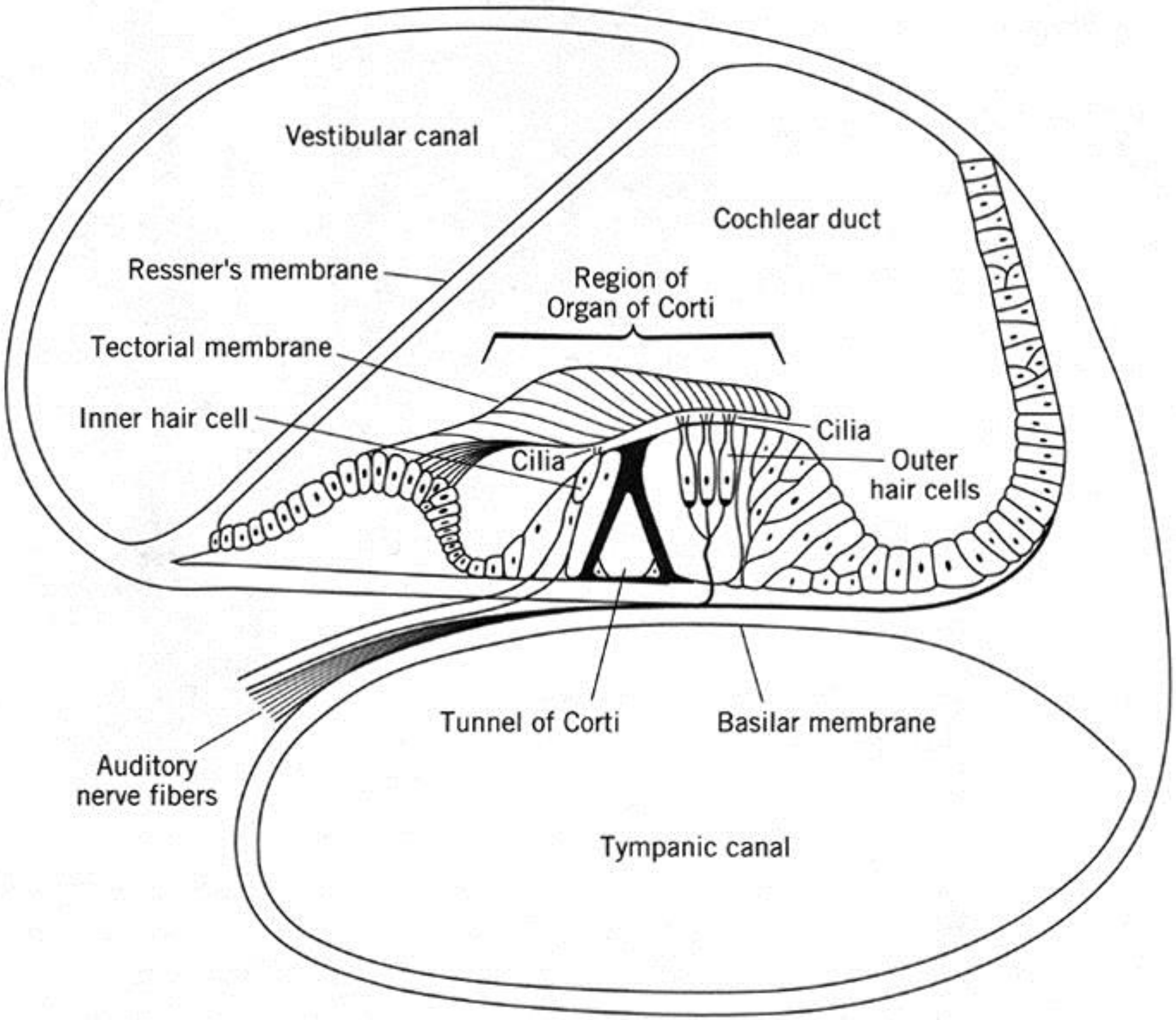


Figure removed due to copyright restrictions.
Please see the video.
Source: Chapter 8, Ochsner, Kevin, and Stephen M. Kosslyn. The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges. Vol. 2. Oxford University Press, 2013.

Cochlear transduction is frequency tuned:

Diagram of fourier transform in the basilar membrane removed due to copyright restrictions. Source: "Transmission of sound within the inner ear" from Hawkins, Joseph, "Human ear: Anatomy" Encyclopedia Britannica, Last updated on February, 24 2017. https://www.britannica.com/science/ear/Transmission-of-sound-by-bone-conduction.

Cross section of cochlea



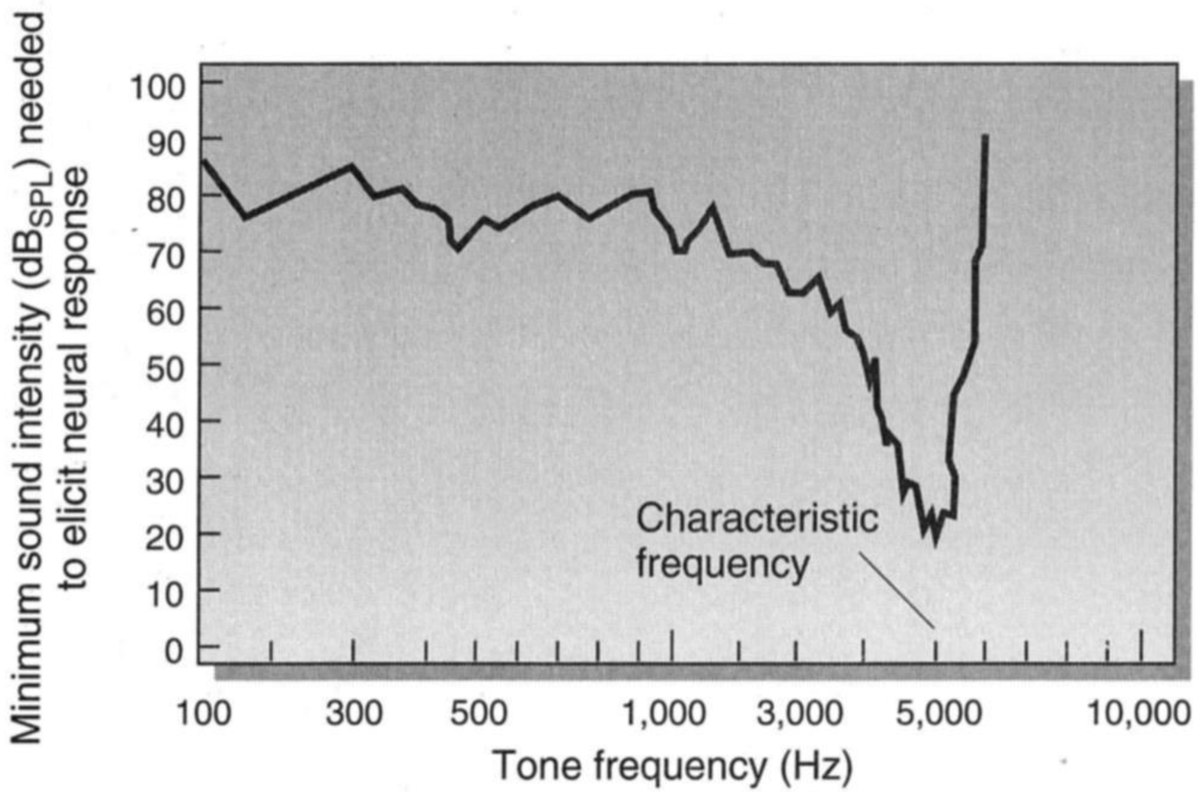
Movement of the basilar membrane causes the hair cells to move against the tectorial membrane, which causes the cilia to bend.

Diagram of the inner ear removed due to copyright restrictions.
Please see the video.

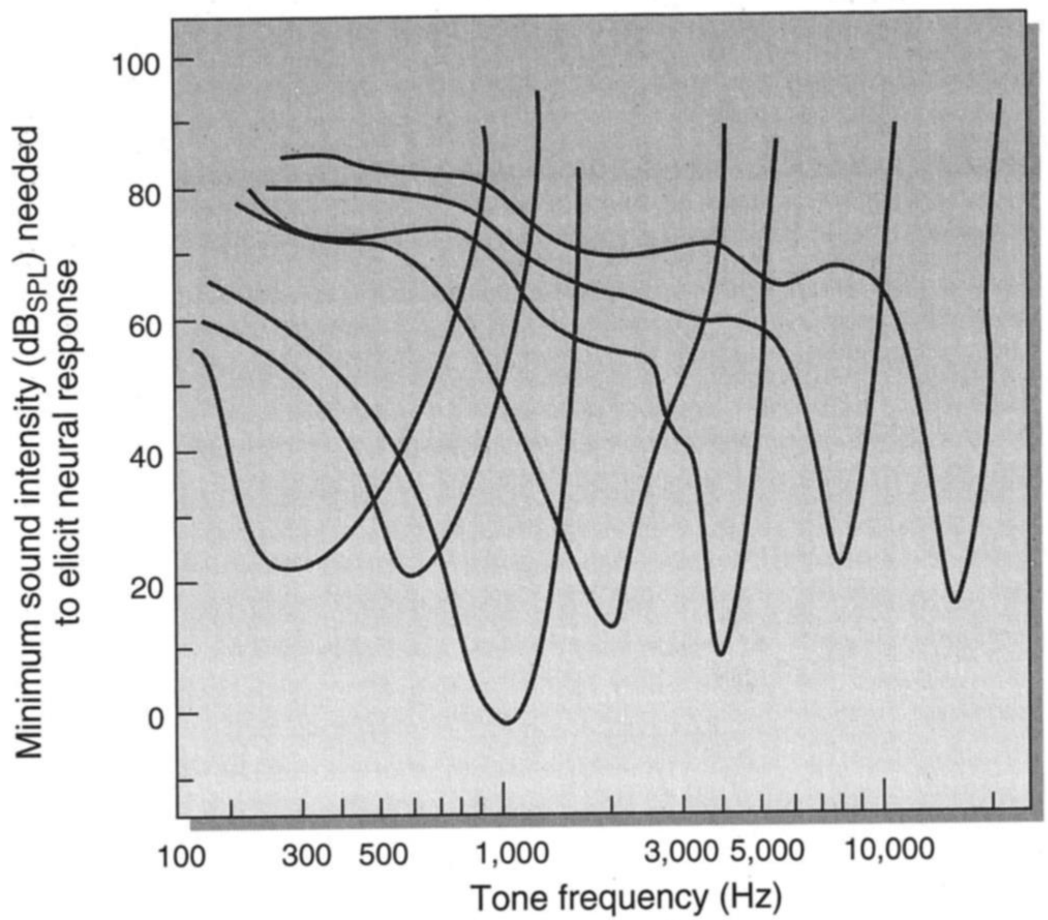
When the cilia bend, the hair cells release neurotransmitter onto synapses with auditory nerve fibers that send signals to the brain.

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.

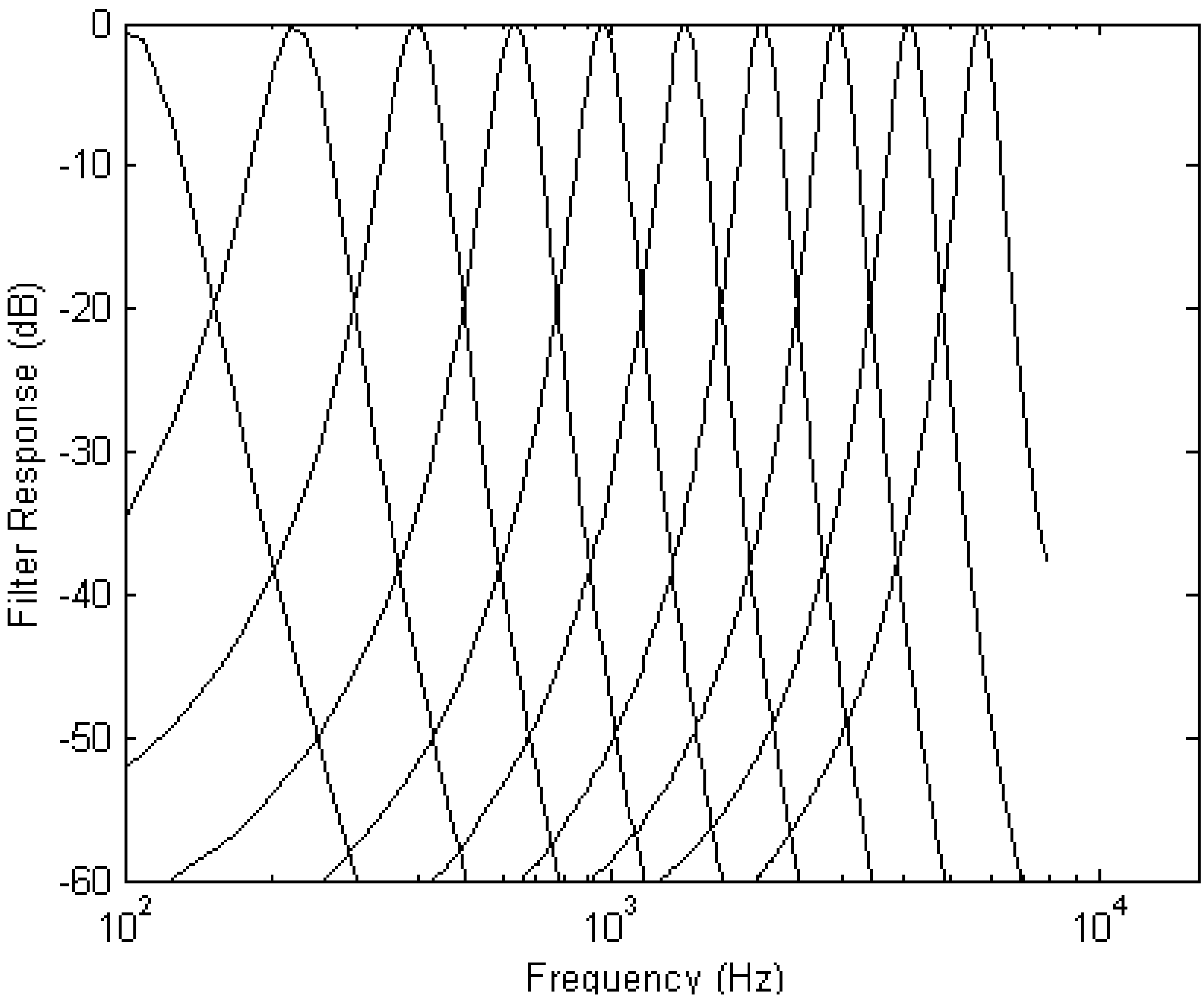
One example:



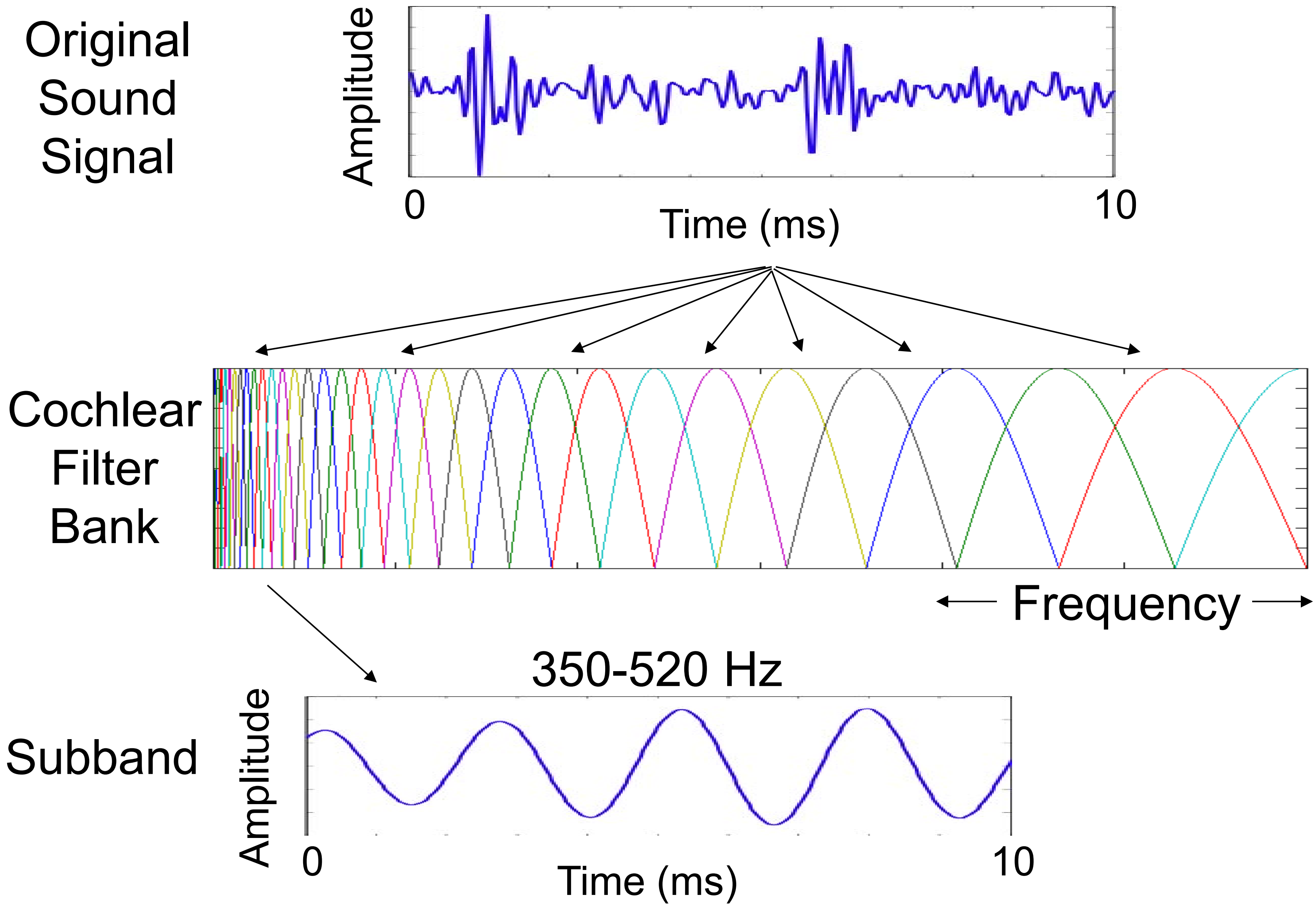
Different nerve fibers (synapsing at different points along the basilar membrane) are tuned to different frequencies:



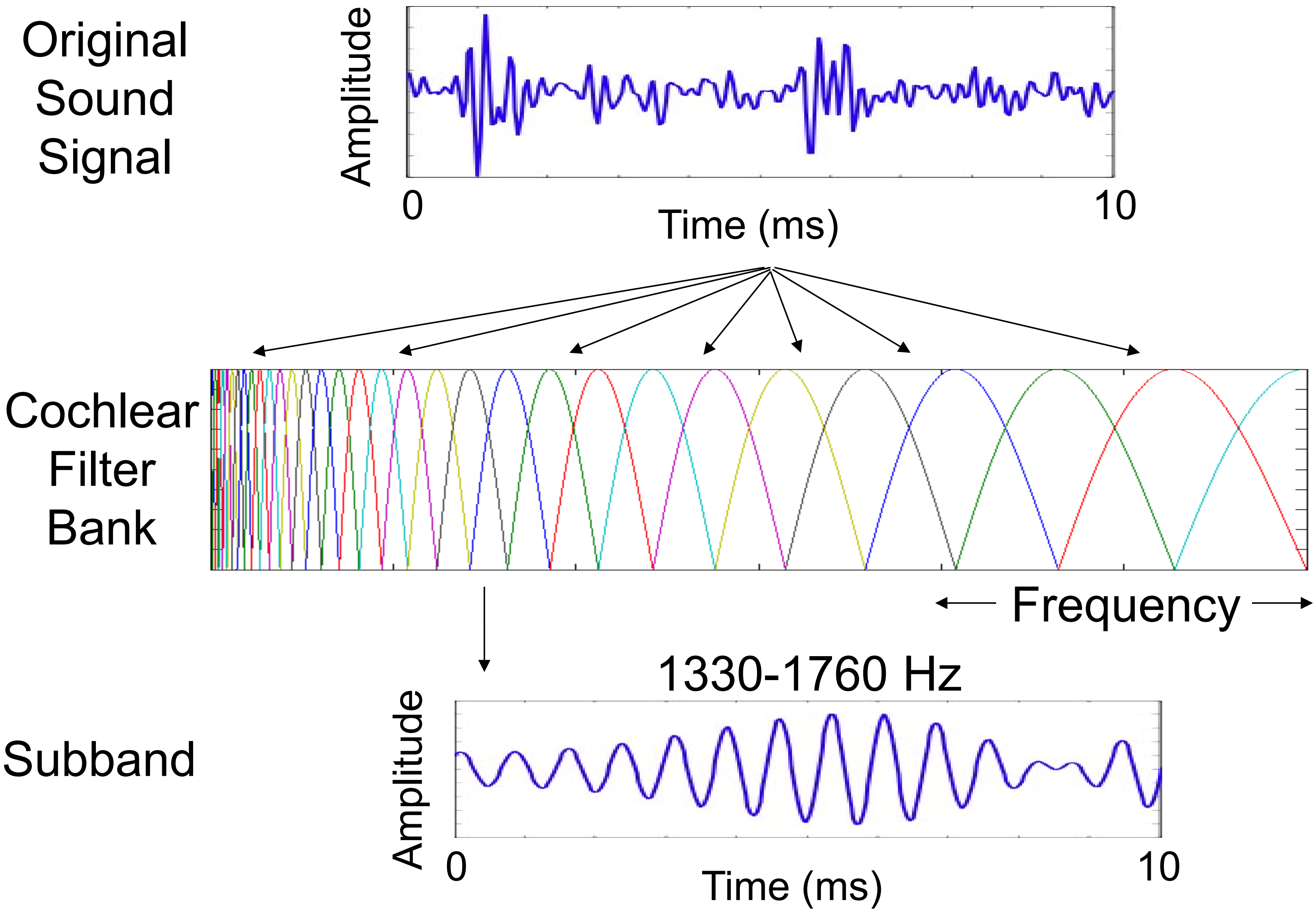
Auditory nerve fibers usually approximated as bandpass filters:



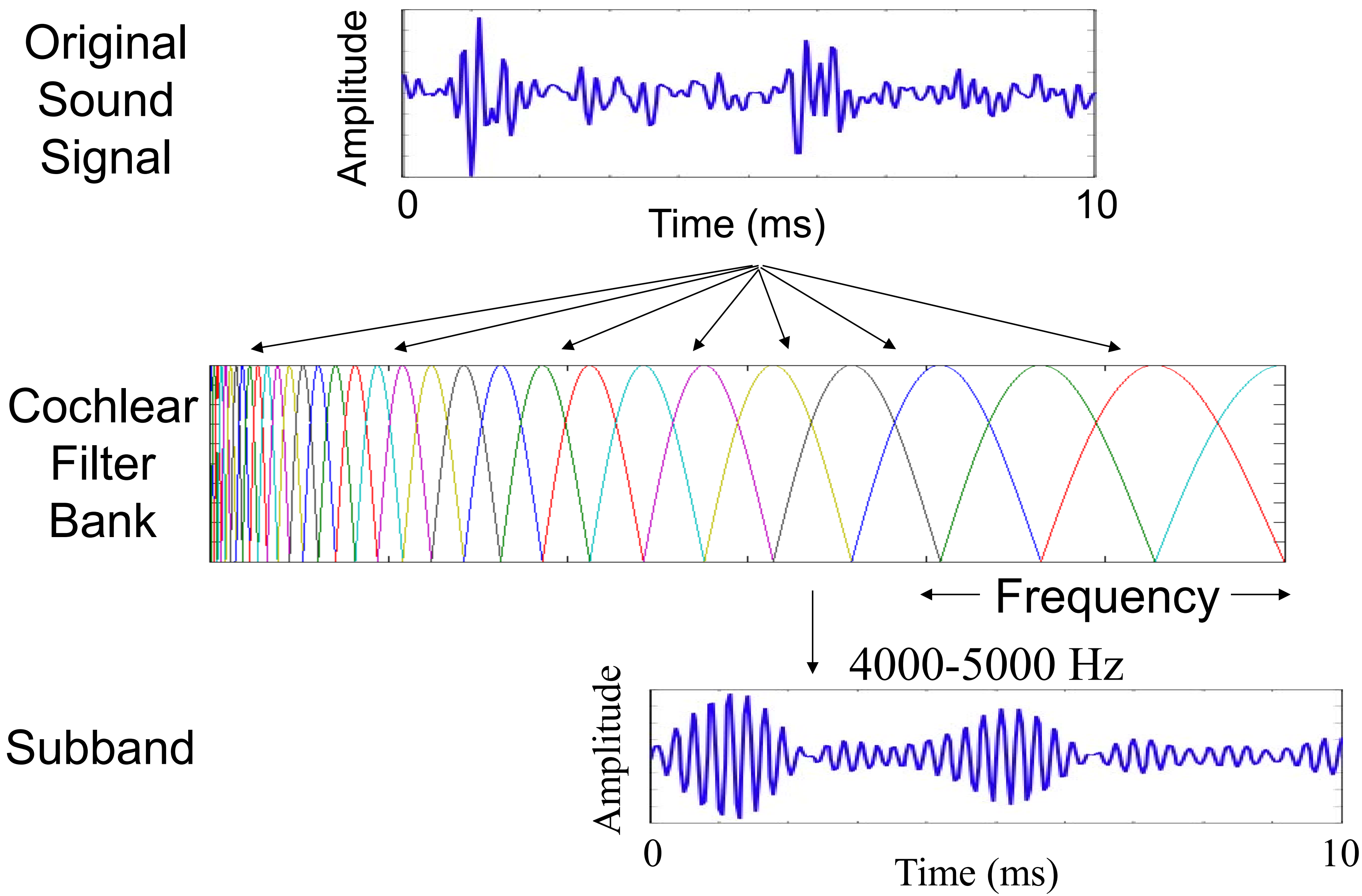
1. Sound signal represented with “subbands”, like auditory nerve:



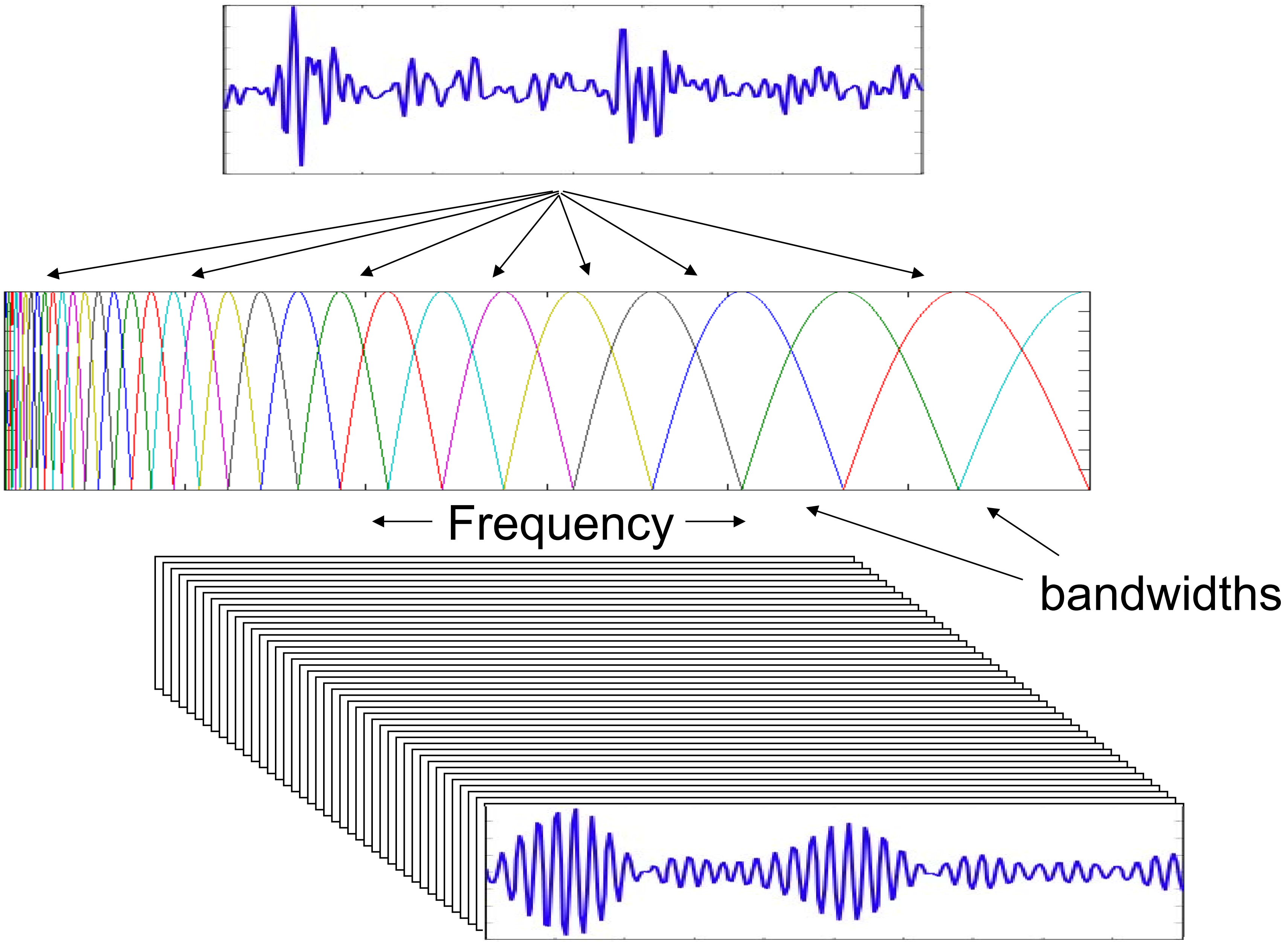
1. Sound signal represented with “subbands”, like auditory nerve:



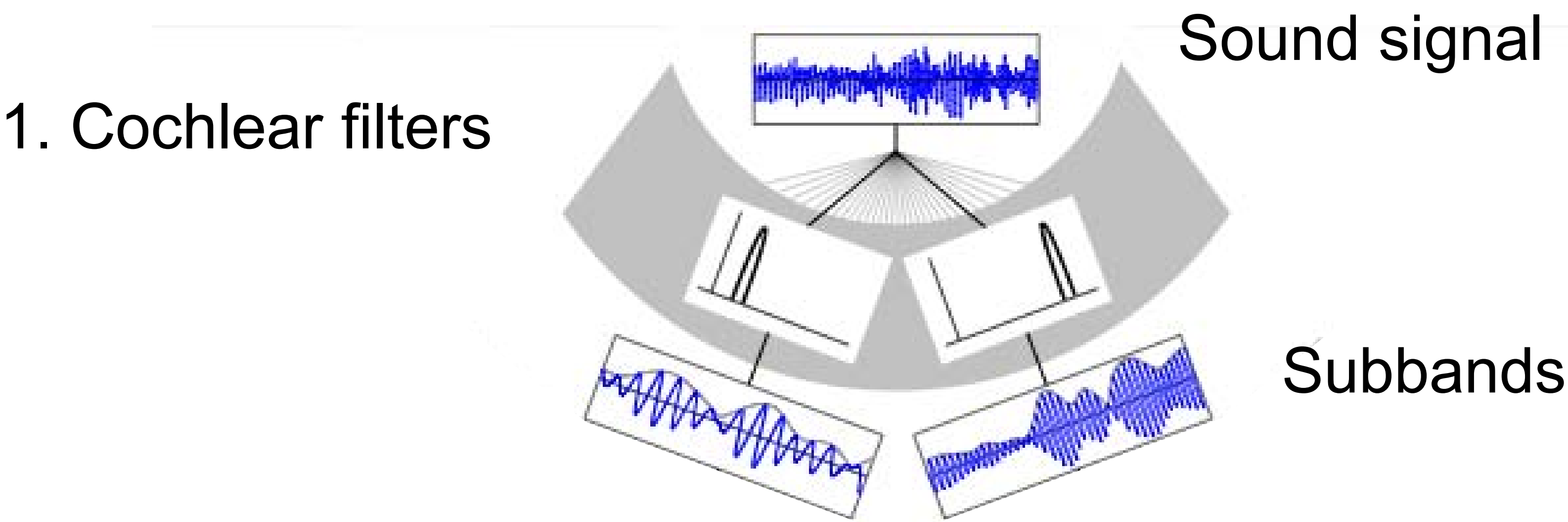
1. Sound signal represented with “subbands”, like auditory nerve:



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

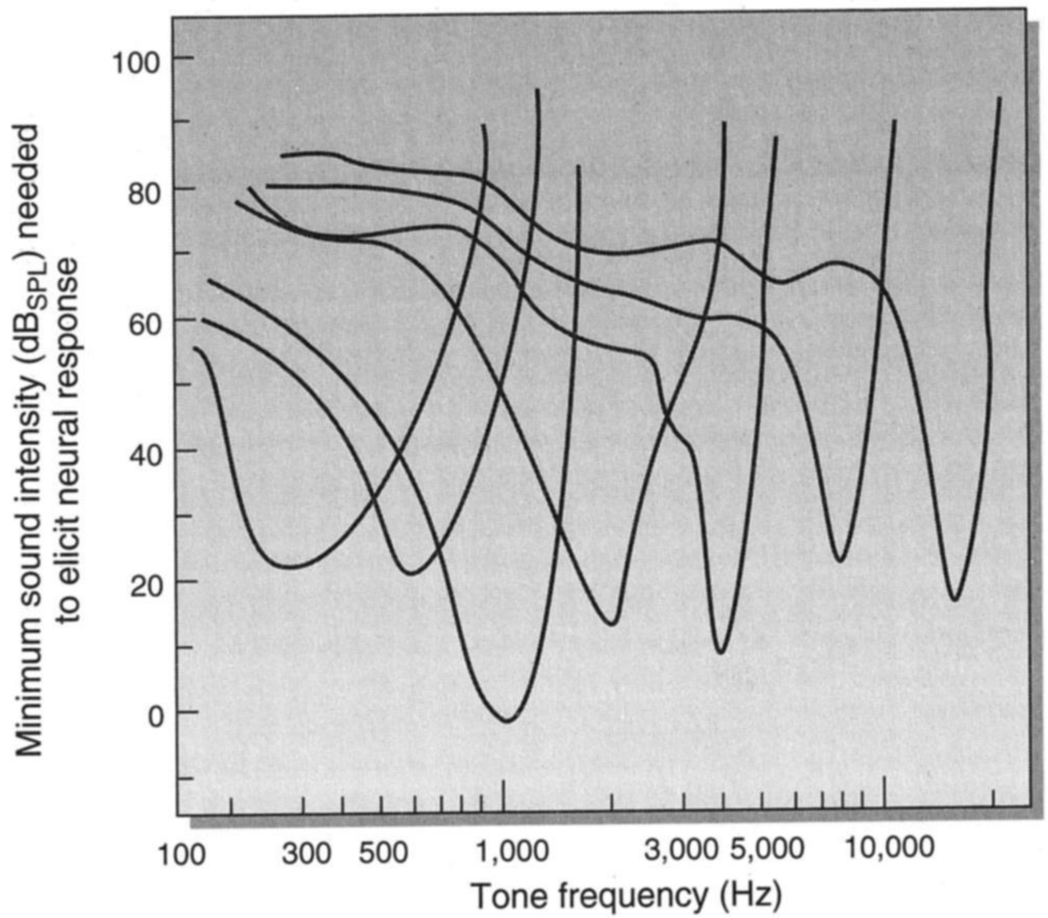


AUDITORY MODEL



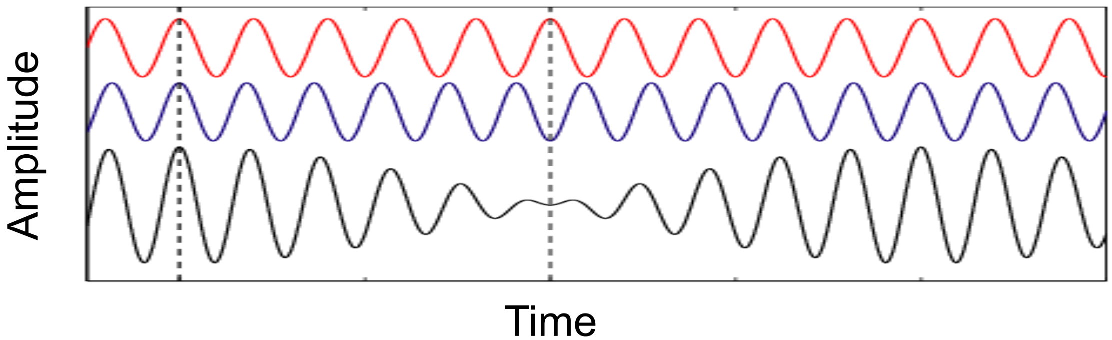
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Frequency selectivity has a host of perceptual consequences.



Frequency selectivity is evident by the ability to “hear out” individual frequency components of a complex tone:

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.

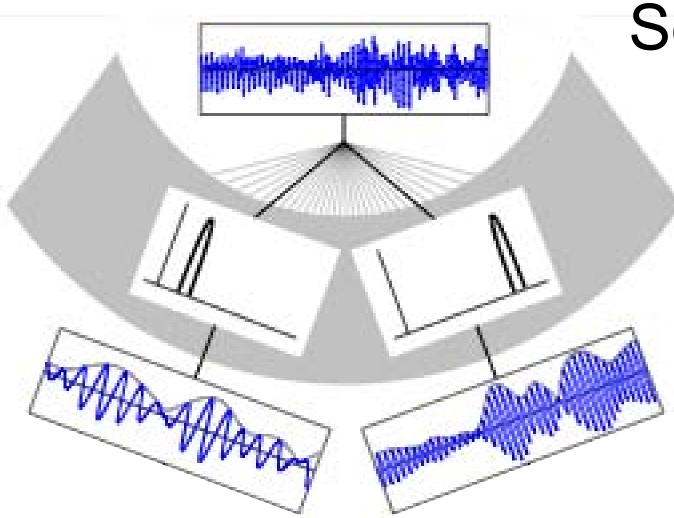
•Perception of beating is constrained by the cochlea:
 -Beats are only heard if two frequency components fall within the filter bandwidth of the cochlea:

1 semitone frequency difference:
3 semitones:
8 semitones:

AUDITORY MODEL

Sound signal

1. Cochlear filters



Subbands

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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)

Figure removed due to copyright restrictions.
Please see the video.
Source: Rose, Jerzy E., Joseph E. Hind, David J. Anderson, and John F. Brugge. "Some effects of stimulus intensity on response of auditory nerve fibers in the squirrel monkey." J. Neurophysiol 34, no. 4 (1971): 685-699.

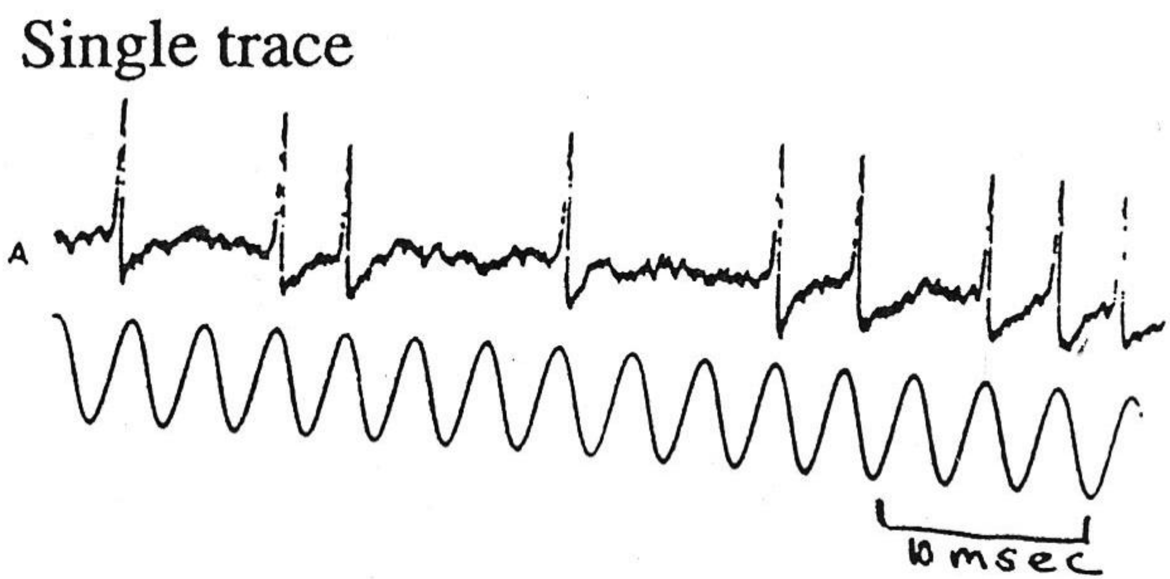
At higher stimulus levels, frequency tuning broadens:

Figure removed due to copyright restrictions.
Please see the video.
Source: Chapter 8, Ochsner, Kevin, and Stephen M. Kosslyn. The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges. Vol. 2. Oxford University Press, 2013.

Neural Coding of Sound

Another important response property of the cochlea:

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

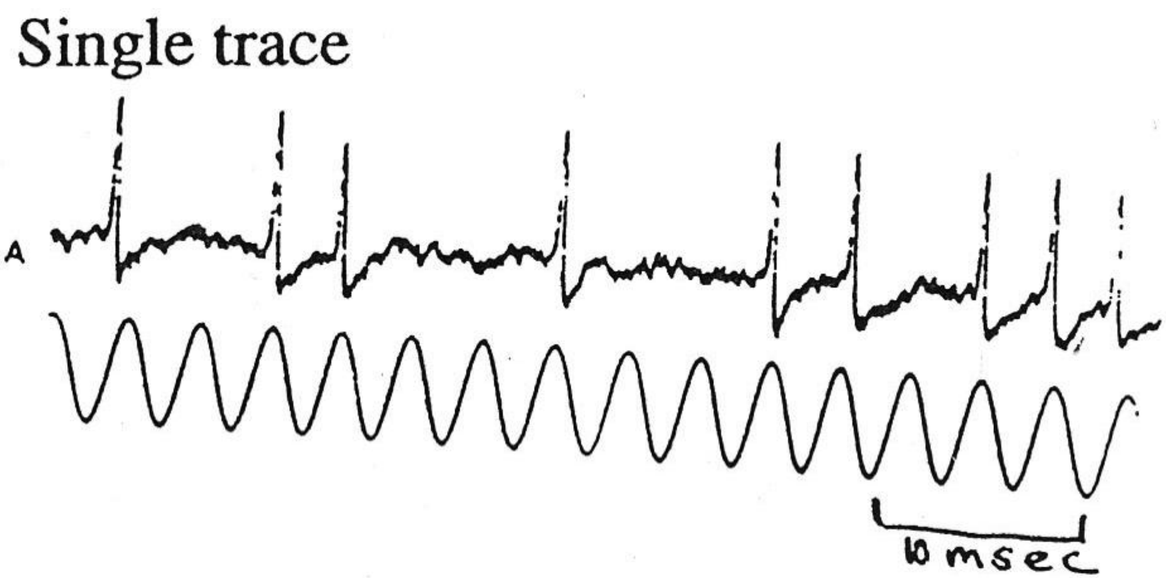


Phase locking occurs for frequencies under ~4kHz (in nonhuman animals – no available data in humans).

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?

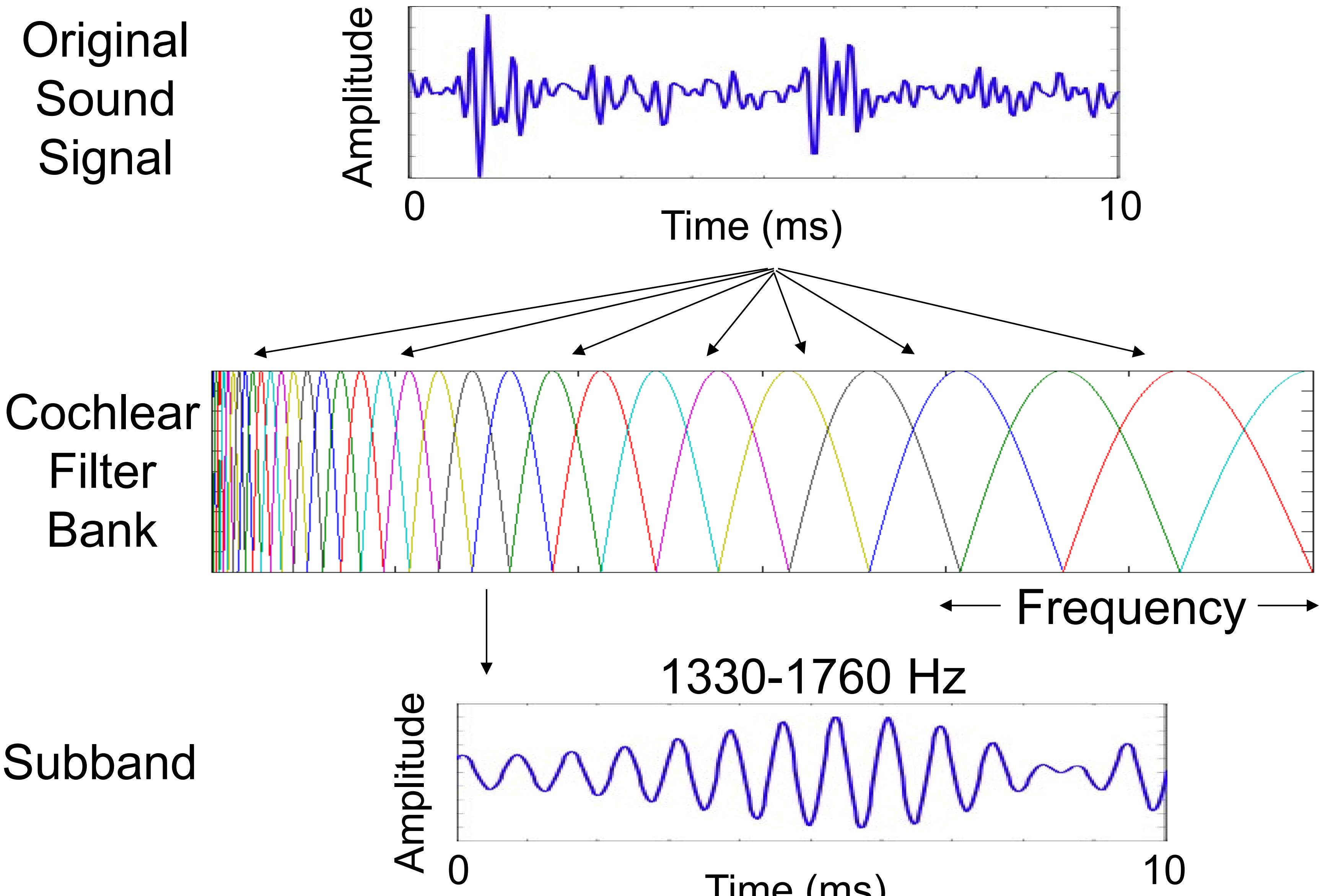
Per eye:

5 million cones
100 million rods
1.5 million optic nerve fibers

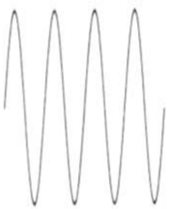
140 million neurons in primary visual cortex

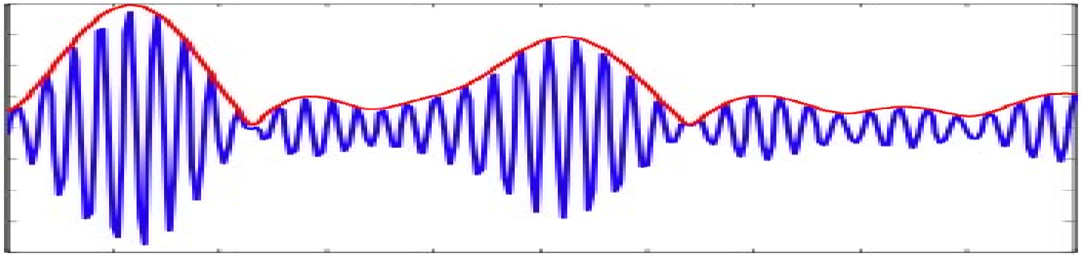
Figure removed due to copyright restrictions.
Please see the video.
Source: Chapter 8, Ochsner, Kevin, and Stephen M. Kosslyn. The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges. Vol. 2. Oxford University Press, 2013.

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

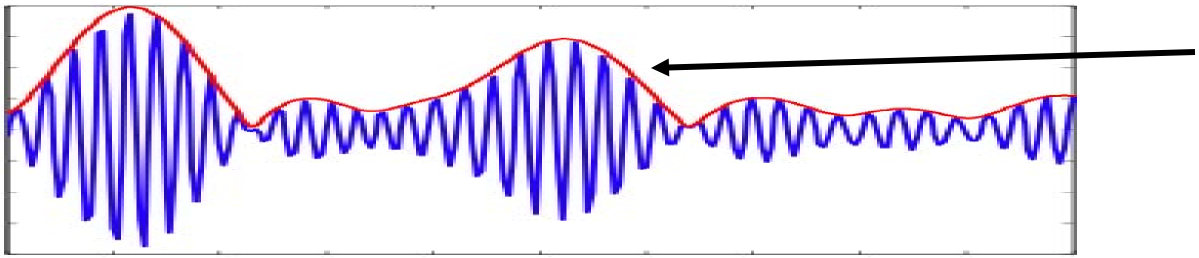


Subbands can be characterized by instantaneous amplitude and phase, loosely mapping onto rate and spike timing in auditory nerve:





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:

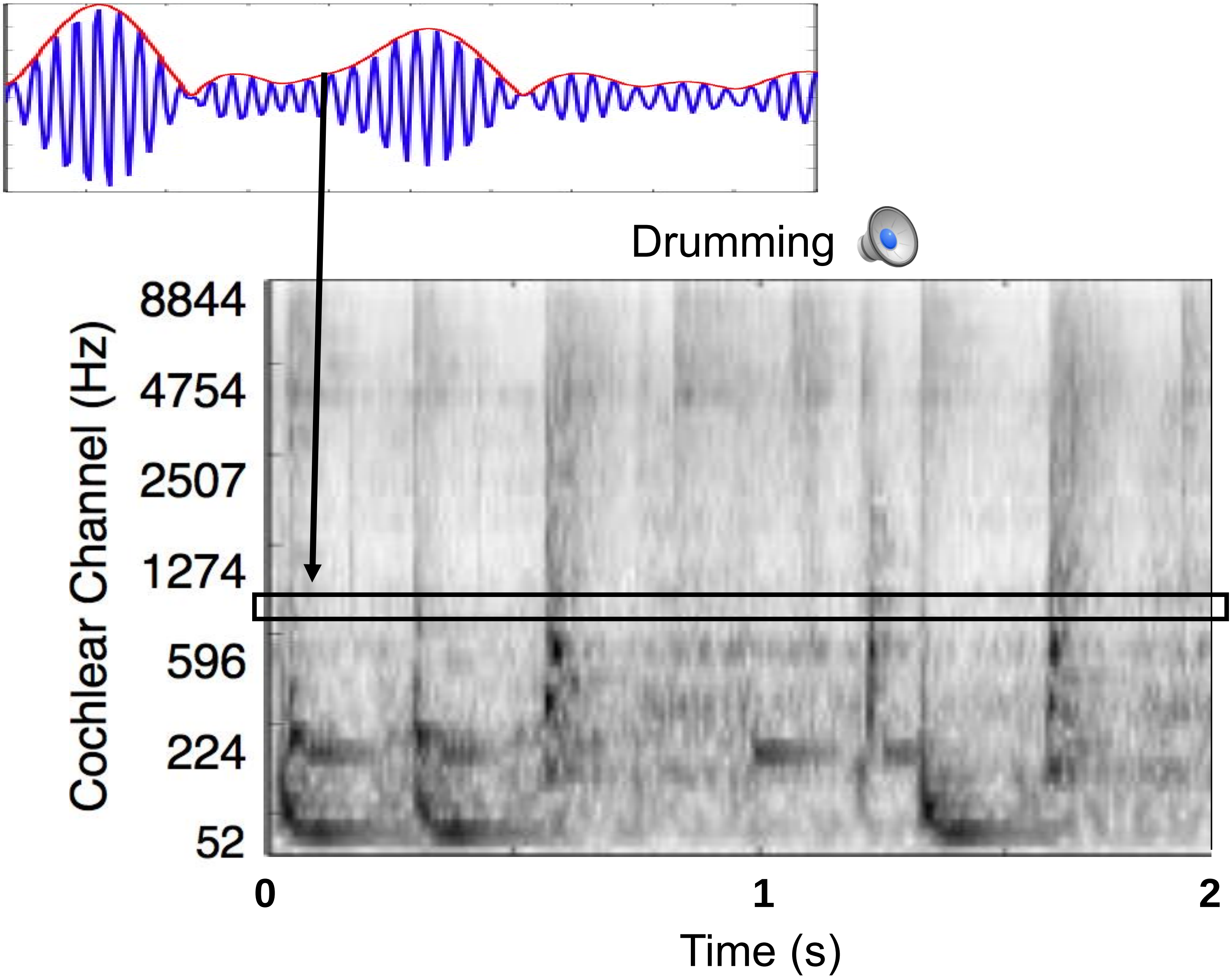


Amplitude modulations are captured by the **envelope**

-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)

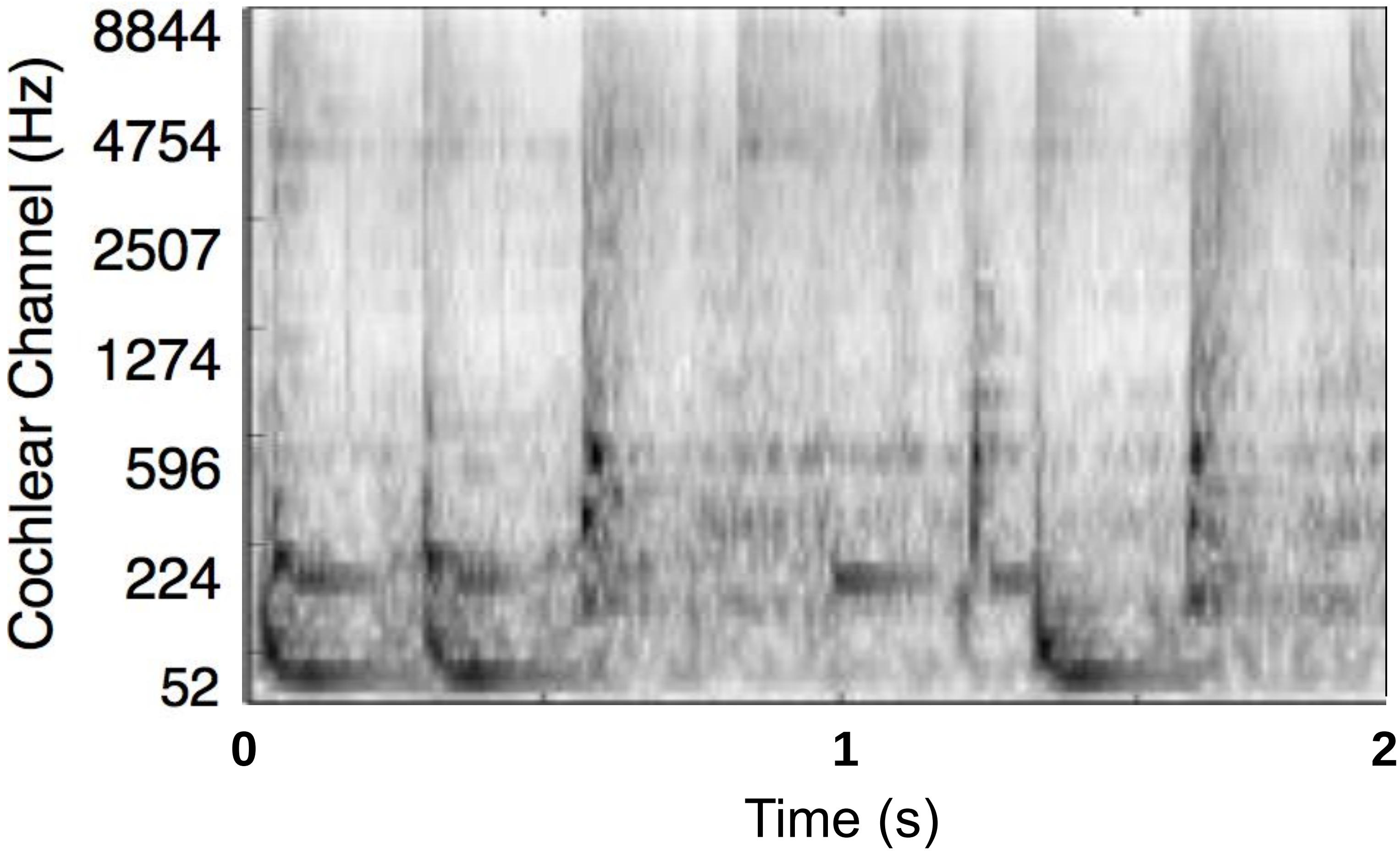
A spectrogram contains the envelope of each subband:



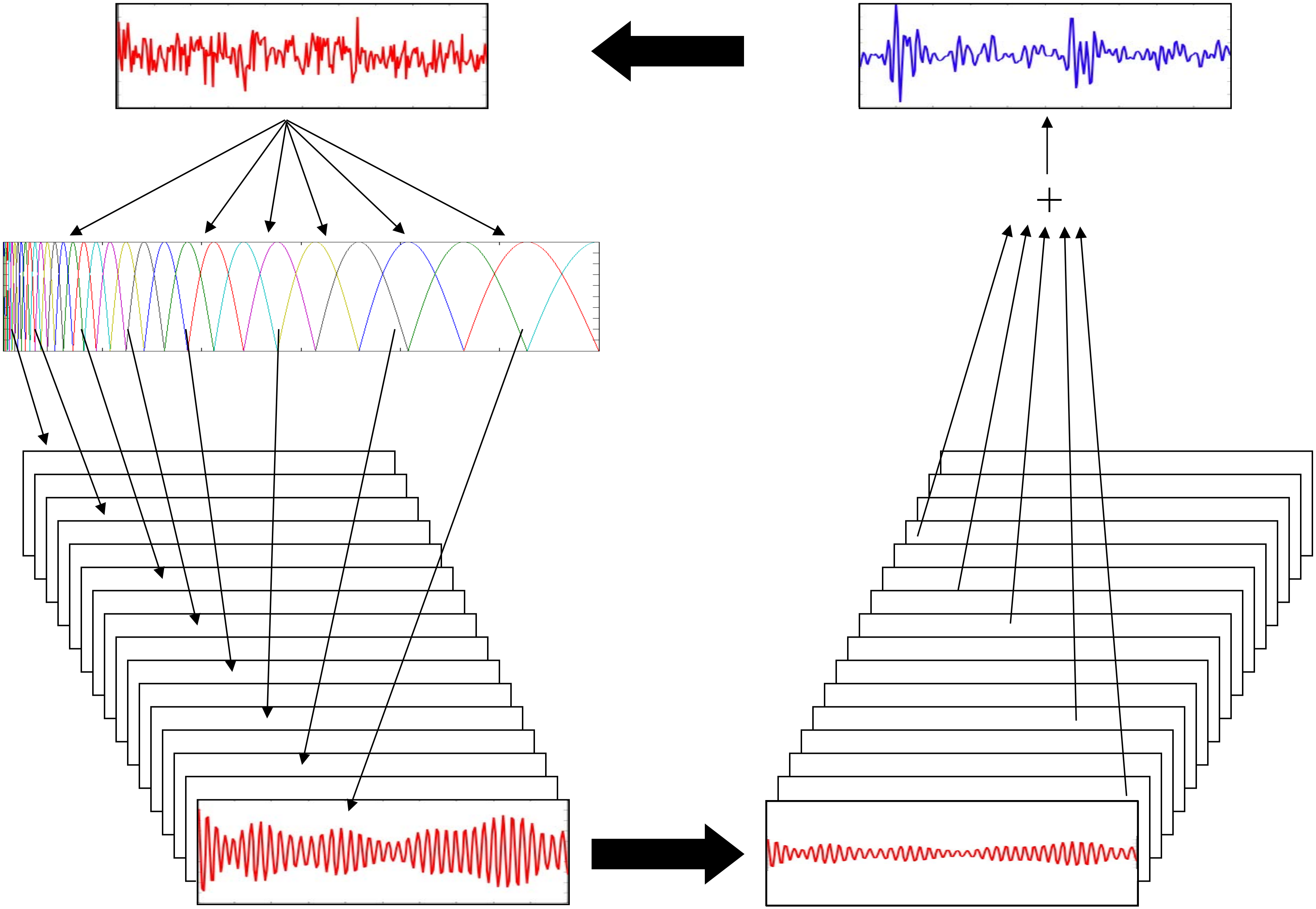
Envelopes often capture all the information that matters perceptually*.

Sounds can be reconstructed just from the envelopes:

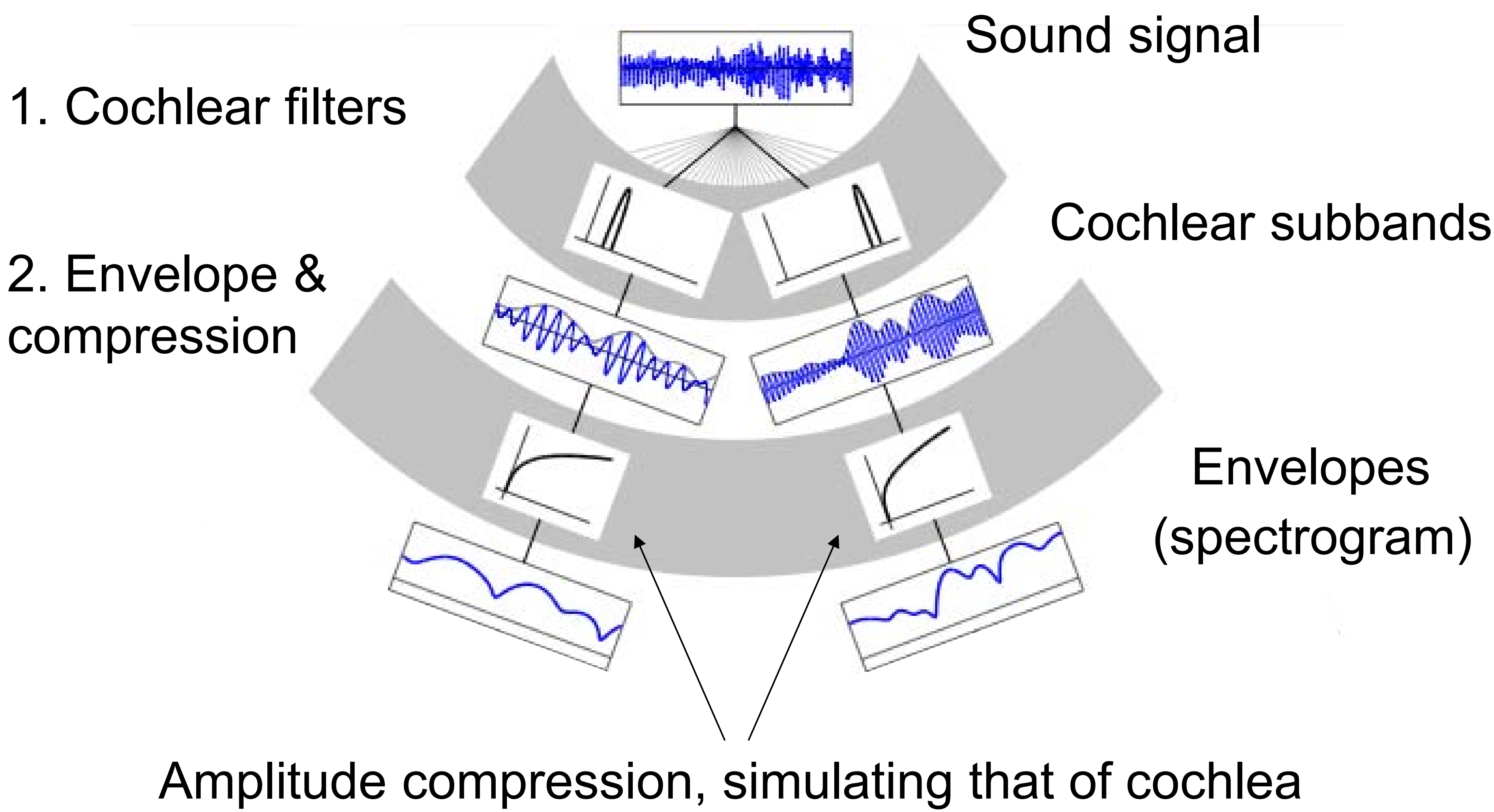
Drumming



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

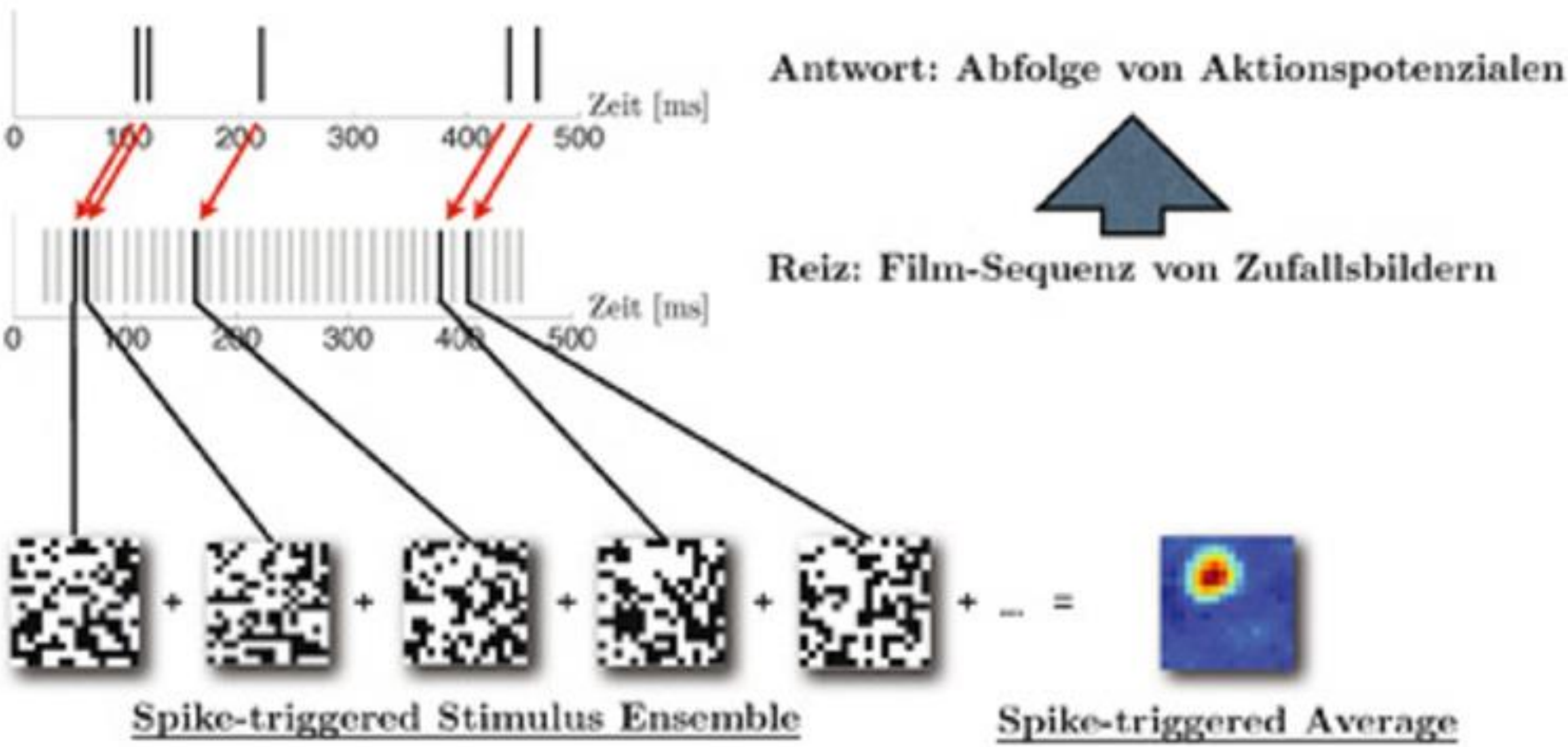


AUDITORY MODEL



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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



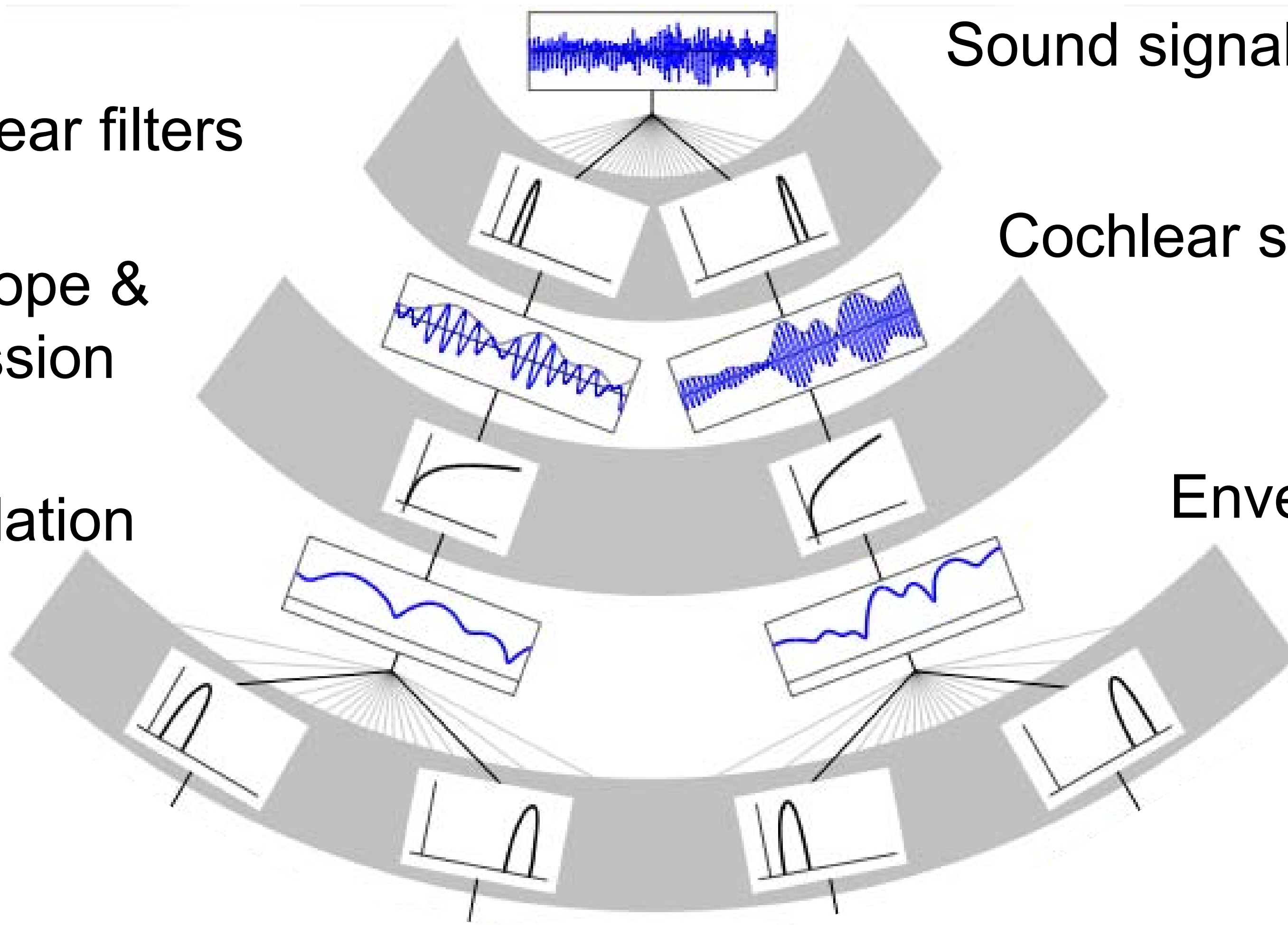
STRF = spectro-temporal receptive field

Derived from methods like the STA applied to stimulus spectrogram.

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

Figure removed due to copyright restrictions.
Please see the video.
Source: Chapter 8, Ochsner, Kevin, and Stephen M. Kosslyn. The Oxford Handbook of Cognitive Neuroscience, Volume 2: The Cutting Edges. Vol. 2. Oxford University Press, 2013.

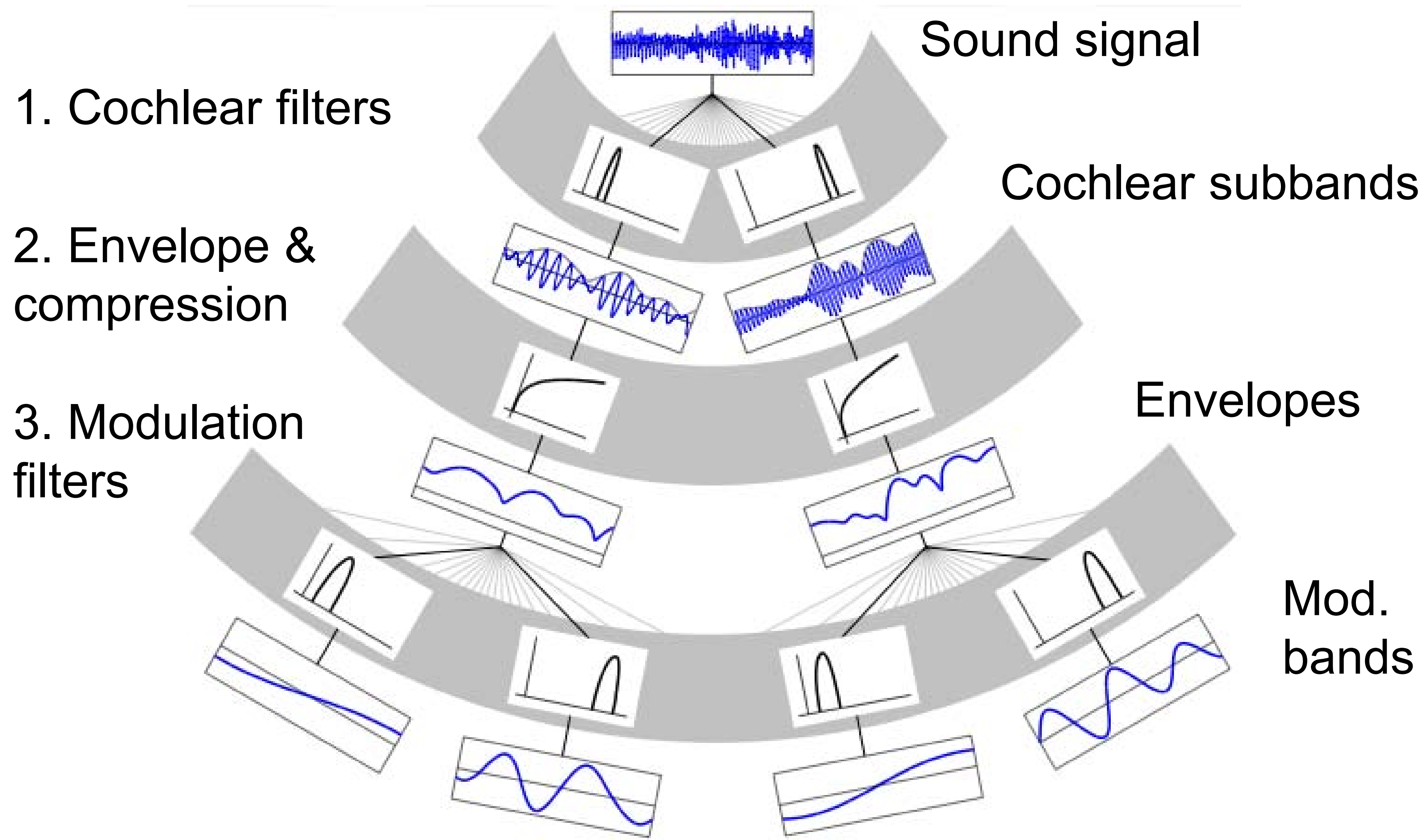
Envelope structure measured with second filter bank:



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

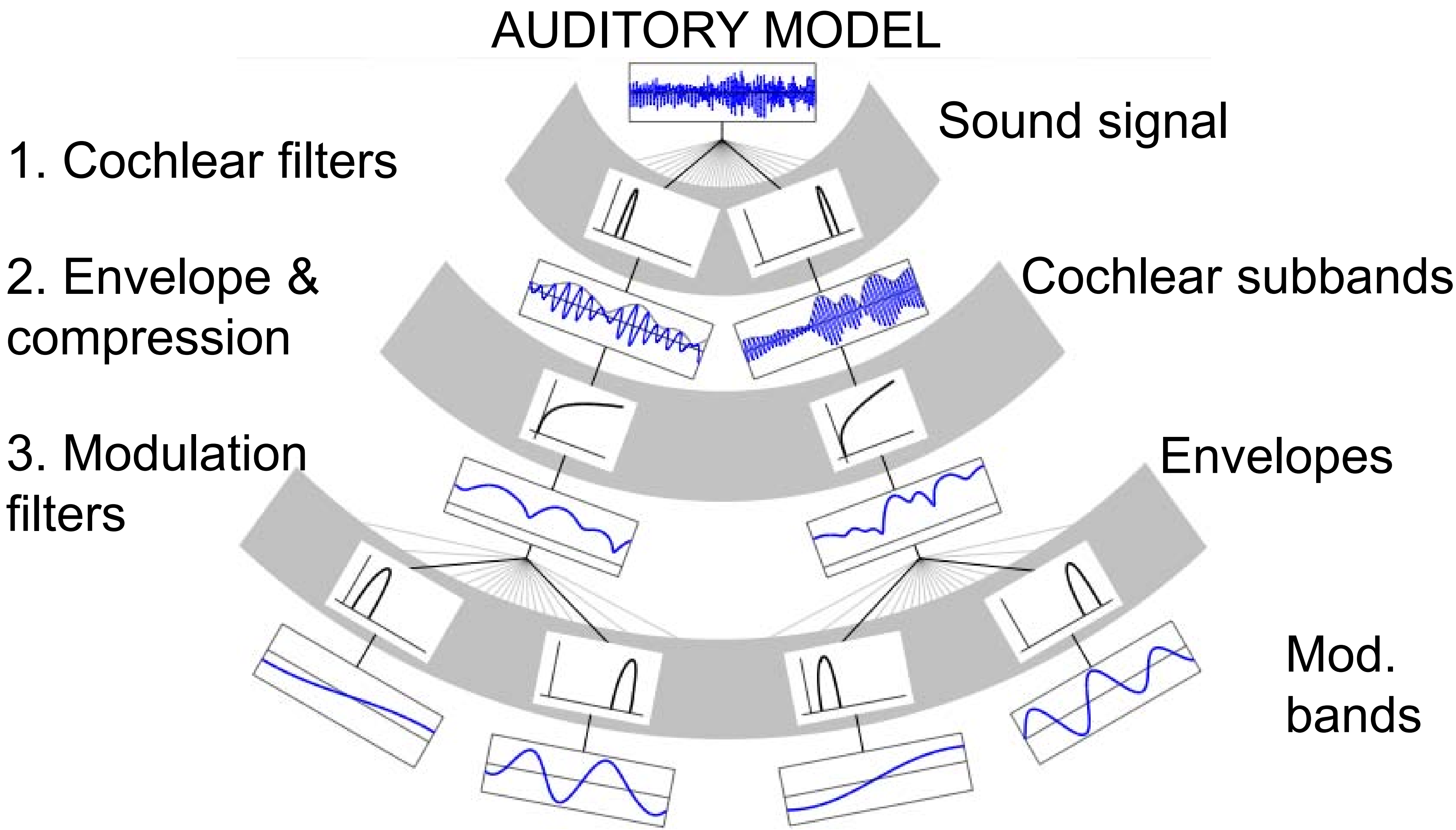
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Envelope structure measured with second filter bank:



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

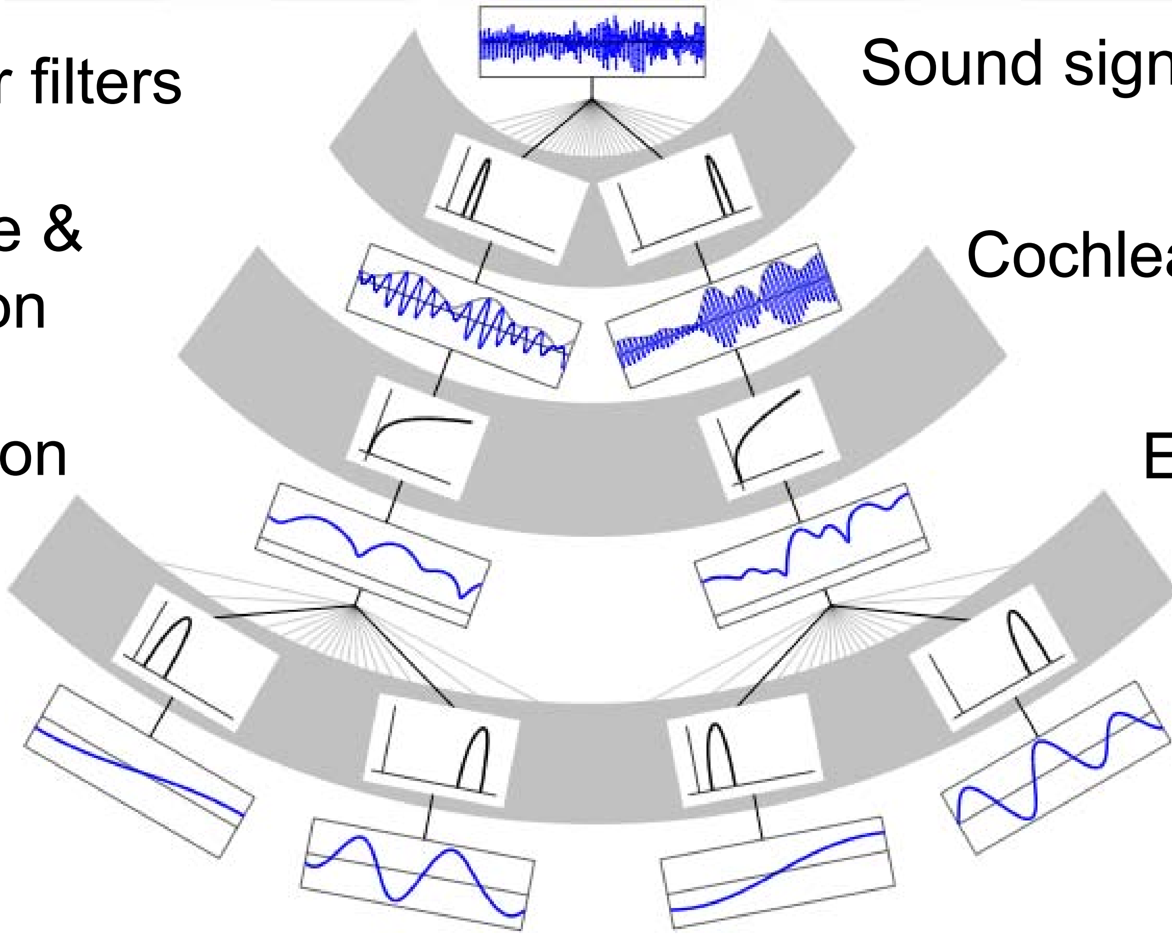
Given these representations, how do we recognize sounds and their properties?

AUDITORY MODEL

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



Sound signal

Cochlear subbands

Envelopes

Mod. bands

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Part 2: Sound Texture

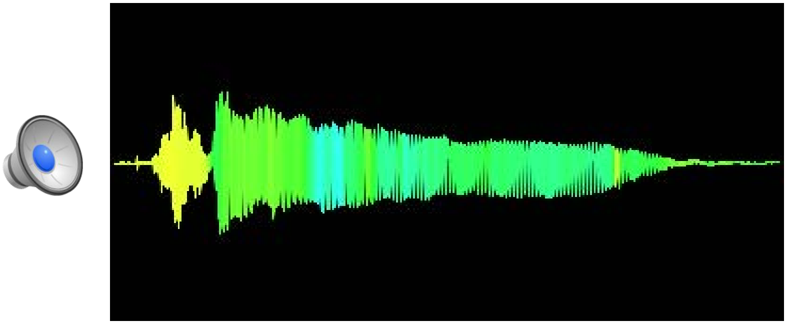
SOUND TEXTURE

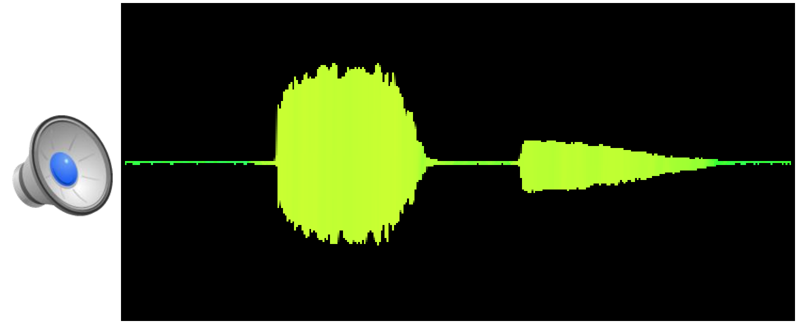
Textures result from large numbers of acoustic events.

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

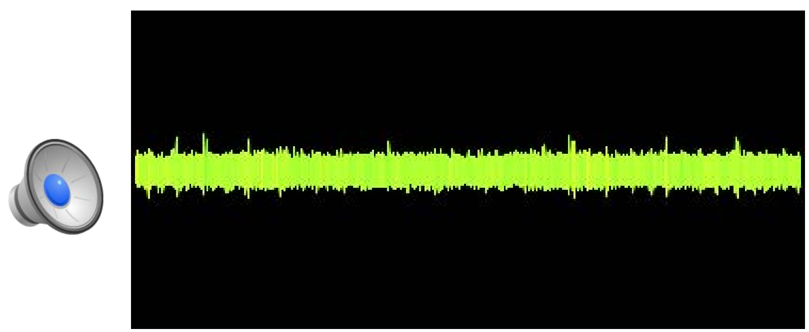
Sound textures are common in the world, but largely unstudied.

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



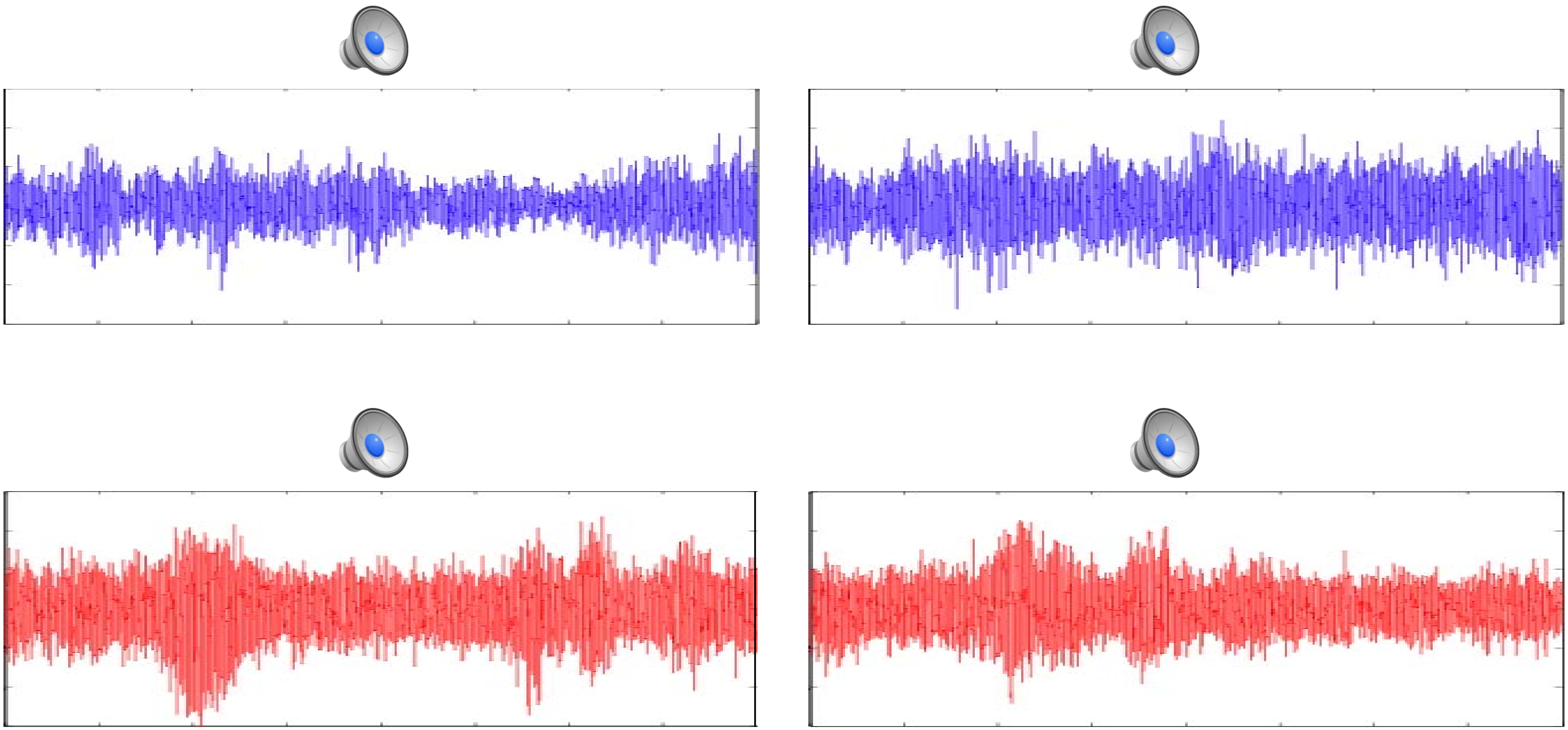


Unlike event sounds, textures are stationary - essential properties do not change over time.



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

How do people represent, recognize sound textures?

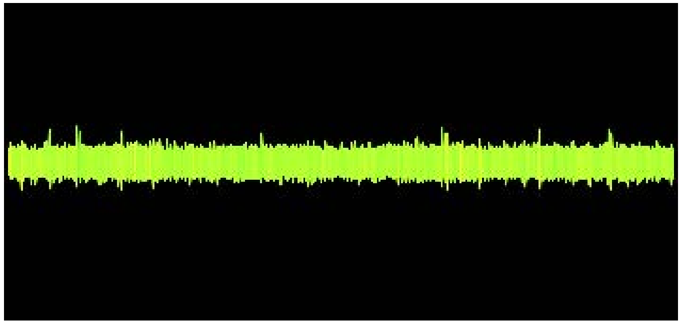


What do you extract and store about these waveforms to recognize that they are the same kind of thing?

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?

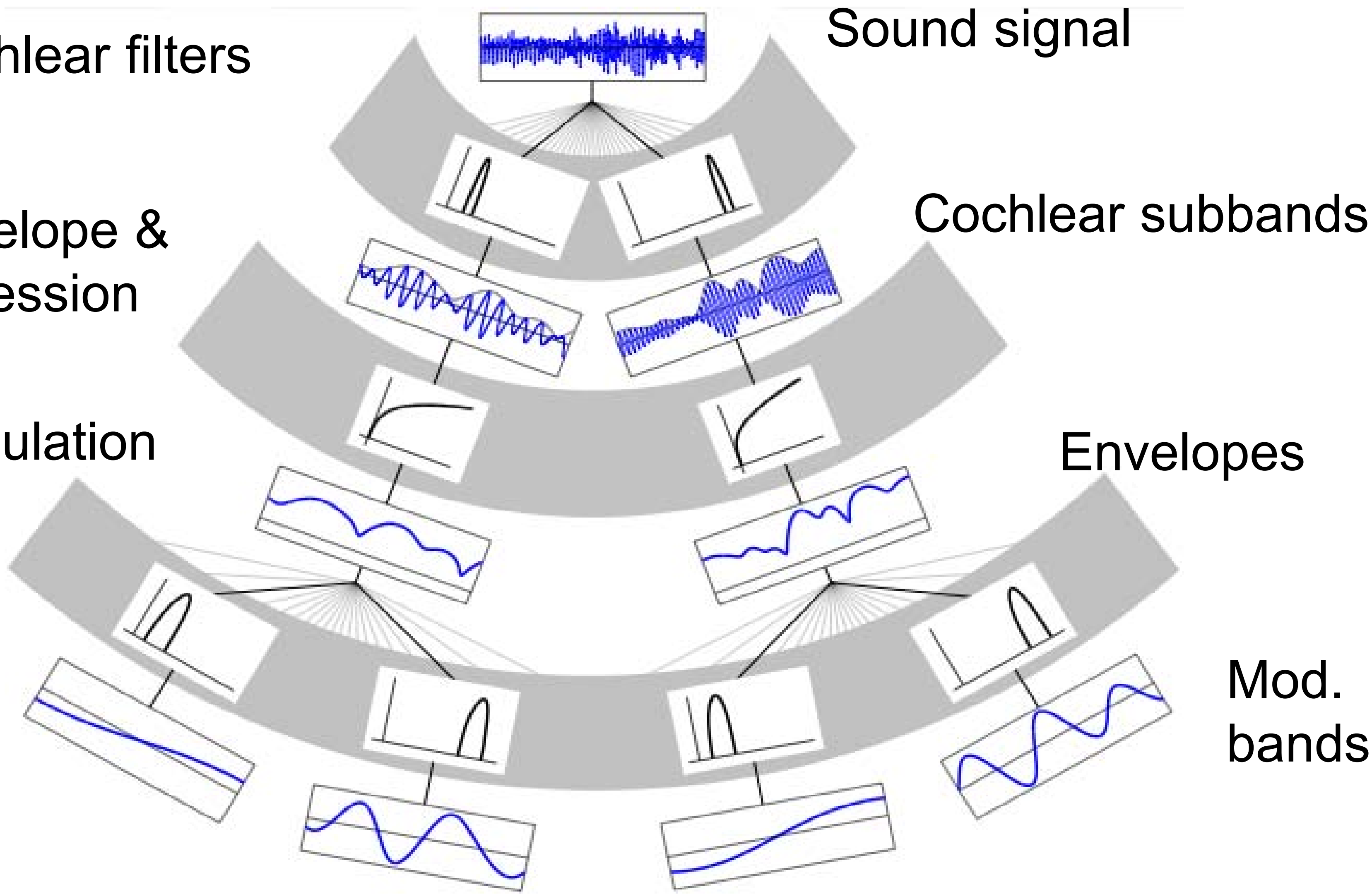


Whatever statistics the auditory system measures are presumably derived from these representations:

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



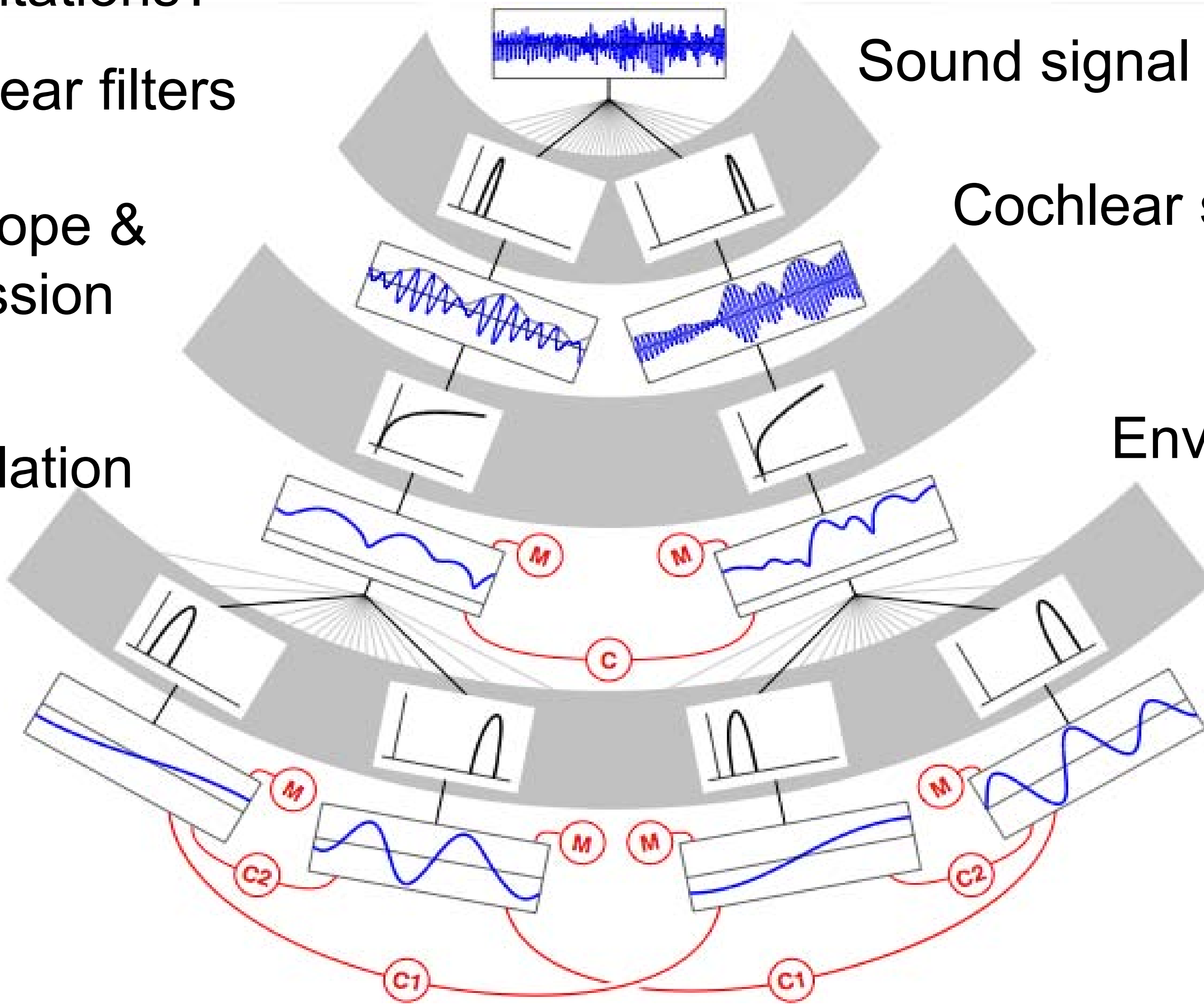
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



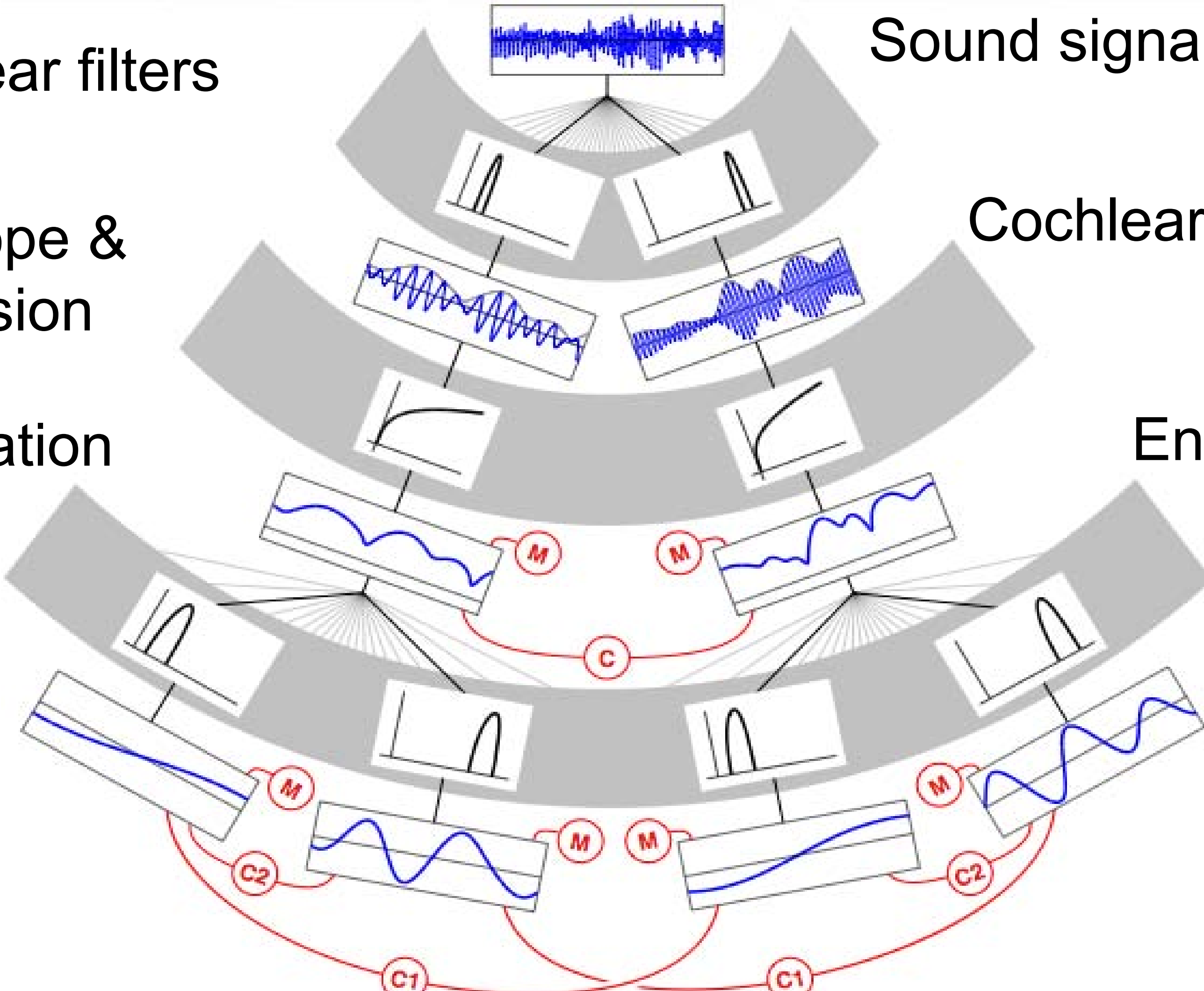
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



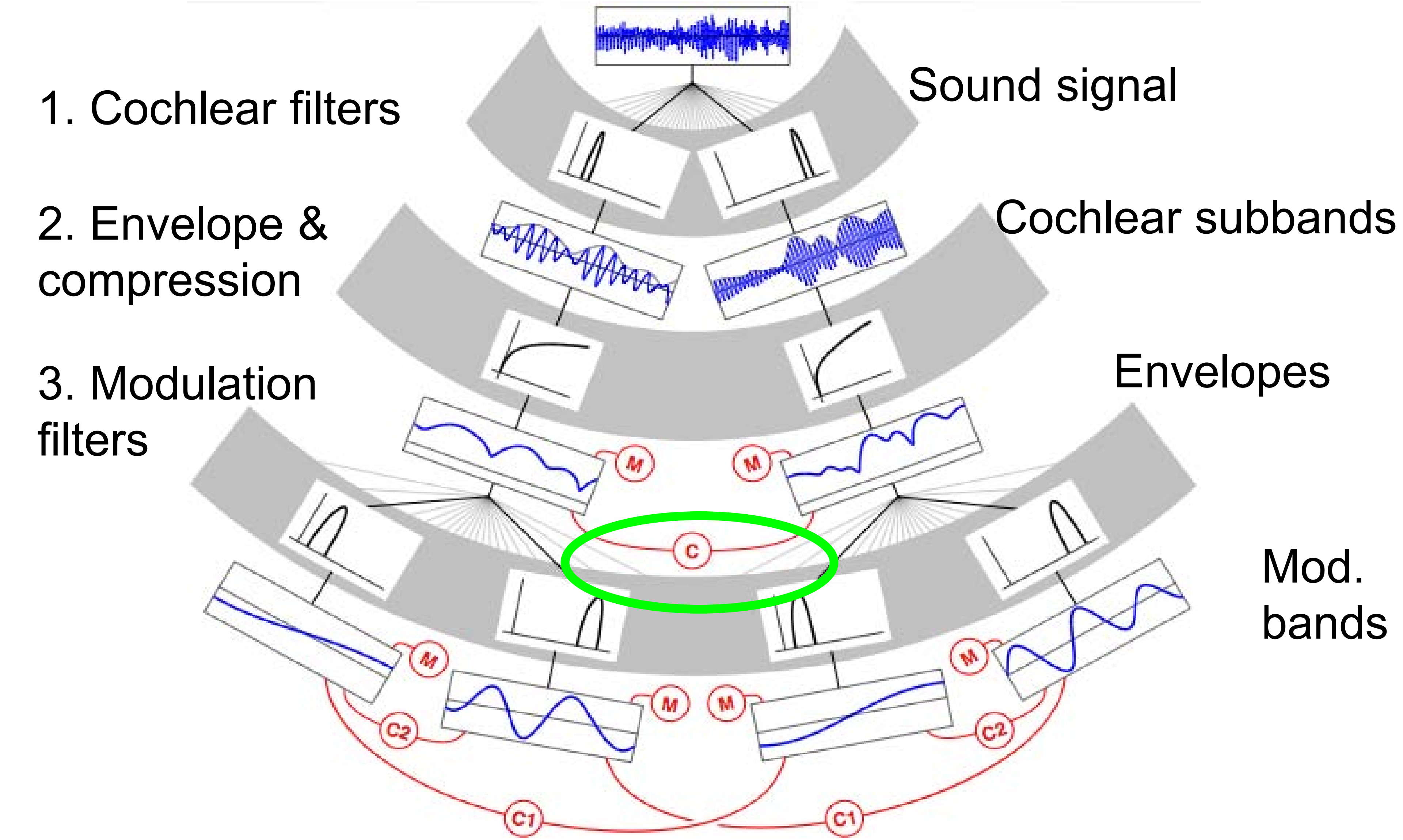
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

•Statistics are not specifically tailored to natural sounds

•Ultimately, would be nice to learn statistics from data (working on it…)

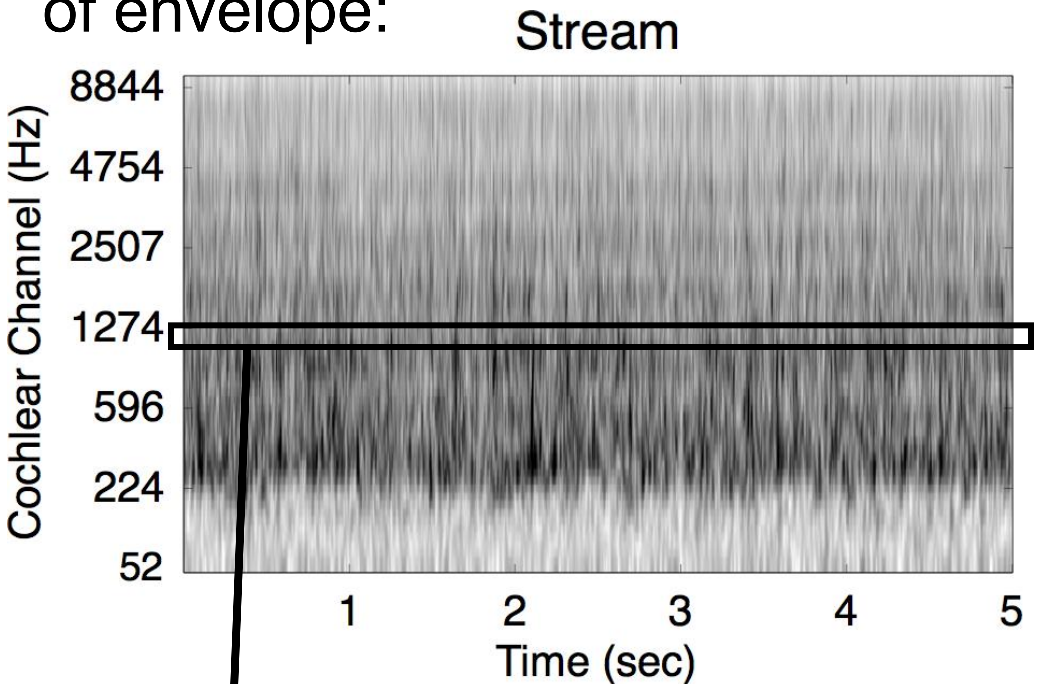
•Simple, involve operations that could be instantiated in neurons

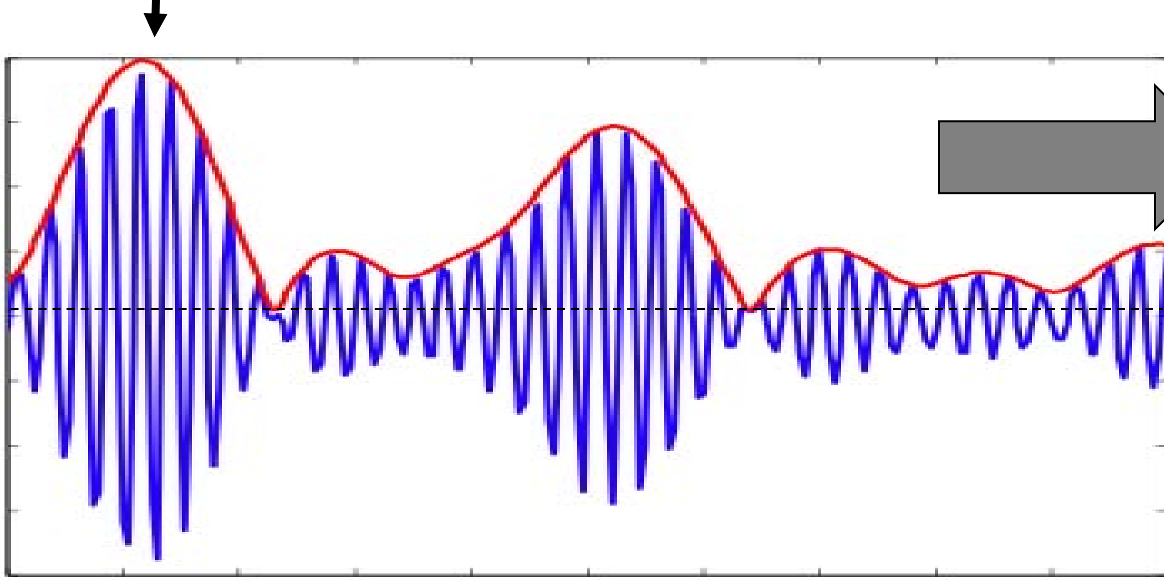
To be useful for recognition, statistics need to give different values for different sounds…

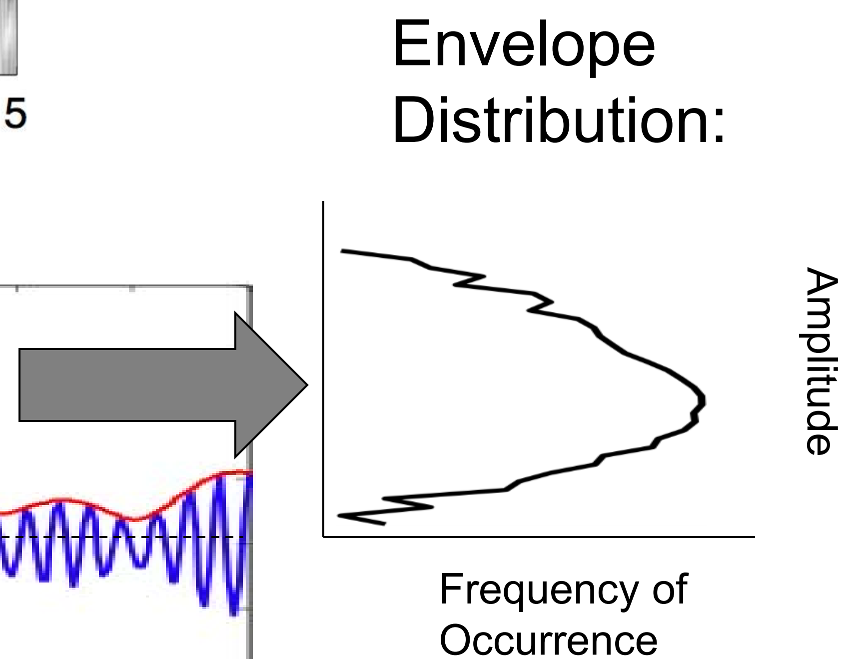


Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

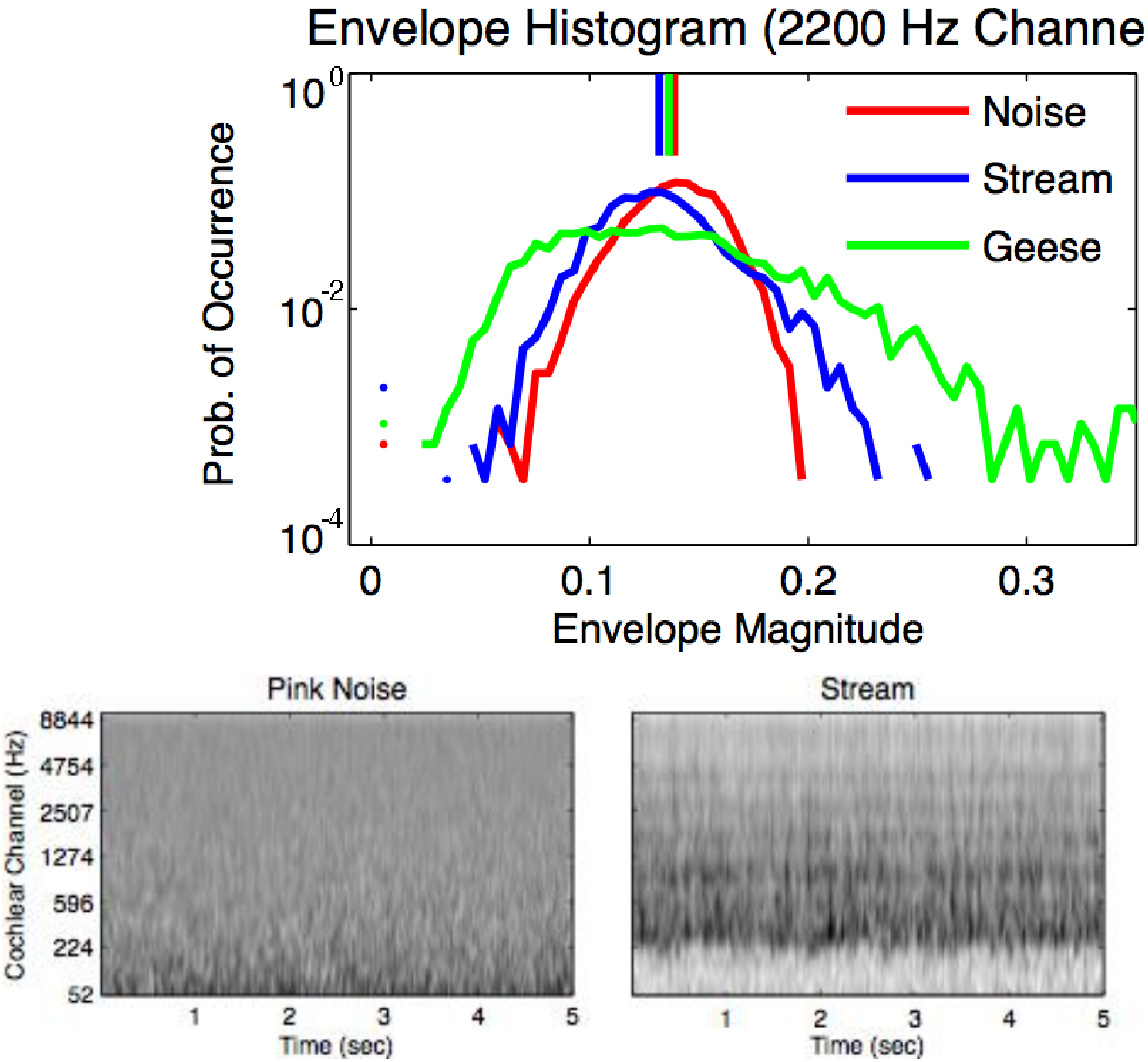
Marginal moments (mean, variance, skew) describe distribution of envelope:







Envelope distributions for natural signals differ from those for noise.



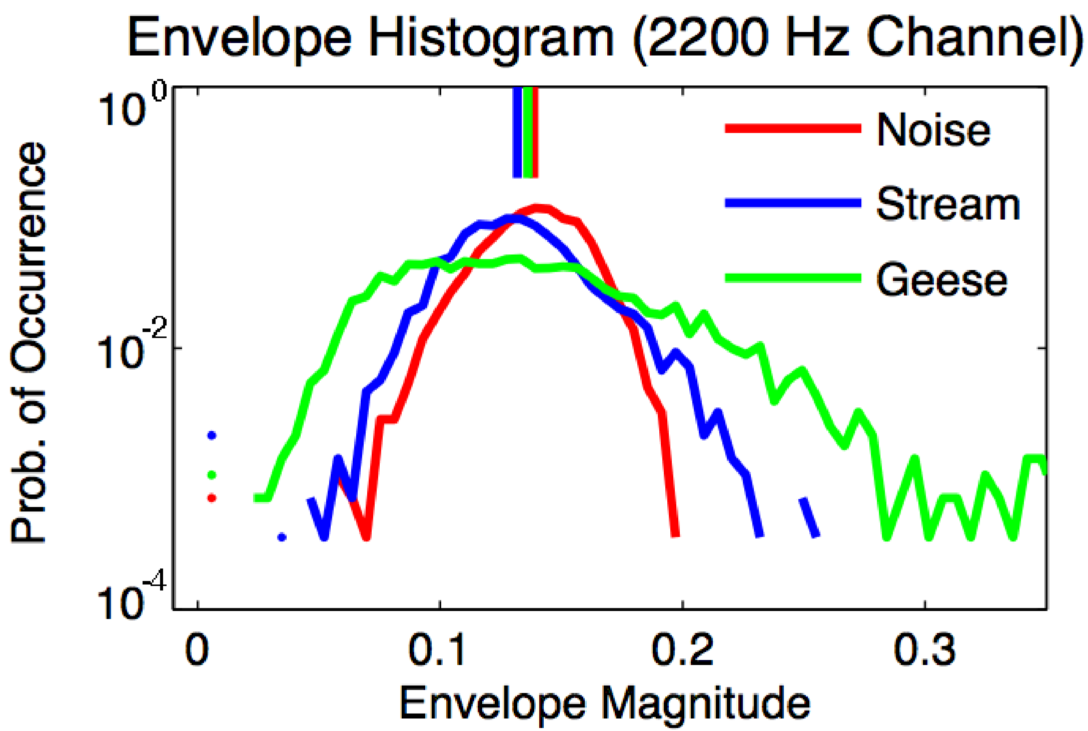
Distributions have similar mean, but different shapes.

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

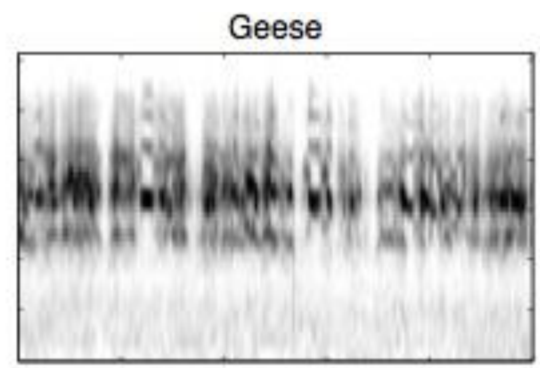
Natural signals are **sparser** than noise.

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).

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

1. Cochlear filters

2. Envelope & compression

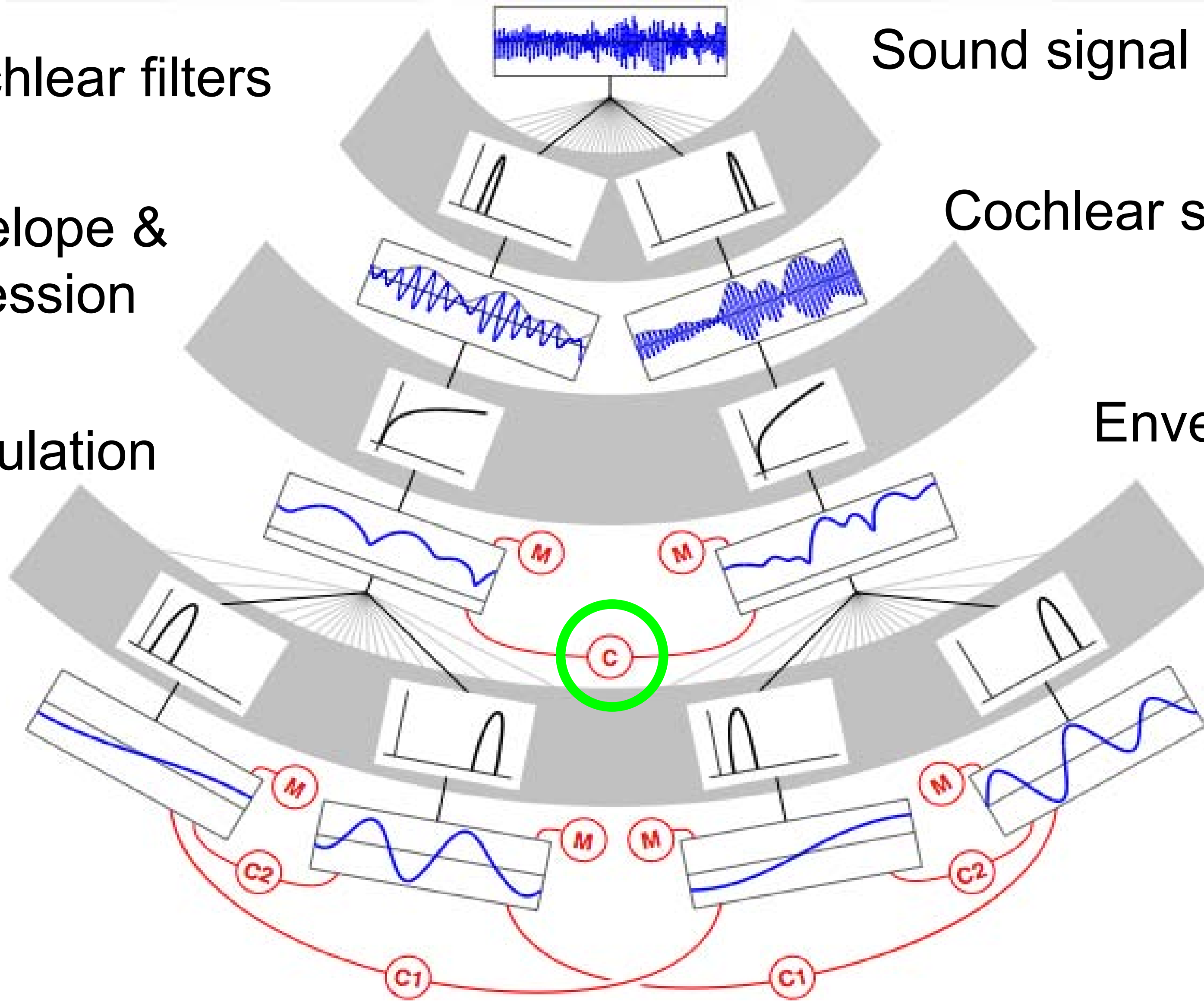
3. Modulation filters

Sound signal

Cochlear subbands

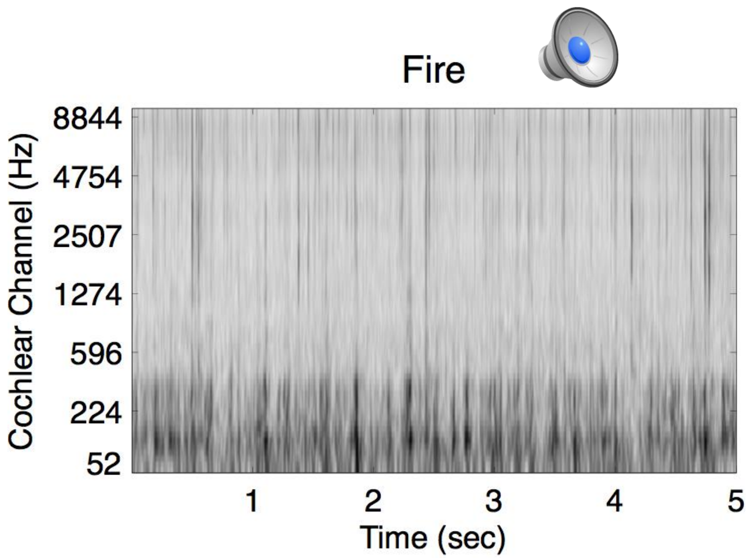
Envelopes

Mod. bands



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

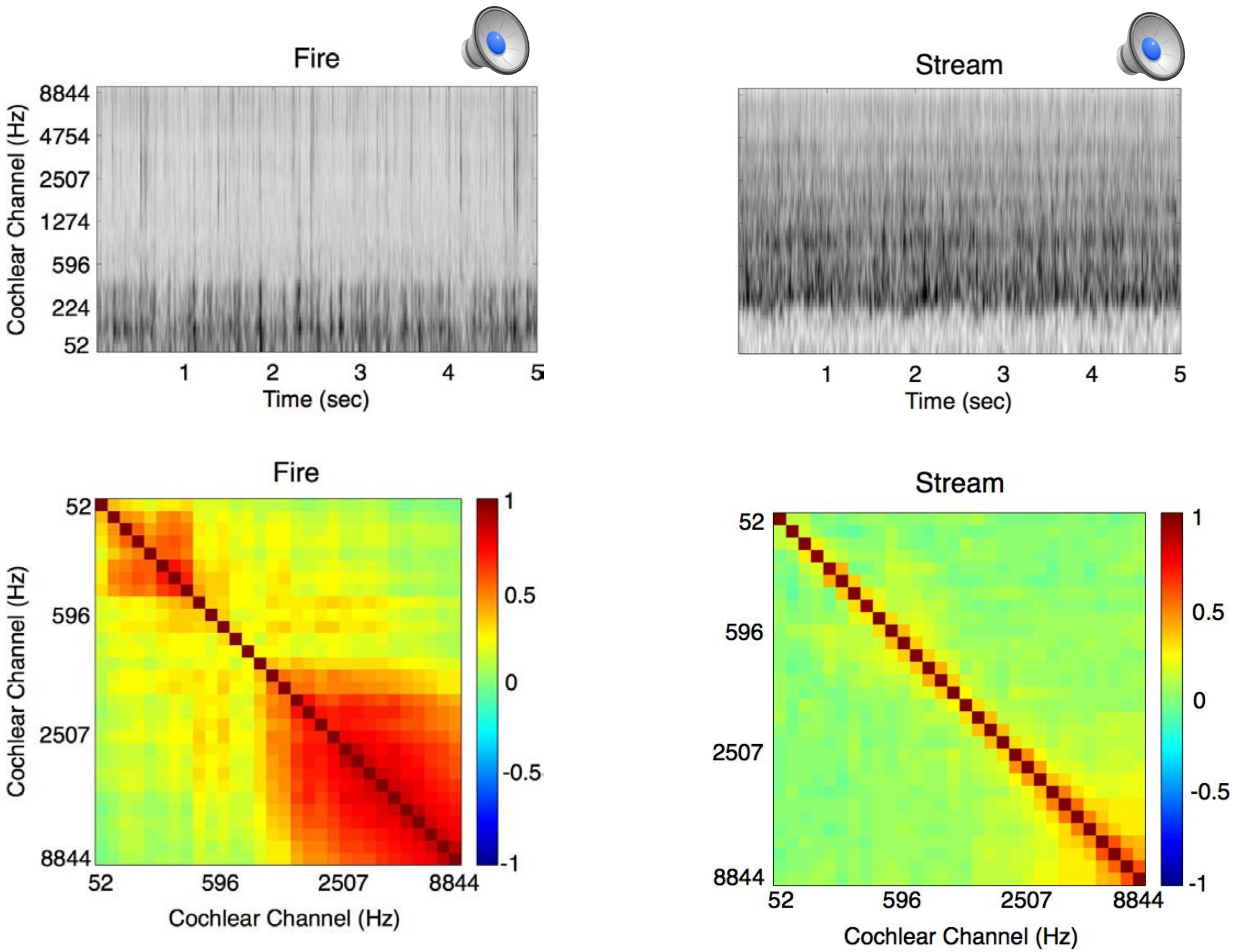
Correlations between envelopes vary across sounds.



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Broadband events induce dependencies between channels.

Correlations reflect broadband events (crackles, claps):

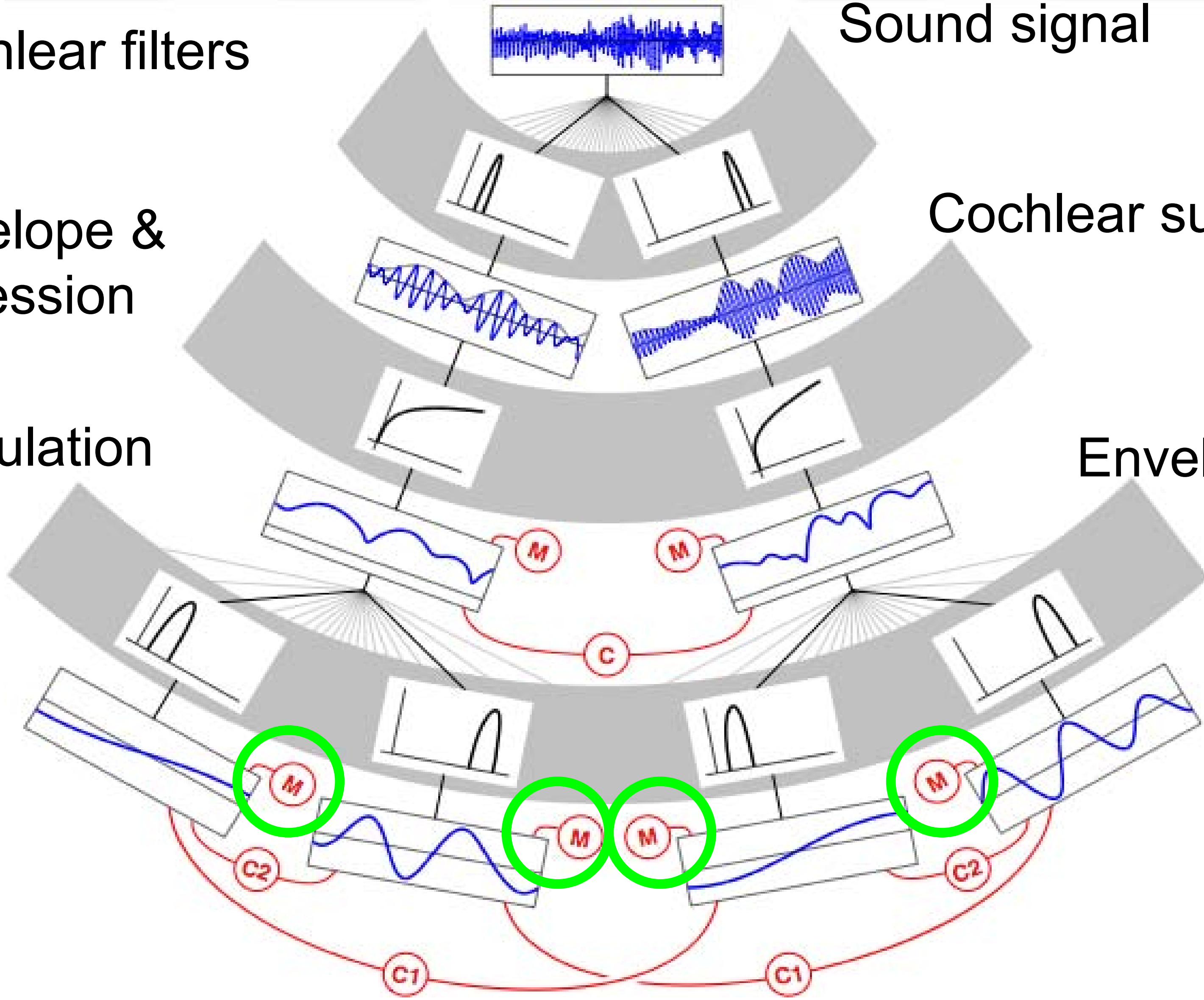


Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

1. Cochlear filters

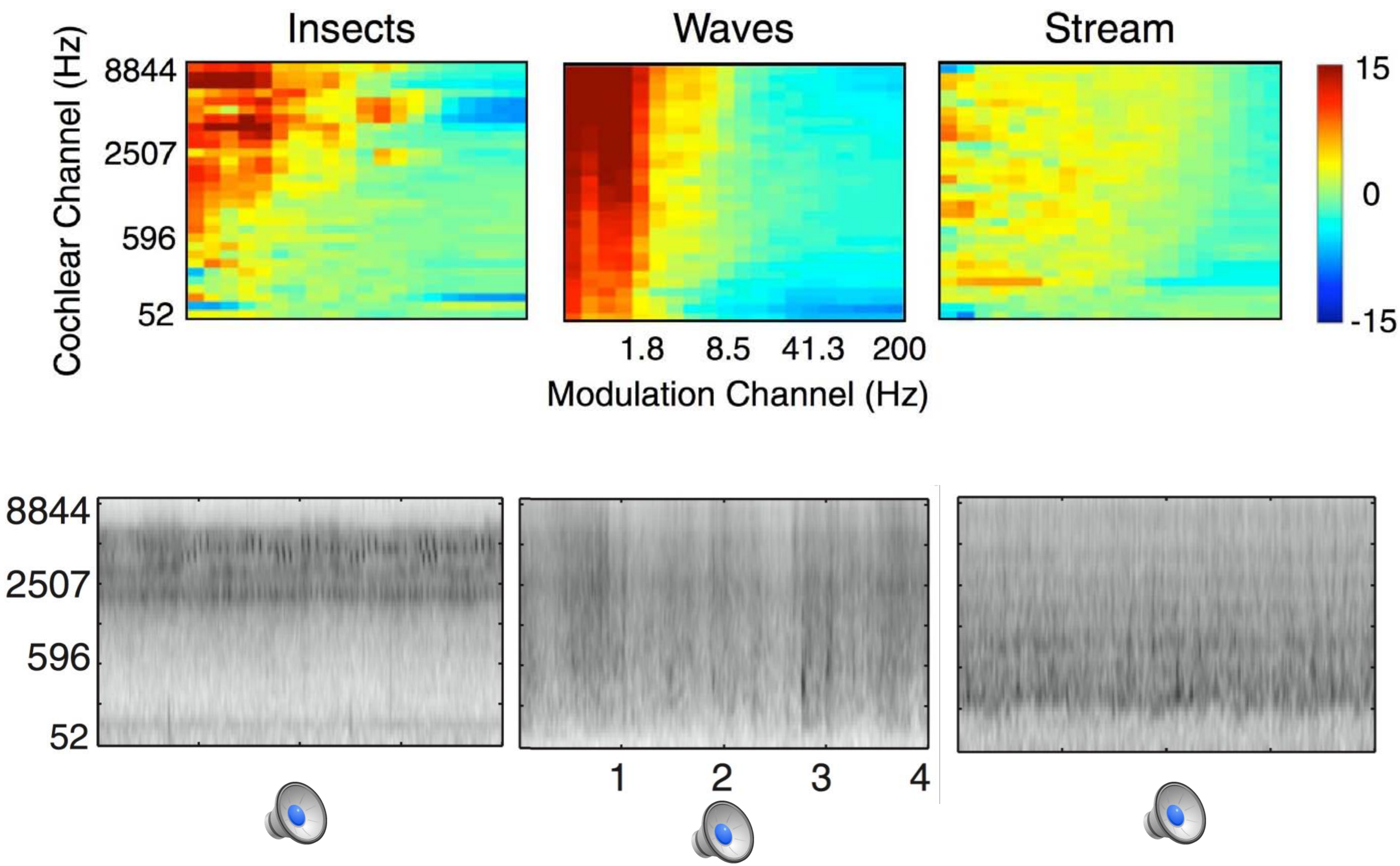
2. Envelope & compression

3. Modulation filters

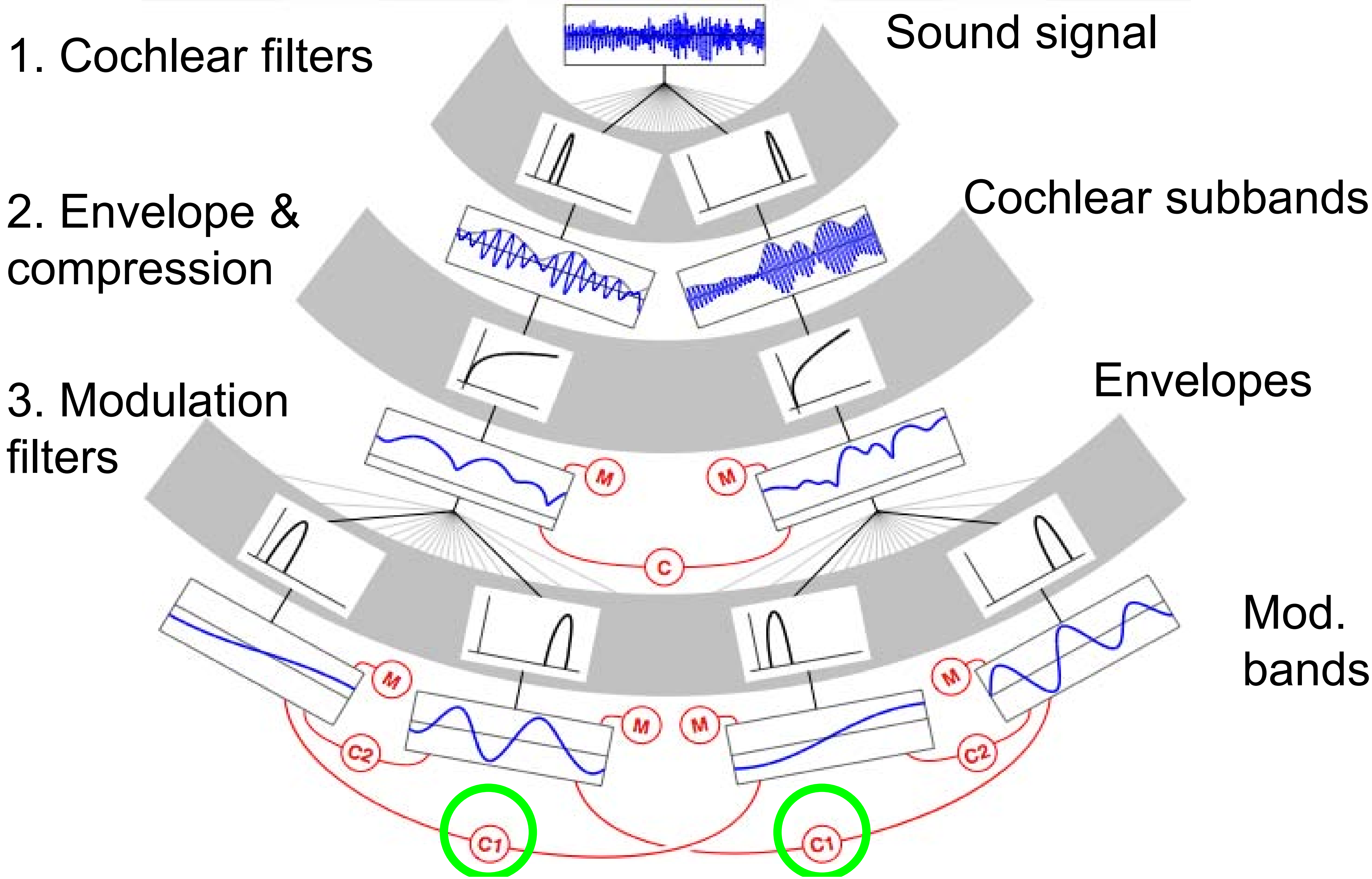


Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

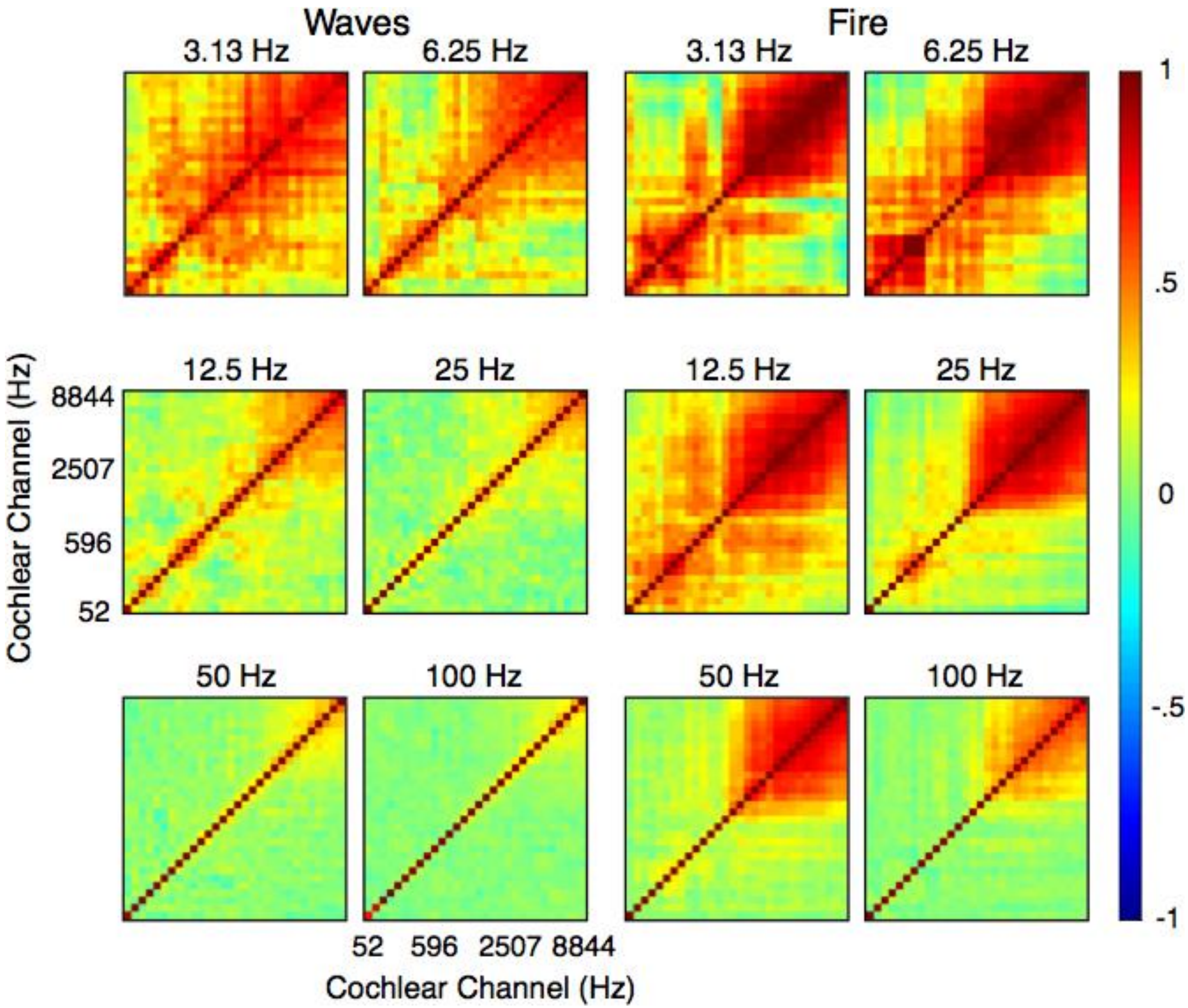
Textures vary in distribution of modulation power:



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

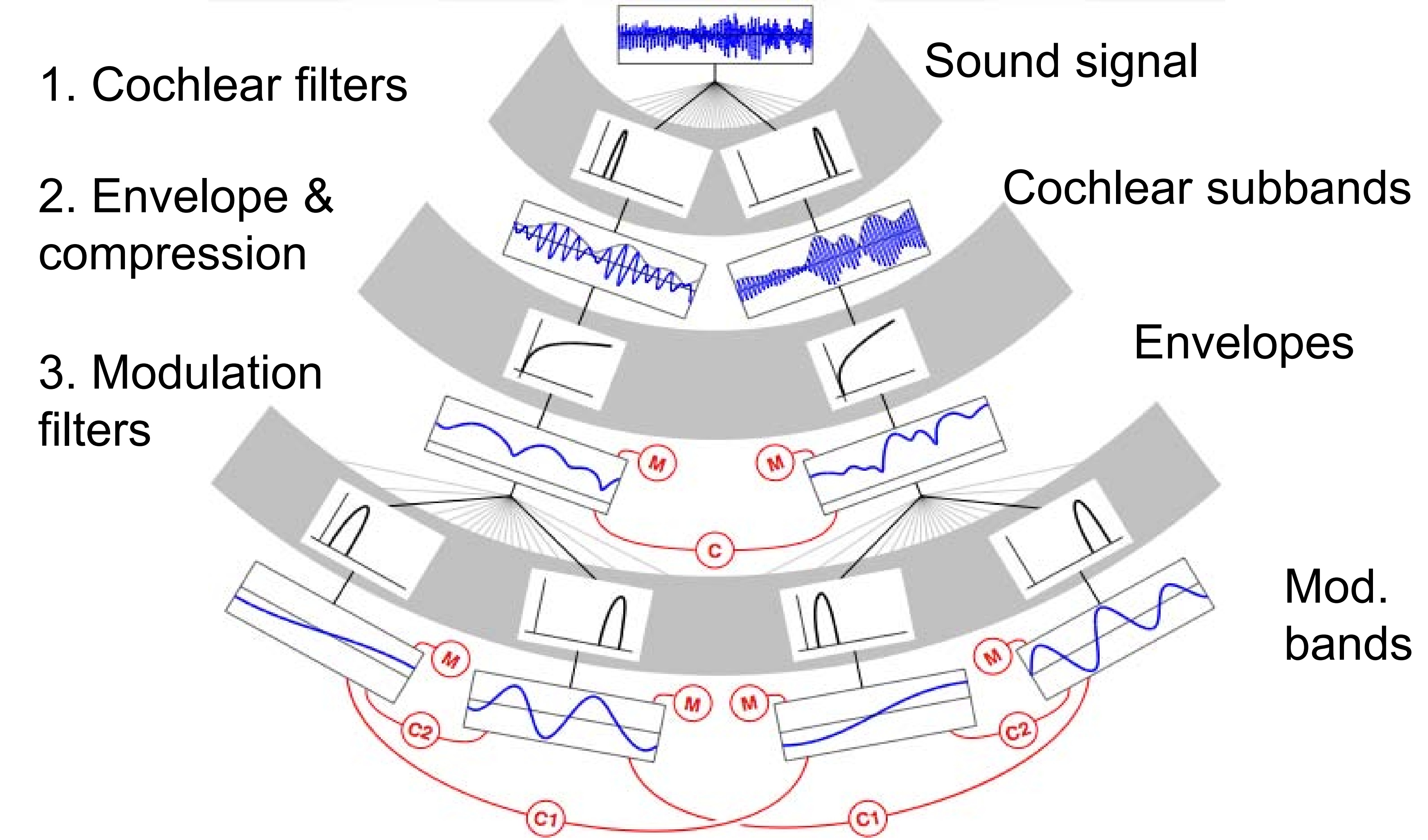


Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

These statistics capture variation across sound...



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Can they account for perception of real-world textures?

Key Methodological Proposal:

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

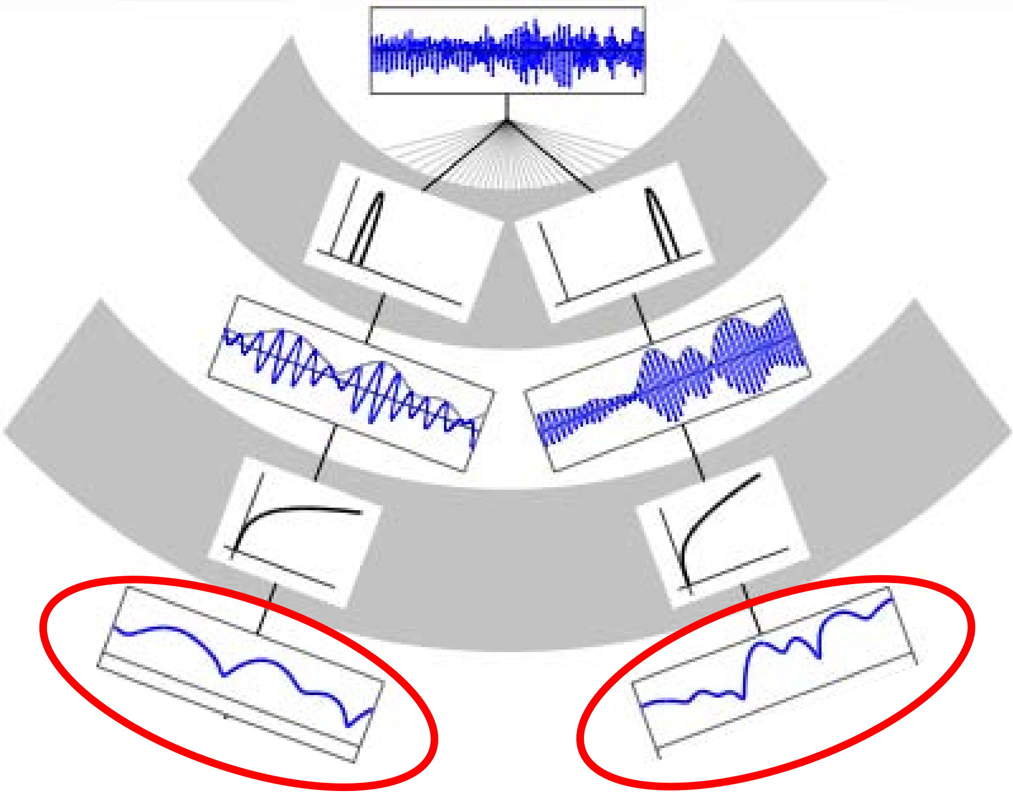
- If your brain represents sounds with a set of measurements, then:

    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)

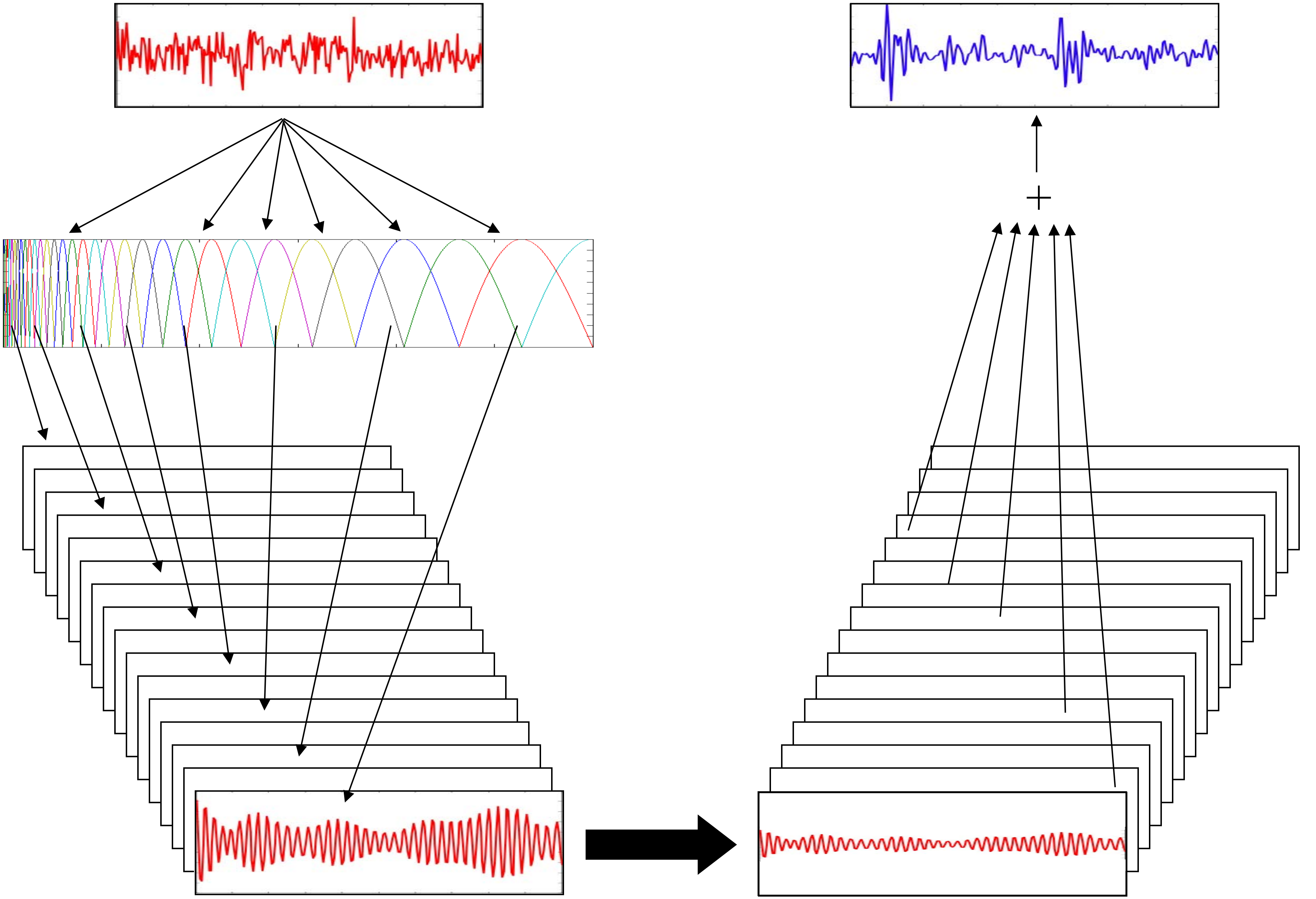


Envelopes

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

- 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?

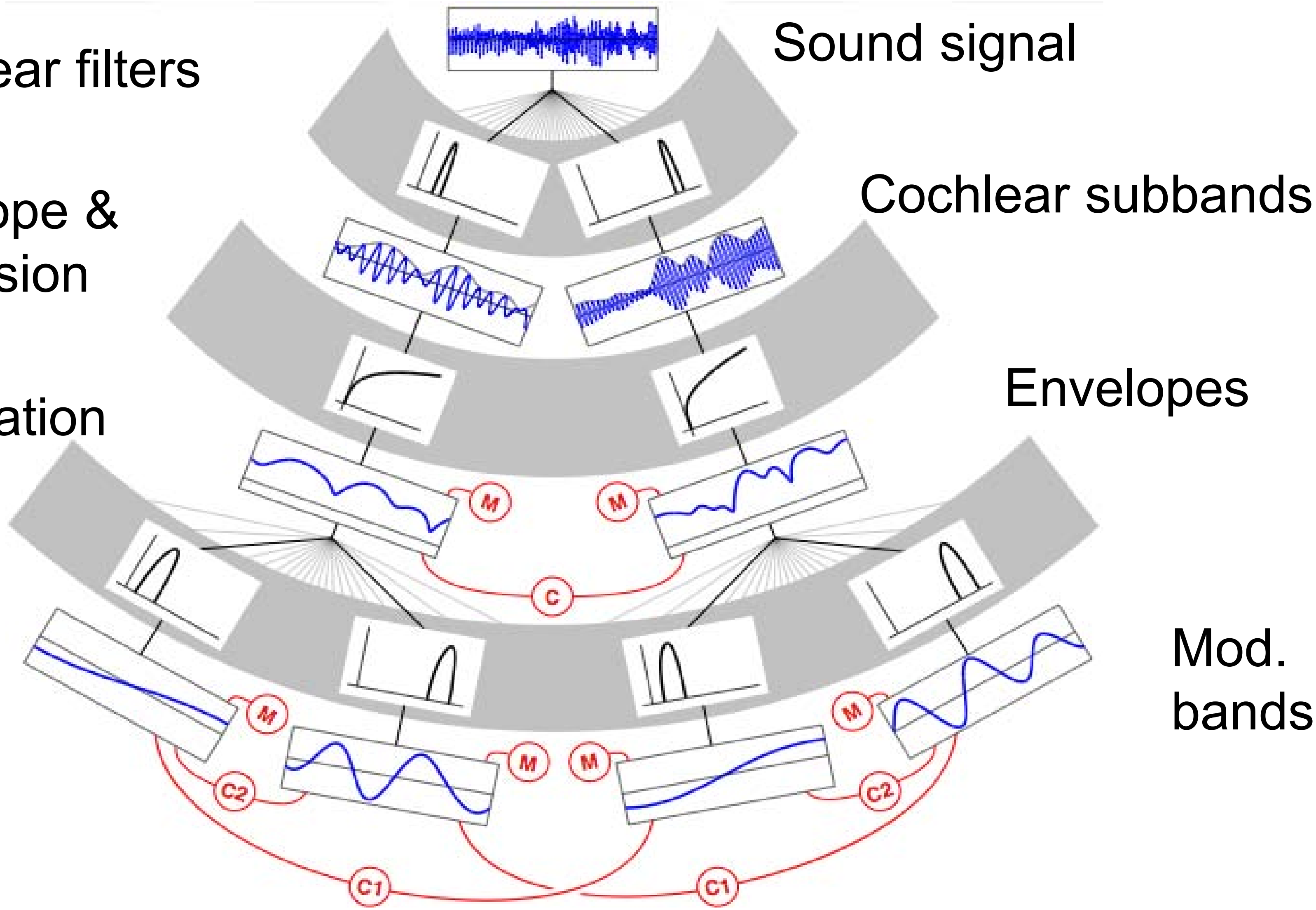
	Original	Power
Rain		
Stream		
Bubbles		
Fire		
Applause		

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

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

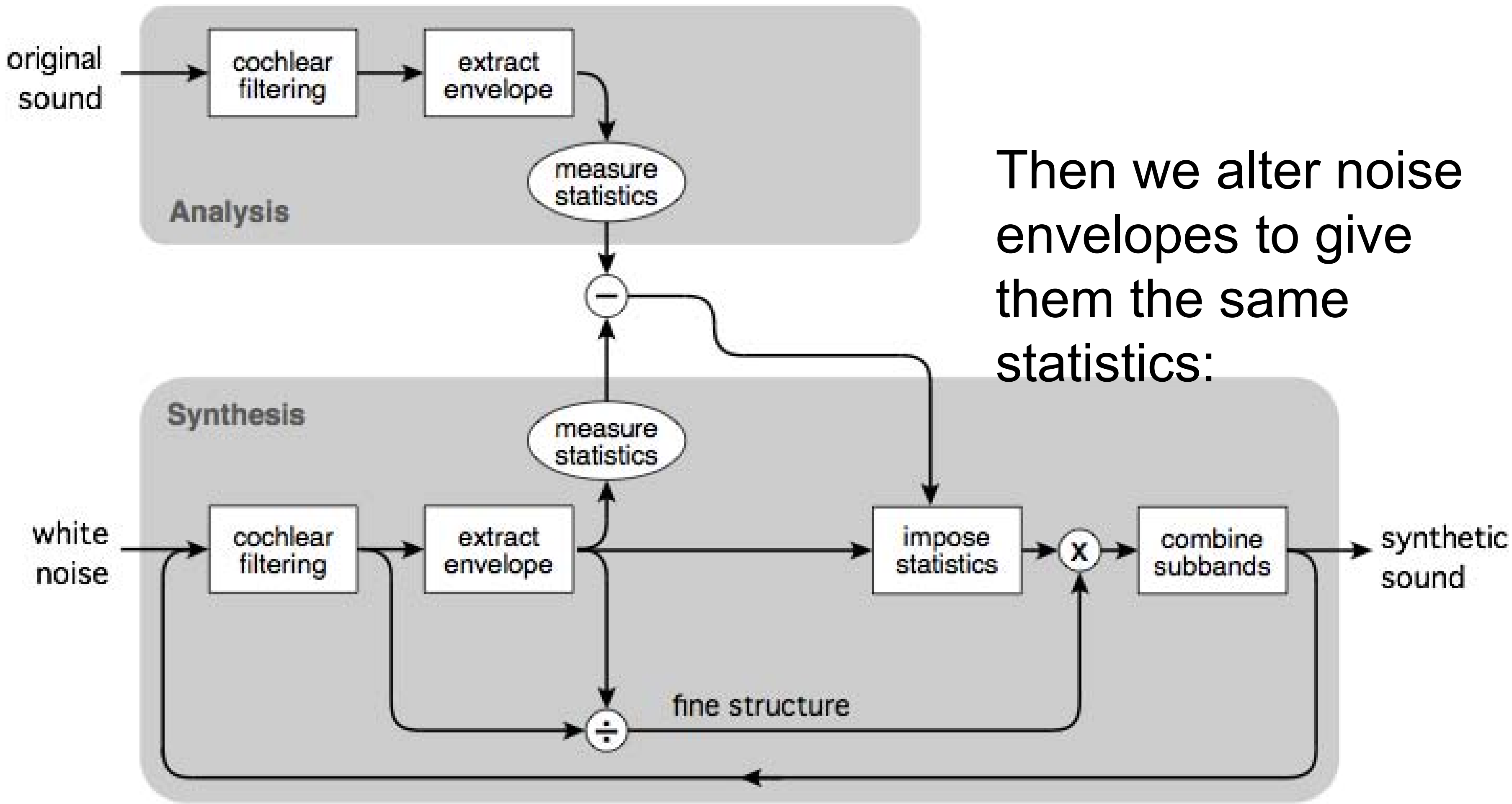
Will additional simple statistics do any better?

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



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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

Source: McDermott, Josh H., and Eero P. Simoncelli."Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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:

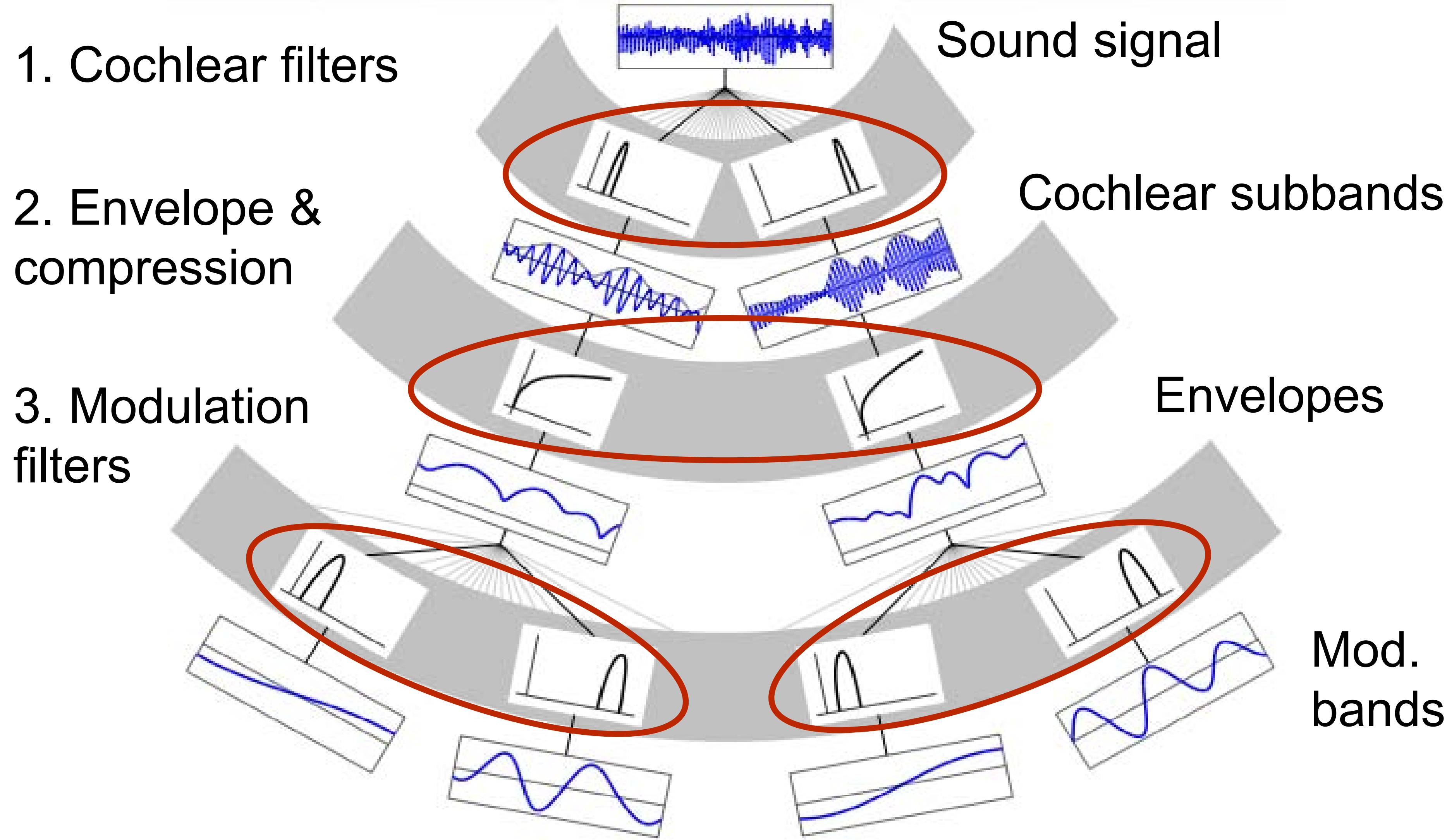
	Original	Power		All Stats
Rain				
Stream				
Bubbles				
Fire				
Applause				
Wind				
Insects				
Birds				
Crowd				

Also works for many “unnatural” sounds:

	Original	Power		All Stats
Rustling Paper				
Train				
Helicopter				
Jackhammer				

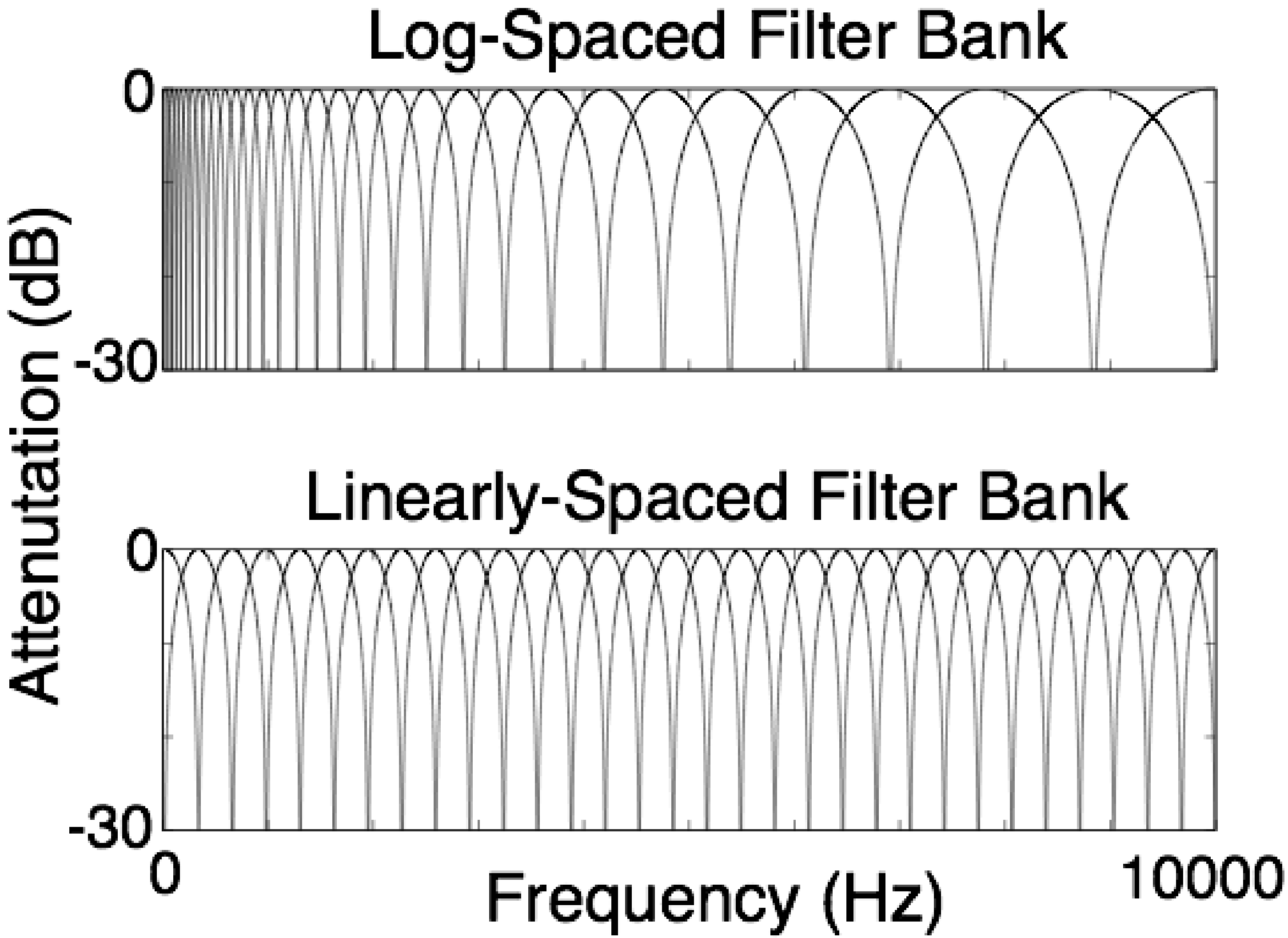
Success of synthesis suggests these statistics could underlie representation and recognition of textures.

Will any set of statistics do?
What if we measure statistics from model deviating from biology?



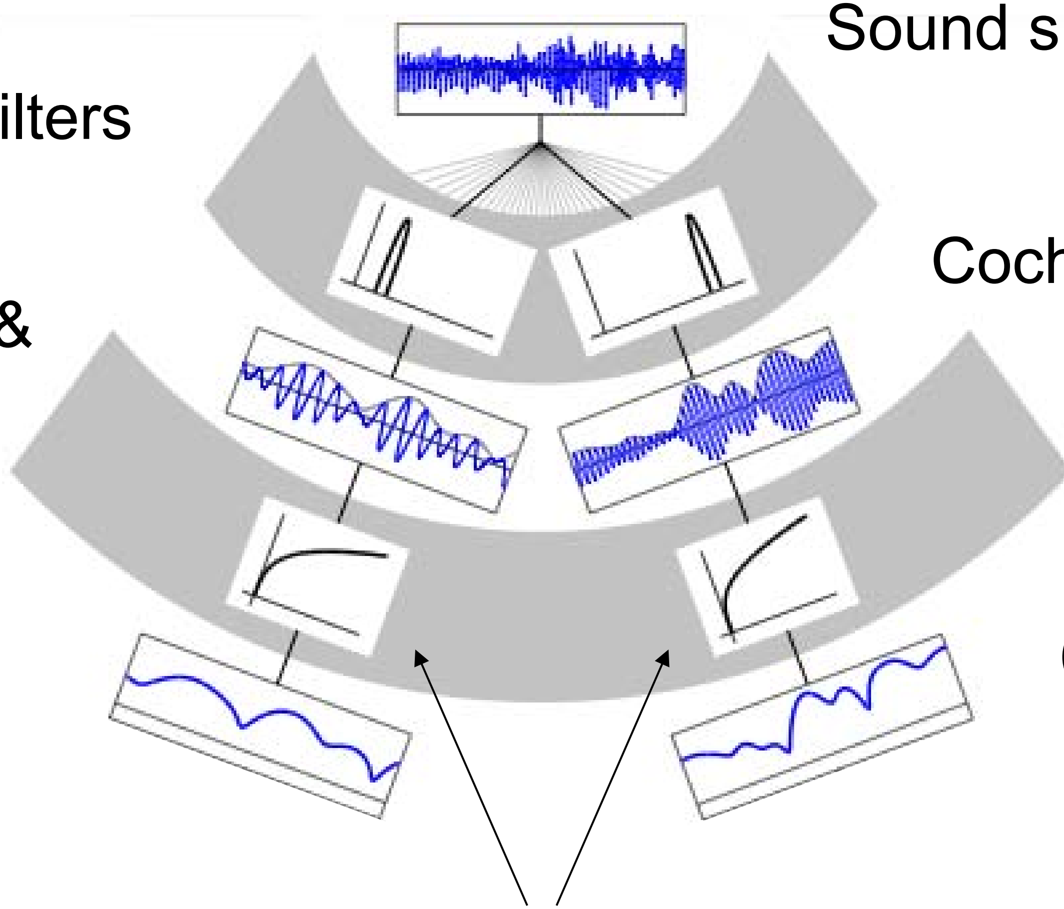
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Will any set of statistics do?
What if we measure statistics from model deviating from biology?



Source: McDermott, Josh H., and Eero P. Simoncelli."Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

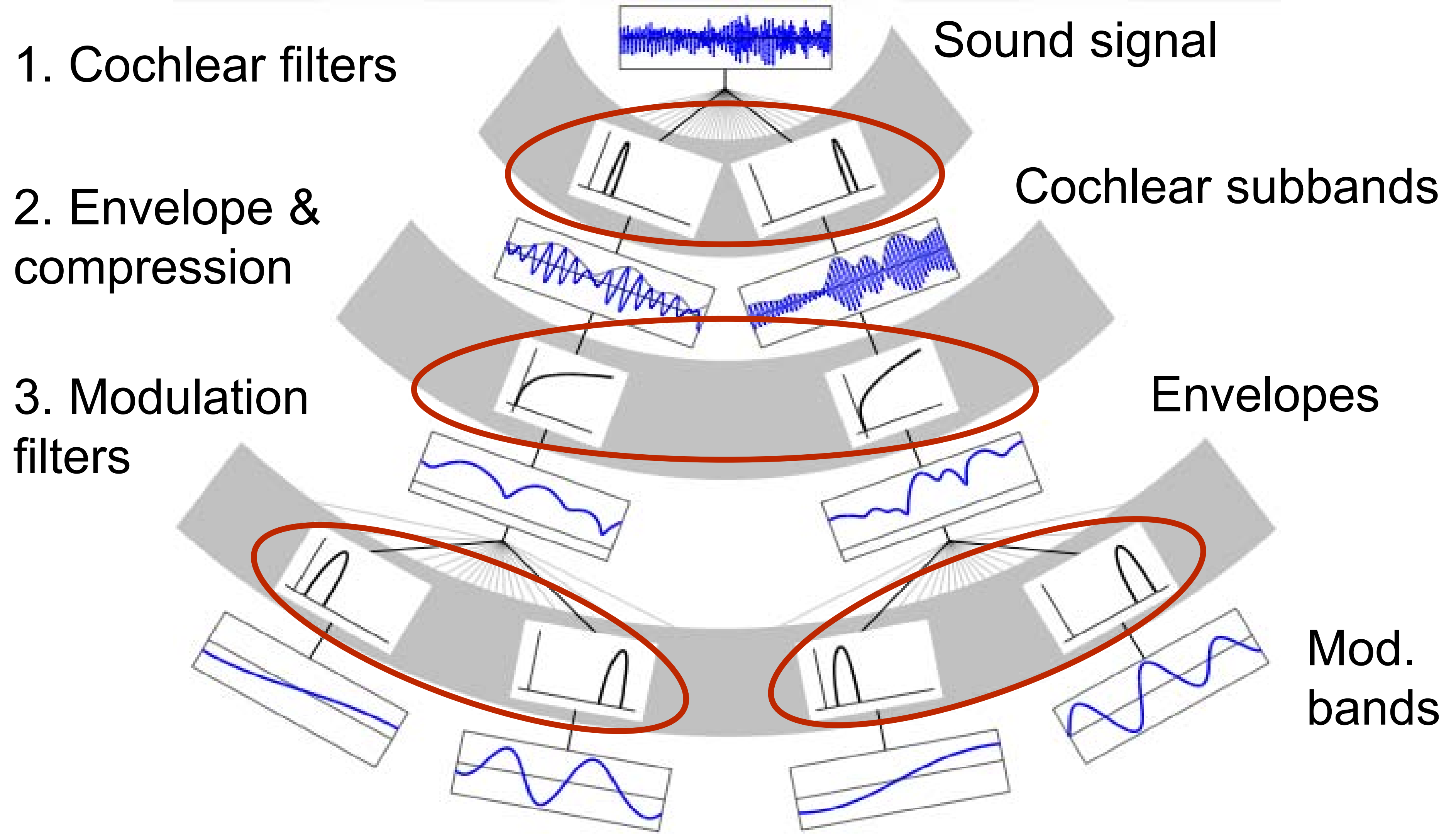
AUDITORY MODEL



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

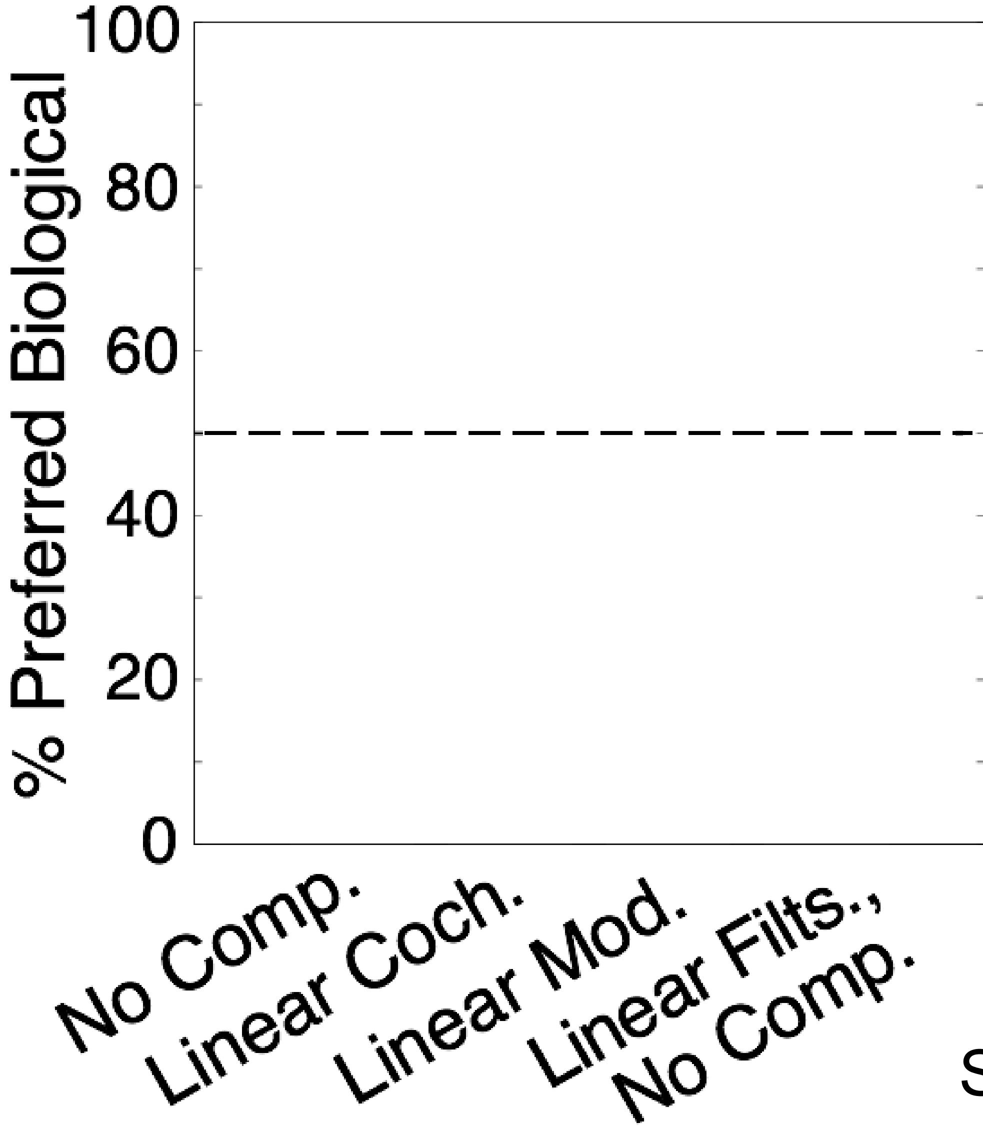
Amplitude compression, simulating that of cochlea

Will any set of statistics do?
What if we measure statistics from model deviating from biology?



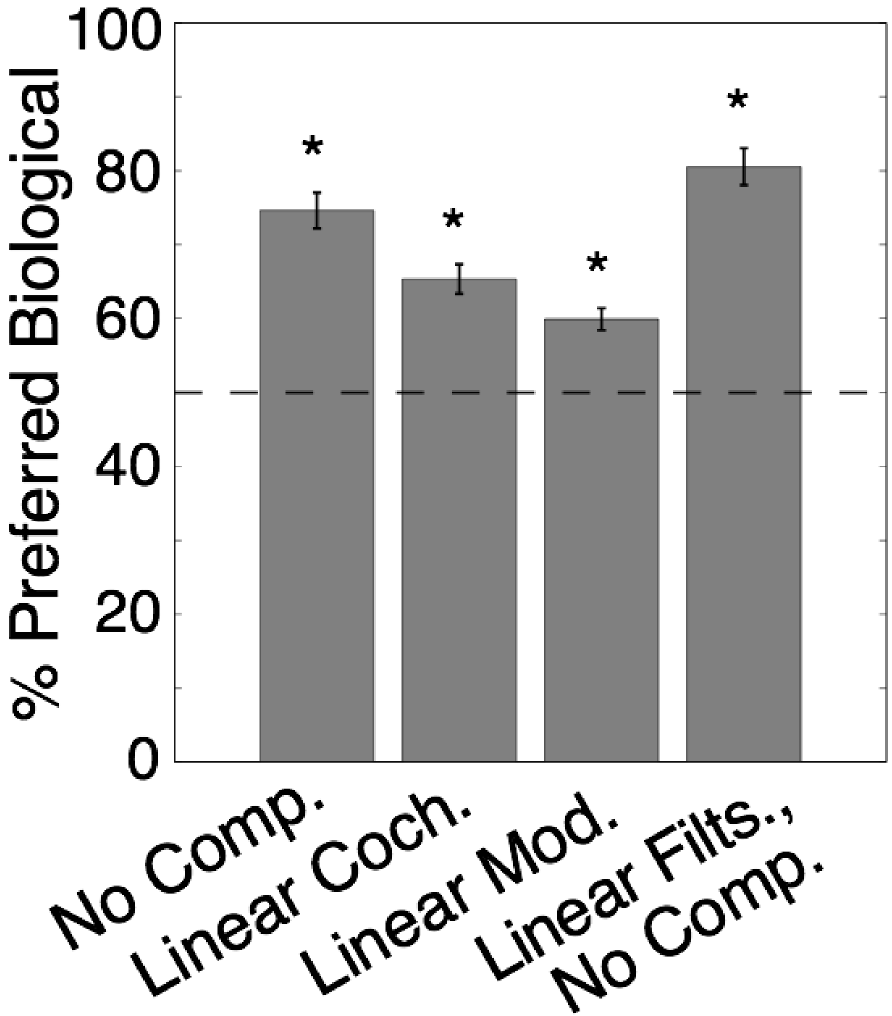
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Experiment: Original - Synth1 - Synth2
Which synthetic version sounds more realistic?



McDermott & Simoncelli, Neuron, 2011

Biologically inspired model is crucial - altering either filter bank, or compression, degrades synthesis:



Crowd Noise:

Biological

Non-biological

Helicopter:

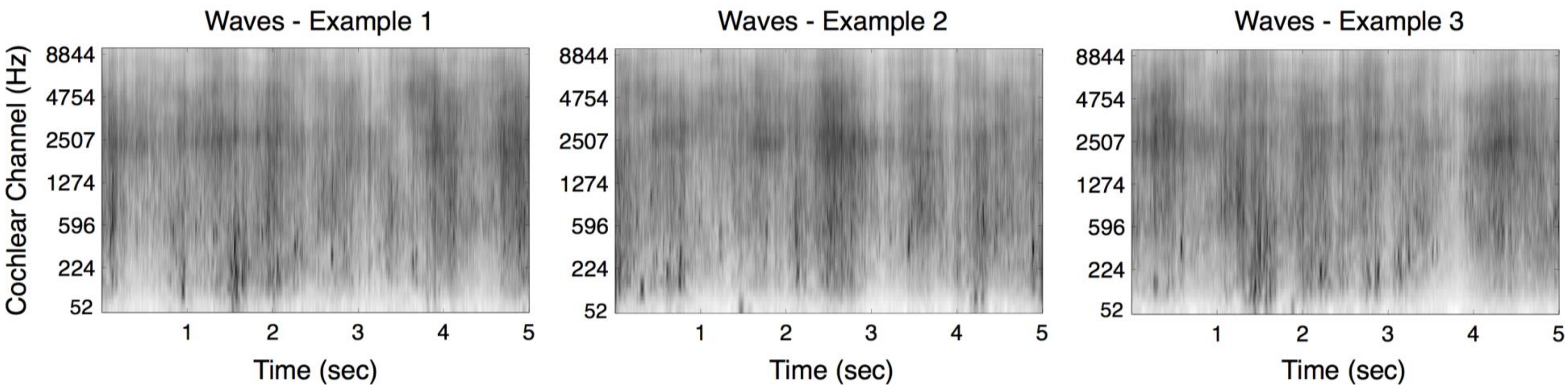
Biological

Non-biological

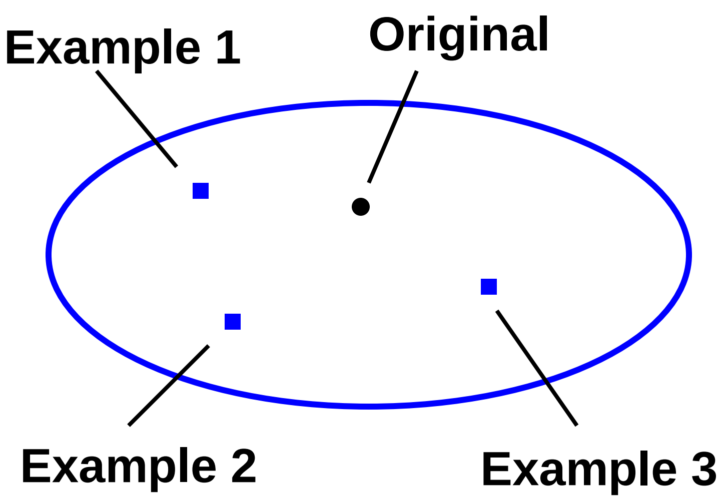
McDermott & Simoncelli, Neuron, 2011

Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

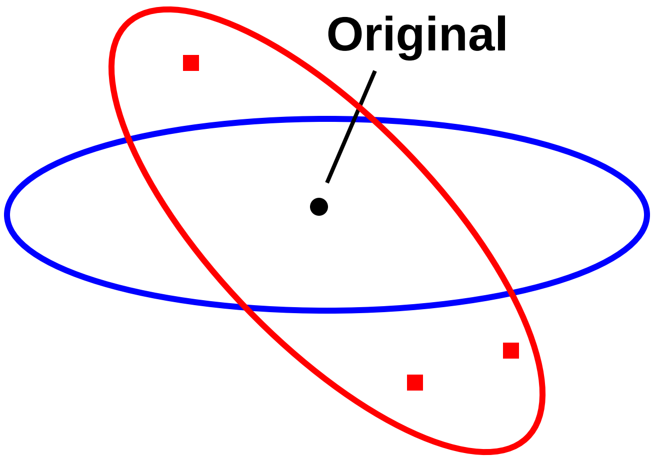
Procedure is initialized with noise, so produces a different sound signal every time, sharing only statistical properties:



- Statistics define a class of sounds that include the original and many others.
- If the statistics measure what the brain is measuring, the samples should sound like another example of the original sound.

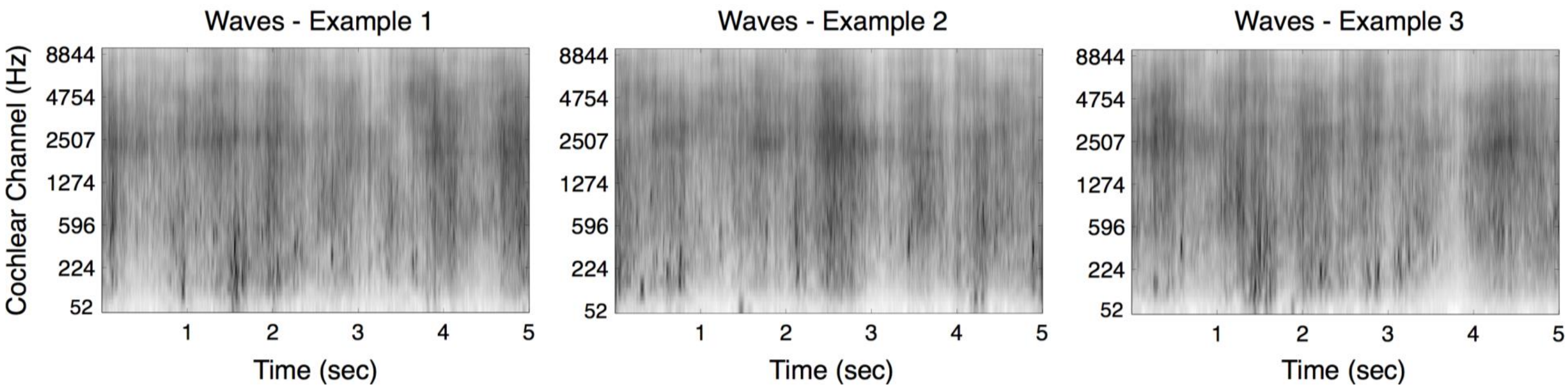


Statistics of non-biological model define a different class of sounds:



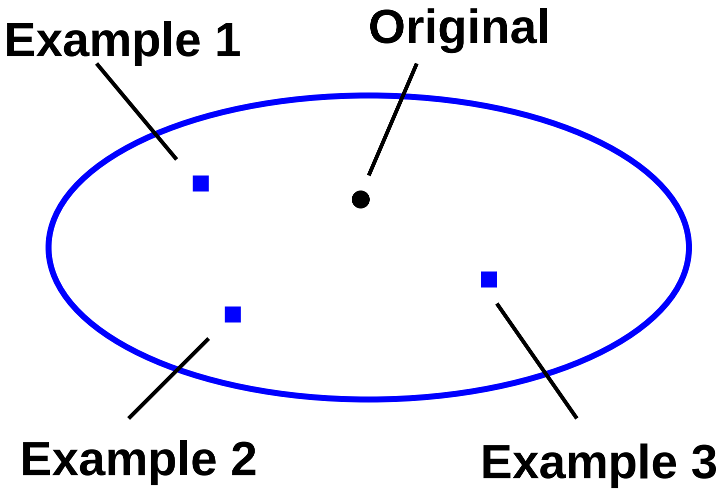
The sounds in the non-biological class don't sound like the original, because they are not defined with the measurements the brain is making.

Many sound signals have the same statistics:



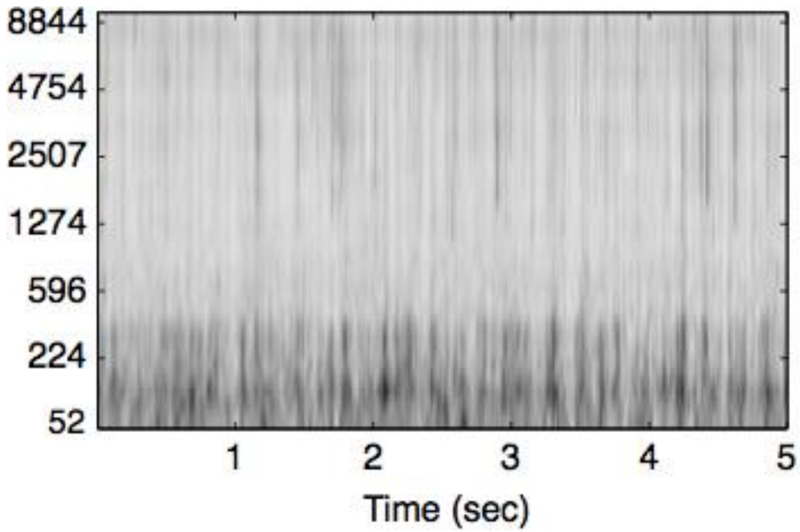
- Time-averages provide invariance.

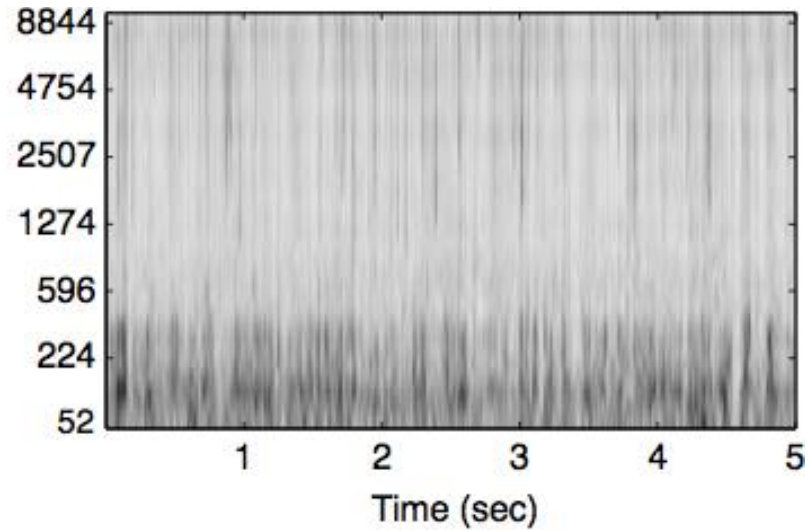
- Many sound signals have the same statistics…



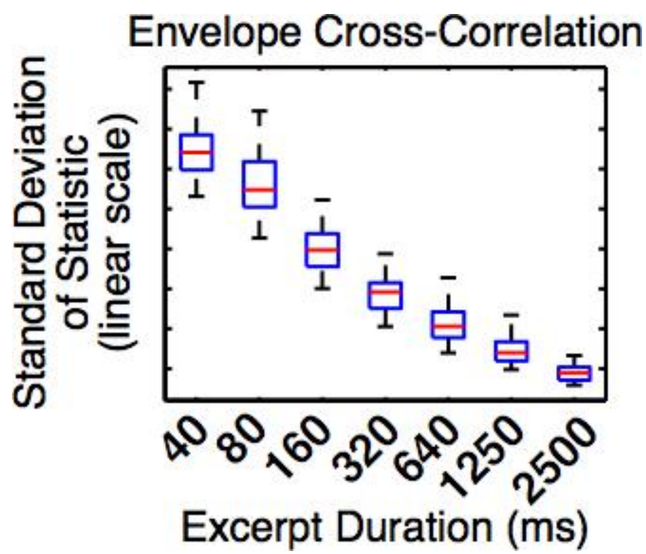
Interesting possibility:

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

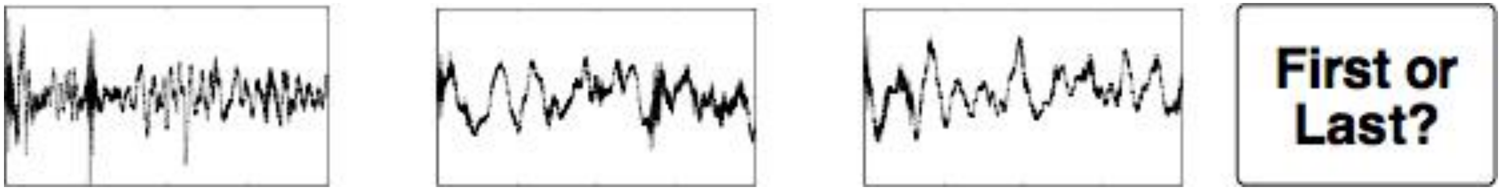




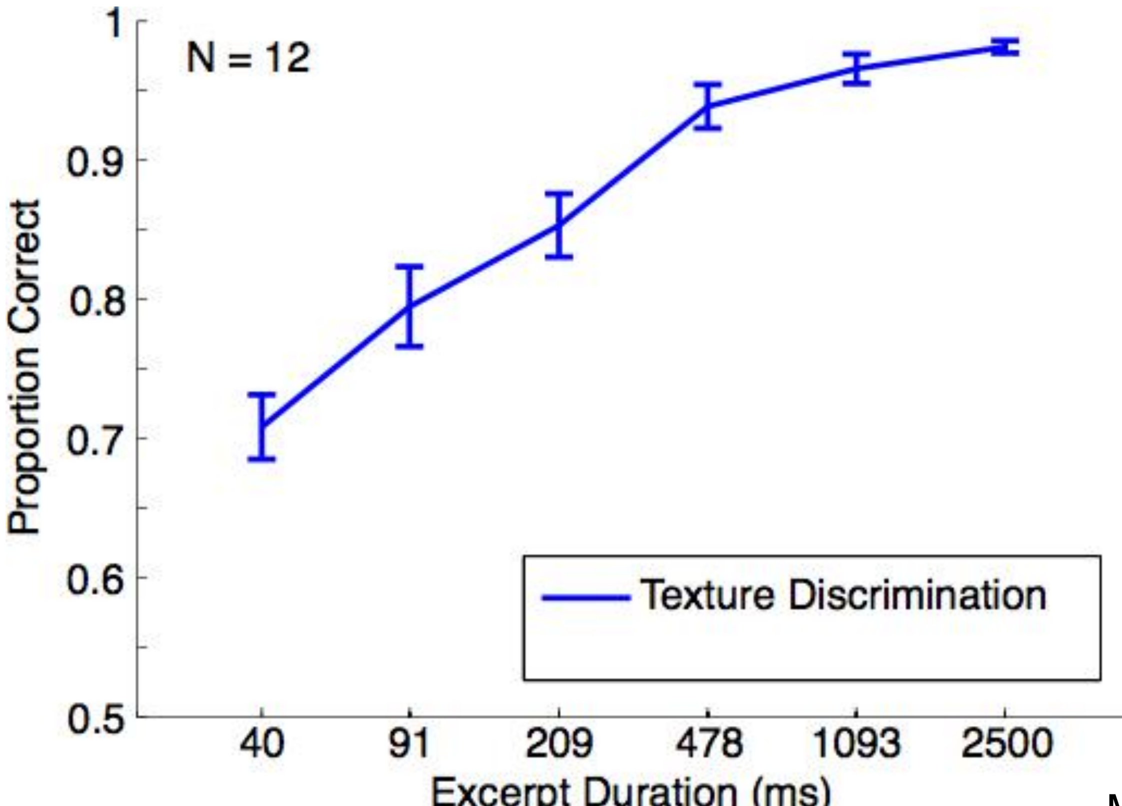
But only for long excerpts:



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



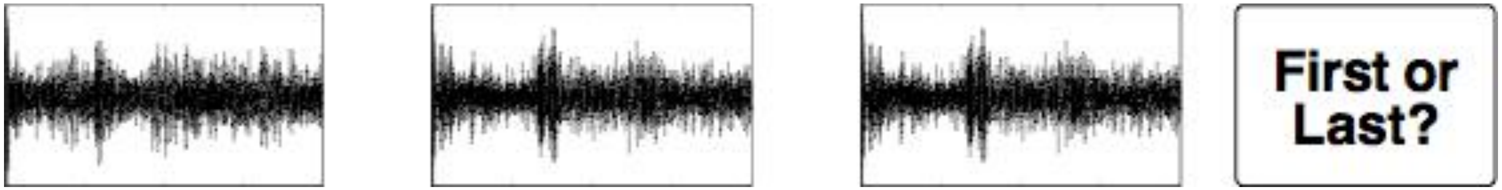
Which sound was produced by a different source?



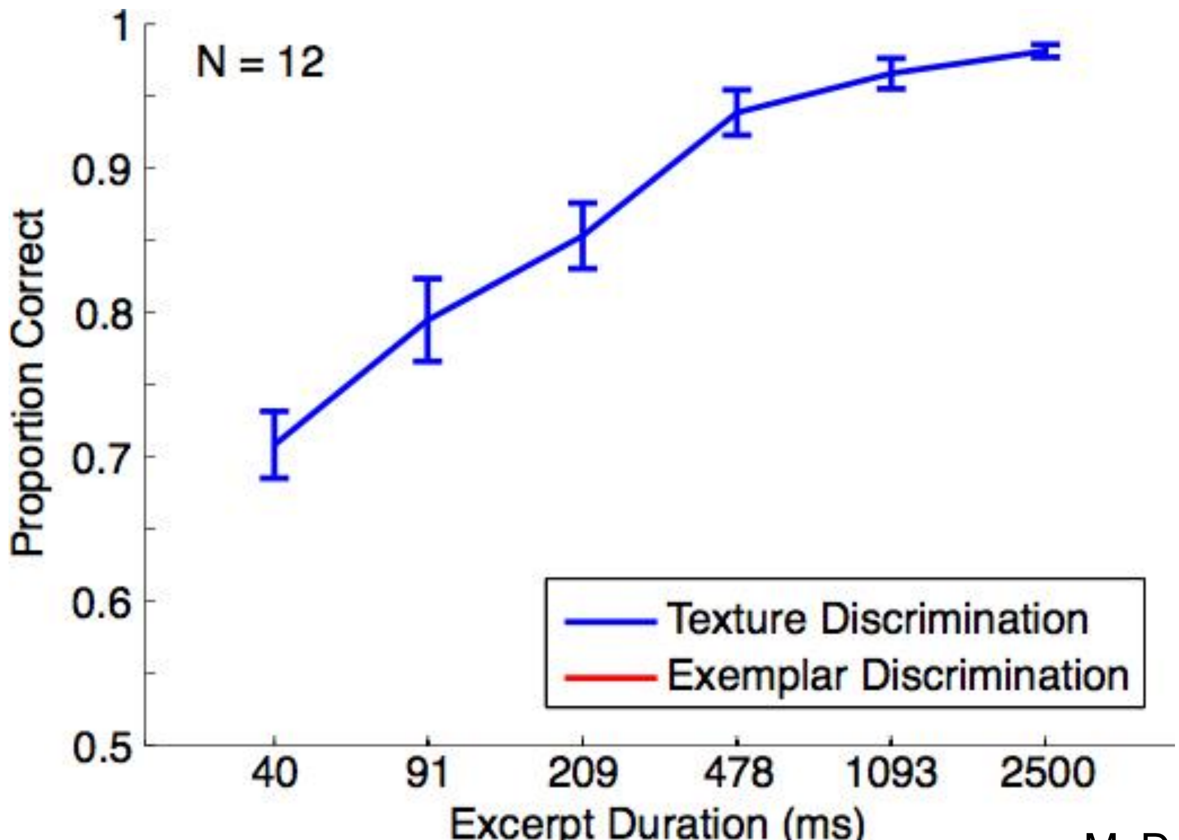
McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

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



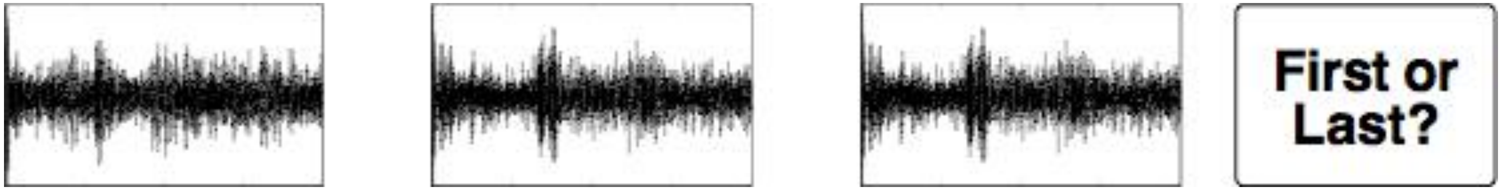
Which sound was different from the other two?



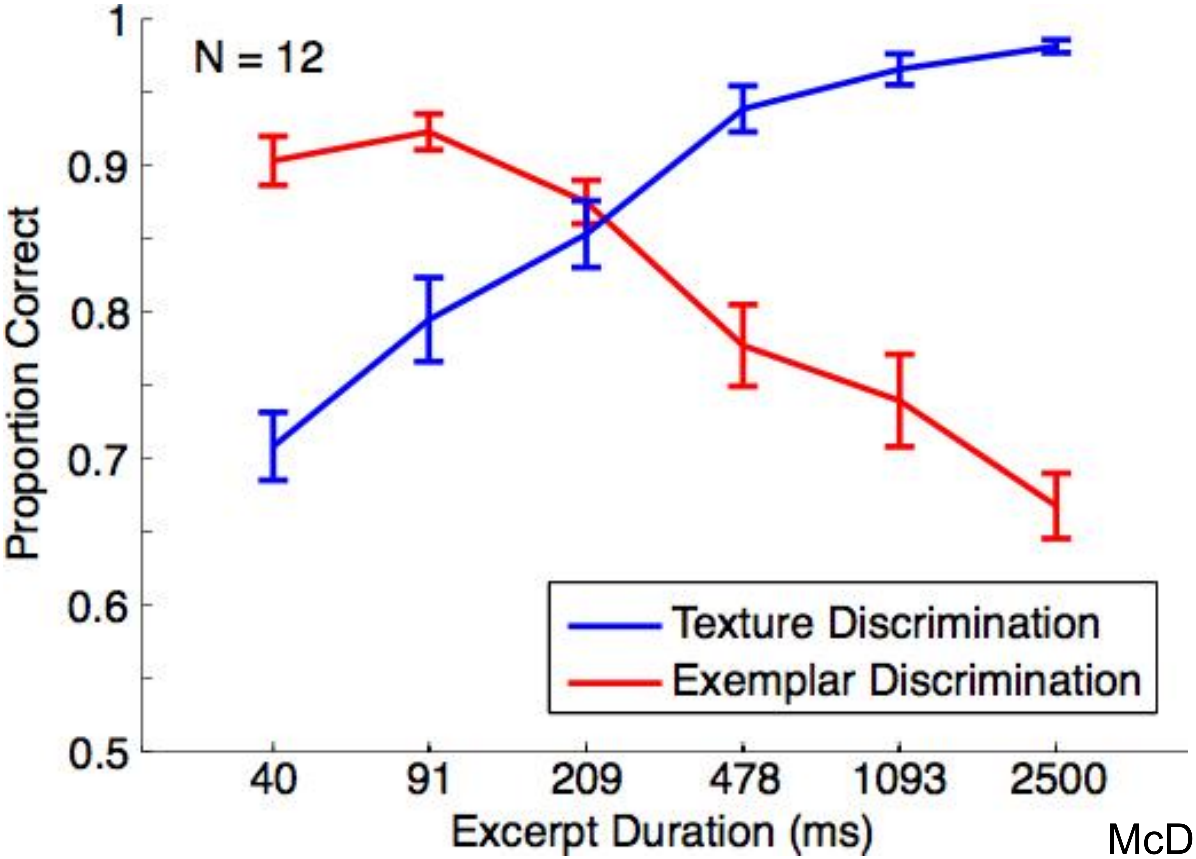
McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

Although information increases w. duration, performance declines!



Which sound was different from the other two?

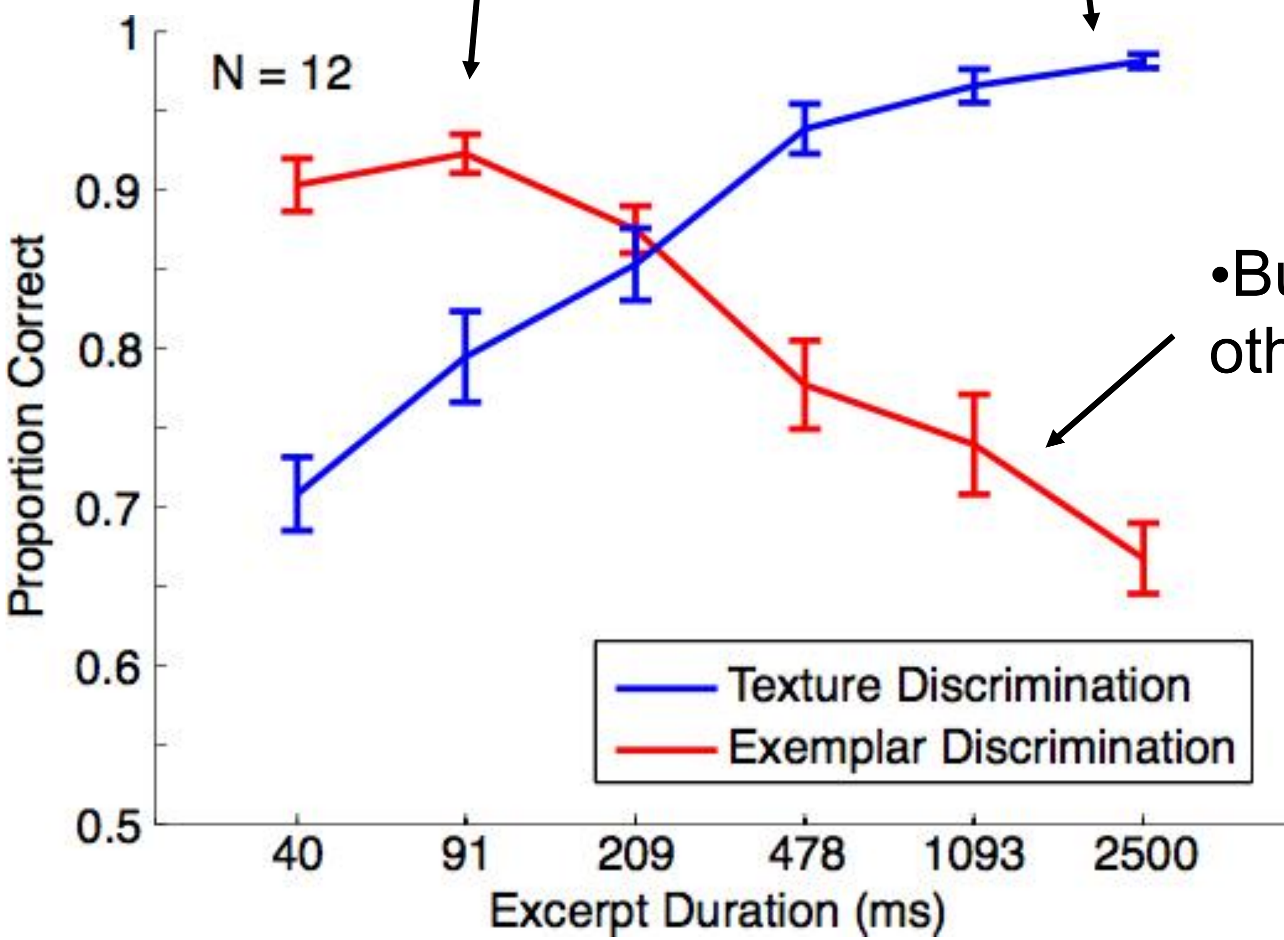


McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

•Textures contain discriminable detail…

•Details are apparently accrued into statistics…



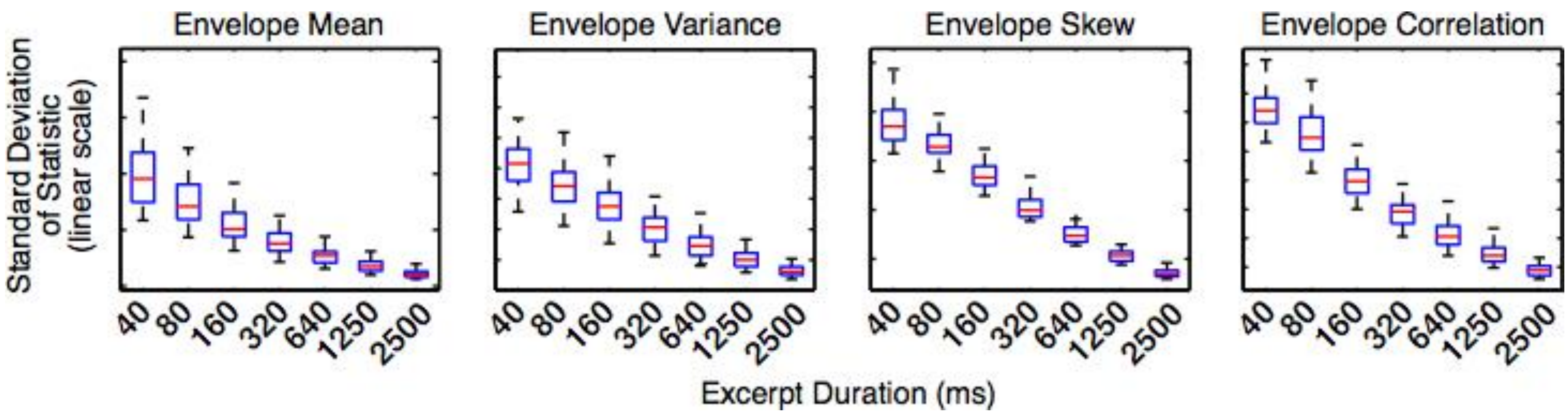
•But details are not otherwise retained.

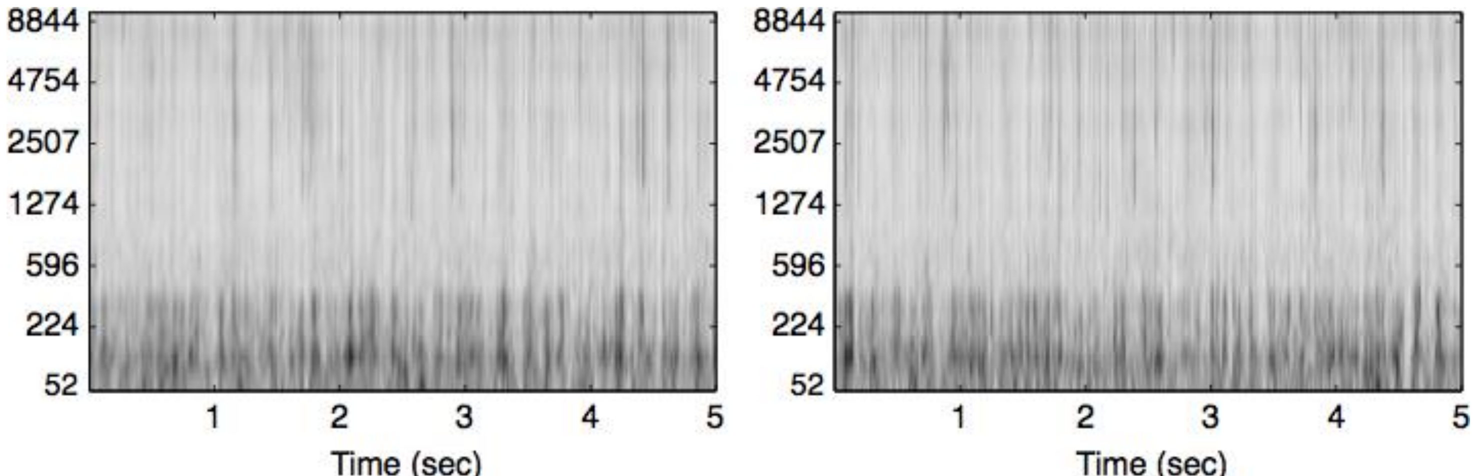
McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

Results suggest representation of time-averaged statistics.

(because statistics converge as duration increases)

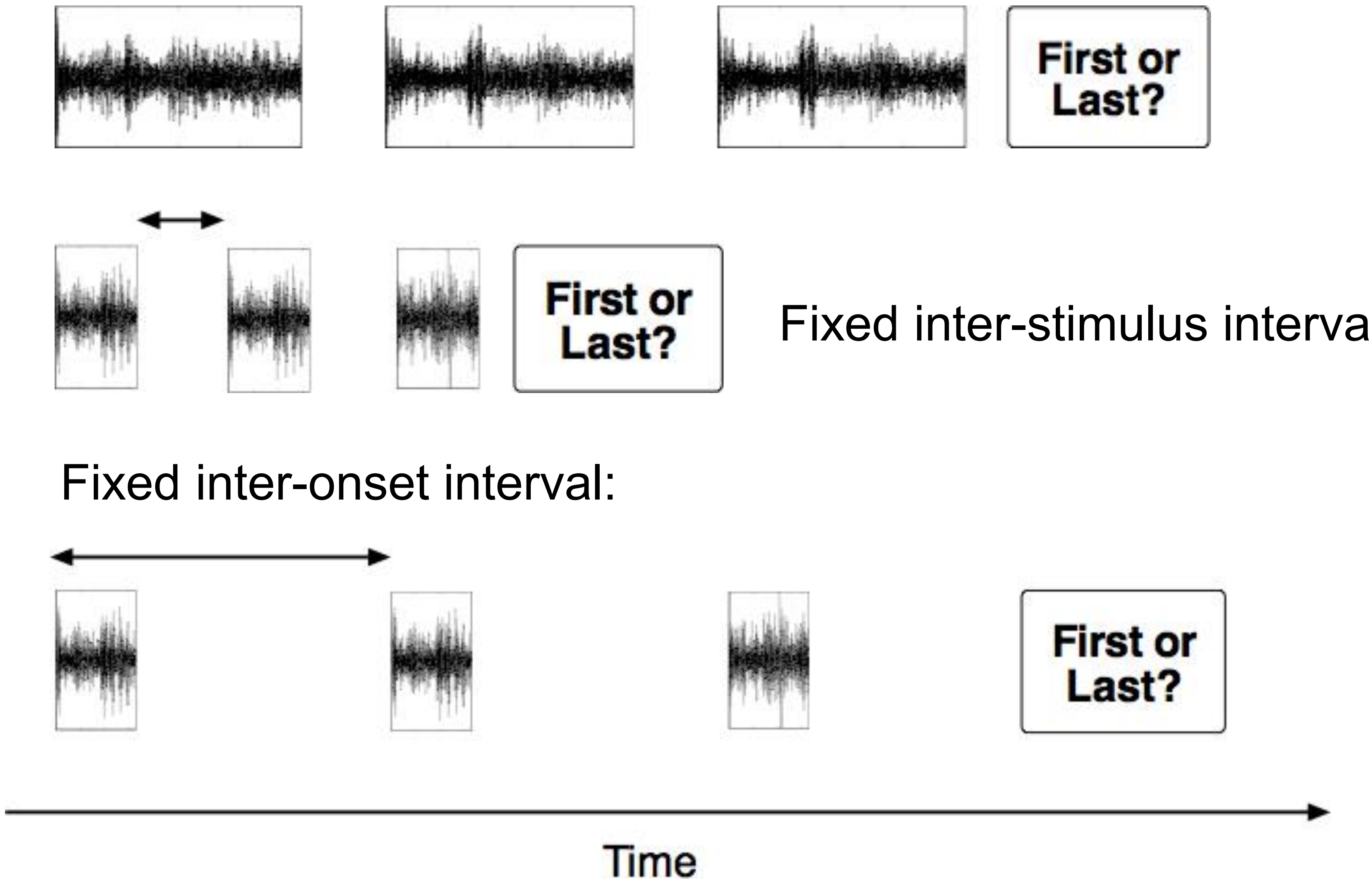




McDermott et al. 2013

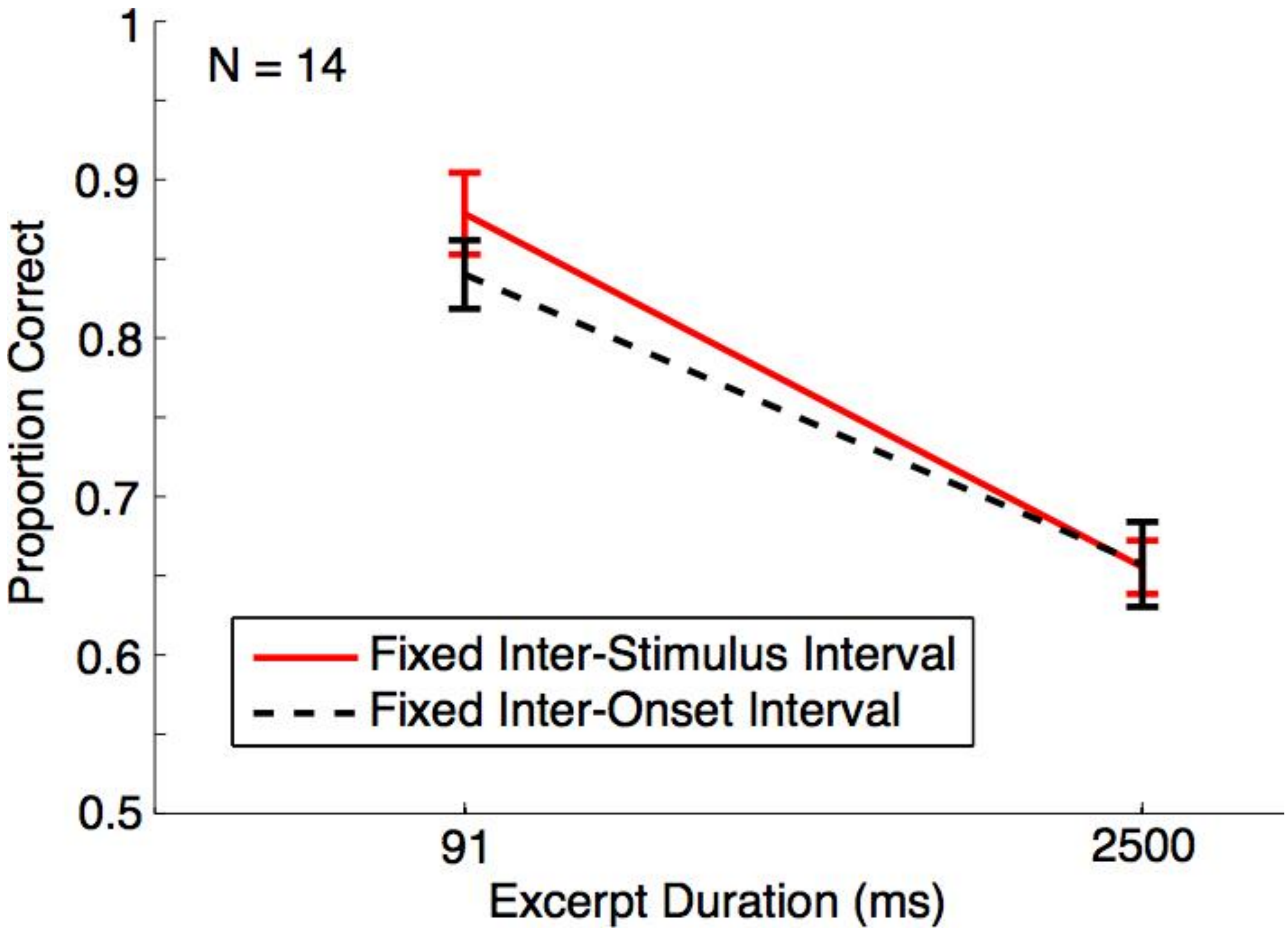
Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

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



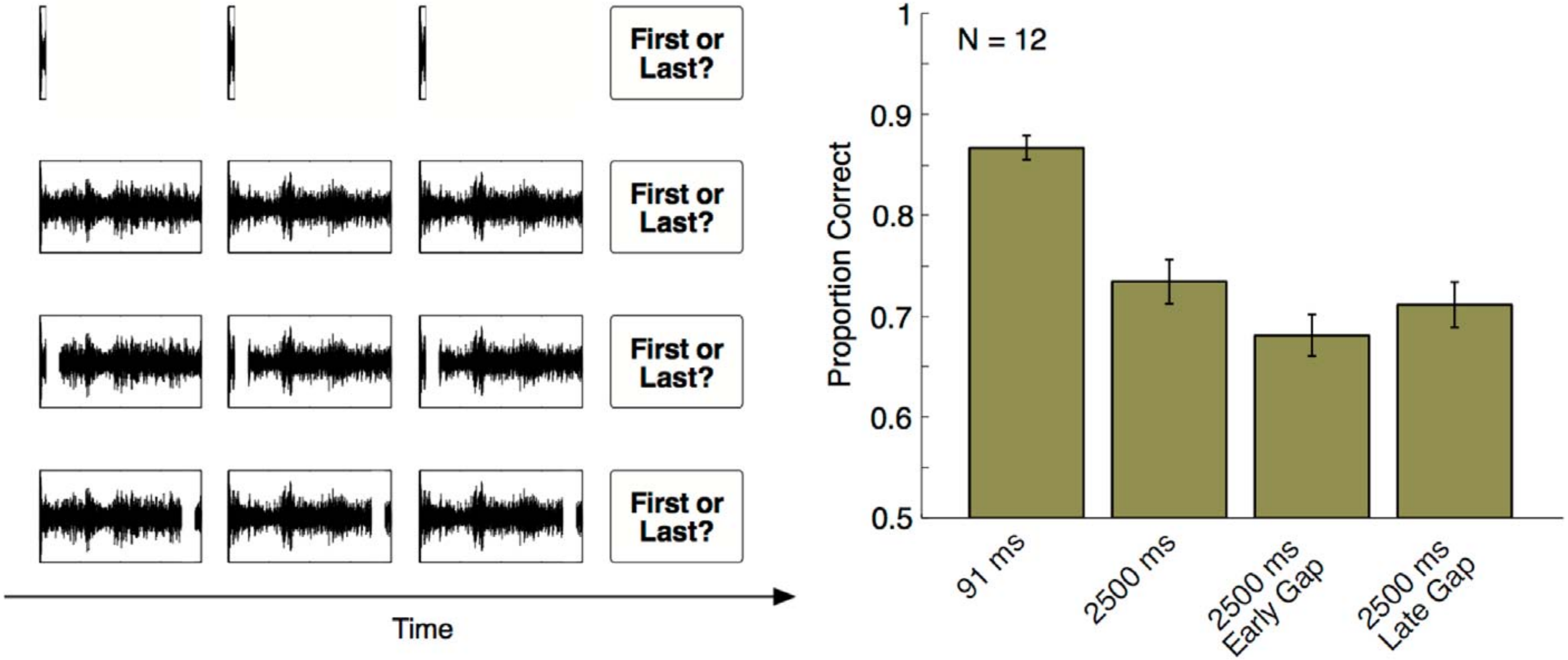
Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

Time delay per se has little effect:



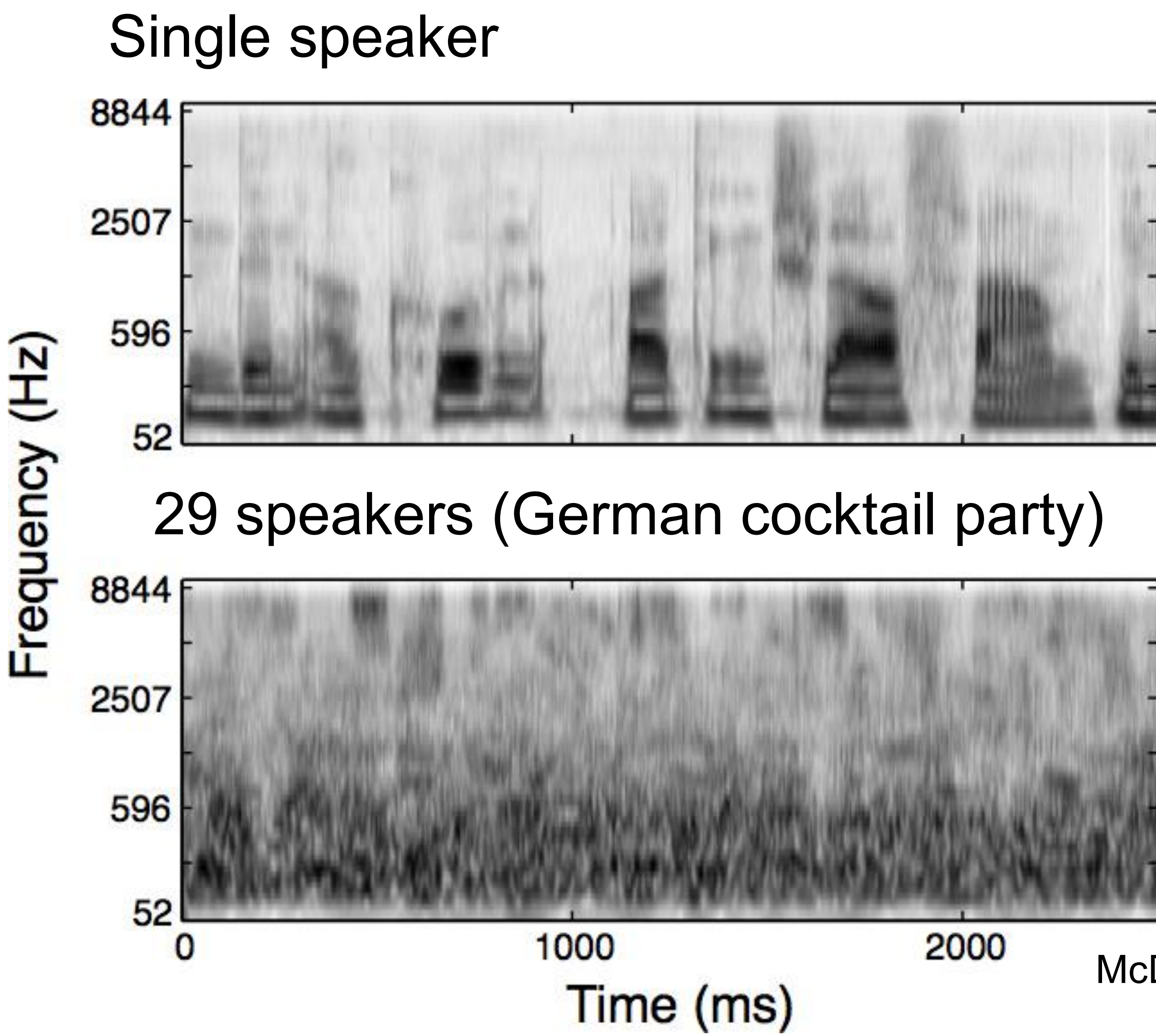
Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

Cueing subjects to beginning or end has little effect:



Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

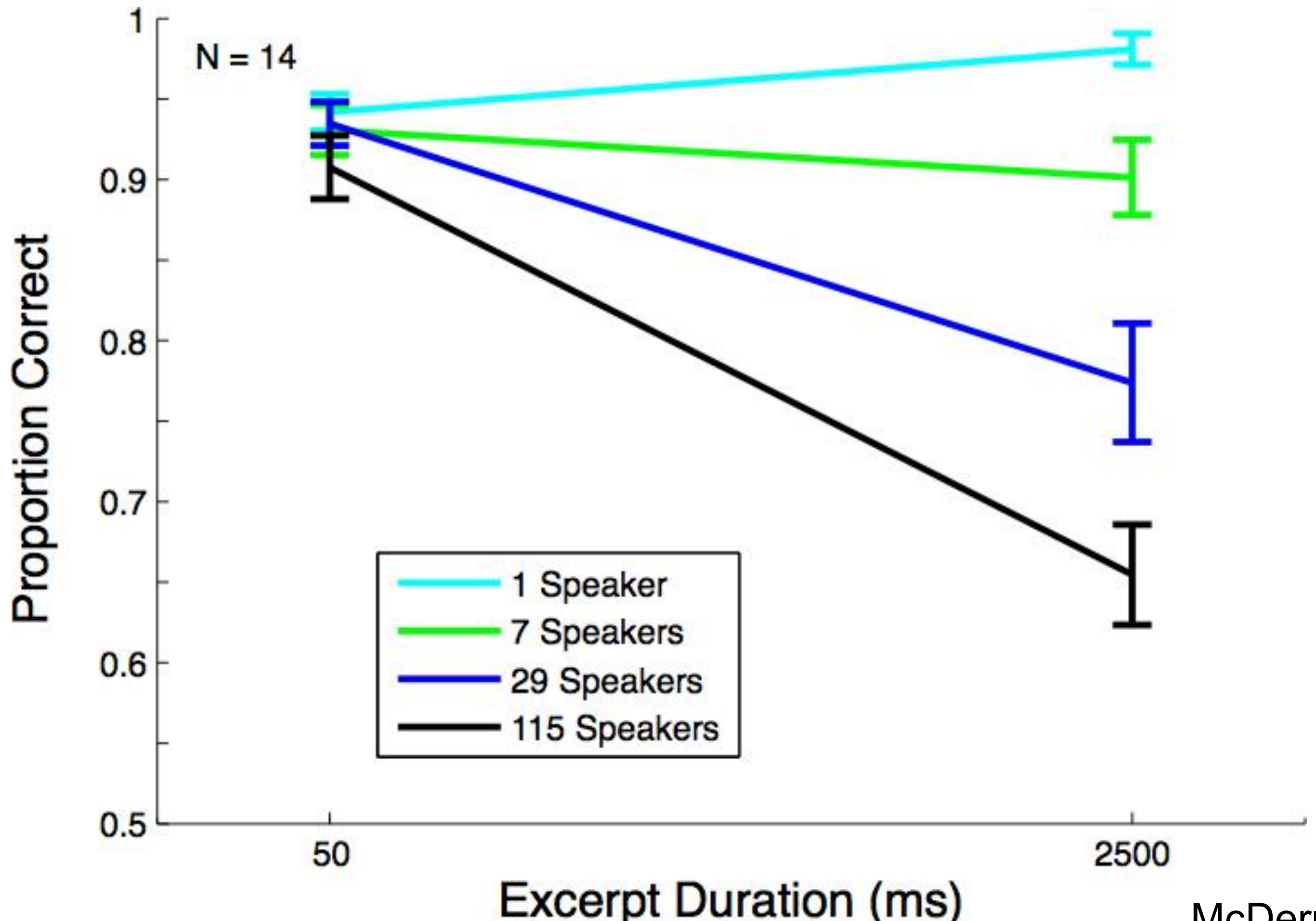
Textures are normally generated from superposition of sources…



McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

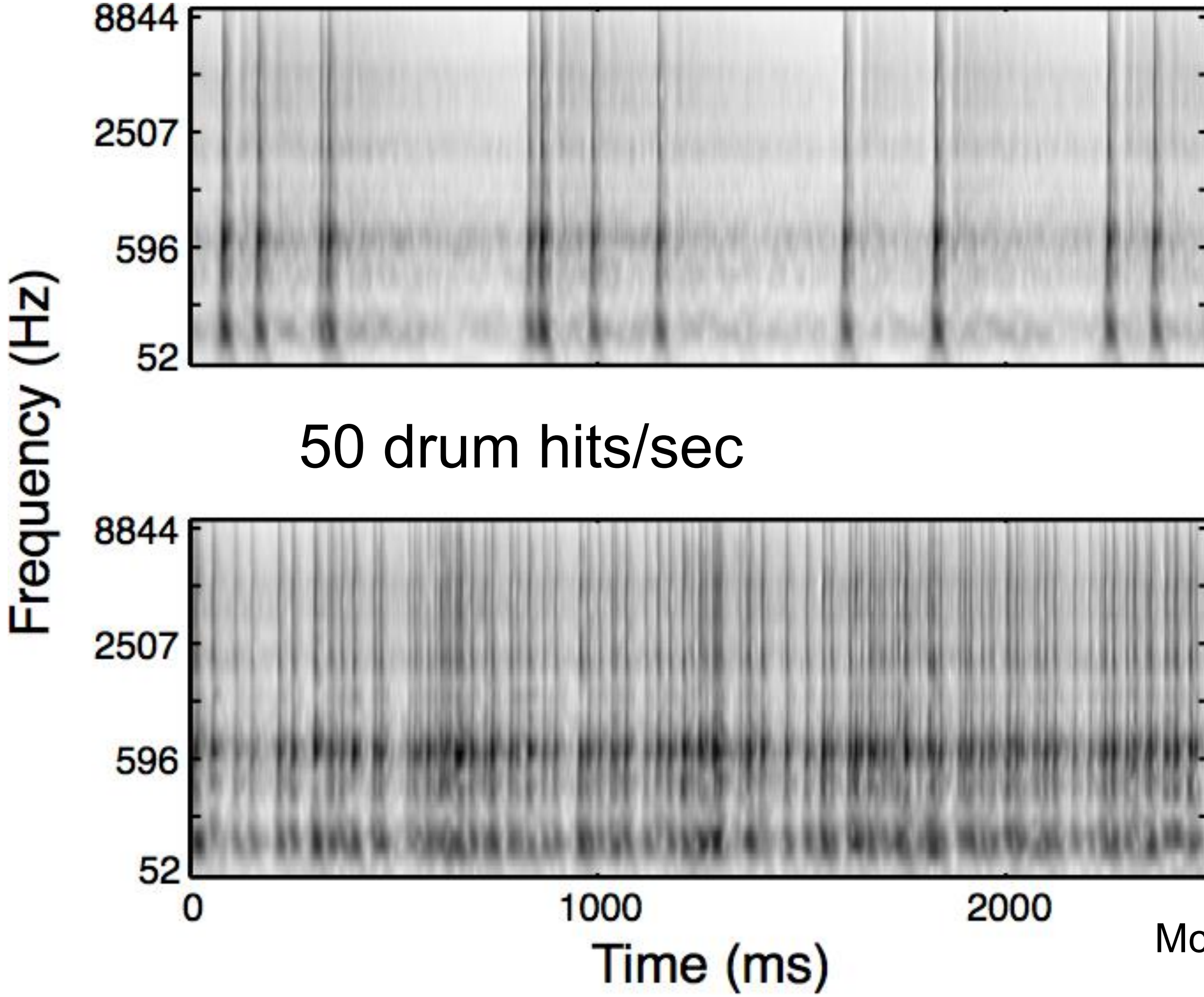
Impairment at long durations is specific to textures, not present for single sources:



McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

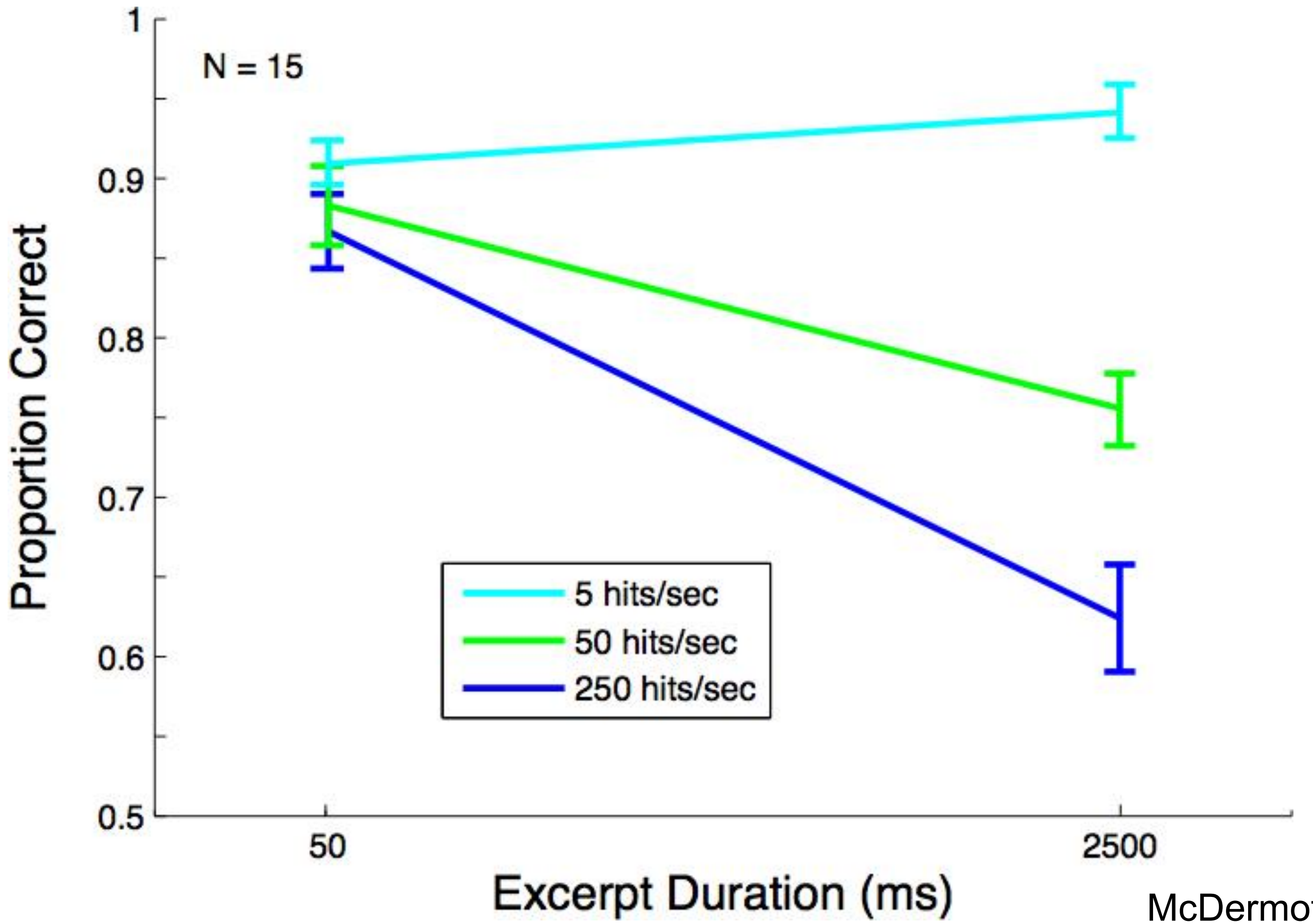
5 drum hits/sec



McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

Impairment at long durations again specific to dense textures.

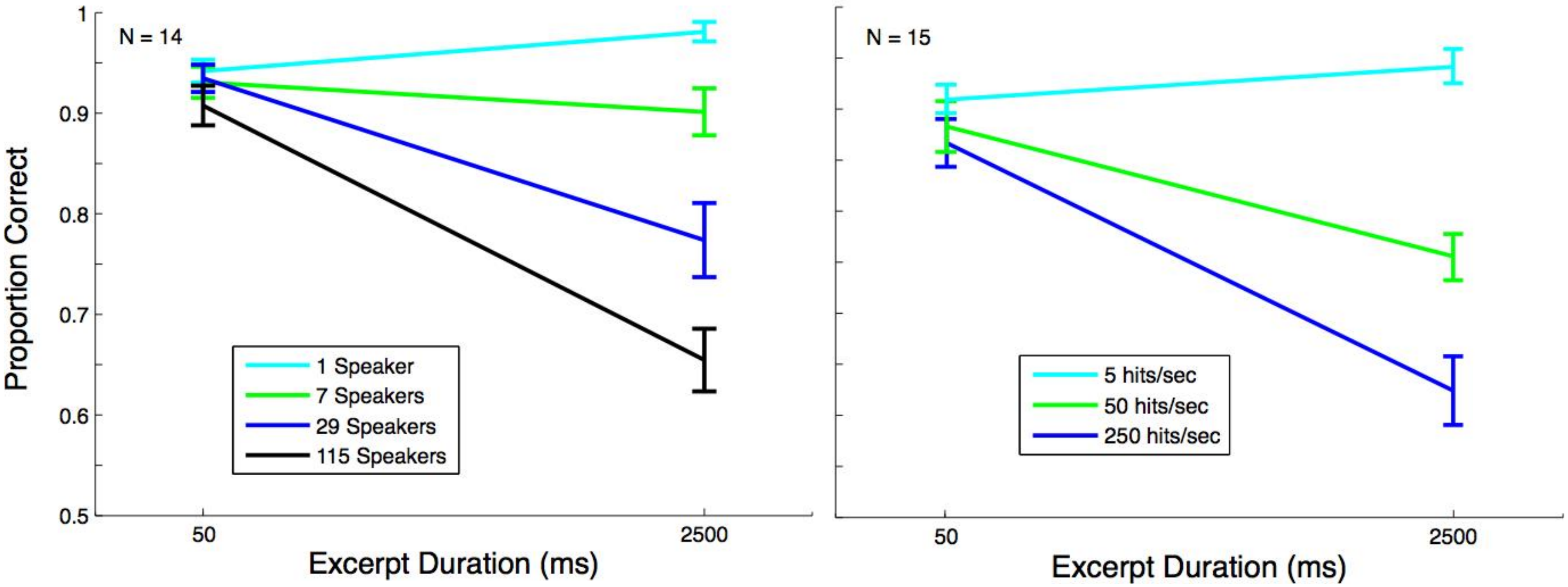


McDermott et al. 2013

Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

- High performance with short excerpts indicates that all stimuli have discriminable variation.

- But temporal detail is not retained when signals are texture-like.

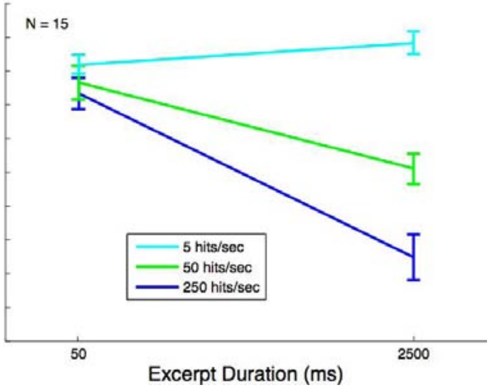


Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

McDermott et al. 2013

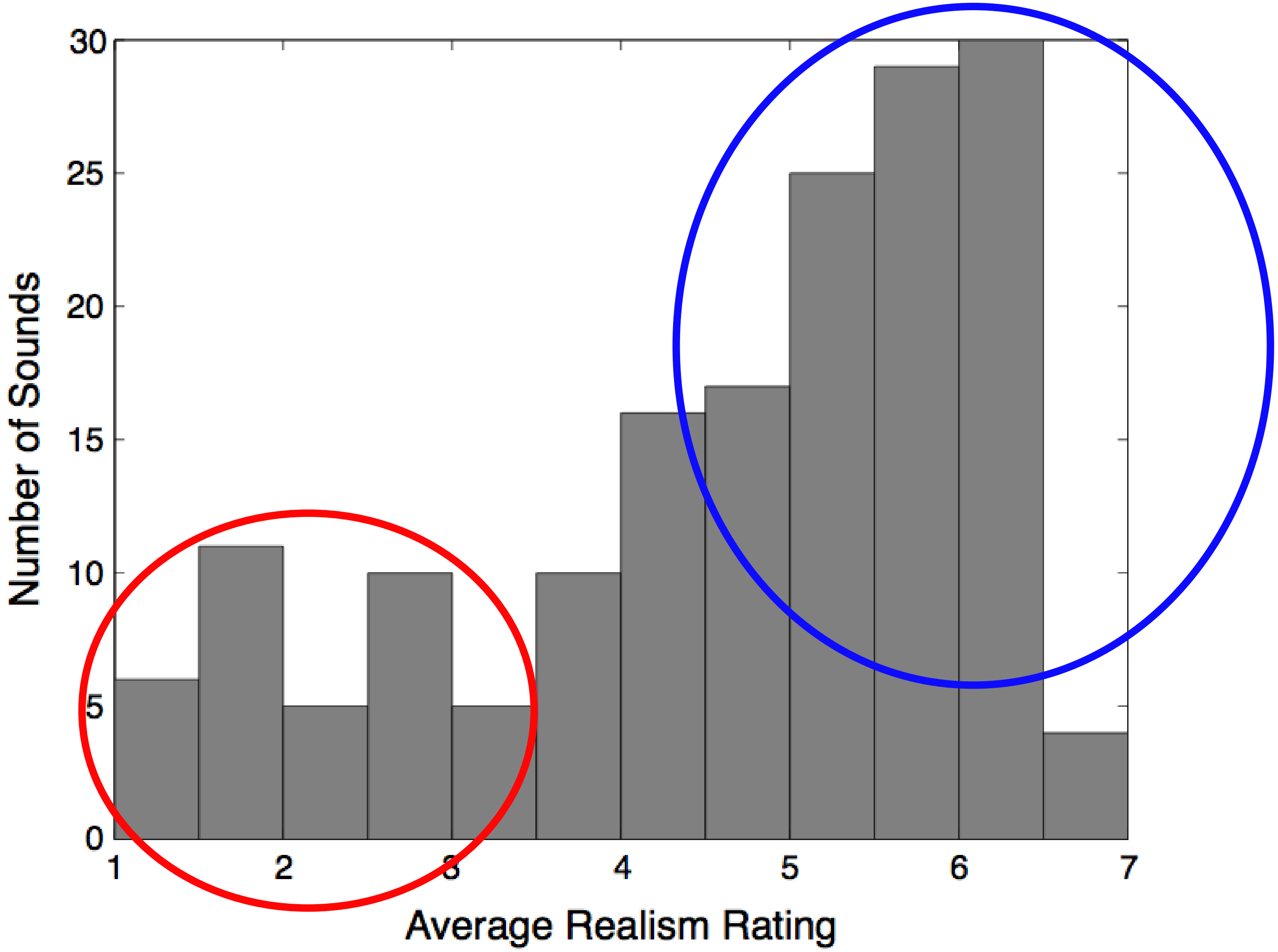
A Speculative Framework:

- Sounds are encoded both as sequences of features, and with statistics that average information over time.
- Feature encodings are sparse for typical natural sound sources, but dense for textures.
- Memory capacity places limits on number of features that can be retained.
- Sound is continuously and obligatorily encoded
  - When memory capacity for feature sequences is reached, memory is overwritten by incoming sound, leaving only statistics



Source: McDermott, Josh H., Michael Schemitsch, and Eero P. Simoncelli. "Summary statistics in auditory perception." Nature neuroscience 16, no. 4 (2013): 493-498.

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



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

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

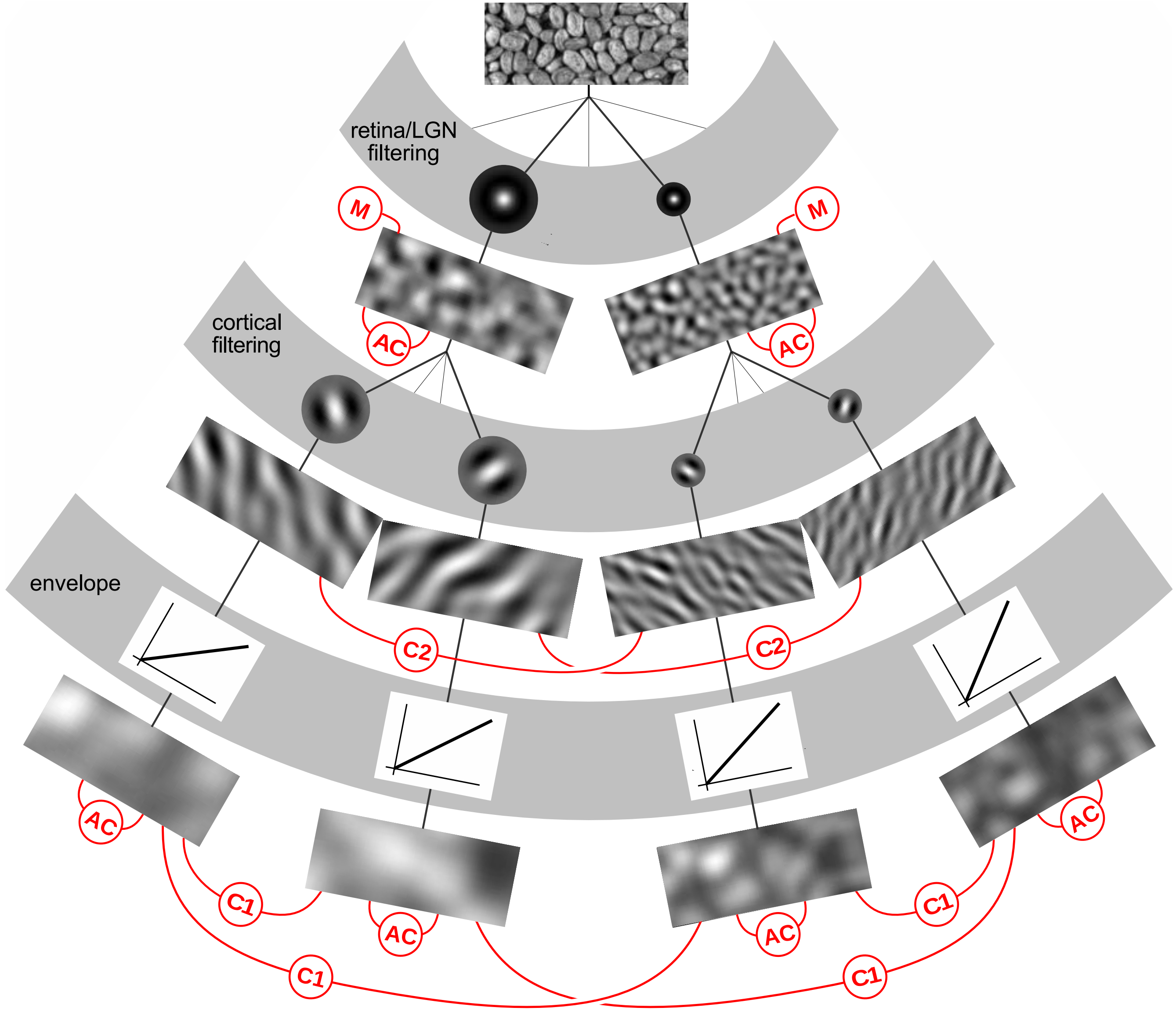
- Many natural sounds may be recognized with relatively simple statistics of early auditory representations
  - -simplest statistics (spectrum) are not that informative
  - -slightly more complex statistics are quite powerful
  - -for textures of moderate length, statistics may be all we retain

OPEN QUESTIONS

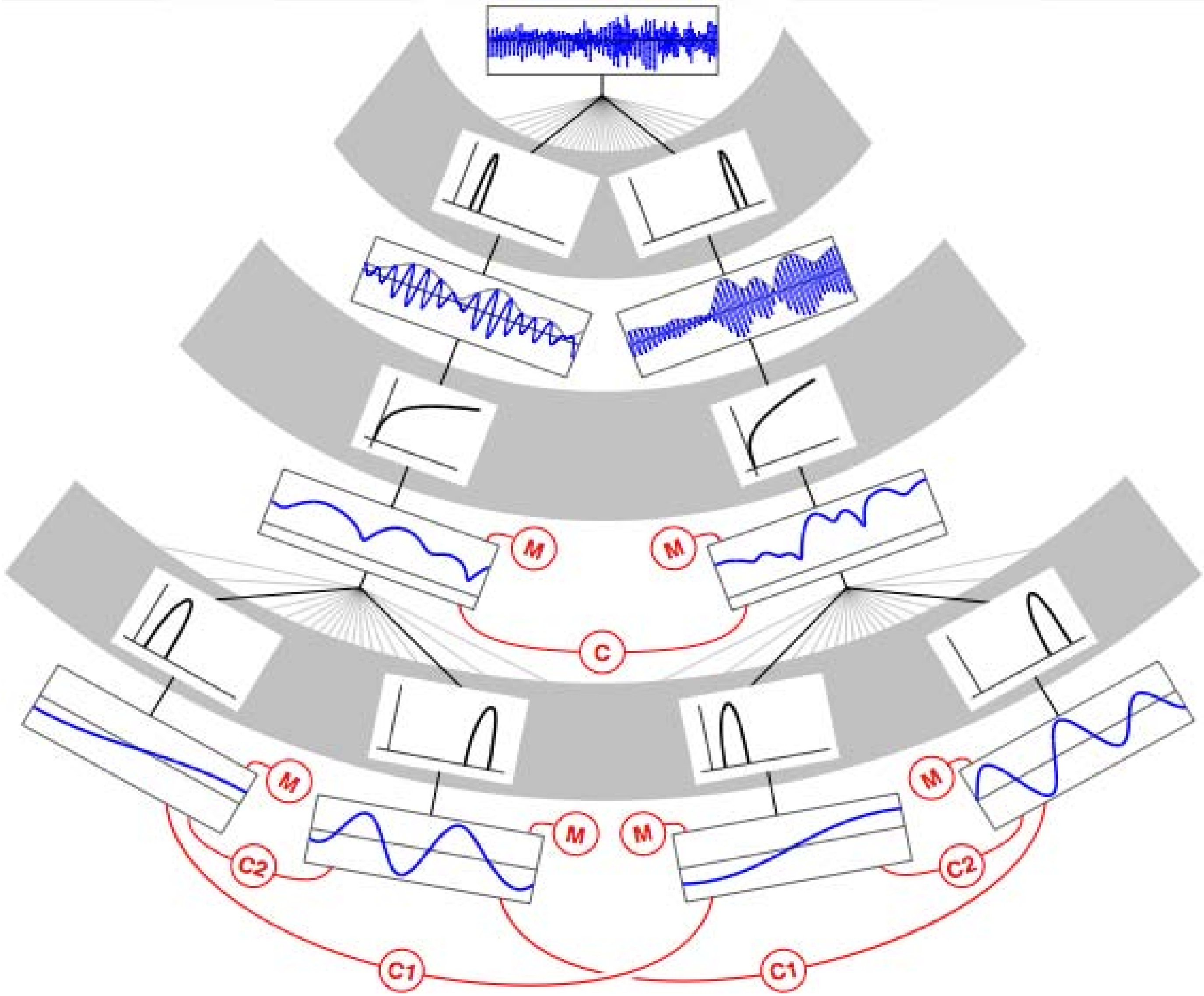
- Locus of time-averaging?
  - Presumptive integration windows of several seconds are long relative to typical timescales in auditory system…

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

- What statistics are needed to account for synthesis failures?



Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.



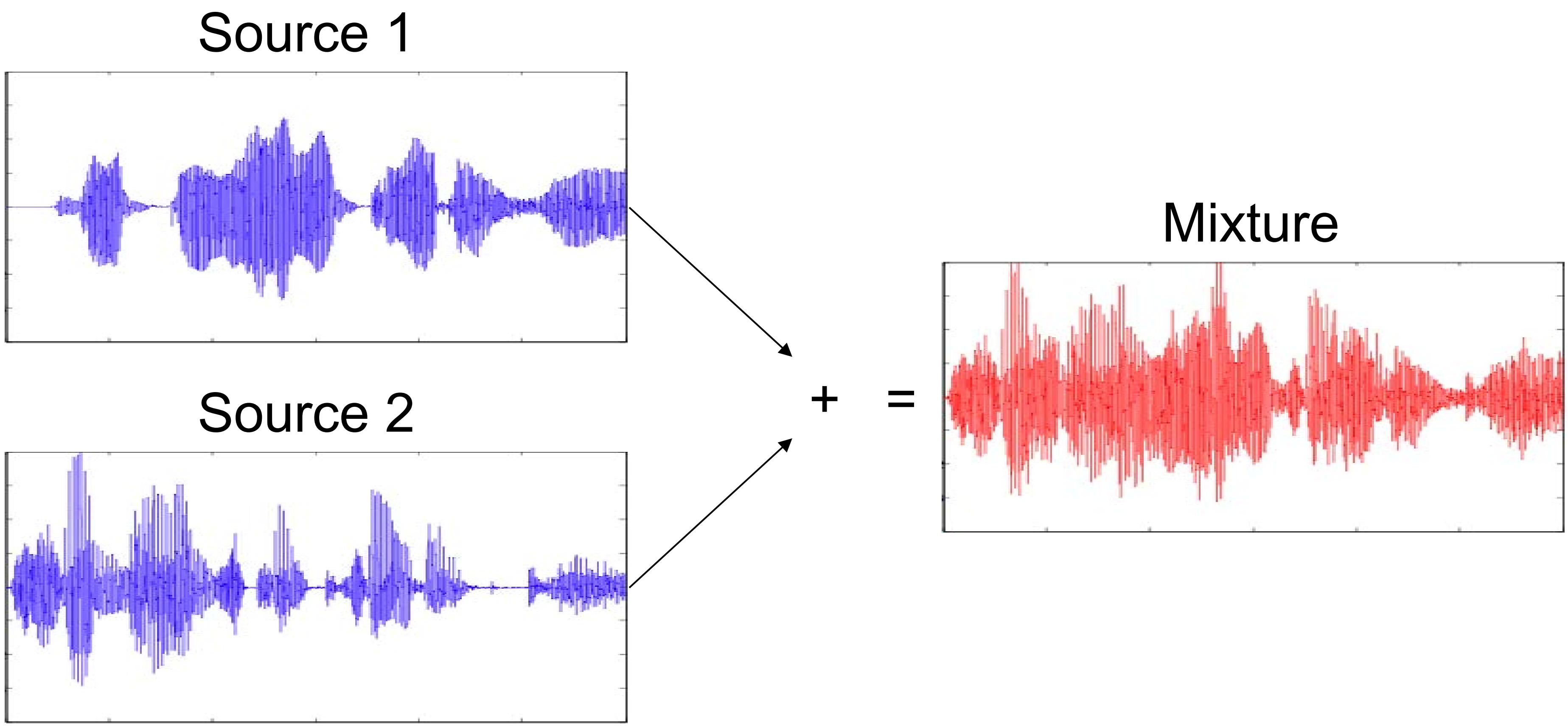
Source: McDermott, Josh H., and Eero P. Simoncelli. "Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis." Neuron 71, no. 5 (2011): 926-940.

Part 4: Auditory Scene Analysis

Auditory Scene Analysis

- Process of inferring events in the world from sound

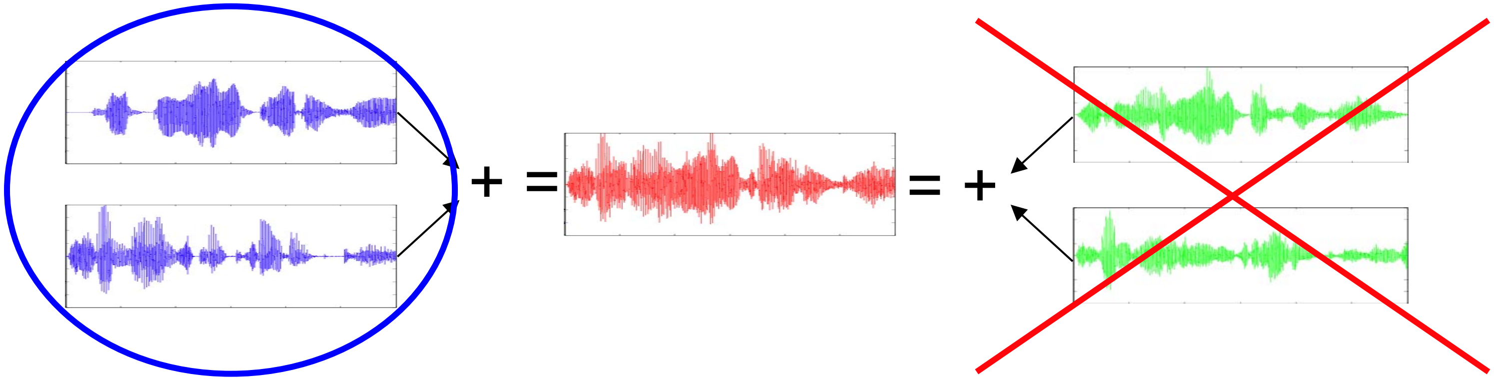
Cocktail party problem: ear receives mixture of sources:



The listener is usually interested in individual sources, which must be inferred from the mixture.

Sound segregation is *ill-posed*:

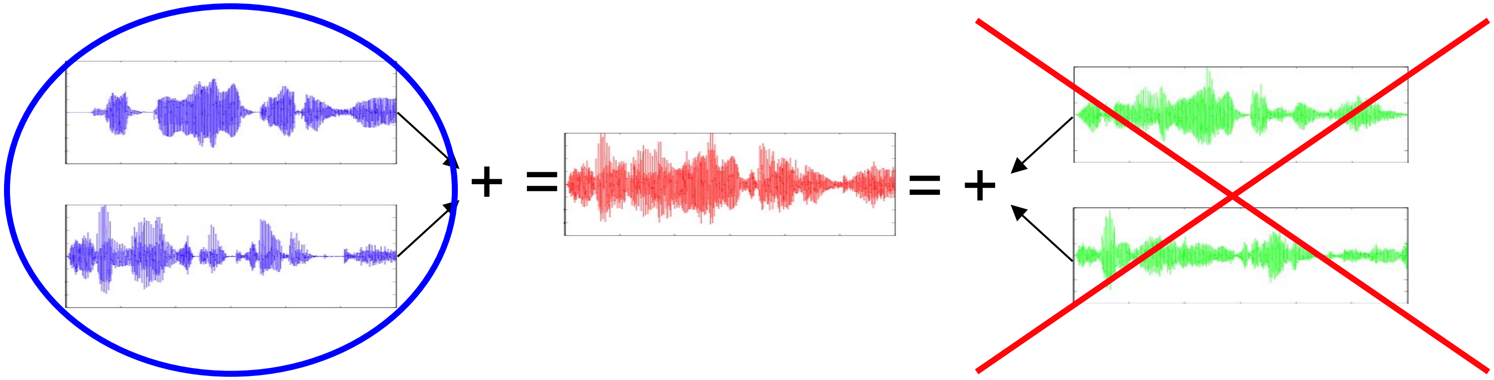
Many sets of possible sounds add up to equal the observed mixture:



The brain must choose the correct set over the other possibilities.

X + Y = 17

How do we manage to hear?



Can only solve with **assumptions** about sources…

Can only make assumptions about sources if real-world sound sources have some degree of regularity.

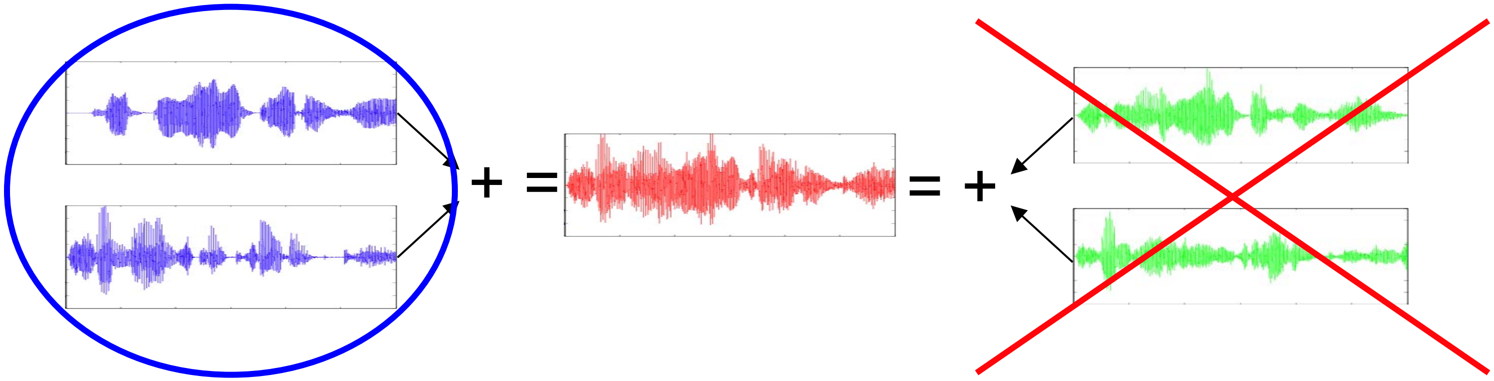
Real-world sounds are not random.

Real-world sounds

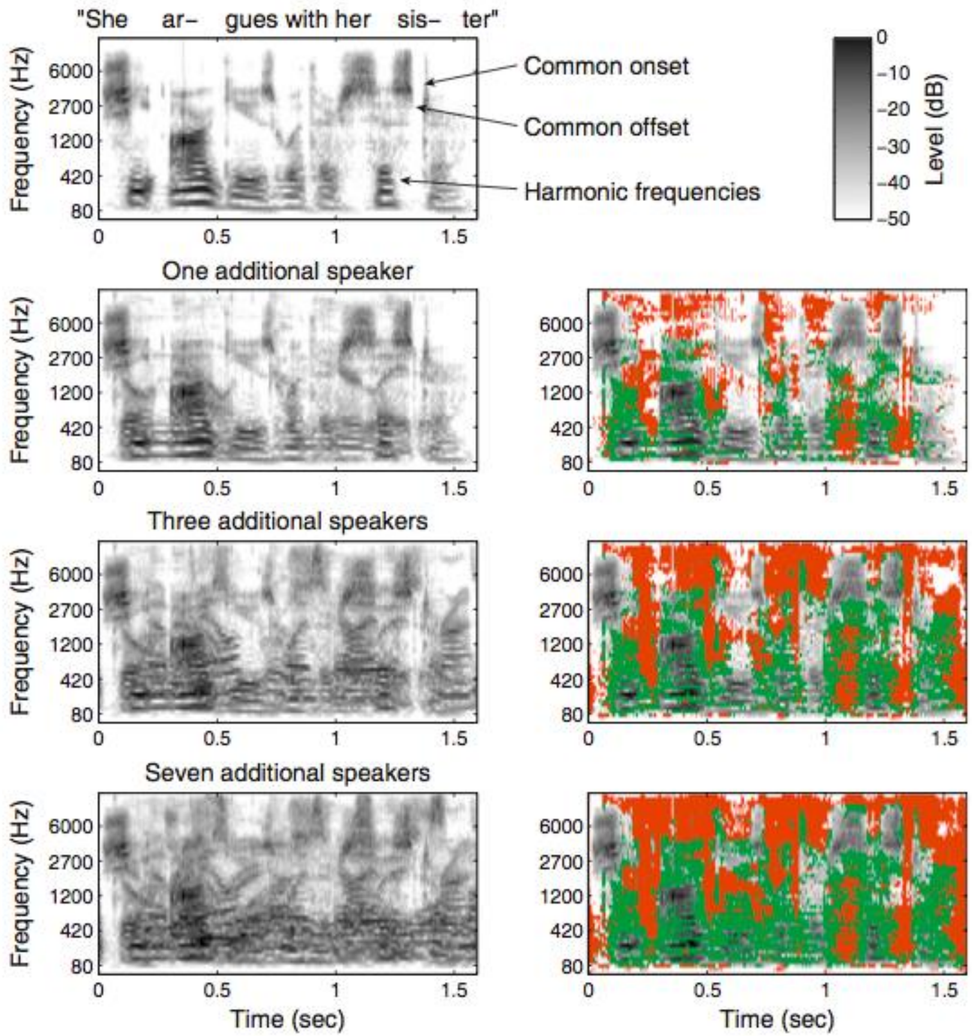
Fully random sounds

→ Real-world sounds are a very small portion of all possible sounds.

We rely on the regularities of natural sounds in order to hear.



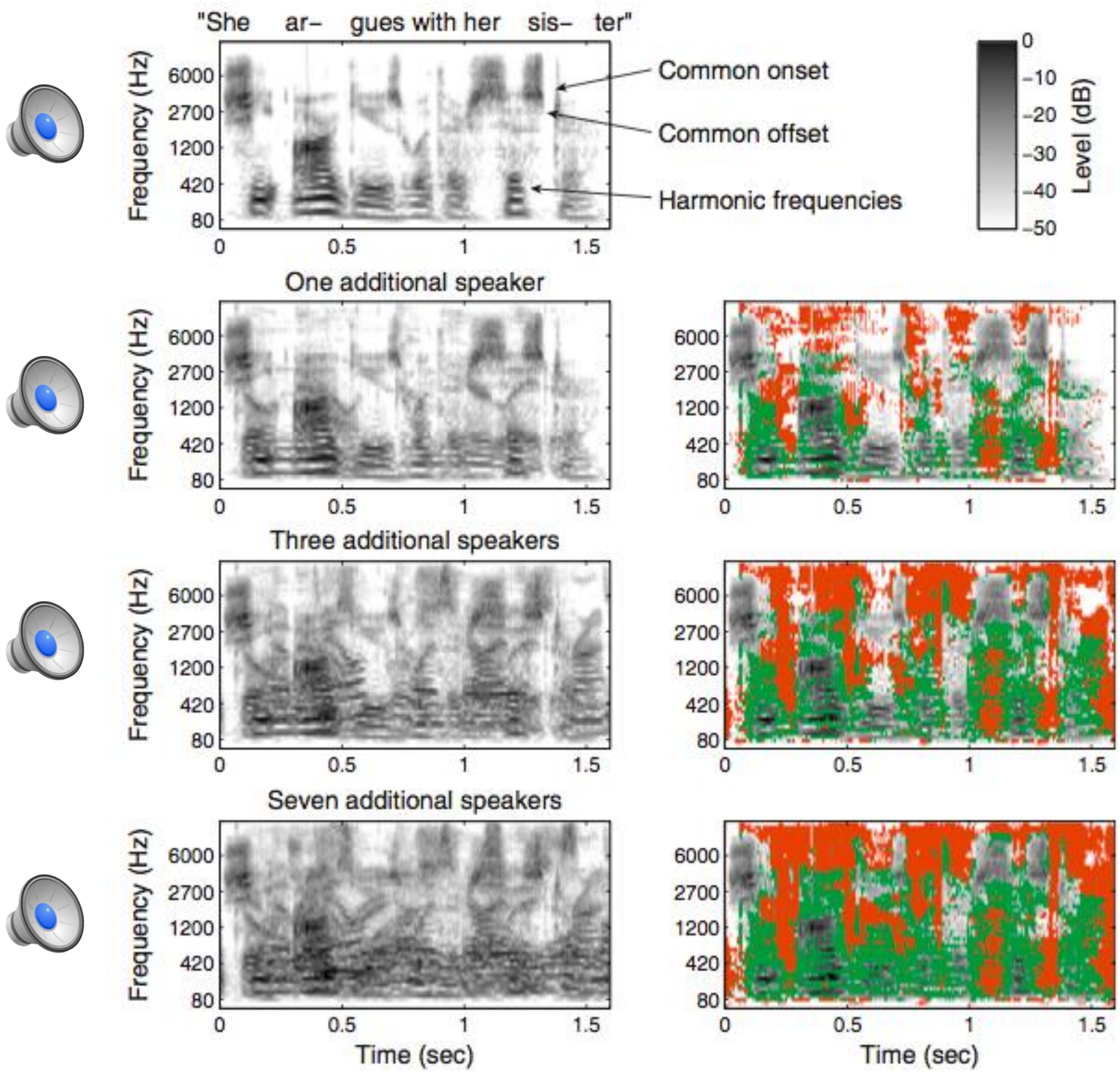
One intuitive view of inferring a target source from a mixture:
1) Determine grouping of observed sound elements.



Red:
Other sources
Green:
Masked

Courtesy of Elsevier, Inc., http://www.sciencedirect.com. Used with permission.
Source: McDermott, Josh H. "The cocktail party problem." Current Biology 19, no. 22 (2009): R1024-R1027.

One intuitive view of inferring a target source from a mixture:
2) Estimate parts of source that are masked.

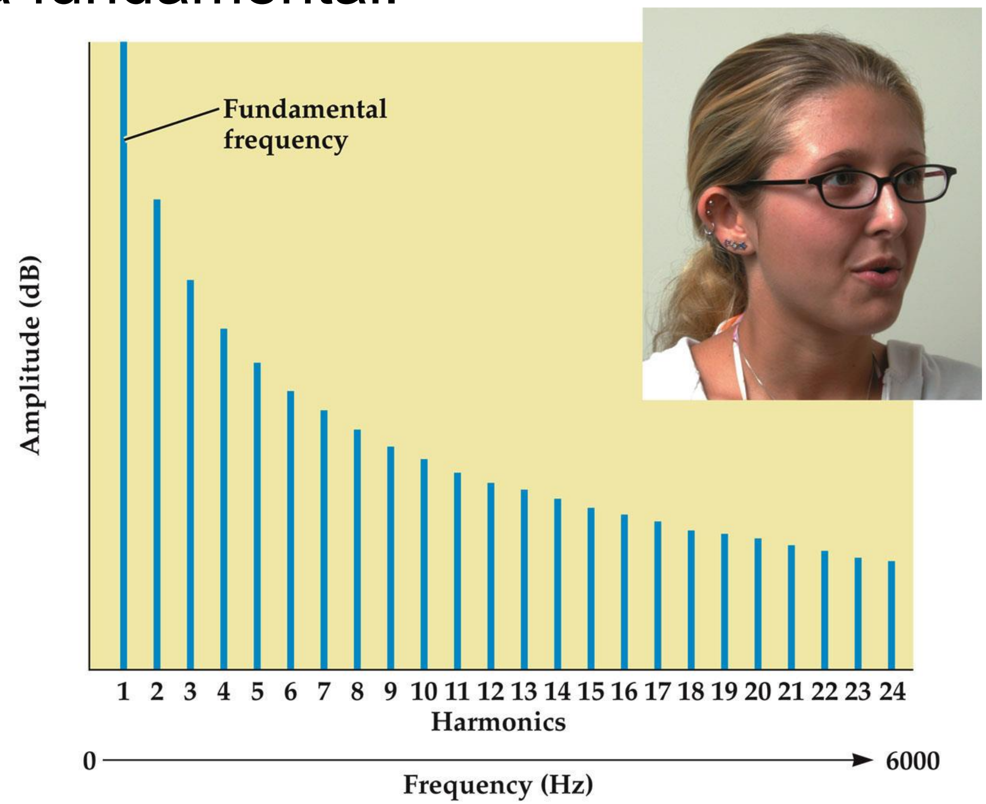


Red:
Other sources
Green:
Masked

Source: McDermott, Josh H. "The cocktail party problem." Current Biology 19, no. 22 (2009): R1024-R1027.

One example of a regularity that could be used to group sound: harmonic frequencies

Voices and instruments produce frequencies that are harmonics (multiples) of a fundamental:

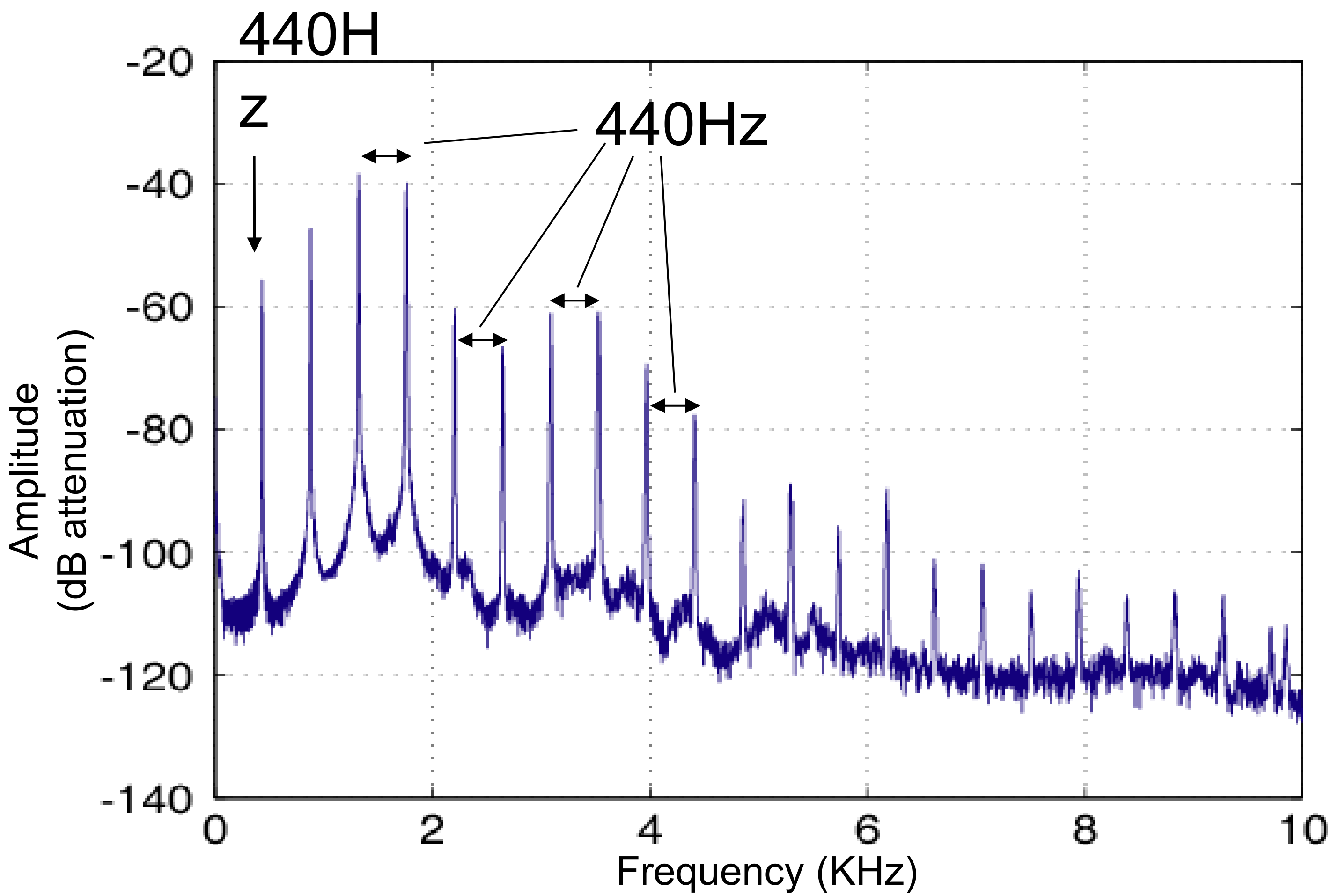


SENSATION & PERCEPTION 3e, Figure 10.13
© 2012 Sinauer Associates, Inc.

Source: Wolfe, Jeremy M., Keith R. Kluender, Dennis M. Levi, Linda M. Bartoshuk, Rachel S. Herz, Roberta L. Klatzky, Susan J. Lederman, and Daniel M. Merfeld. Sensation & perception. Sunderland, MA: Sinauer, 2006.

A440 on an oboe:

F0, 2F0, 3F0, 4F0, 5F0, 6F0, 7F0, …



Source: McDermott, Josh H., and Andrew J. Oxenham. "Music perception, pitch, and the auditory system." Current opinion in neurobiology 18, no. 4 (2008): 452-463.

Harmonicity is a key property of vocal, instrument sounds.

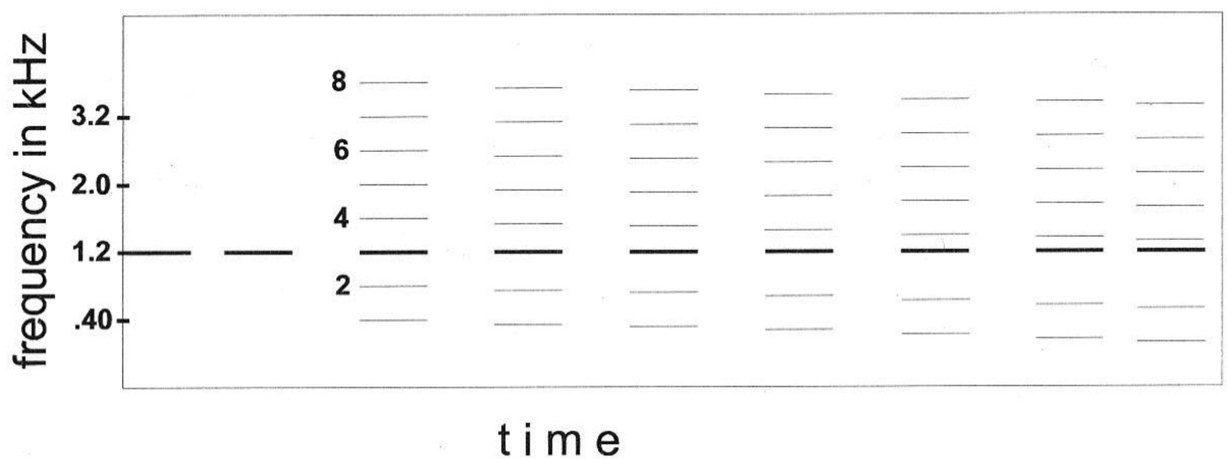
When air is blown through the vocal cords, they open and close at regular time intervals, generating a periodic series of sound pulses:

Image of moving vocal chords removed due to copyright restrictions.
Please see the video.

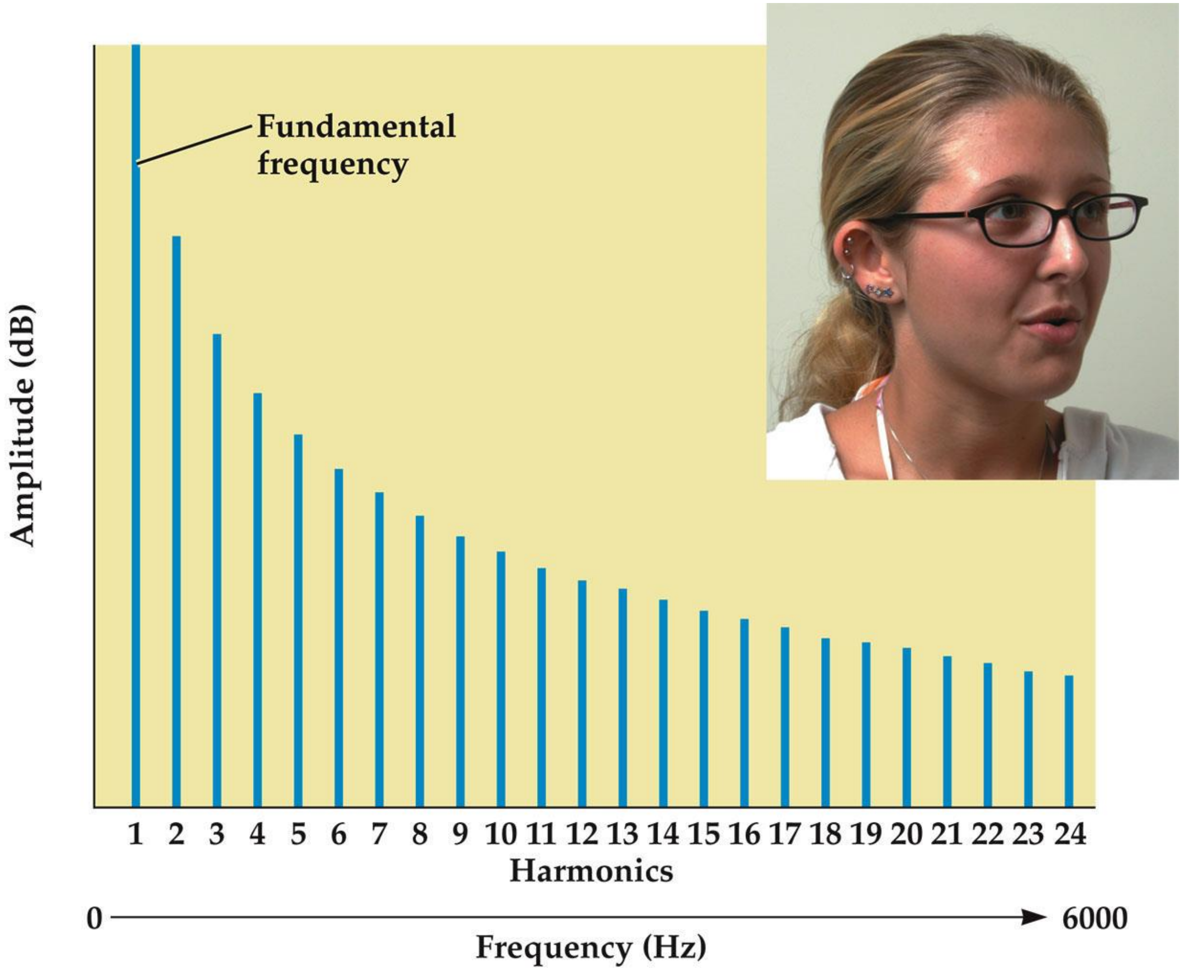
Classic evidence for harmonicity as a grouping cue:

When one frequency is not harmonically related to a bunch of others, it segregates perceptually.

18 Isolation of a frequency component based on mistuning.



The pitch of a sound is inferred collectively from its harmonics.

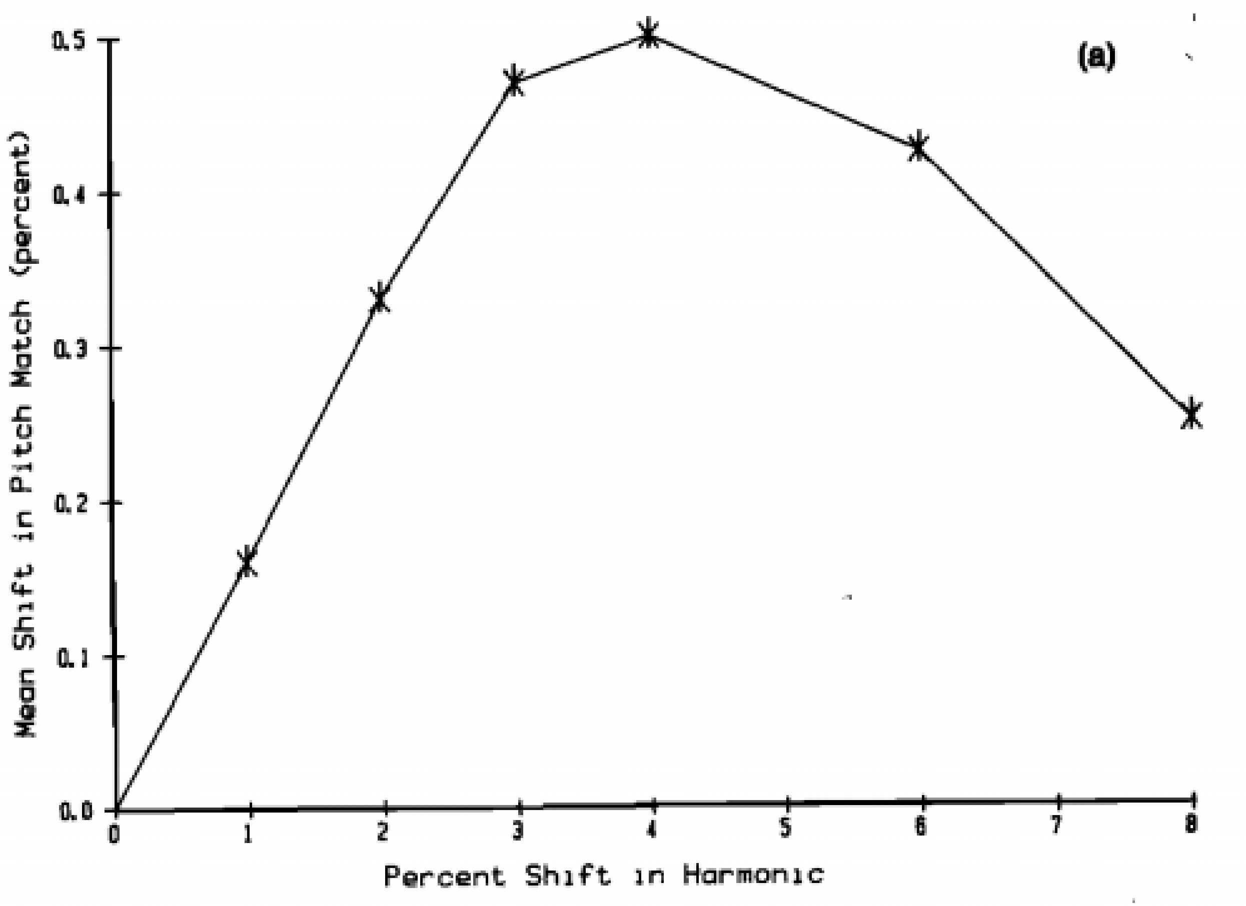


SENSATION & PERCEPTION 3e, Figure 10.13
© 2012 Sinauer Associates, Inc.

Source: Wolfe, Jeremy M., Keith R. Kluender, Dennis M. Levi, Linda M. Bartoshuk, Rachel S. Herz, Roberta L. Klatzky, Susan J. Lederman, and Daniel M. Merfeld. Sensation & perception. Sunderland, MA: Sinauer, 2006.

For small mistunings, pitch of complex is shifted, but effect is reduced for larger mistunings:

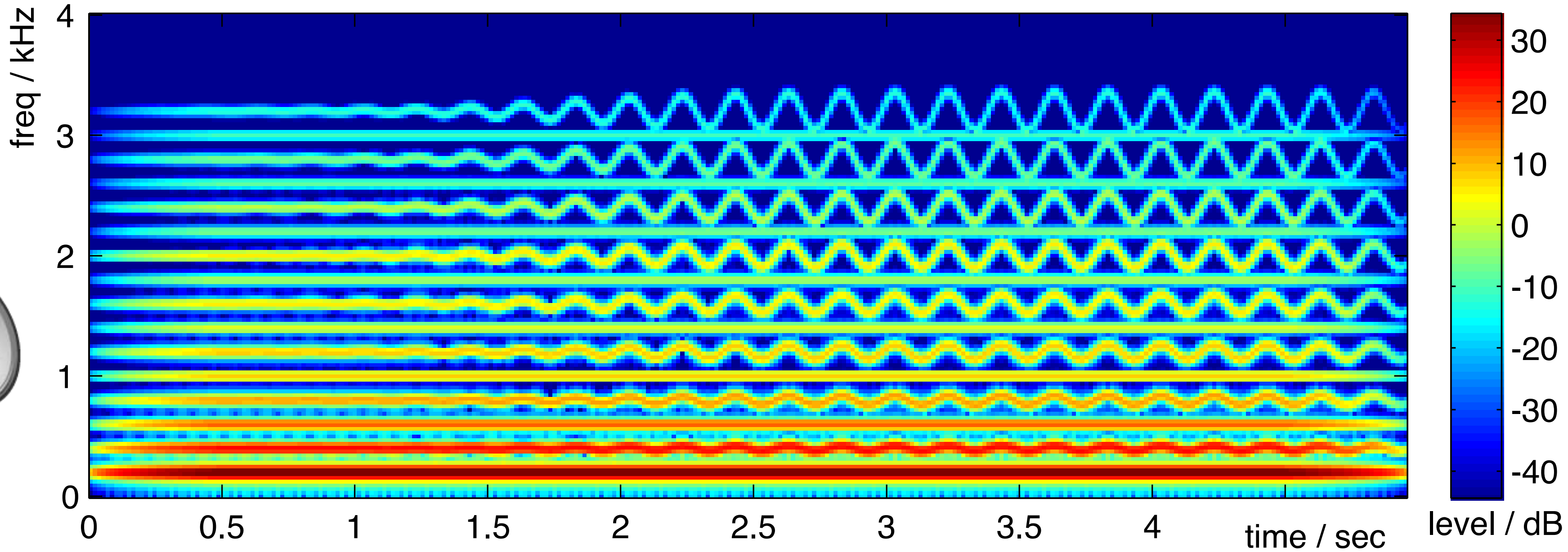
Task: match the pitch of the mistuned complex with a normal complex



Source: Moore, Brian CJ, Brian R. Glasberg, and Robert W. Peters. "Thresholds for hearing mistuned partials as separate tones in harmonic complexes." The Journal of the Acoustical Society of America 80, no. 2 (1986): 479-483.

Moore et al., 1986

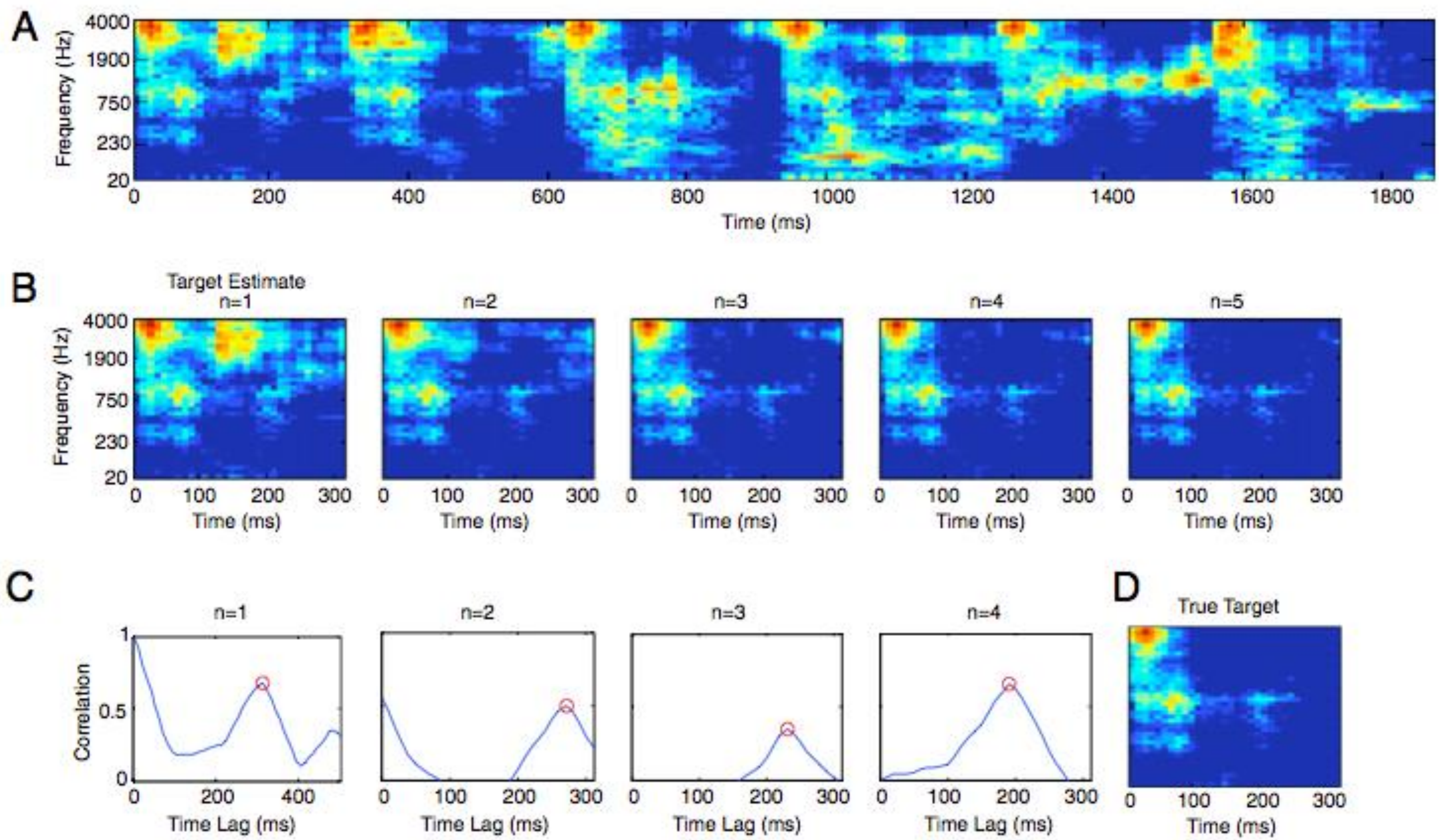
The Reynolds-McAdams Oboe



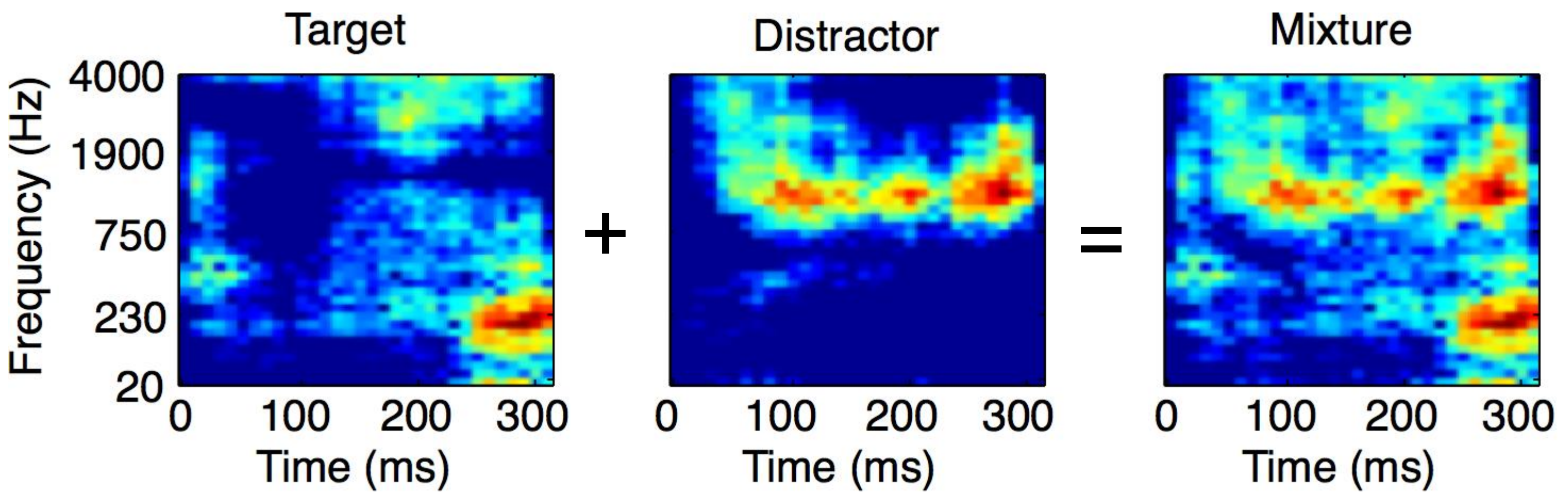
Evidence for grouping/segregation via common frequency modulation?

Could be mediated via harmonicity alone…

Another potential grouping cue: repetition



Can repetition be used to segregate sounds?



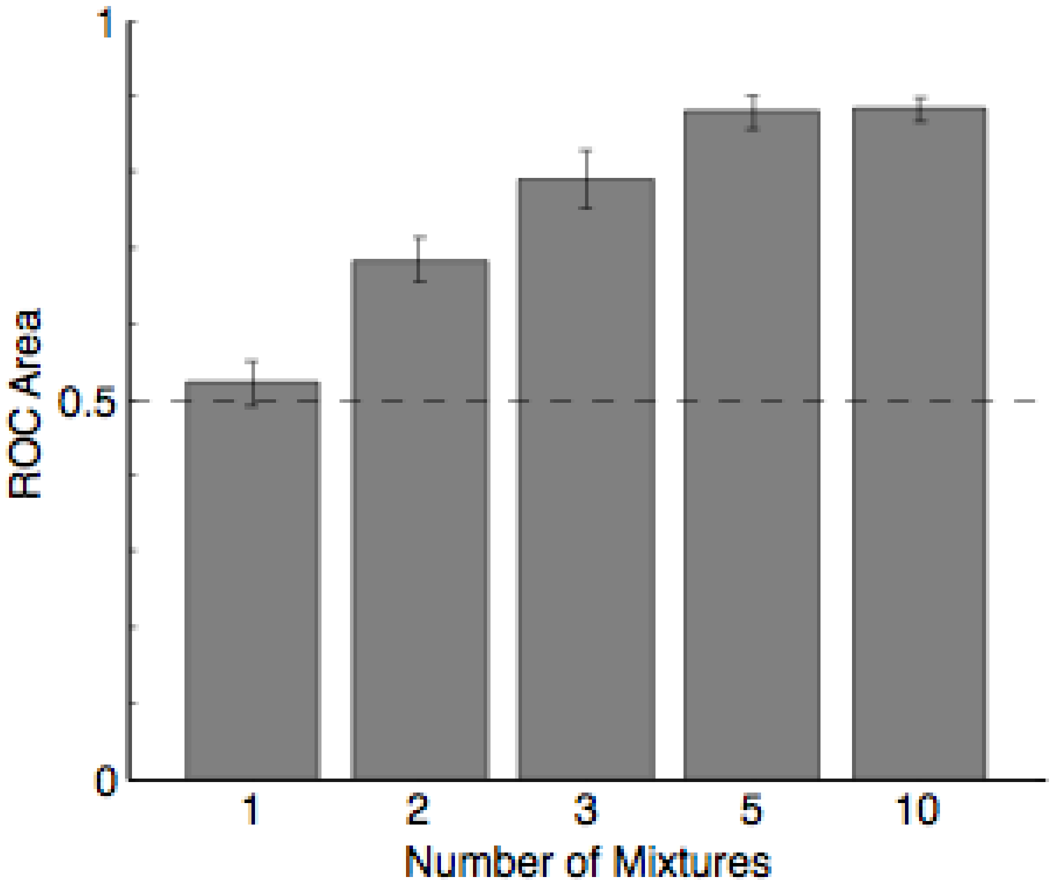
Present mixture, then probe sound:

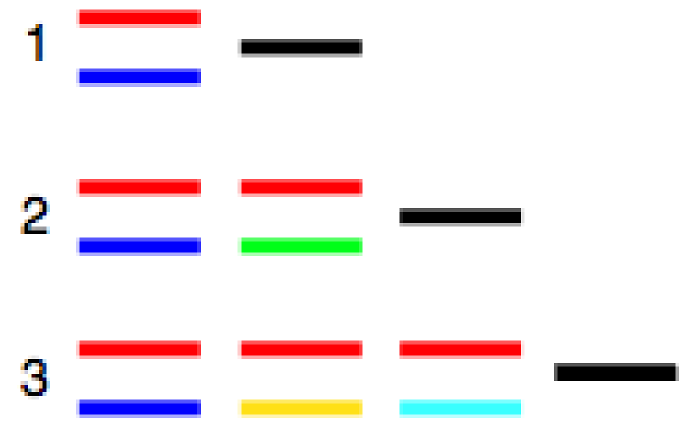
Was the probe one of the sounds in the mixture?

Sounds have some structure, but not enough to produce segregation of a single mixture.

McDermott, Wrobleski & Oxenham, PNAS 2011

Listeners can identify sound sources from multiple mixtures:





McDermott, Wrobleski & Oxenham, PNAS 2011

Summary of Generic Grouping Cues

- Common onset
- Co-modulation
- Harmonicity
- No clear role for frequency modulation (common fate)
- Weak role for spatial cues in concurrent segregation
- Repetition

Open Questions

- How do grouping cues relate to natural sound statistics?
- Are we optimal given the nature of real-world sounds?
- Are grouping cues learned or hard-wired?
- How important are generic cues relative to knowledge of particular sounds (e.g. words)?

TAKE-HOME MESSAGES

- The brain estimates the causes of the sound signal that enters the ears (usually a mixture of sources).
- "Grouping cues" are presumed to be related to statistical regularities of natural sounds.
    - Harmonicity
    - Common onset
    - Repetition
- The brain infers parts of source signals that are masked by other sources, again using prior assumptions.
- We need a proper theory in this domain to be able to predict and explain real-world performance.