Sound Texture Perception via Statistics of the Auditory Periphery

Josh McDermott

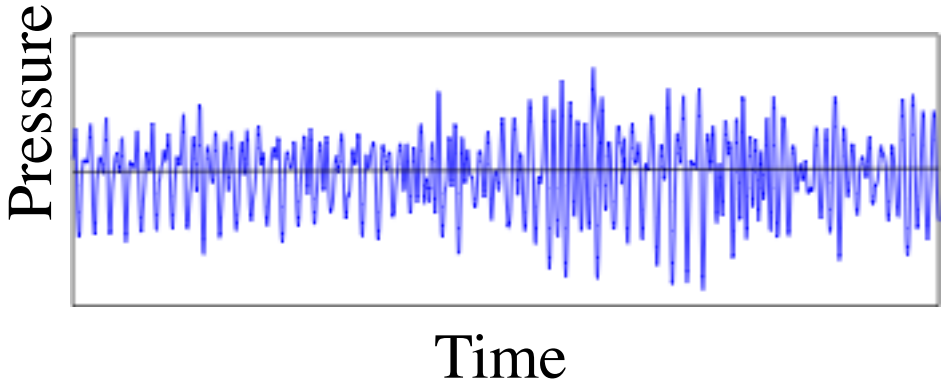
New York University

This talk represents joint work with

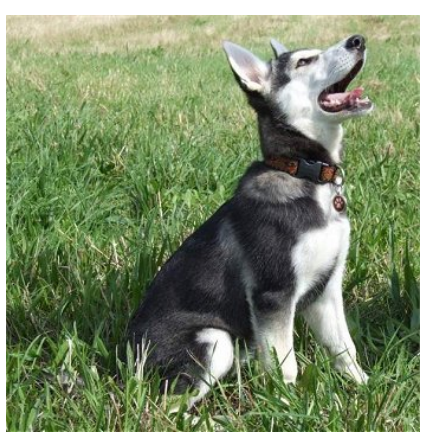
Eero Simoncelli

How do we recognize sounds?

The ear receives a pressure waveform.



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




Perceptual processes must transform sensory input into representations that are useful for behavior, by making things in the world explicit.

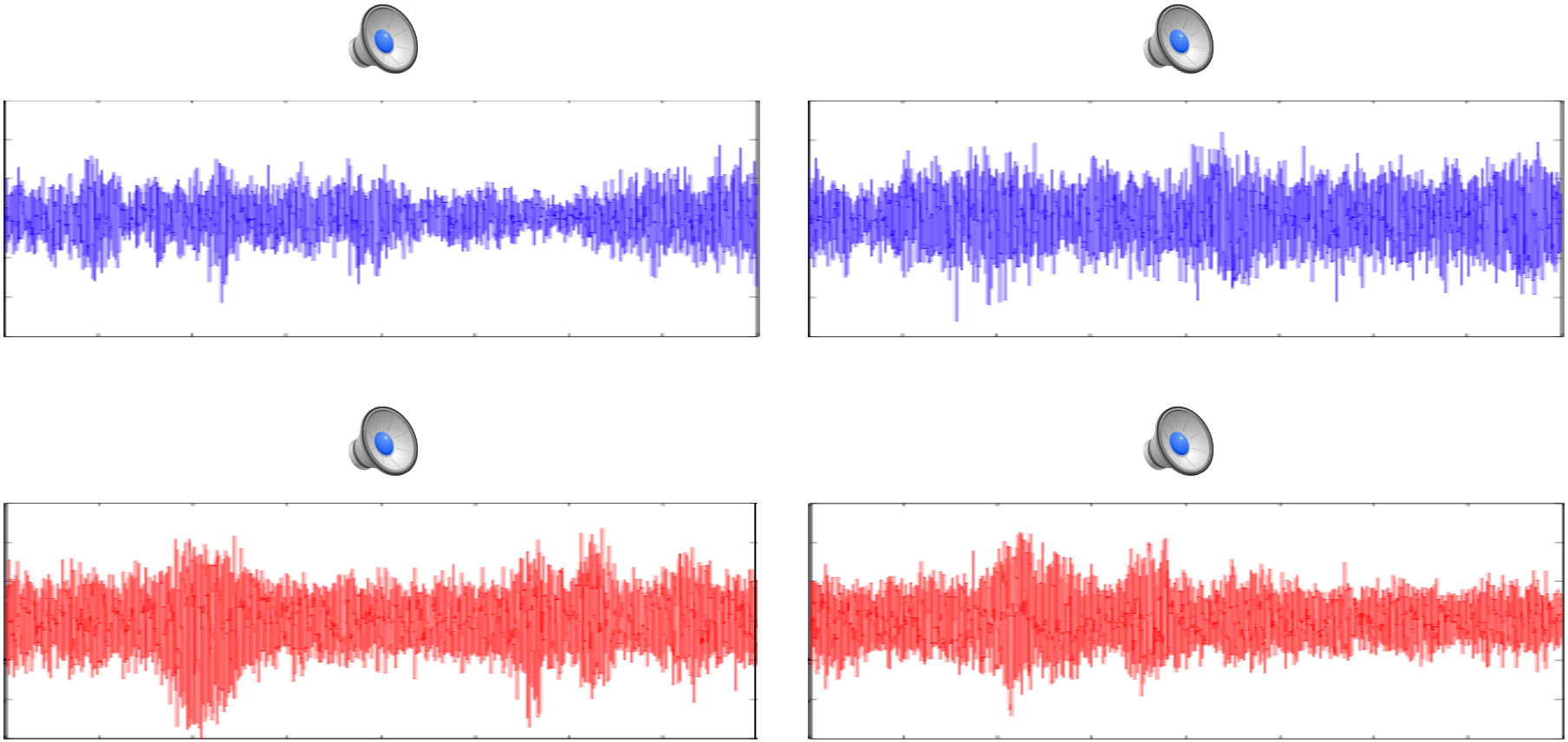
SOUND TEXTURE

Textures result from large numbers of acoustic events.

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

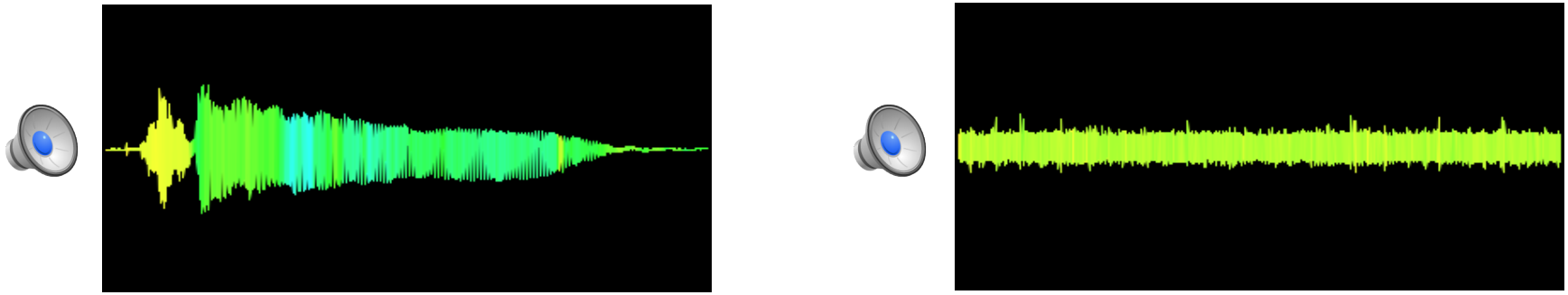
Sound textures are common in the world, but are largely unstudied.

Key problem of recognition: invariance



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

Why texture?



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

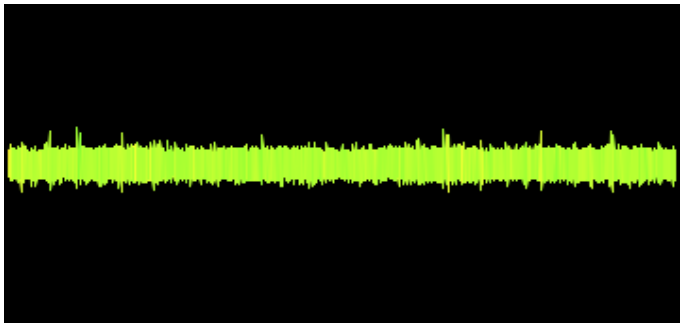
•Stationarity makes textures a good starting point for understanding auditory representation.

Some previous work on modeling sound texture:

Arnaud and Popat 1995
Dubnov et. al 2002
Athineos and Ellis 2003
Lu et. al 2004
Parker and Behm 2004
Zhu and Wyse 2004

Work on environmental sounds often inclusive of texture:

Li et al. 2001
Nordqvist and Leijon 2004
Chu et al. 2009
Verron et al. 2009
Lee et al. 2010



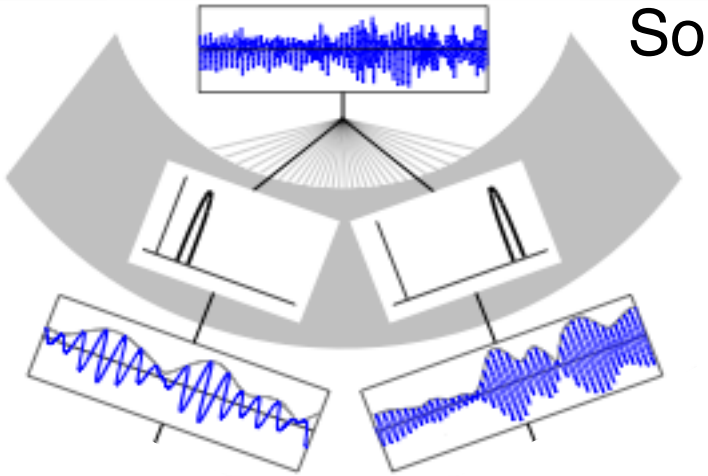
Key Proposal:

•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 summary statistics.

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

1. Cochlear filters



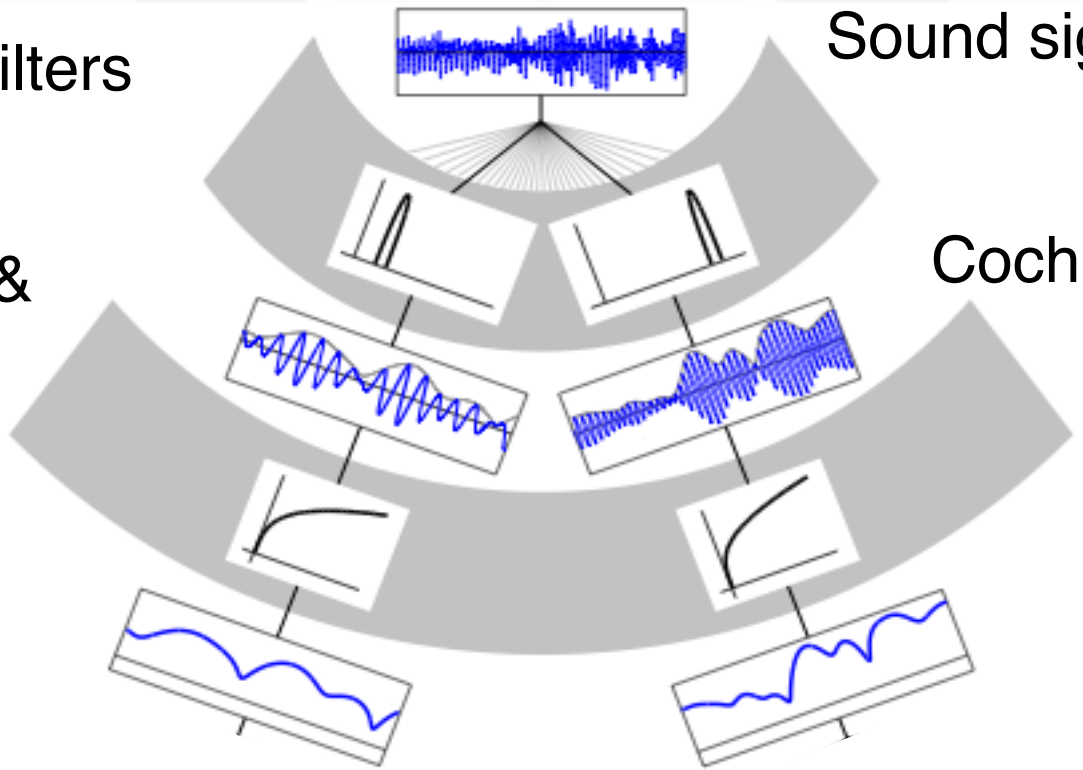
Sound signal

Cochlear subbands

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

1. Cochlear filters

2. Envelope & compression

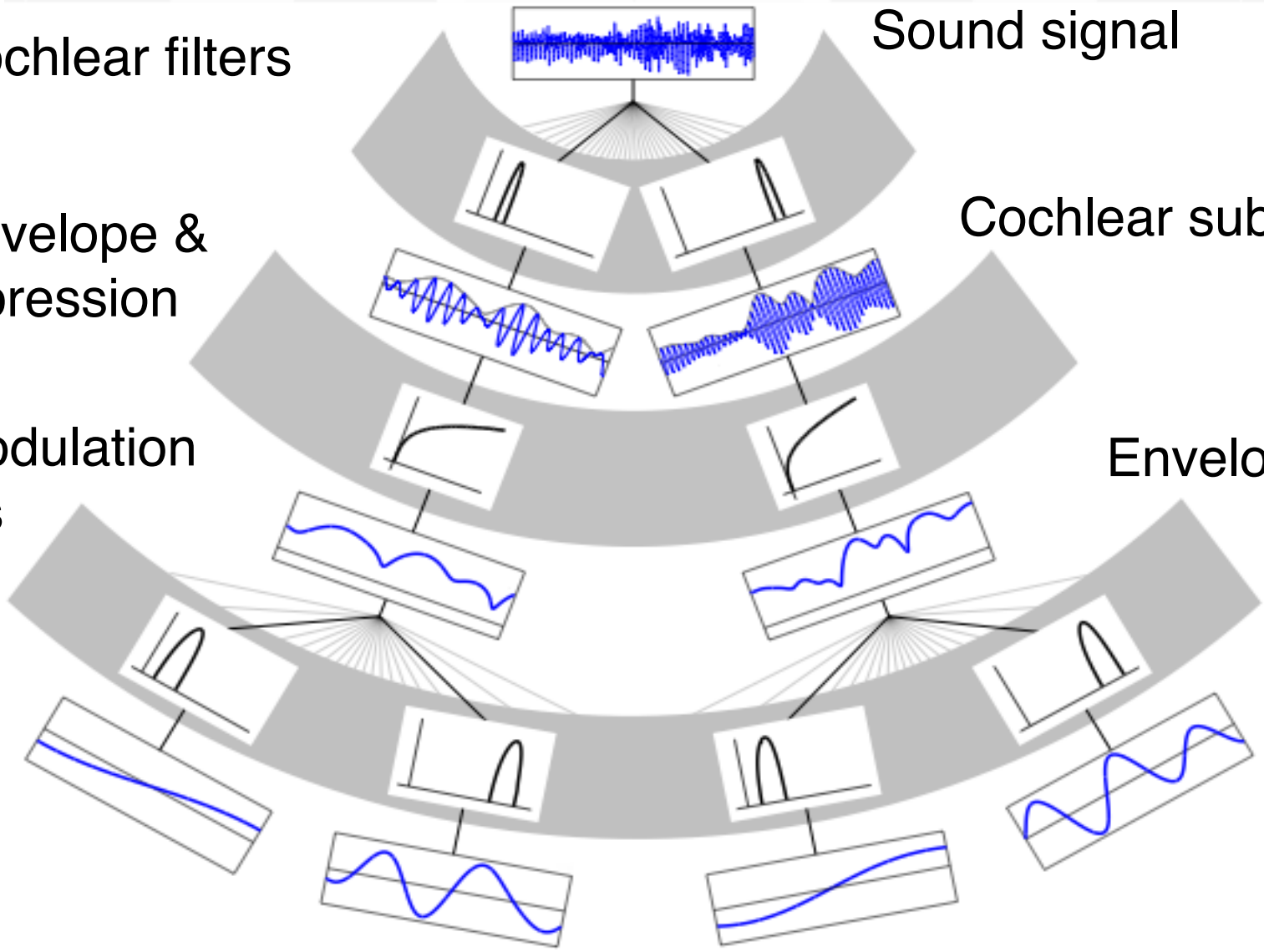


Whatever statistics the auditory system measures are presumably derived from peripheral auditory representations:

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



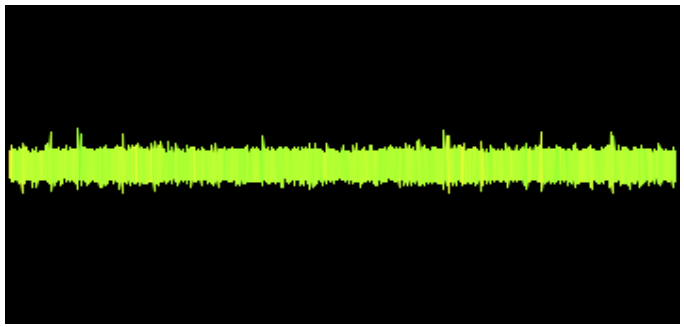
How much of texture perception can be captured with simple summary statistics of these representations?

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.



Basic idea: take an example signal, measure statistics,
synthesize new signals constrained to have same statistics.

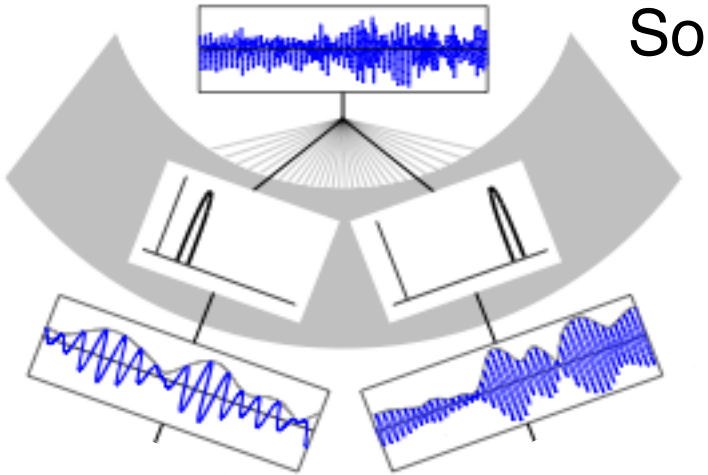
cf. visual texture synthesis:
Heeger and Bergen; Zhu, Wu, & Mumford; Portilla & Simoncelli

cf. Wessel, Risset etc.

How can we synthesize sounds in this way?

Auditory model starts with subband transform:

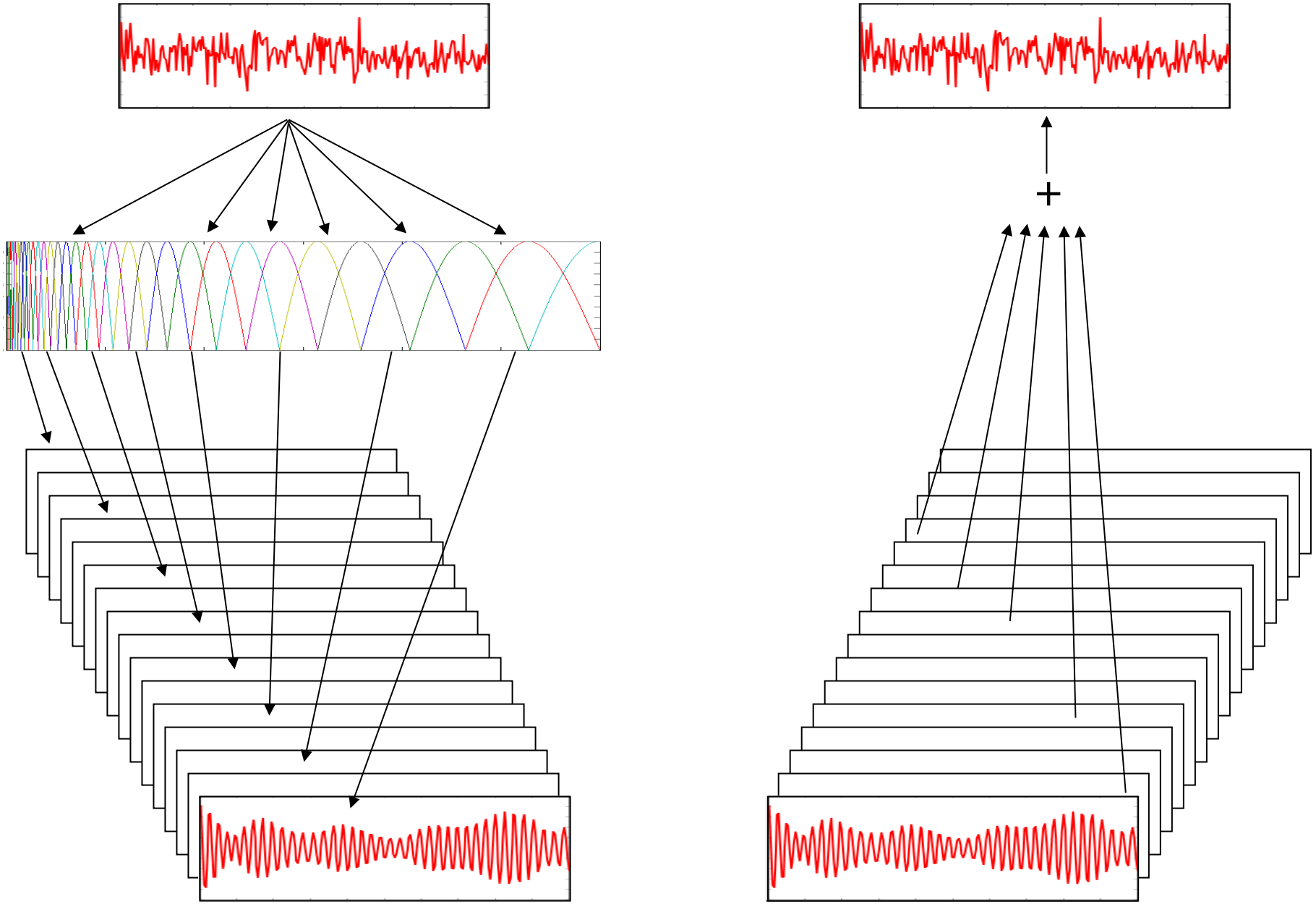
1. Cochlear filters



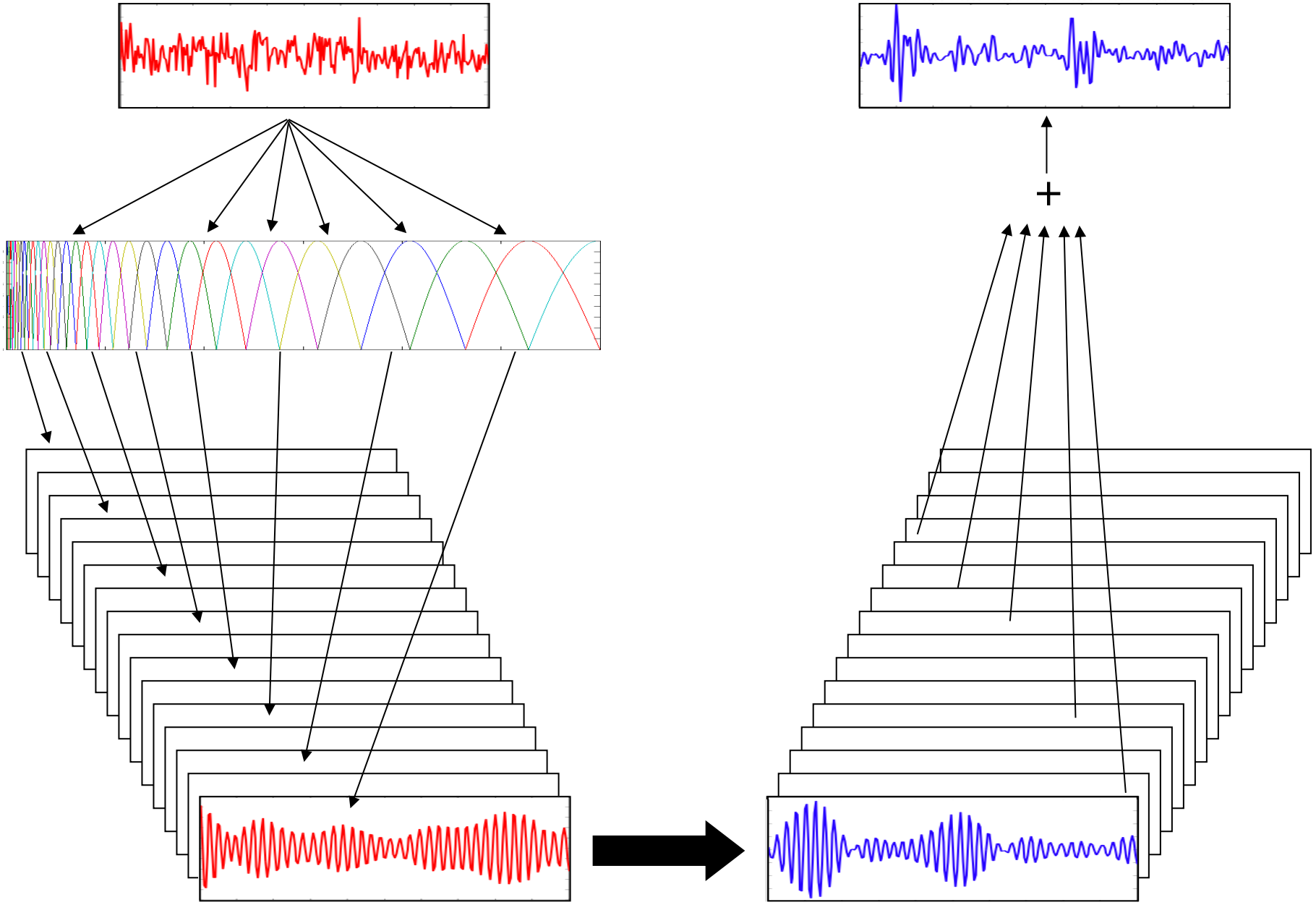
Sound signal

Cochlear subbands

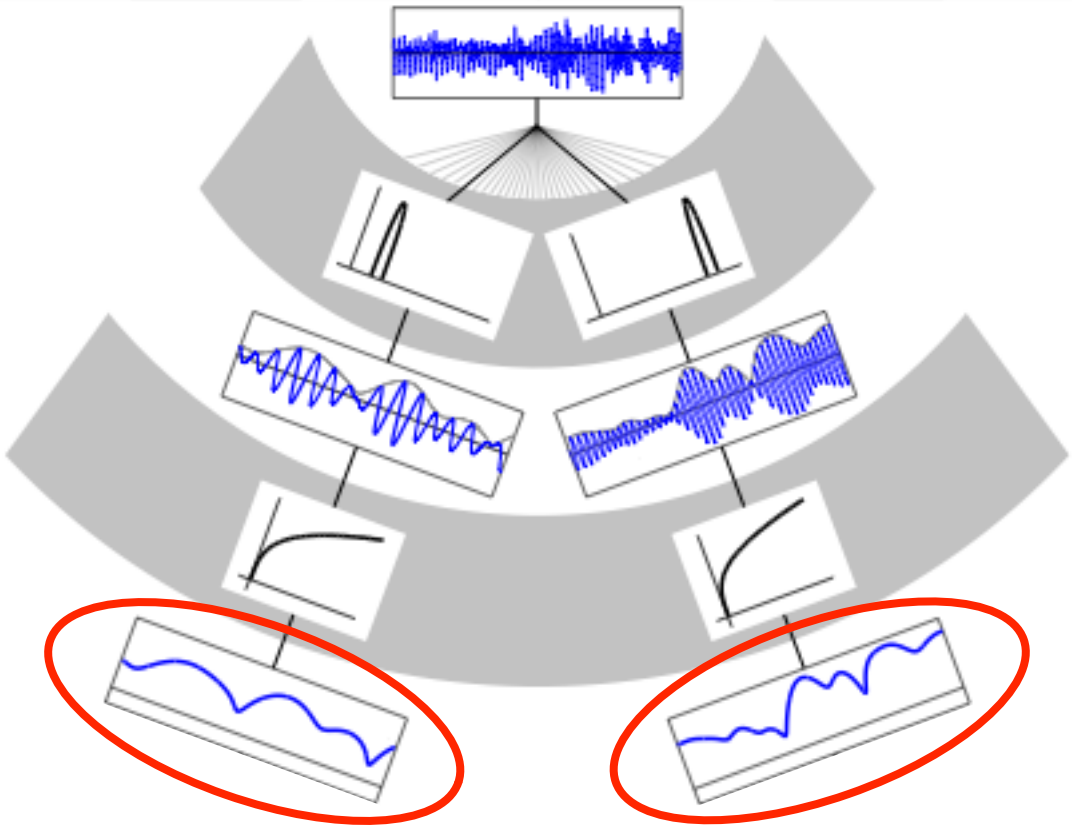
Subband transform can be inverted to regenerate sound signal:



Start with noise, alter noise subbands to have desired stats, resynthesize:



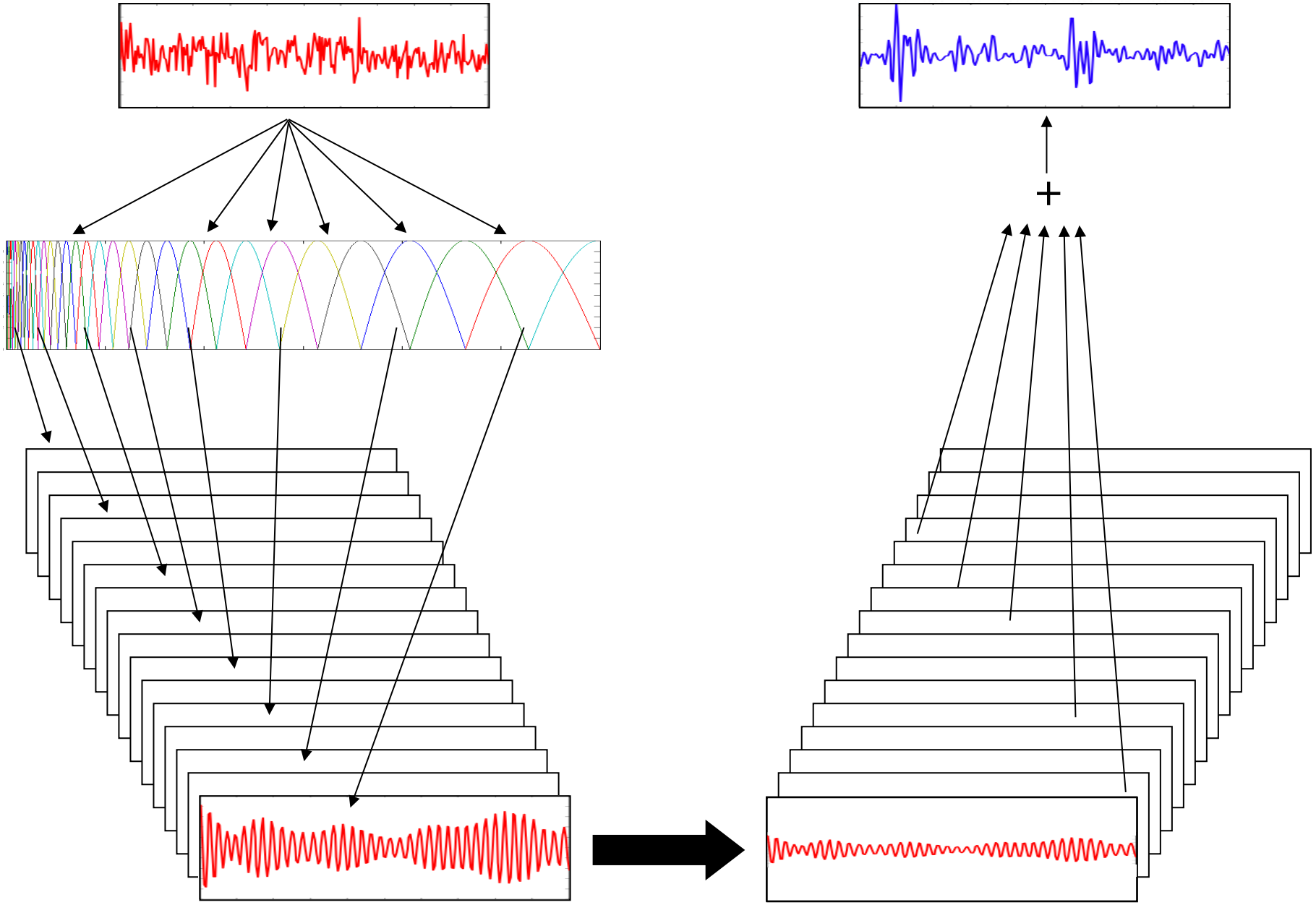
Simple example: test the role of the mean of each cochlear envelope (power spectrum)



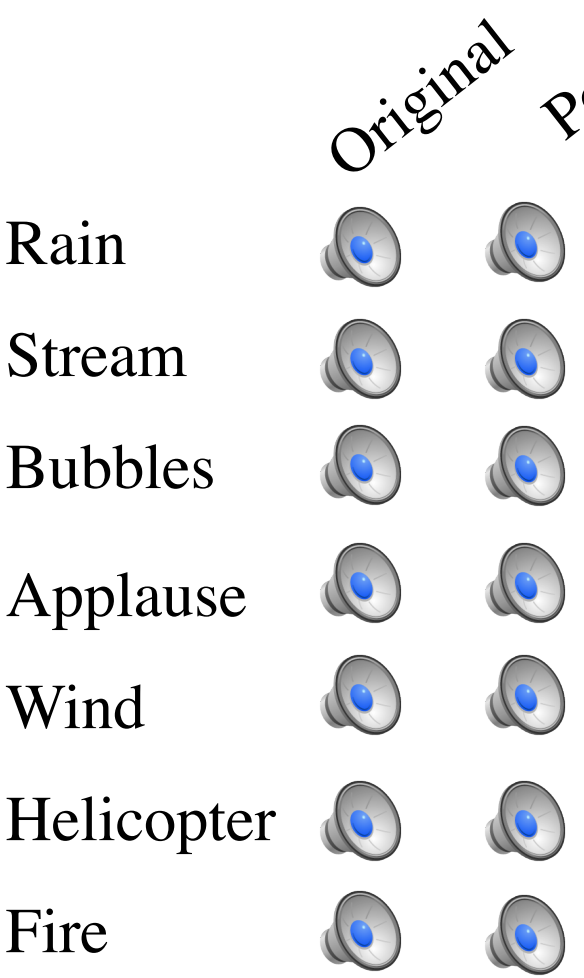
Envelopes

- 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?



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

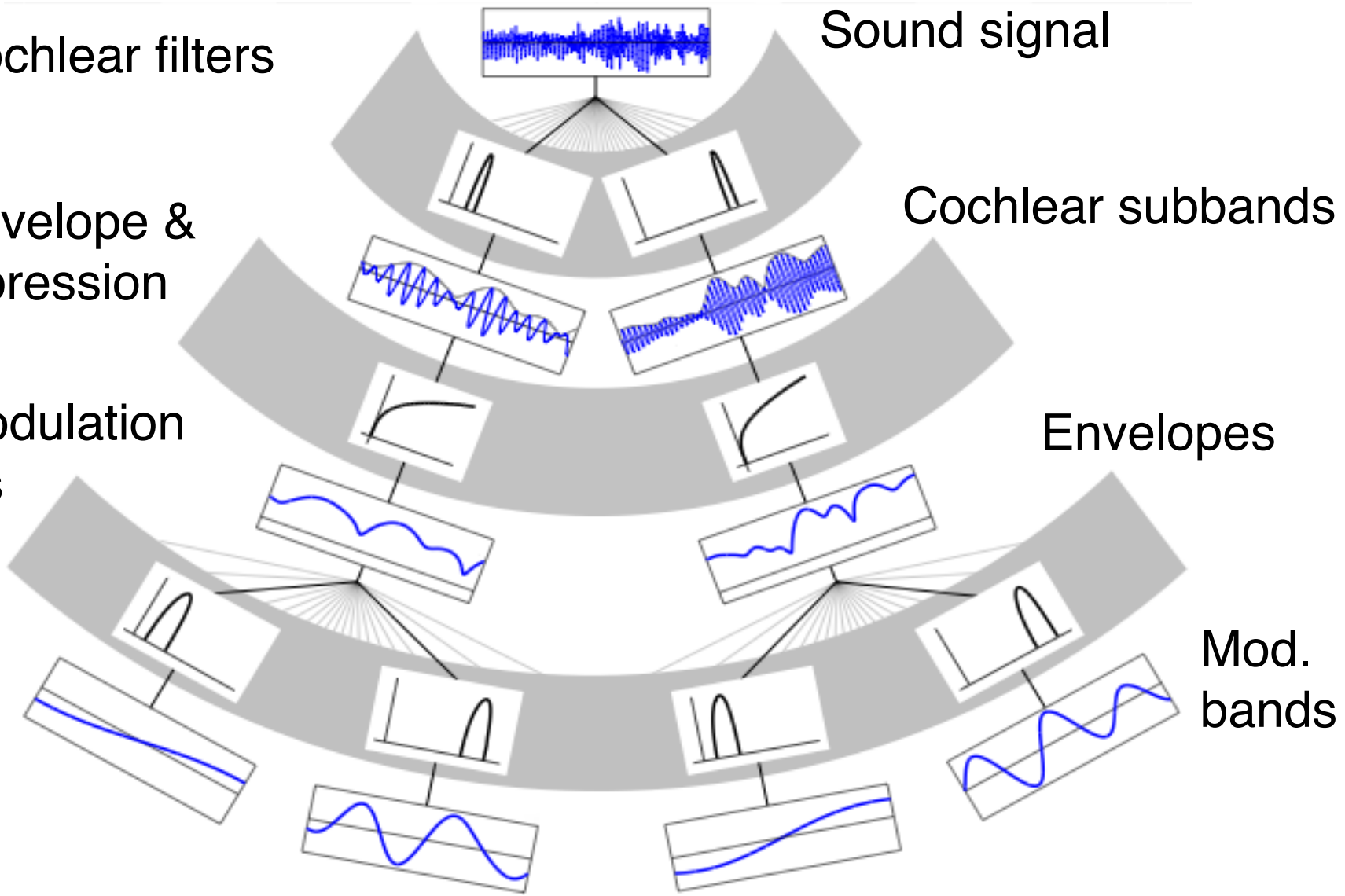
- We aren't simply registering the spectrum when we recognize textures.

Will additional simple statistics do any better?

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



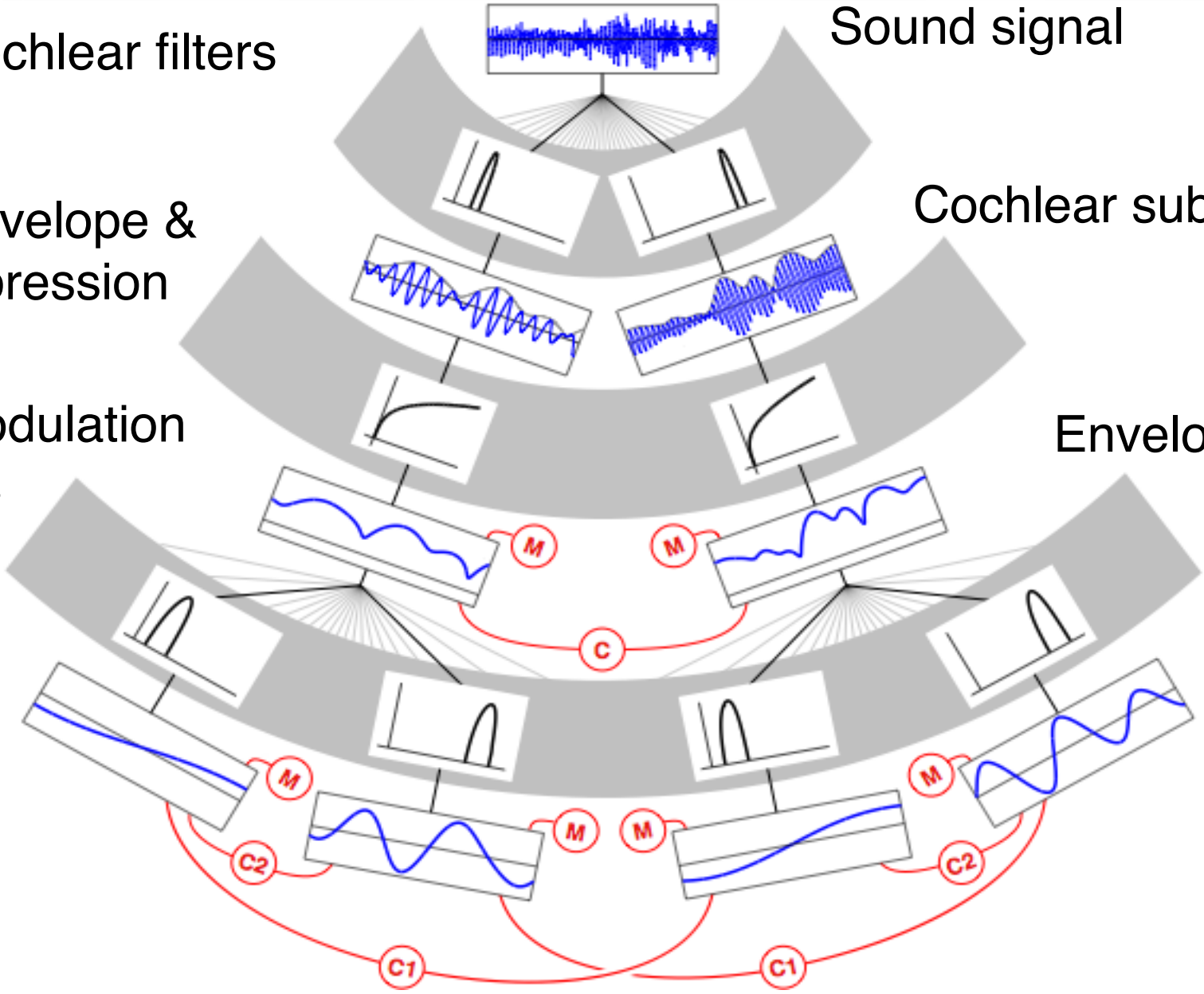
McDermott and Simoncelli, Neuron, 2011

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

1. Cochlear filters

2. Envelope & compression

3. Modulation filters



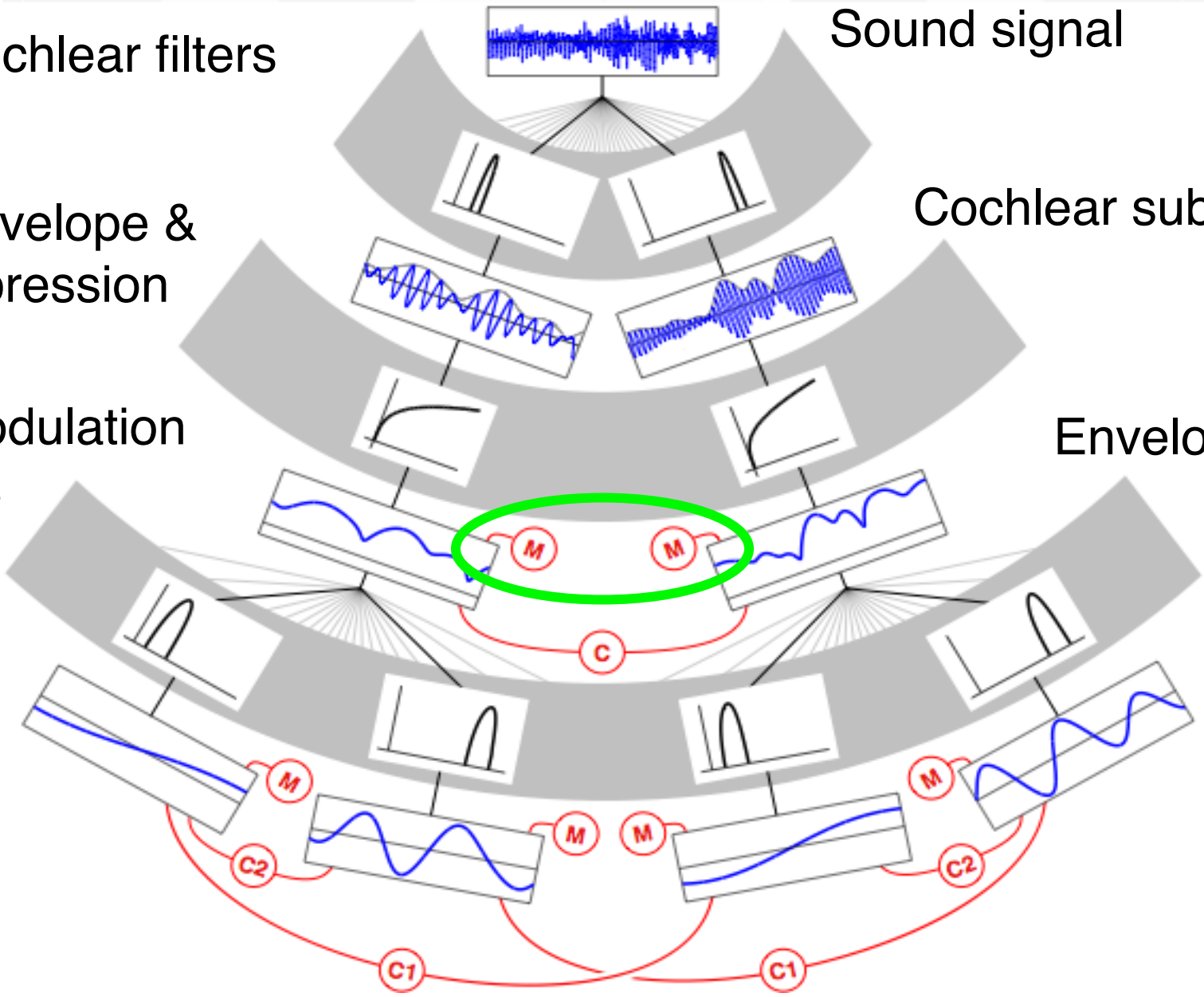
- Statistics are generic
- Not tailored to any specific natural sound
- Simple, easy to measure
- Not obvious that they would account for much of sound recognition
- But maybe a reasonable place to start.

For these statistics to be useful for recognition, at a minimum, they have to yield different values for different types of sounds…

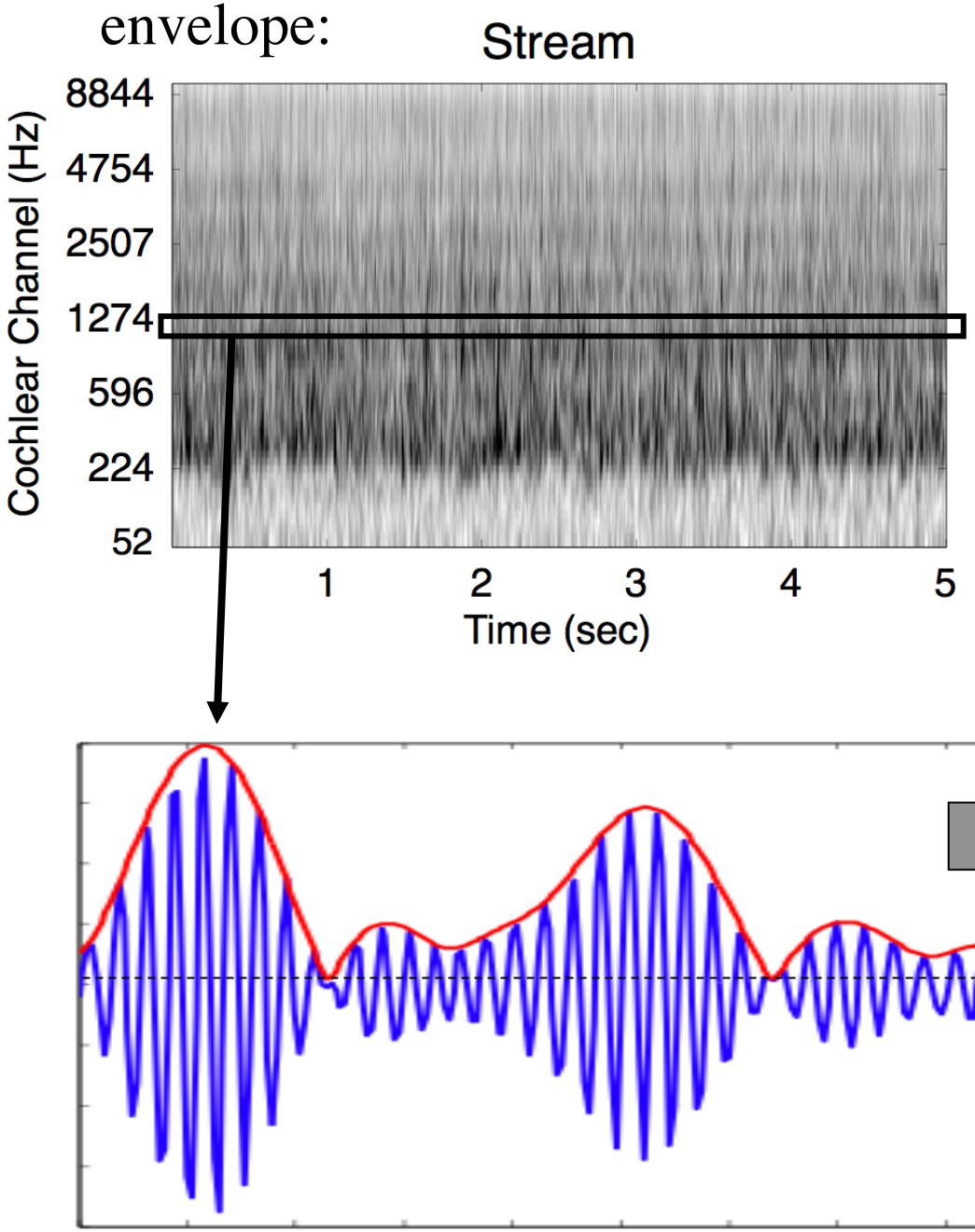
1. Cochlear filters

2. Envelope & compression

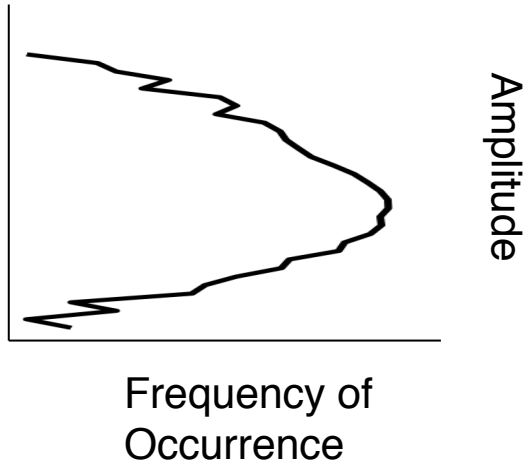
3. Modulation filters



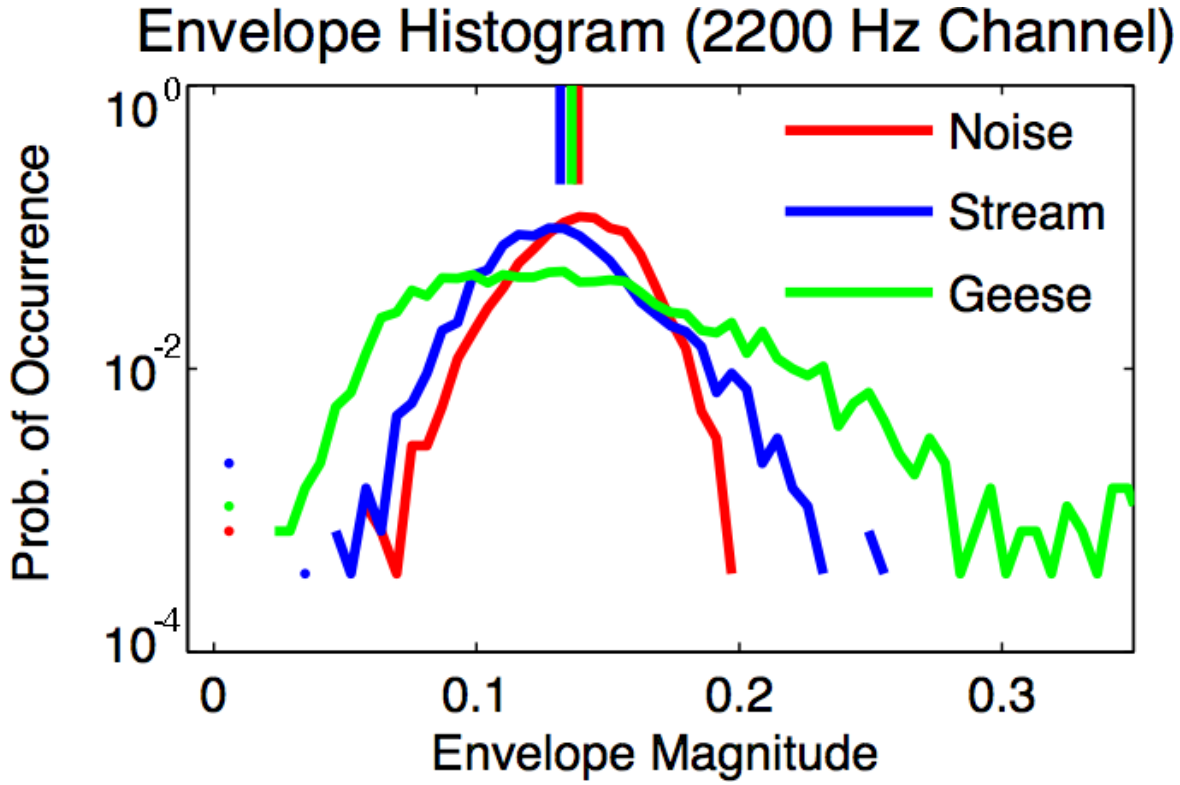
Marginal moments (mean, variance, skew) describe distribution of envelope:



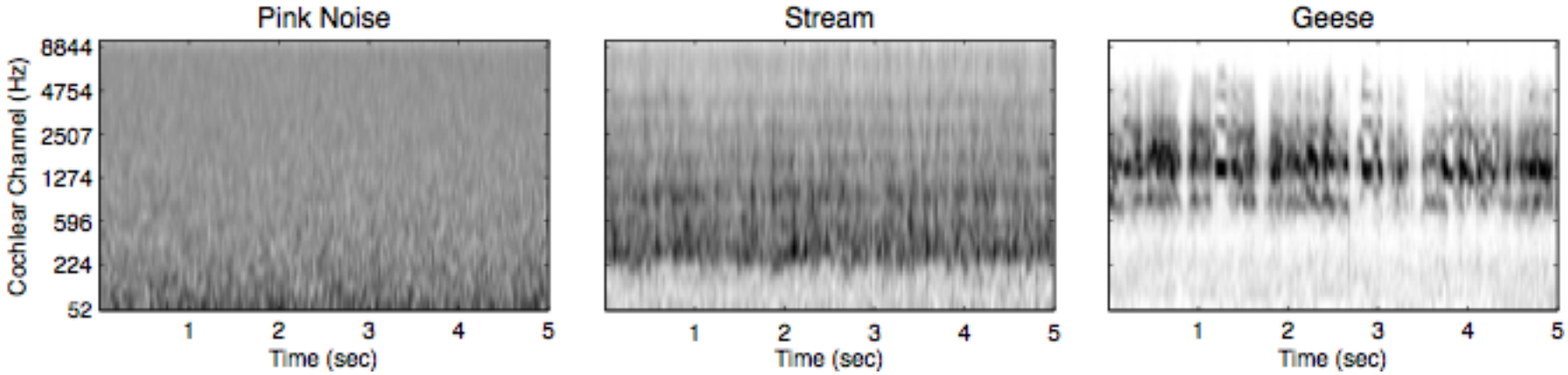
Envelope Distribution:



Envelope distributions for natural signals generally differ from those for noise.



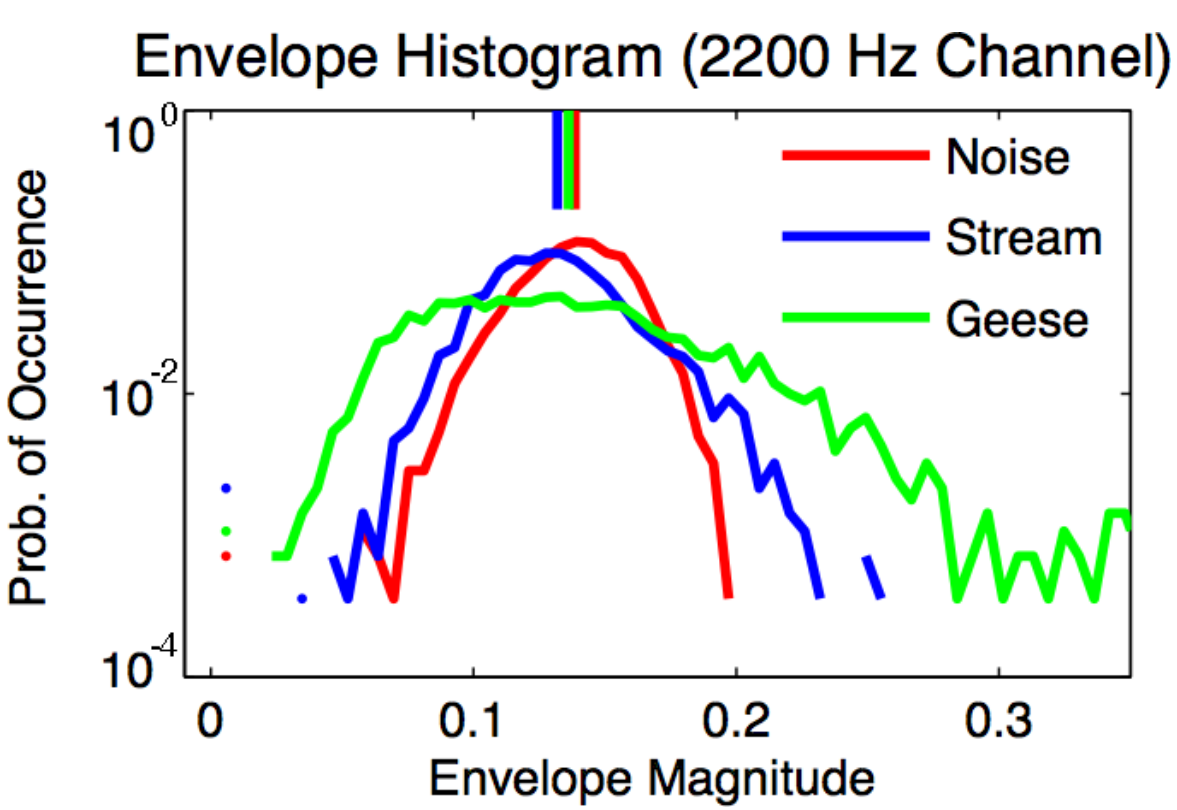
Distributions have similar mean, but different shapes.

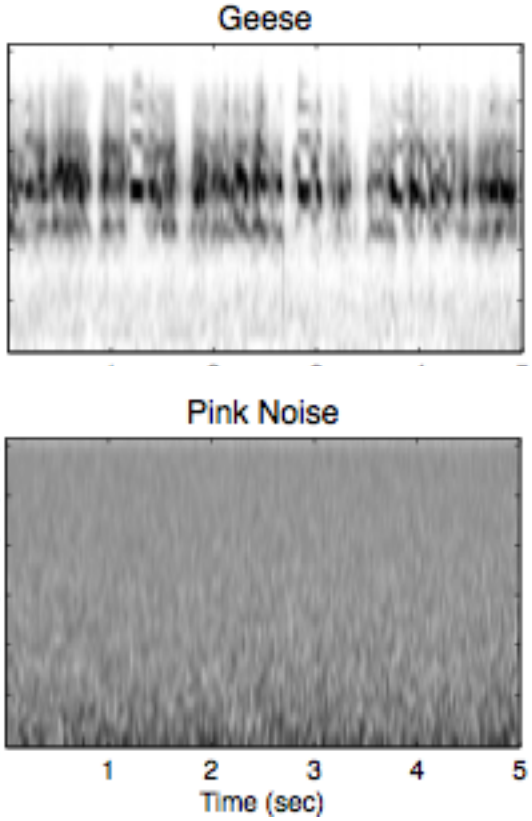


Natural signals are **sparser** than noise.

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

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



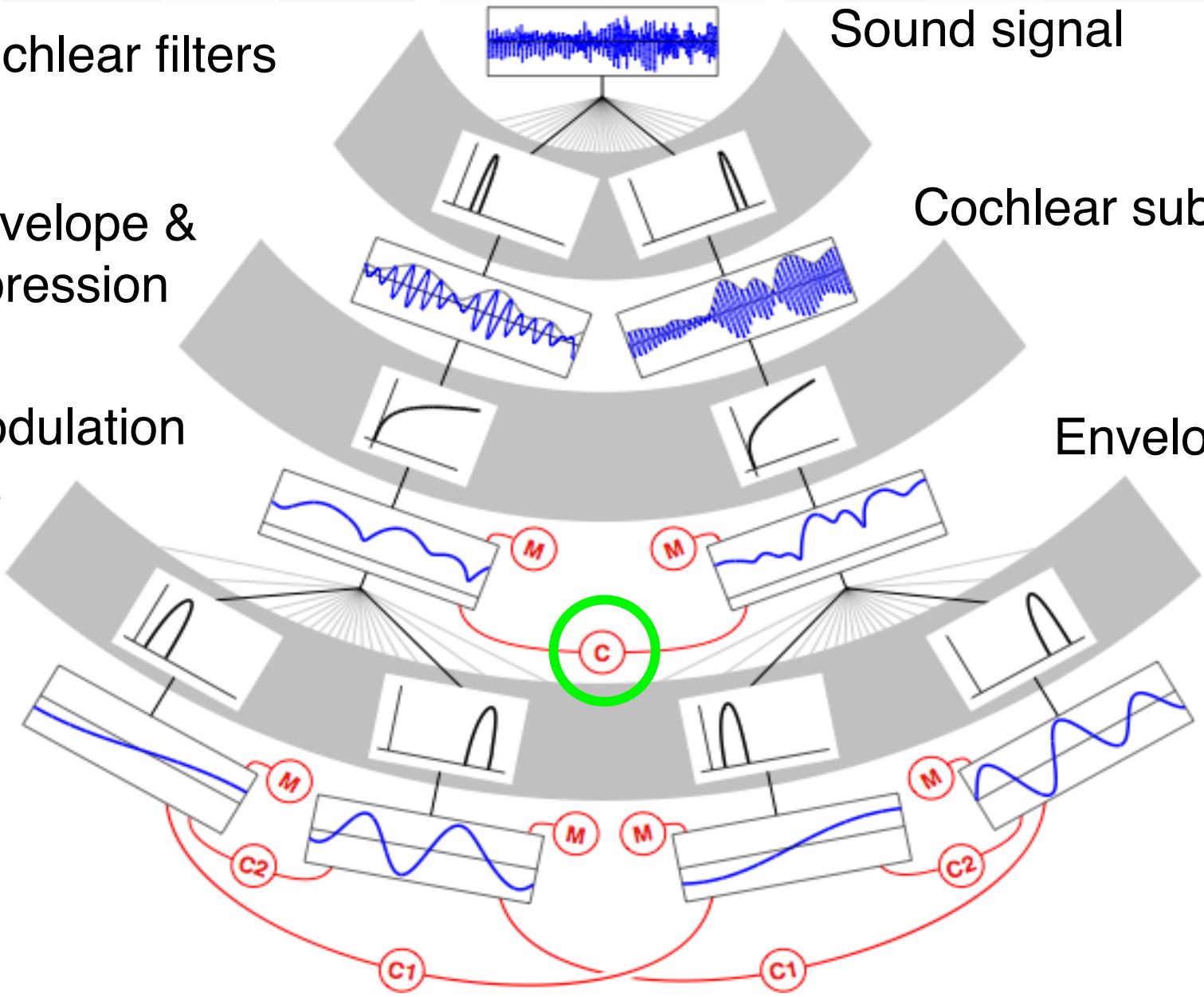


Sparsity reflected in envelope variance, skew.

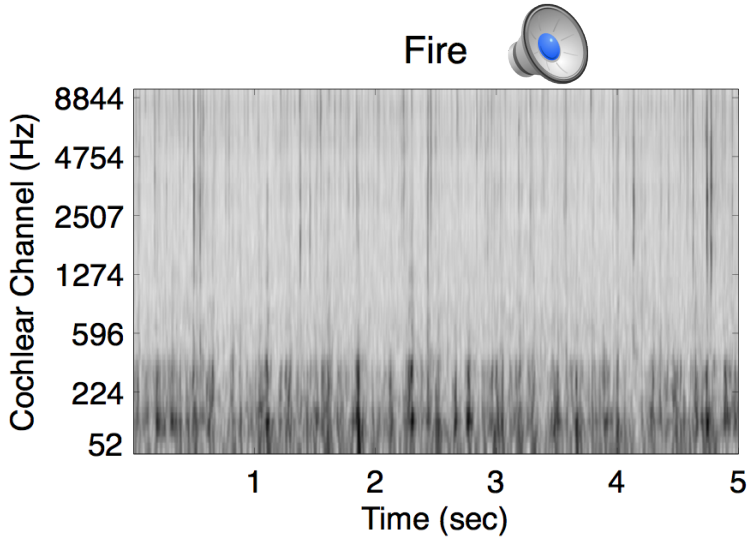
1. Cochlear filters

2. Envelope & compression

3. Modulation filters

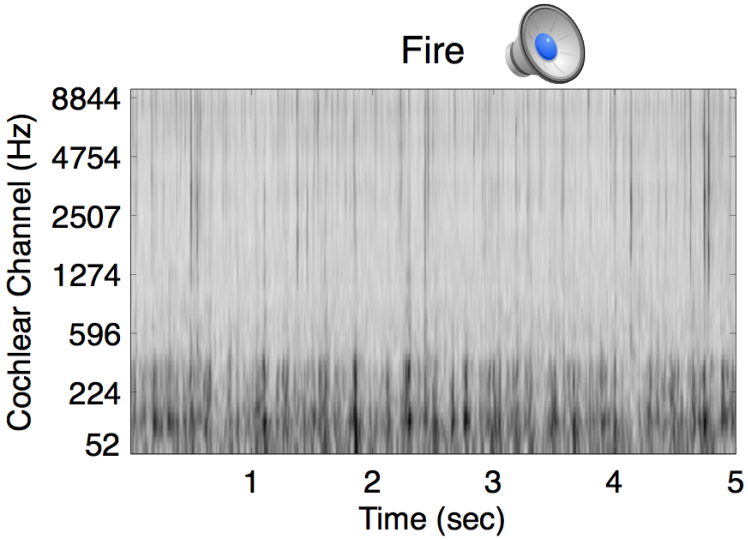


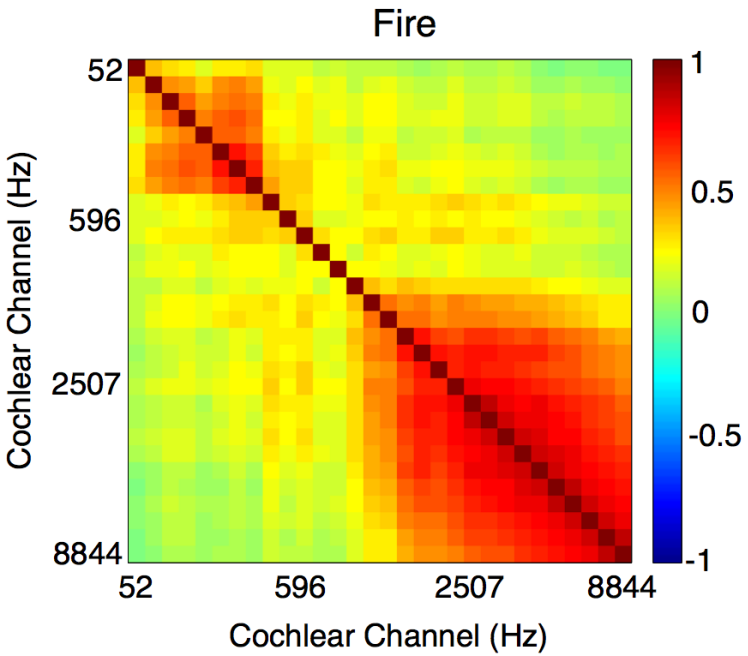
Correlations between cochlear envelopes also vary across sounds.



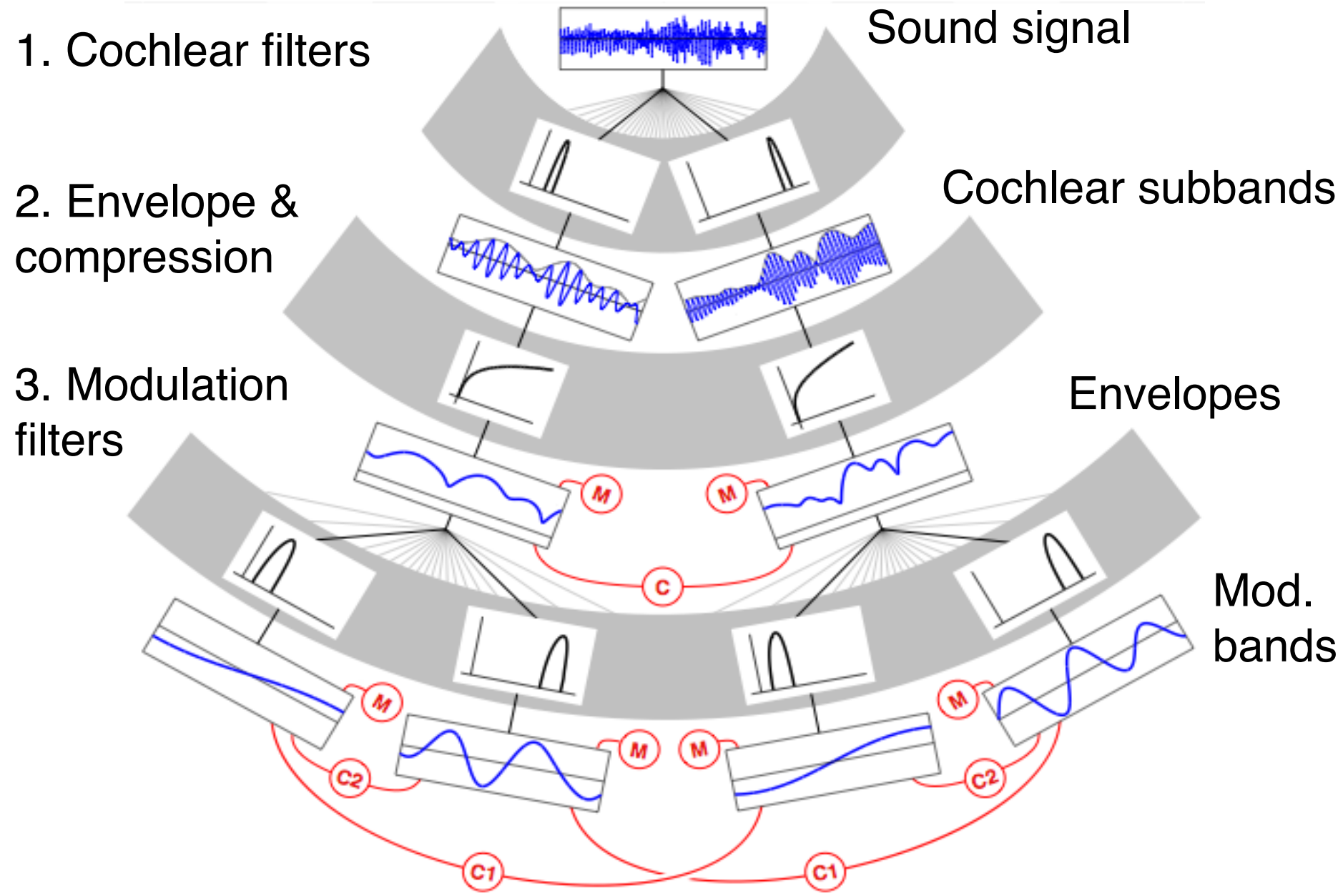
Broadband events induce dependencies between channels.

Correlations reflect broadband events (crackles, claps):





These statistics capture variation across sound.

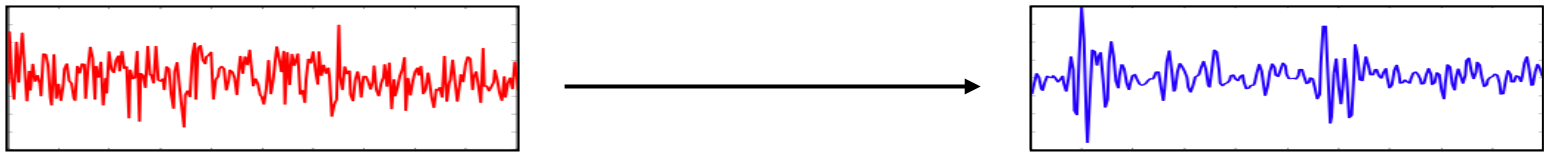


Will they capture the sound of real-world textures?

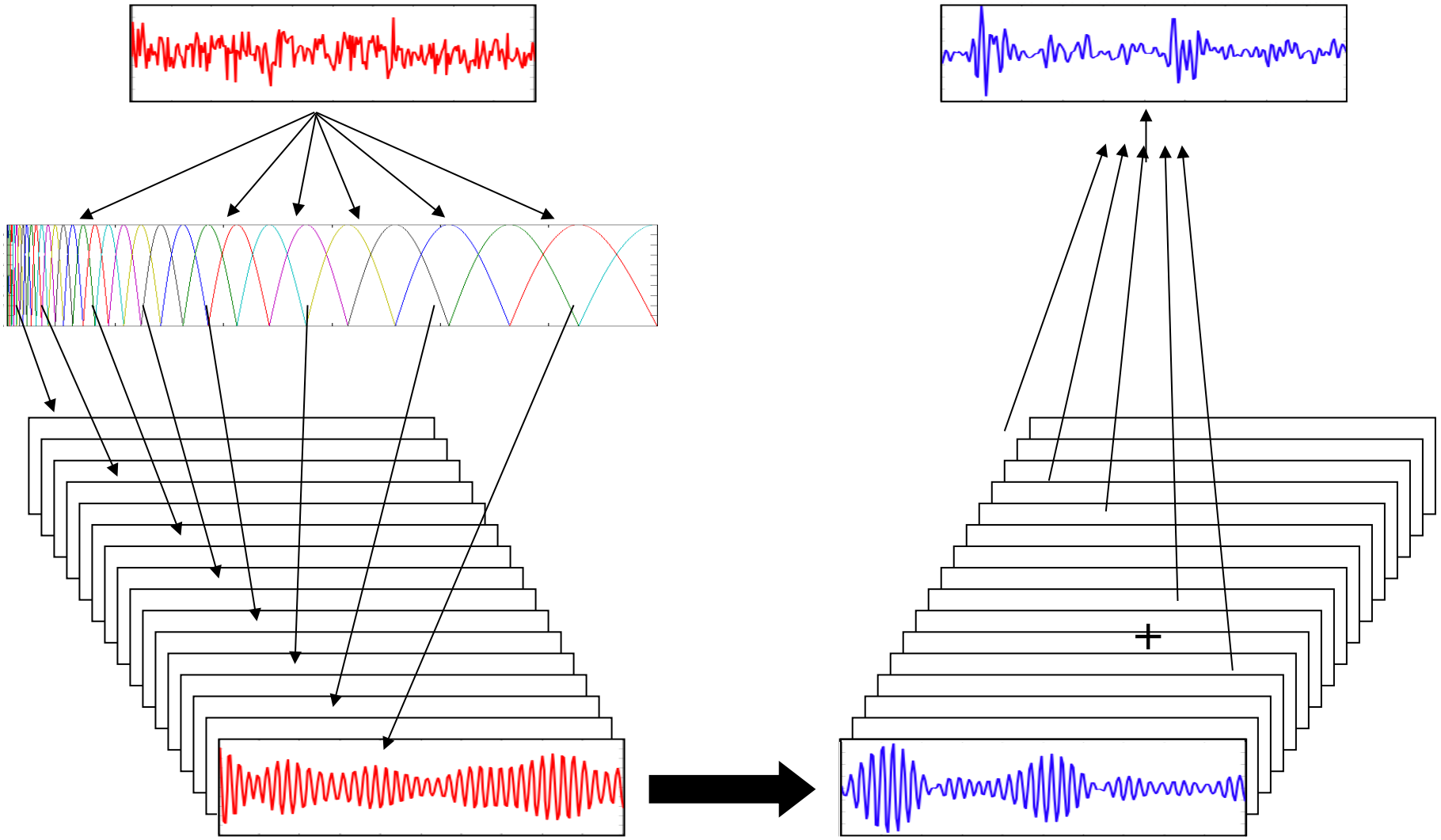
Strategy: synthesize signal constrained only to have the same statistics as some real-world sound.

Start with noise

Adjust it to have desired stats.

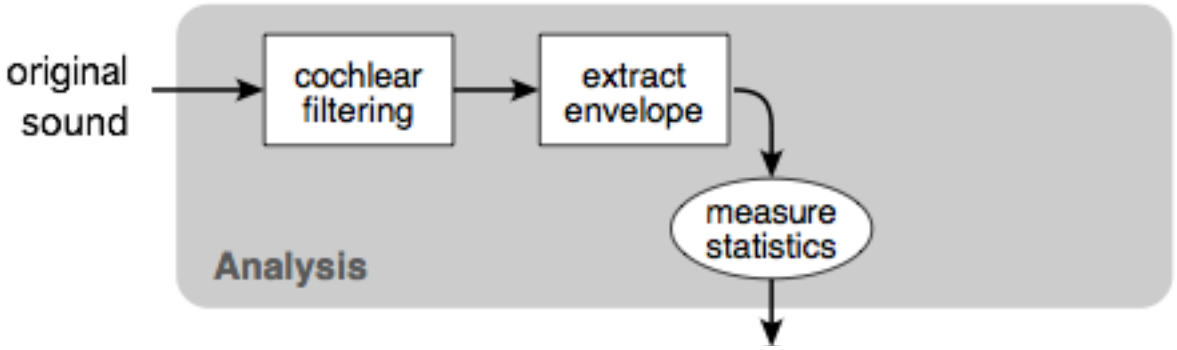


Basic idea: adjust subband envelopes with gradient descent



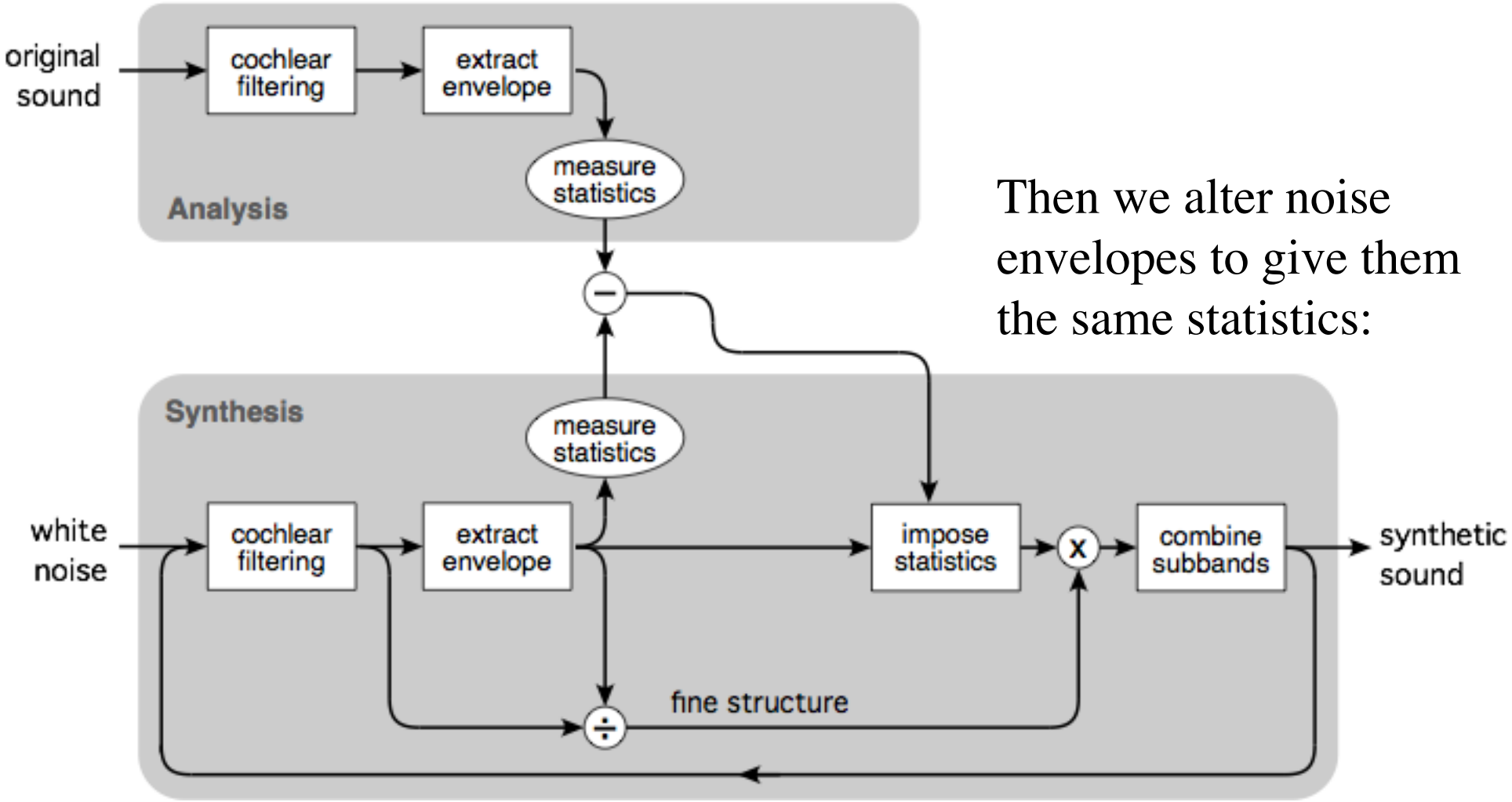
•Compute gradient of statistic w.r.t. envelope, change envelope in gradient direction until statistic matches desired value.

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



McDermott and Simoncelli, Neuron, 2011

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



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

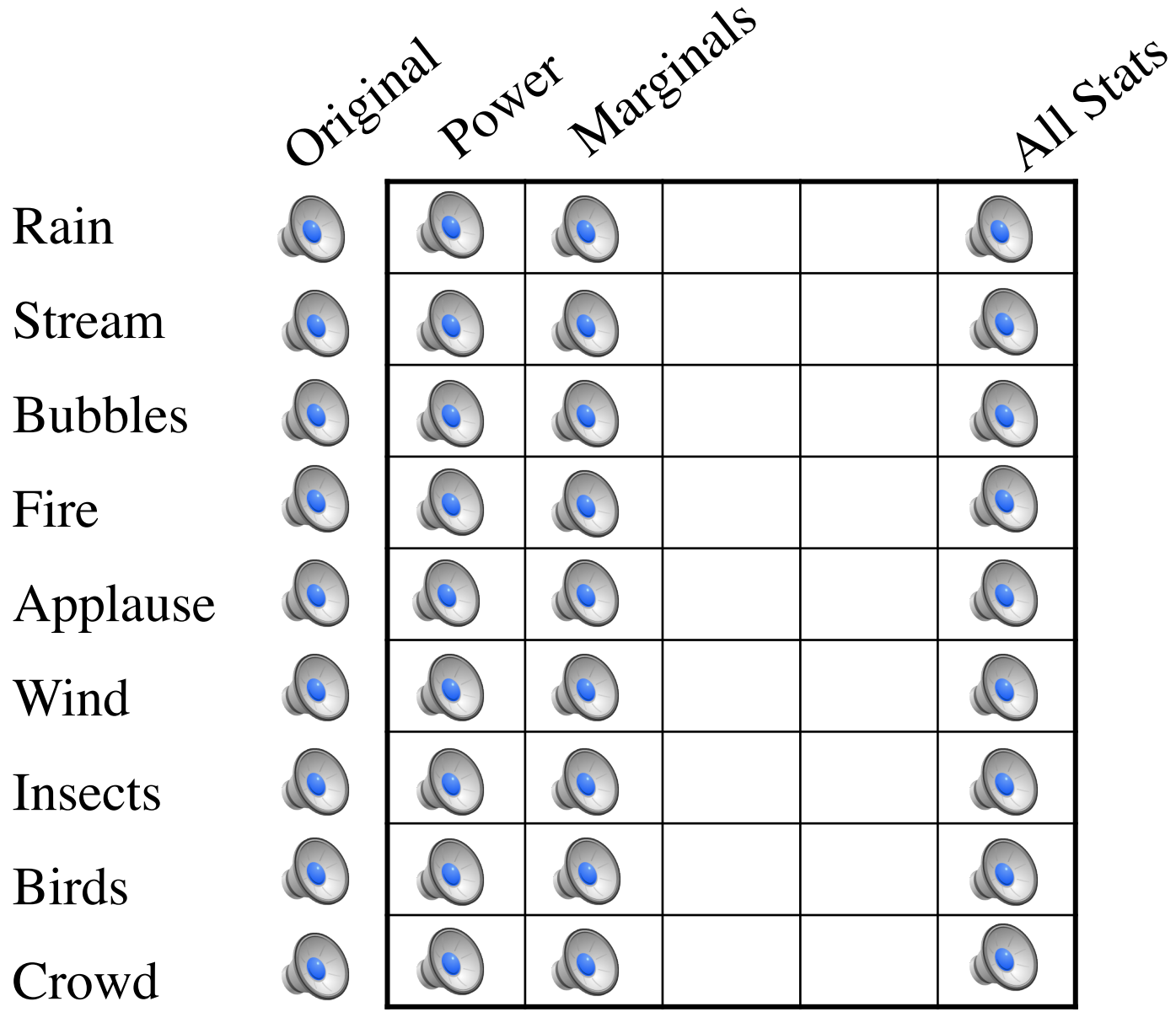
McDermott and Simoncelli, Neuron, 2011

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

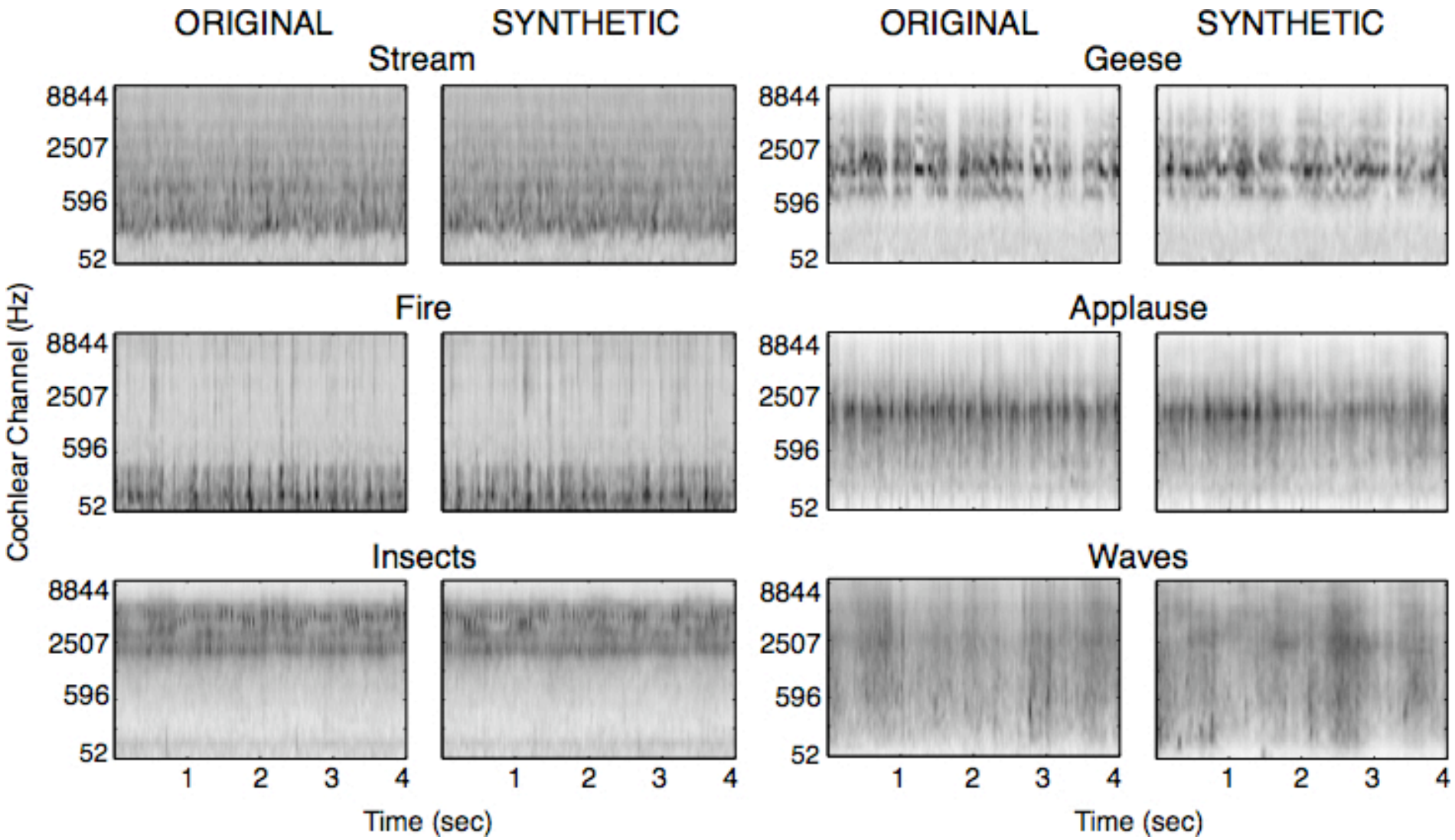
How do they sound?

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

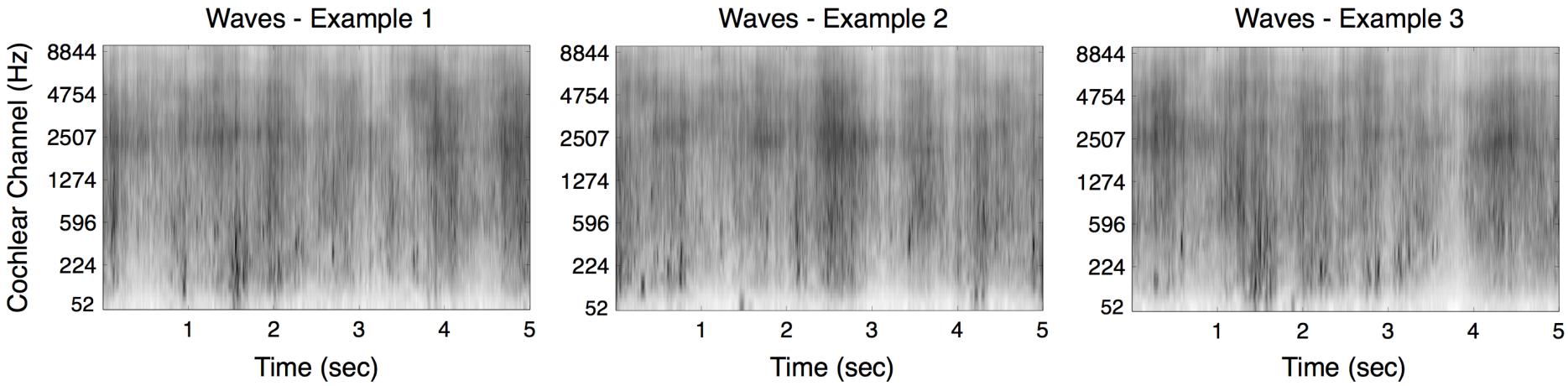
With marginal moments and pairwise correlations, synthesis is often compelling:



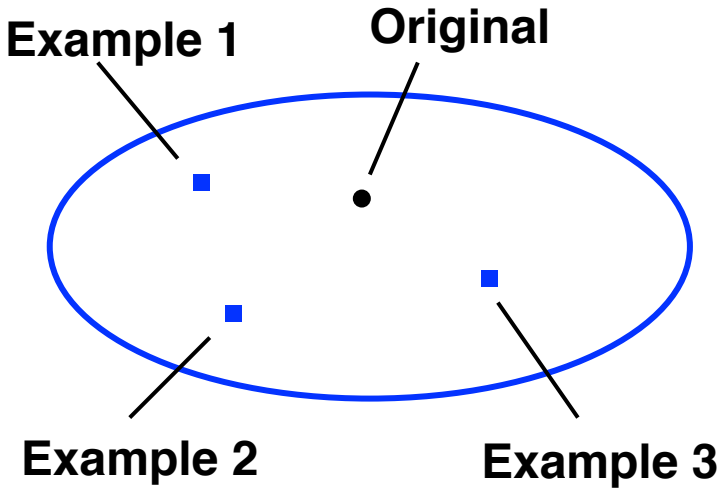
Synthesis does not merely recreate original sound:



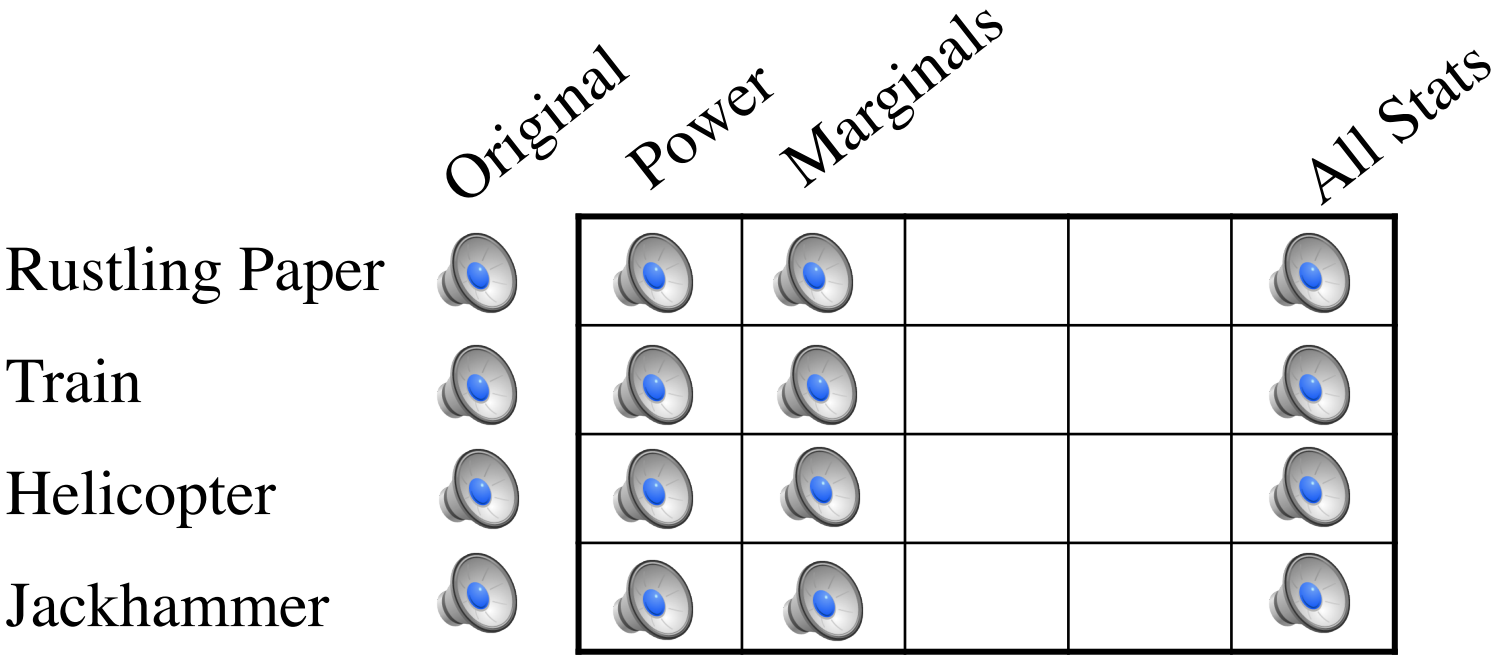
Because procedure is initialized with noise, it produces a different sound signal every time, sharing only statistical properties:



- 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.

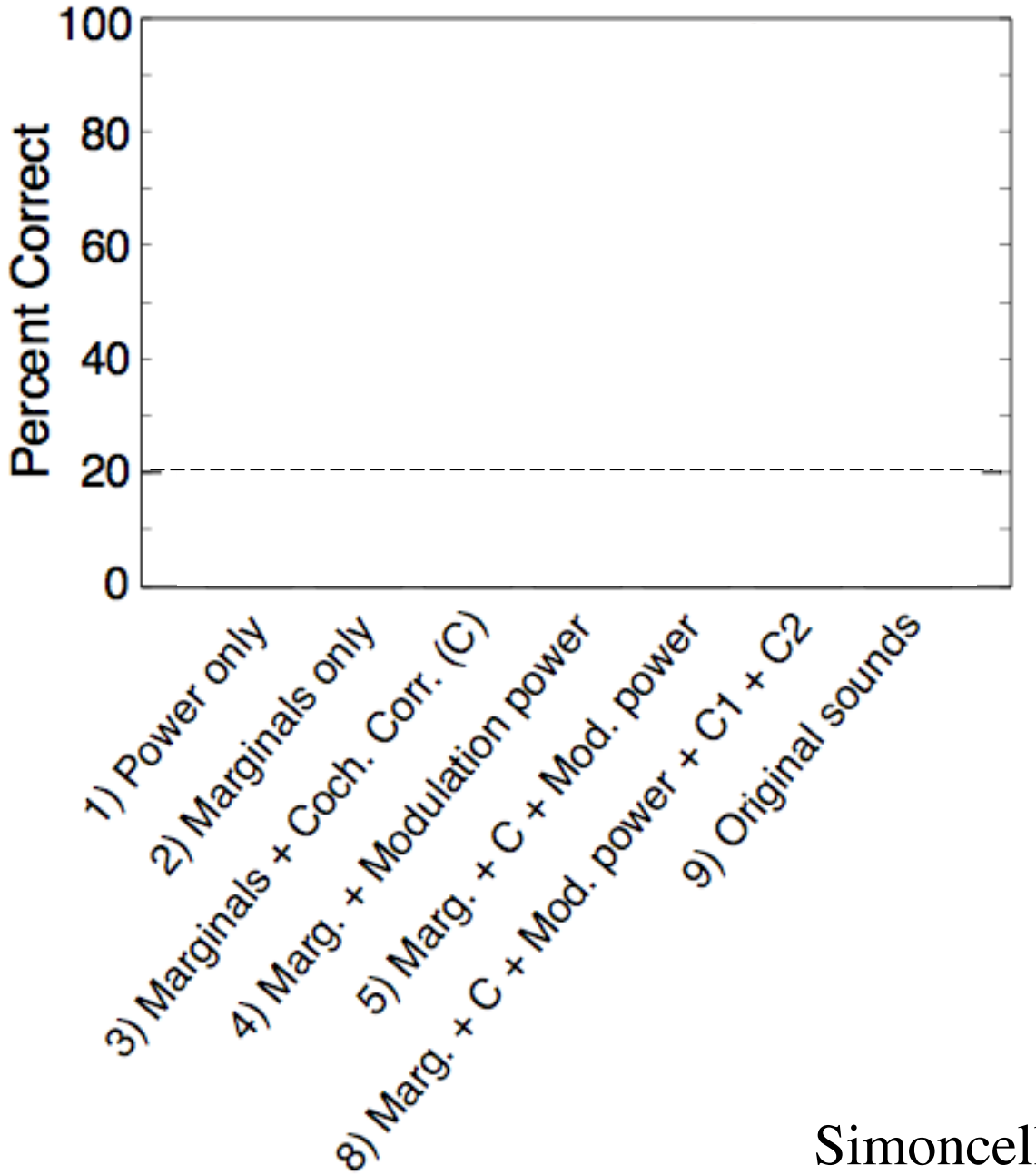


Also works for many “unnatural” sounds:



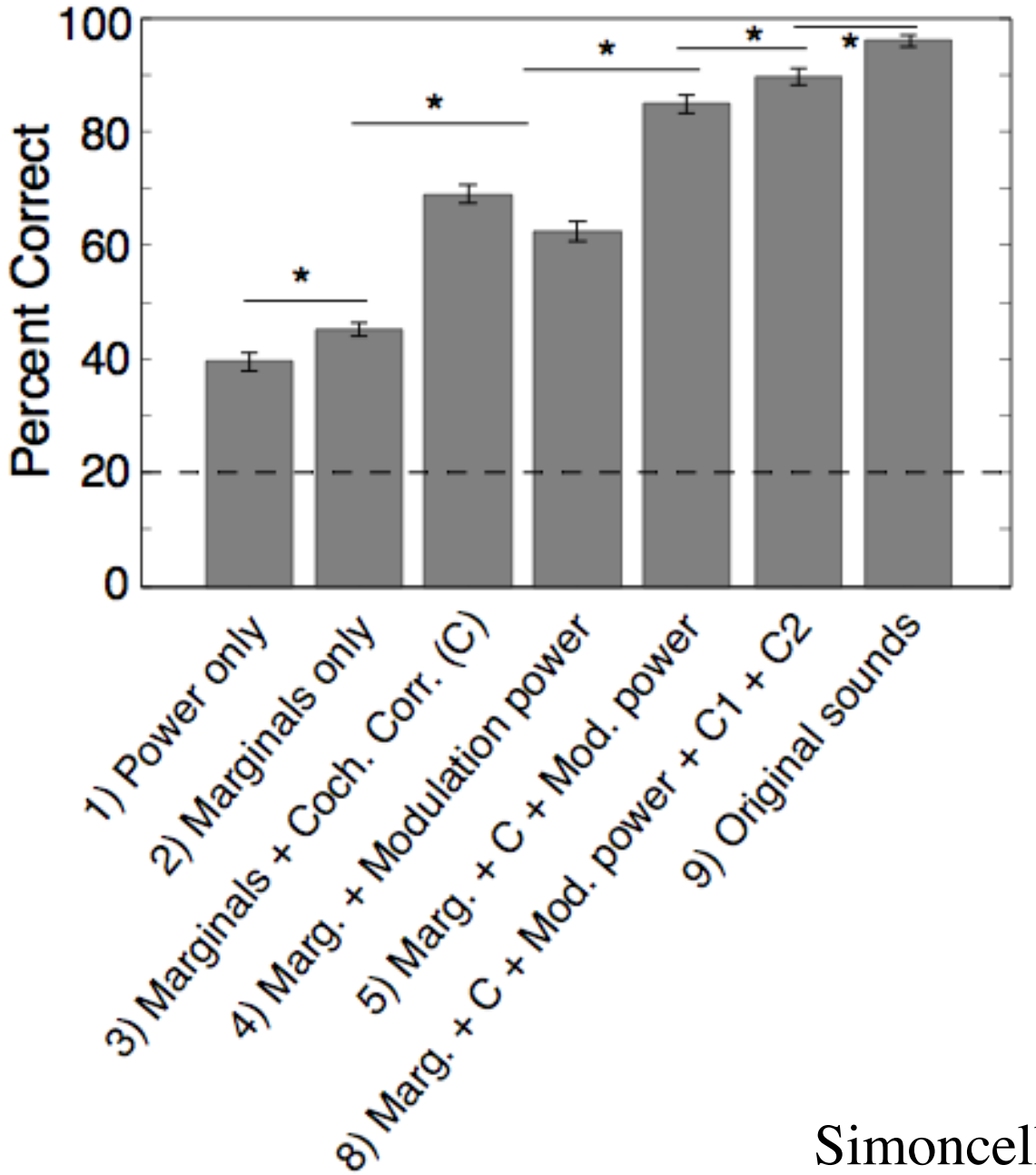
Success of synthesis suggests these statistics could underlie representation and recognition of textures.

Experiment: identify 5 sec sound clip from 5 choices:



McDermott &
Simoncelli, 2011, Neuron

Simple statistics can support recognition of real-world textures.



McDermott &
Simoncelli, 2011, Neuron

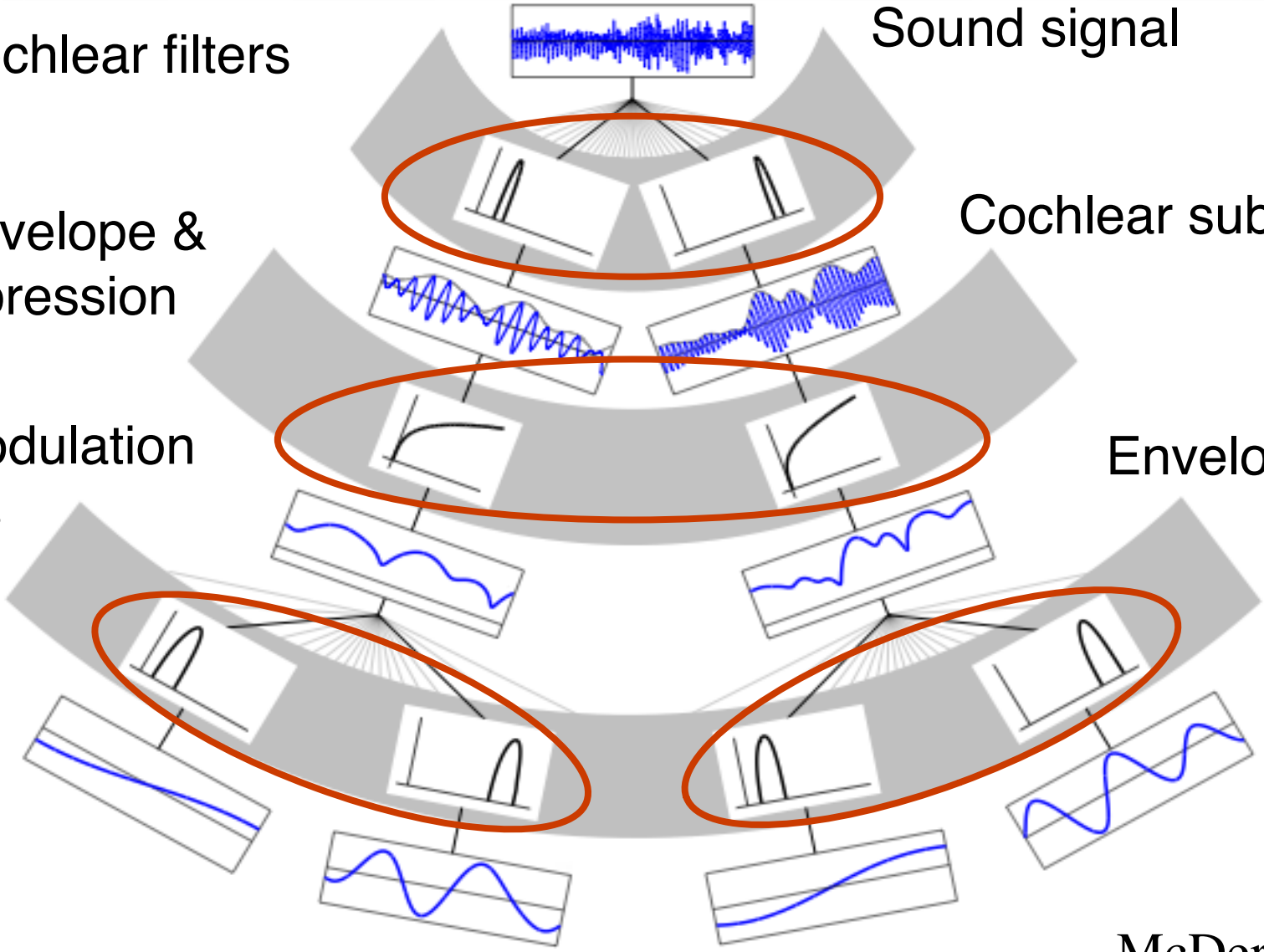
Will any set of statistics do?

What if we measure statistics from model deviating from biology?

1. Cochlear filters

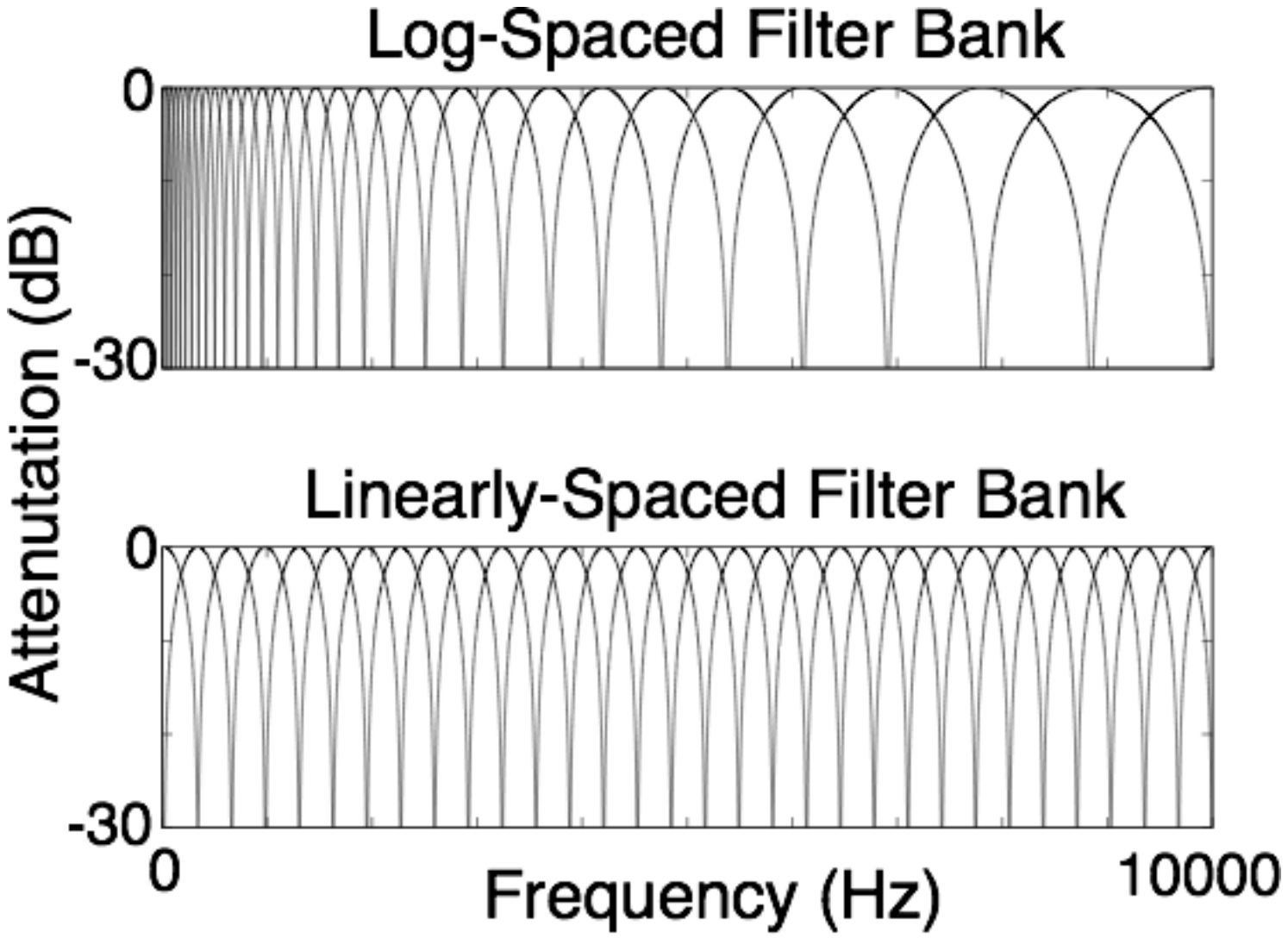
2. Envelope & compression

3. Modulation filters



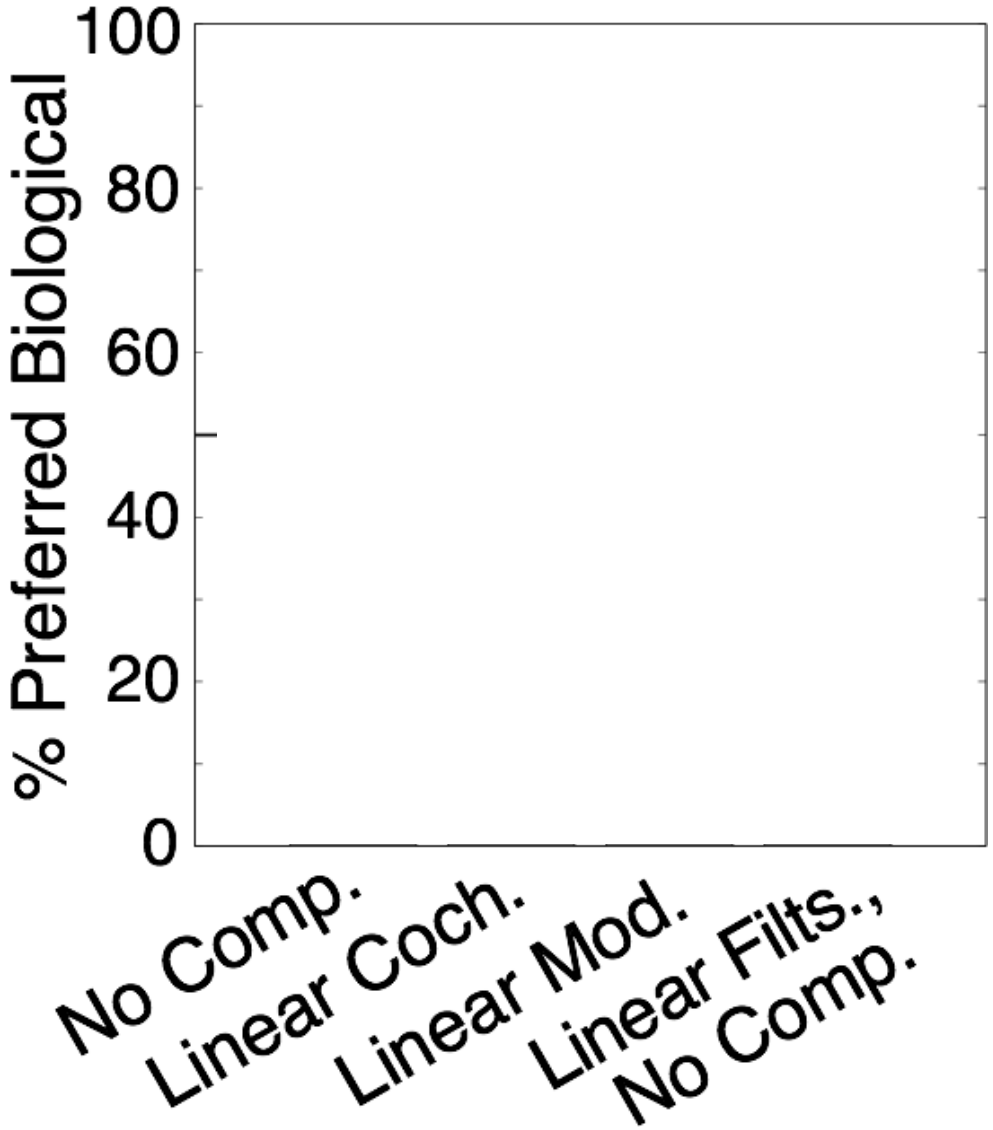
McDermott & Simoncelli, 2011, Neuron

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

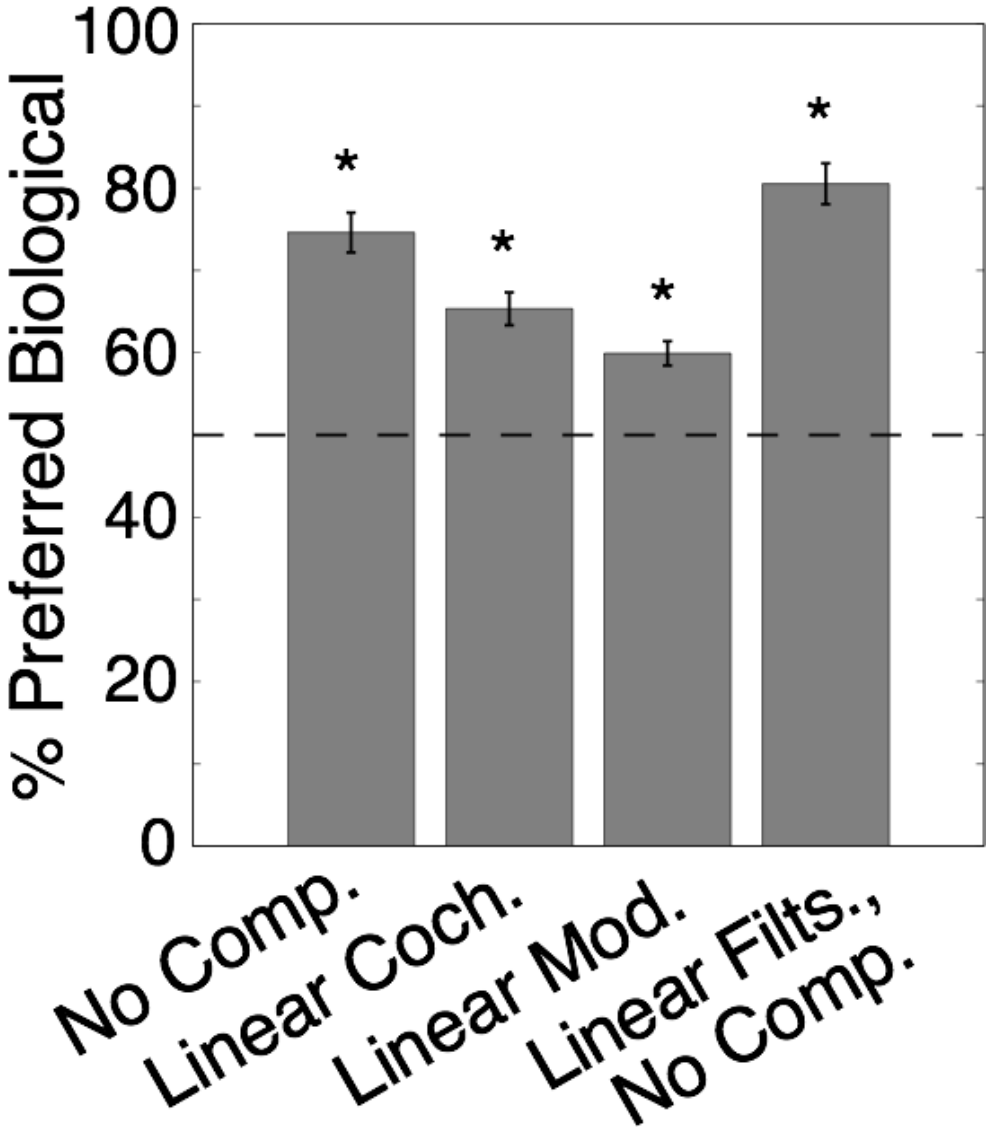


Experiment: Original - Synth1 - Synth2; one version non-biological

Which version sounds more realistic?



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



Crowd Noise:

Biological

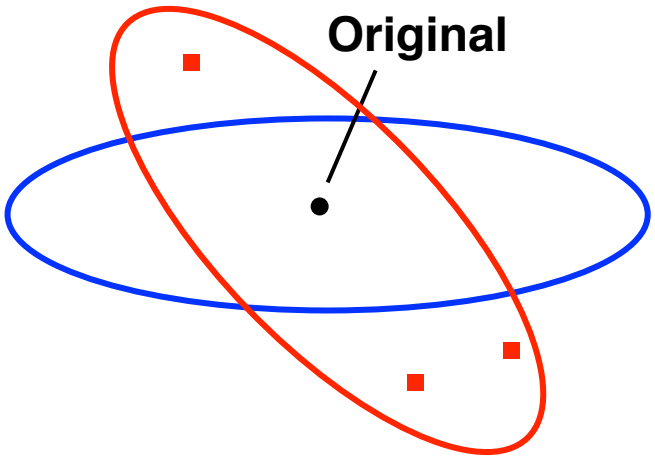
Non-biological

Helicopter:

Biological

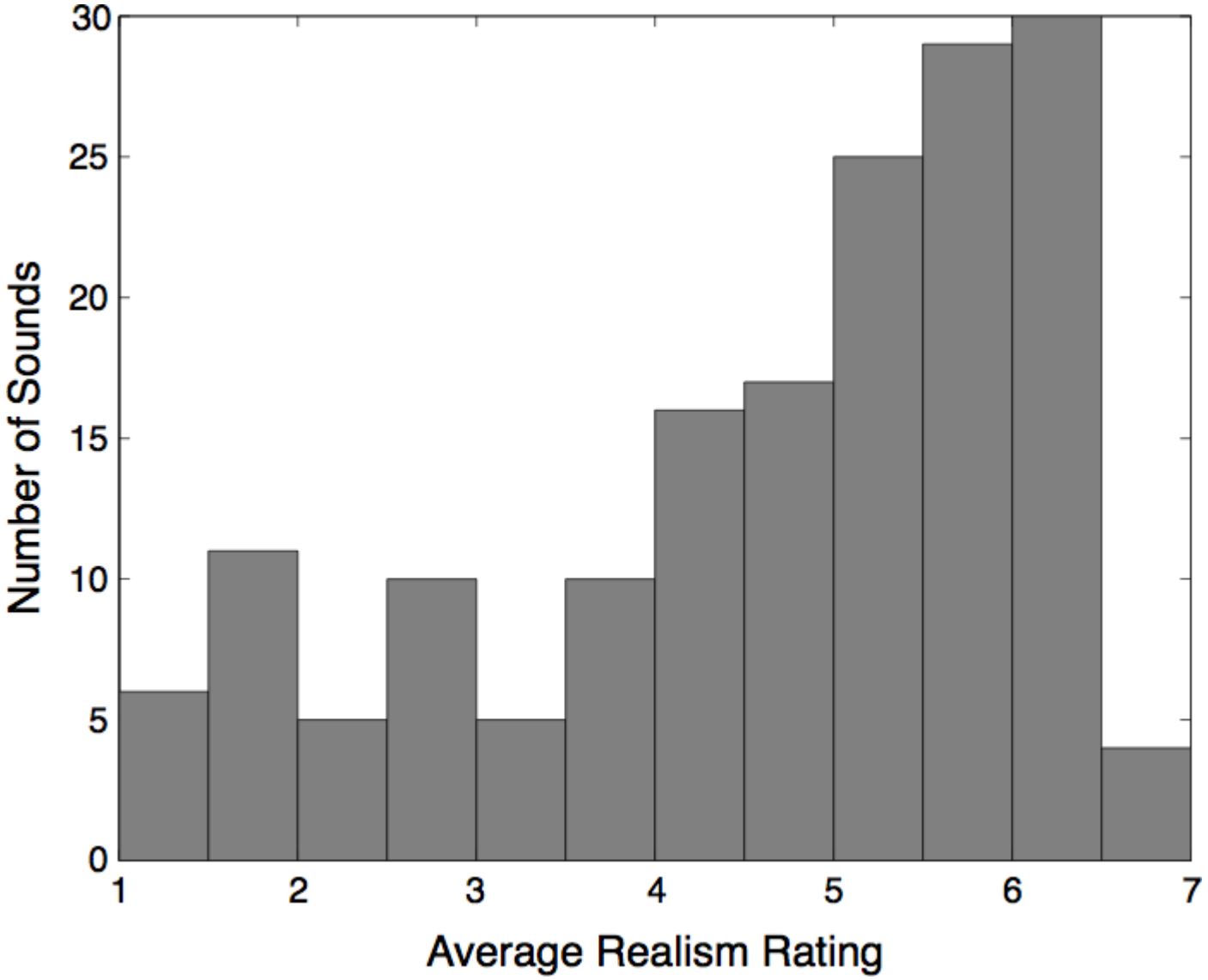
Non-biological

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



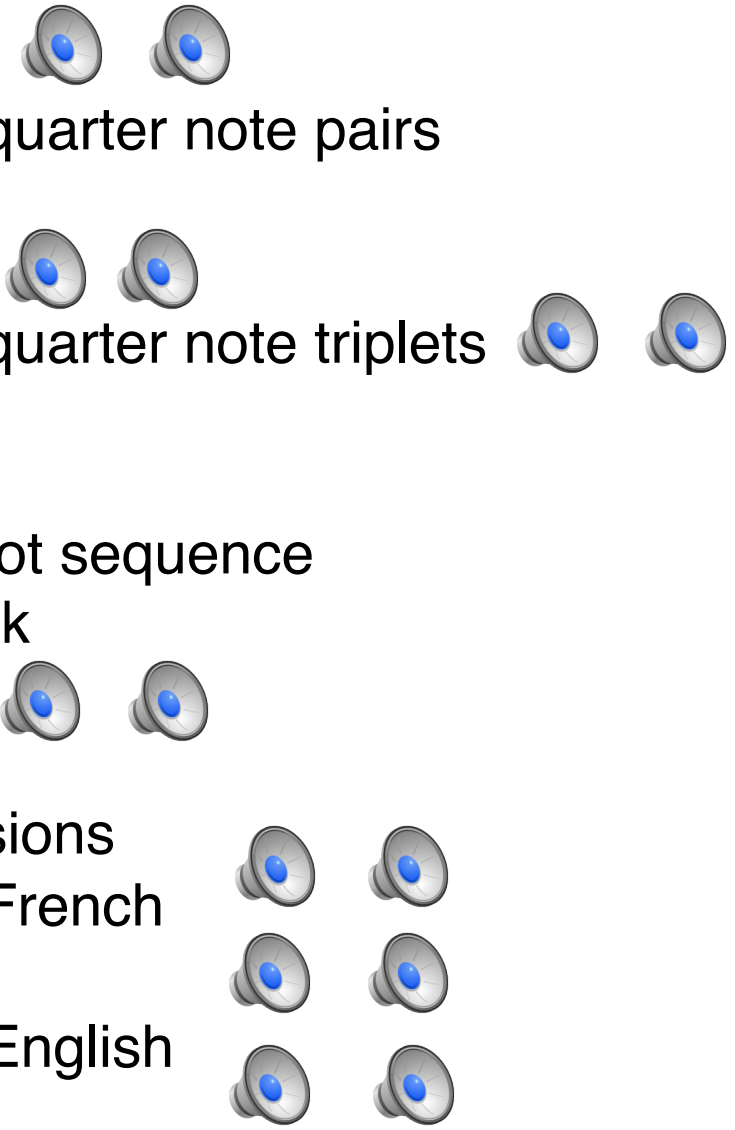
The sounds in the non-biological class don't sound like the original, because they are not defined with the measurements the brain is making.

Listen to original, then synthetic; rate realism from 1-7. (170 sounds)



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



Summary

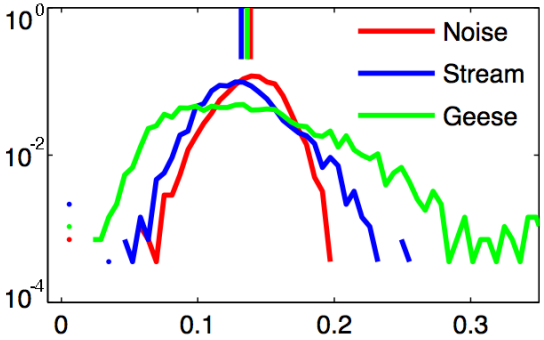
- Naturalistic sounds can be generated from simple statistics of early auditory representations

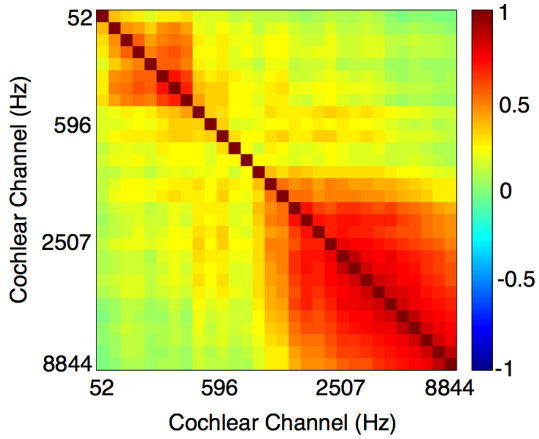
  (marginal moments, pairwise correlations of cochlear, modulation filters)

- Similar statistics may be extracted by auditory system, used for recognition

- Relatively simple time-averaged statistics capture a form of invariance.

- Vocalizations, other sounds, implicate more sophisticated statistics

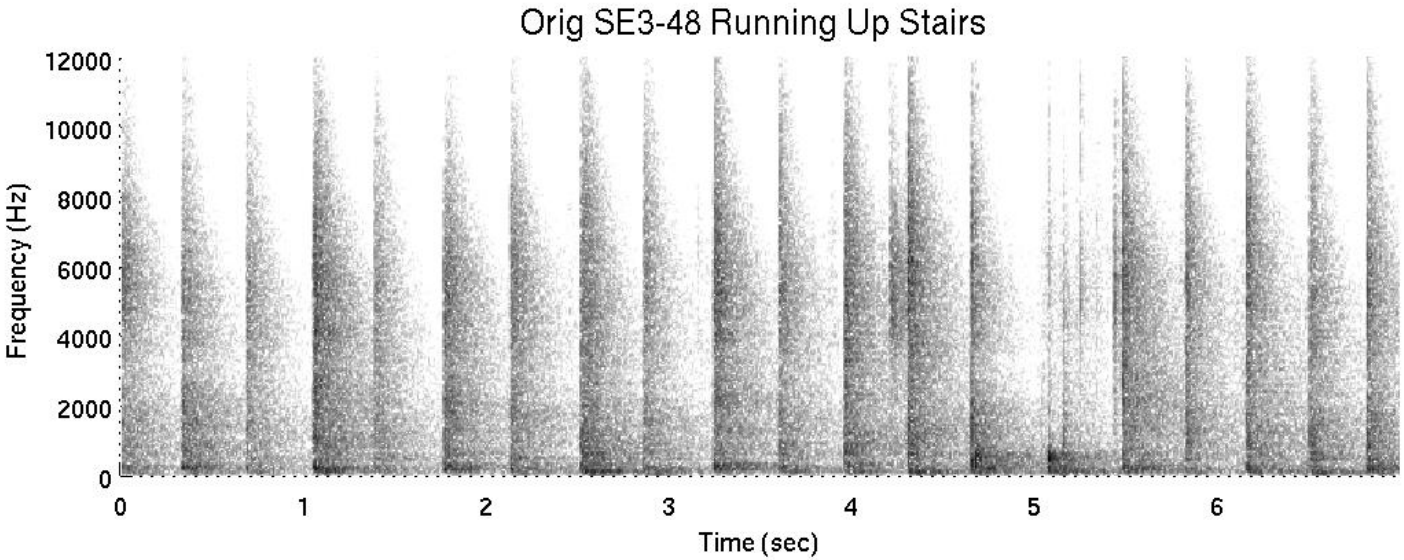


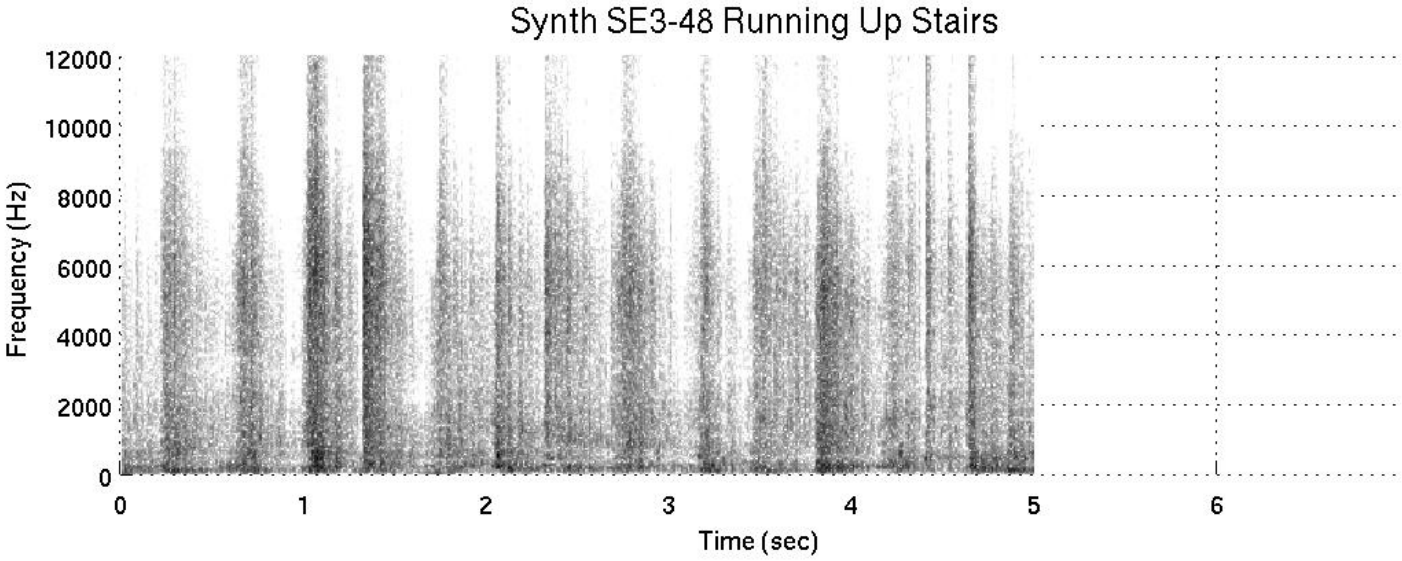


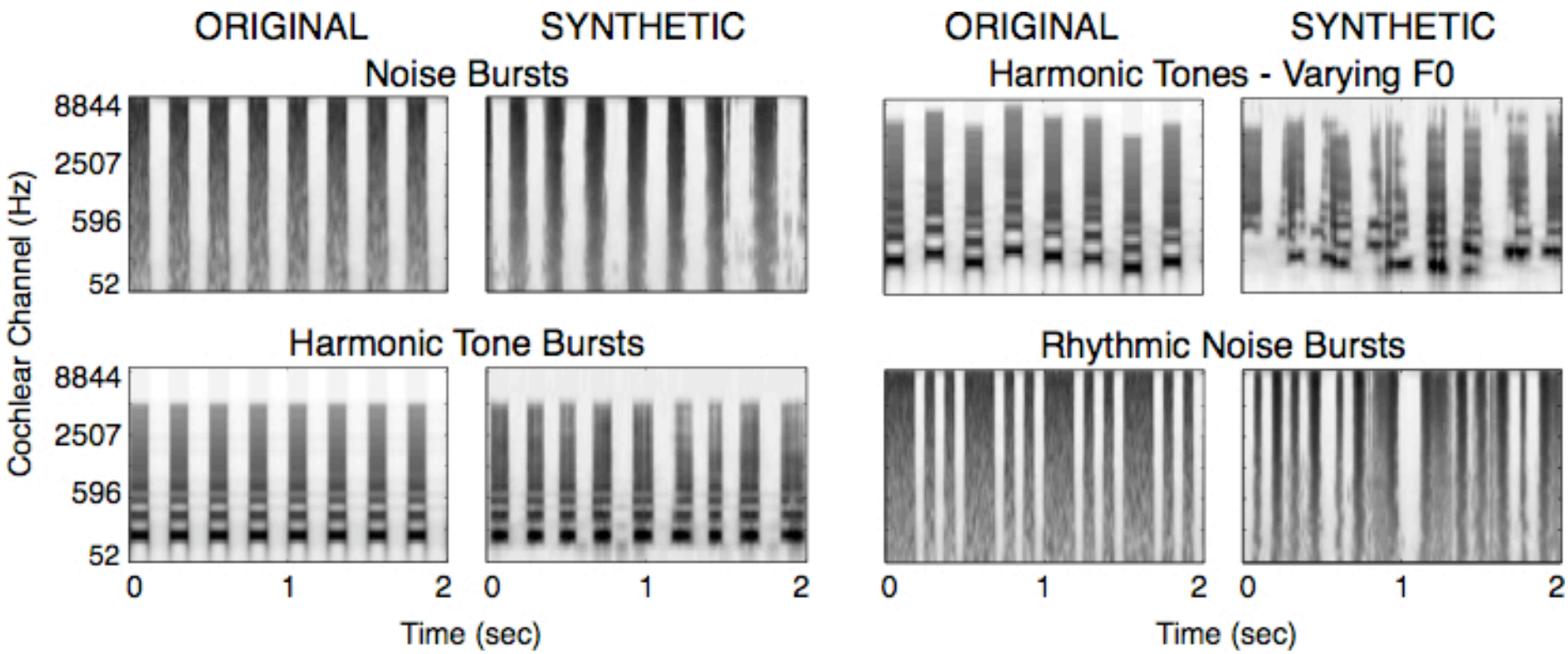
- Synthesis is powerful tool for studying perception.

- Paper just out: McDermott & Simoncelli, 2011 - on my web page

Reverb does not sound right:






ORIGINAL
SYNTHETIC
ORIGINAL
SYNTHETIC
Noise Bursts
Harmonic Tones - Varying F0
Harmonic Tone Bursts
Rhythmic Noise Bursts
Cochlear Channel (Hz)
8844
2507
596
52
0
1
2
Time (sec)

Will texture statistics also be useful for machine recognition?

Potential application: video soundtrack classification

Please check the following annotation results, Task# 640

Replay/Pause

Tick **all the lines** that apply to this video.

Human Activities

- ☐ Person walking
- ☐ Person running
- ☑ Person squatting
- ☐ Person showing the motion of standing up
- ☐ Making/fixing/assembling stuffs with hands (hands visible)
- ☐ Person batting baseball

Object/Scene Categories

- ☐ Indoor Kitchen
- ☑ Outdoor with grass or trees visible
- ☑ Baseball field
- ☐ Crowd (a group of more than 3 people)
- ☐ Close up cakes

Audio Events

- ☐ Speech comprehensible
- ☐ Music
- ☑ Cheering
- ☑ Clapping

Audio Categories

Choose the item best describes the video.

- ◉ Outdoor rural
- ○ Outdoor urban
- ○ Indoor quiet
- ○ Indoor noisy
- ○ Other
- ○ Blank

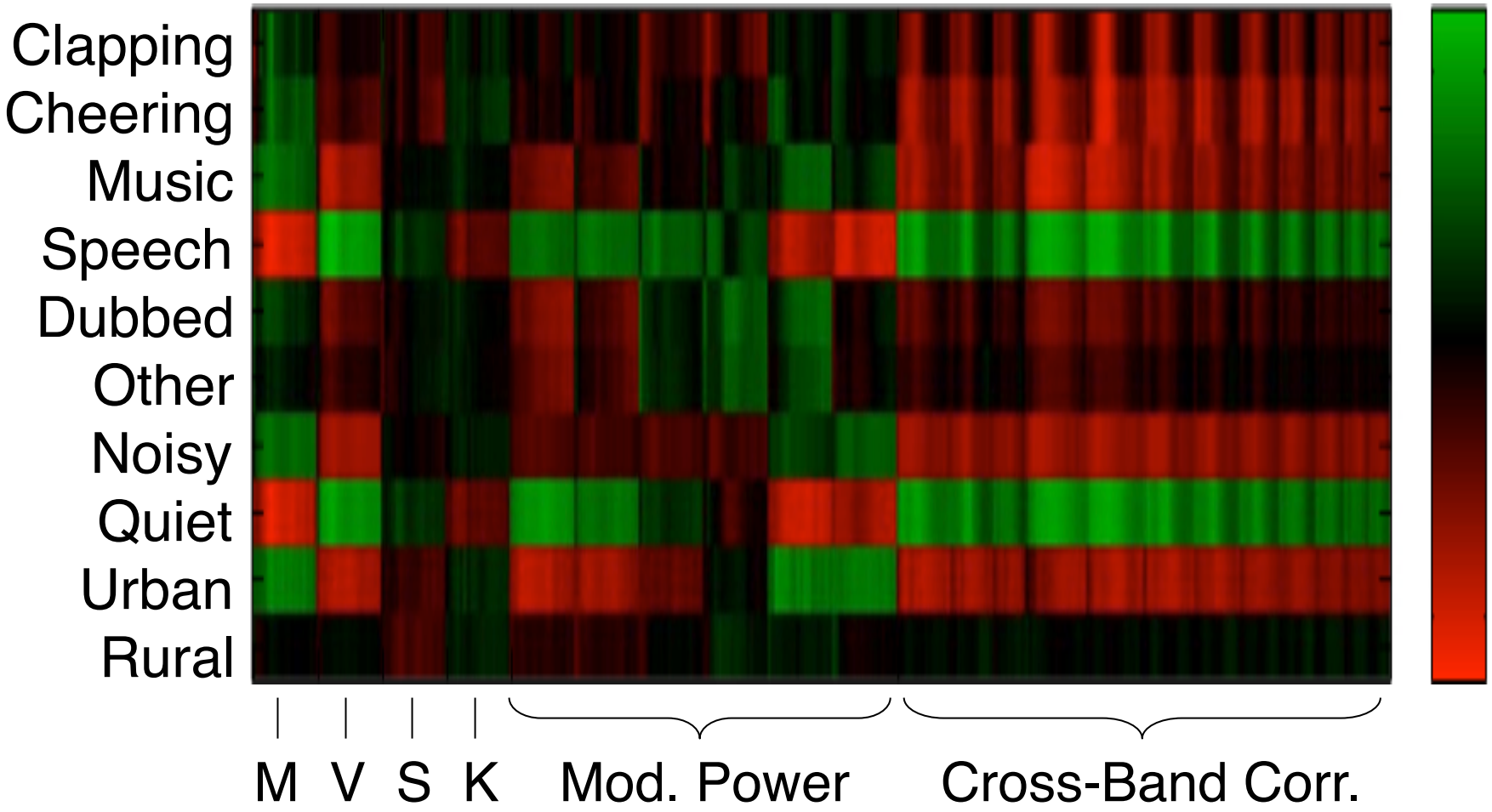
Choose the item best describes the video.

- ◉ Original audio
- ○ Dubbed audio

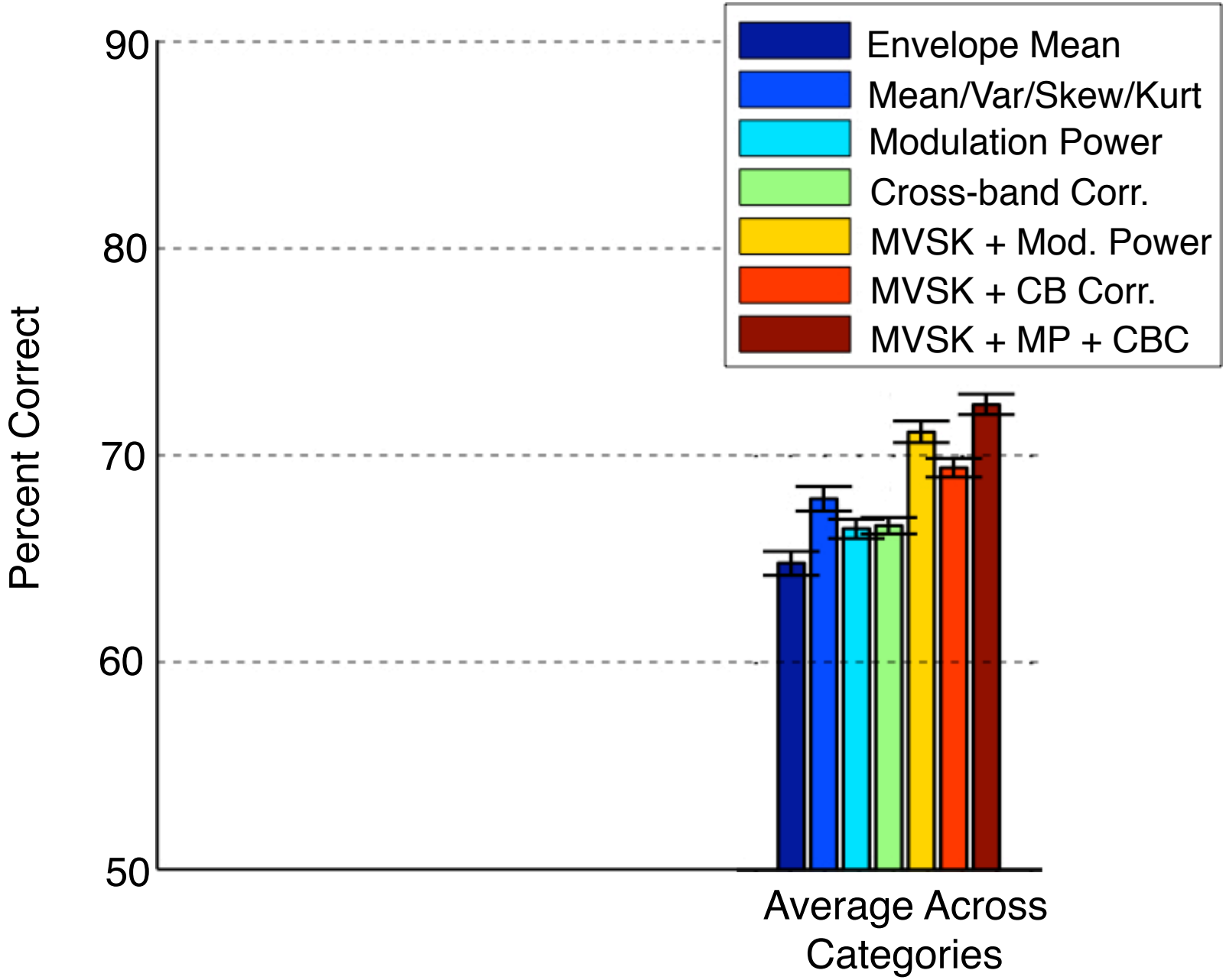
courtesy Yu-Gang Jiang

Texture statistics can be used as features for SVM classification.

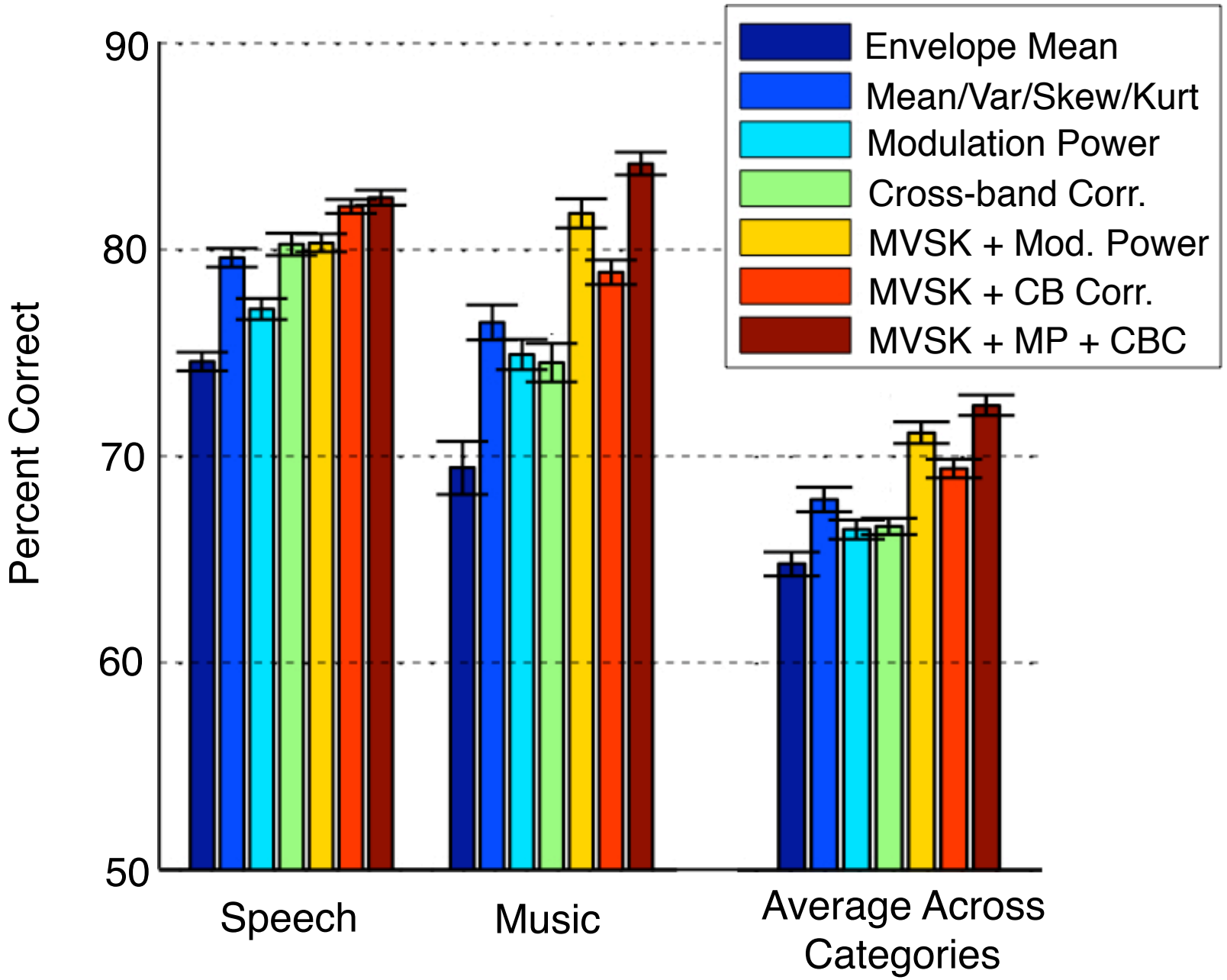
On average, the statistics have different values for different labels:



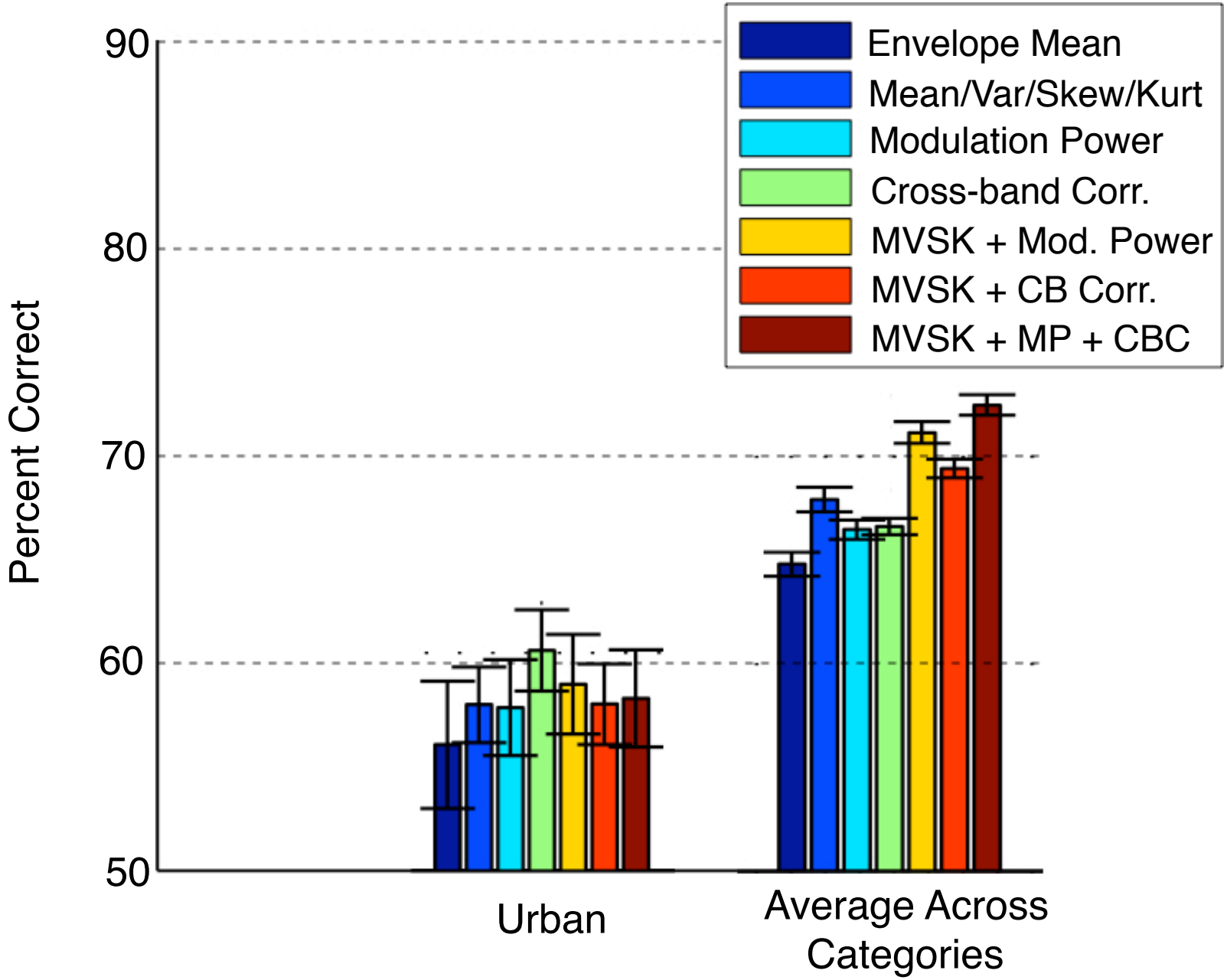
Classification improves as statistics are added; performance is modest.



Some statistics benefit particular classes more than others:



Performance is poor for acoustically heterogeneous labels:

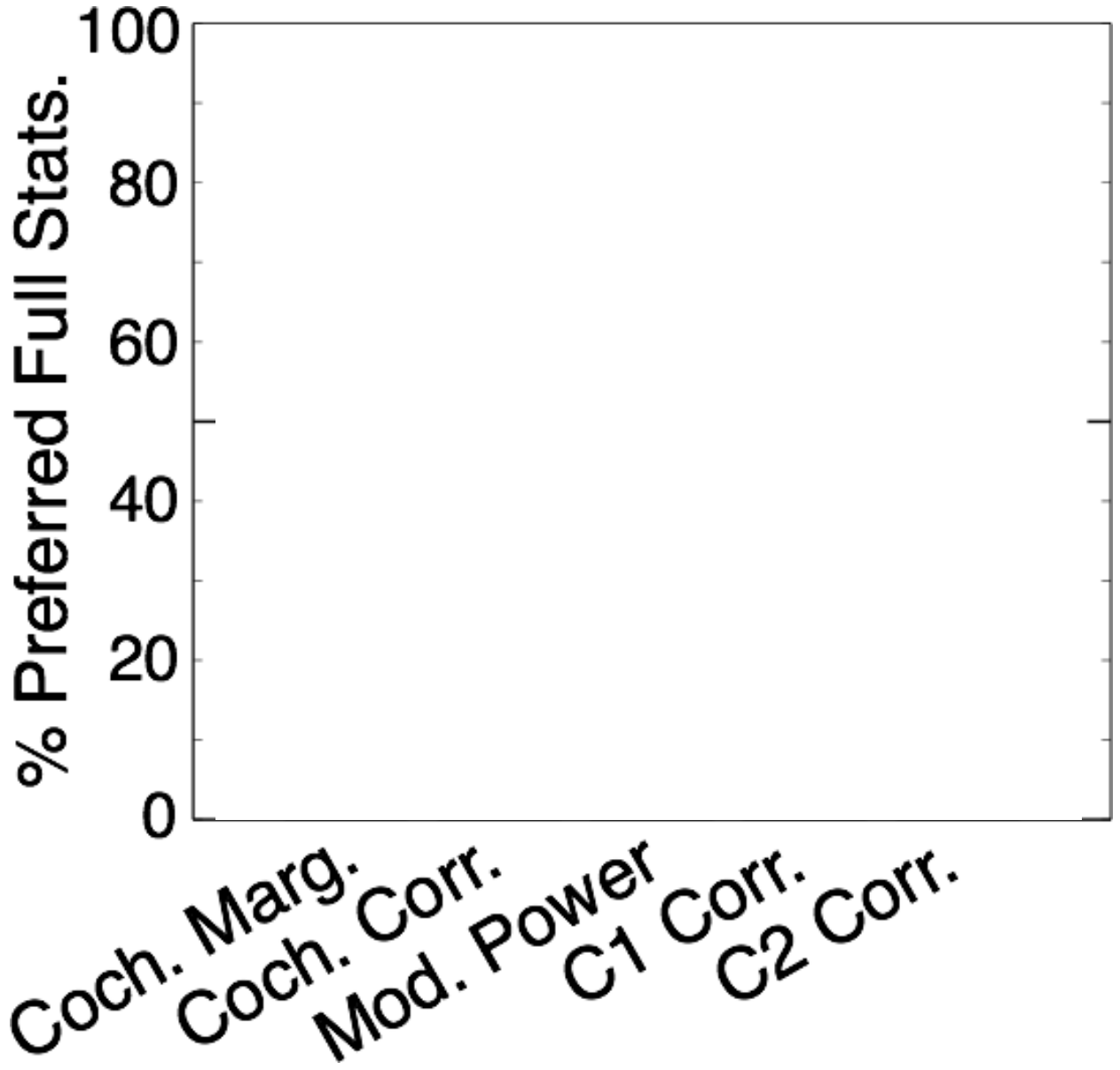


•Task not ideally suited to testing texture recognition.

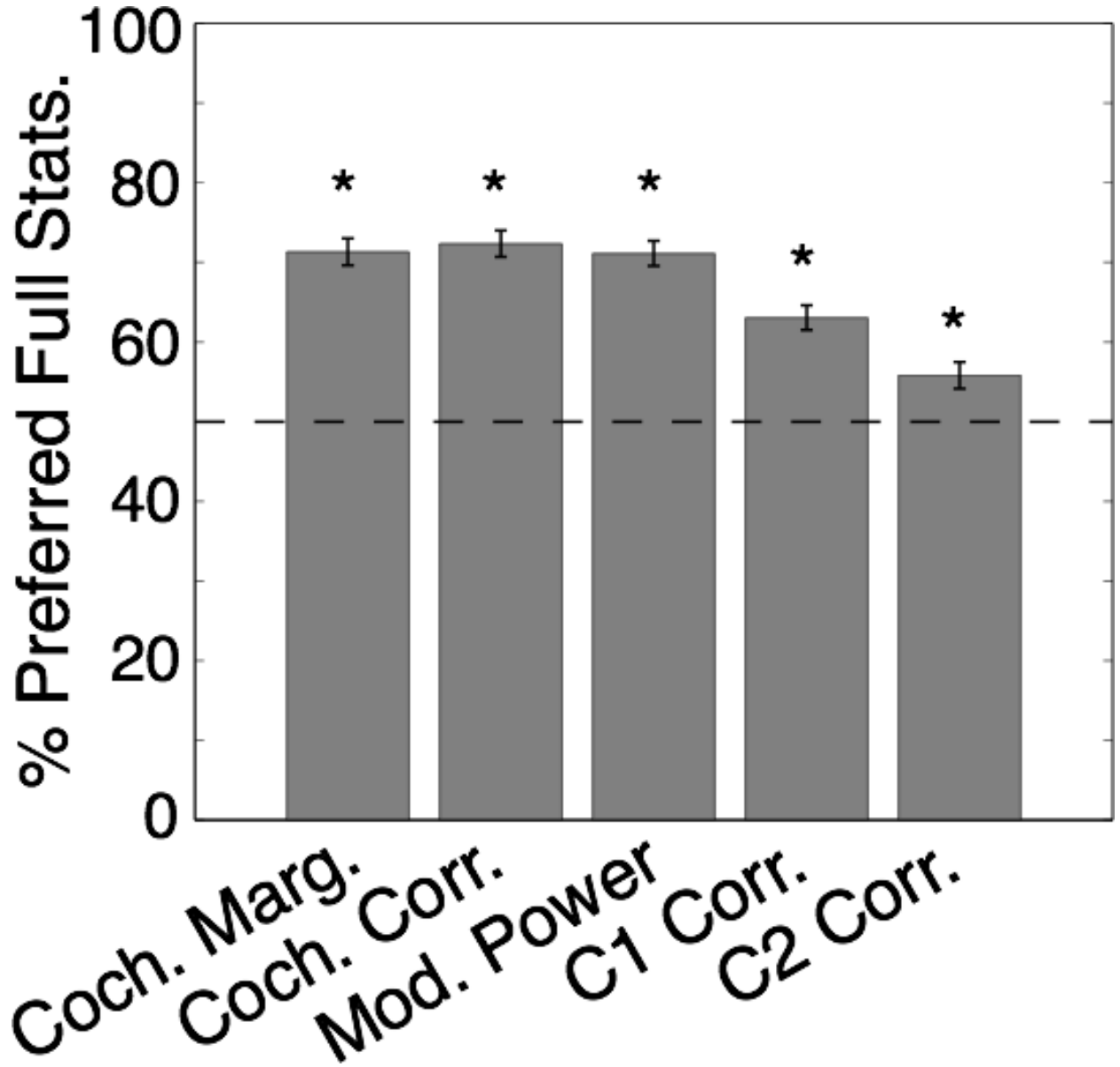
•To use statistics for classifying semantic categories (e.g. urban), probably have to recognize particular textures (e.g. traffic, crowd noise), then link textures to categories.

Experiment: Original - Synth1 - Synth2; one version lacks one statistic

Which is more realistic?

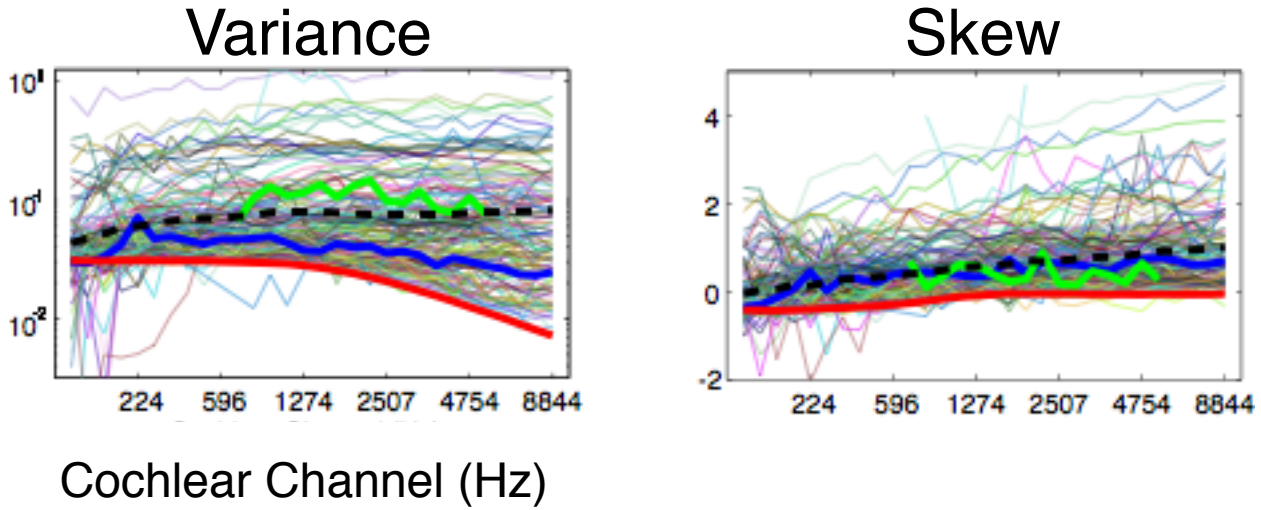


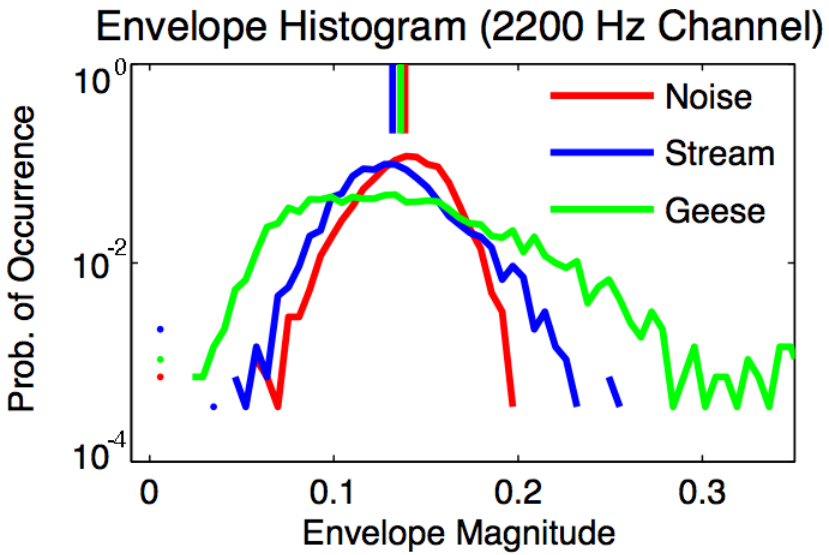
Omitting any class of statistic produces noticeably poorer synthesis:



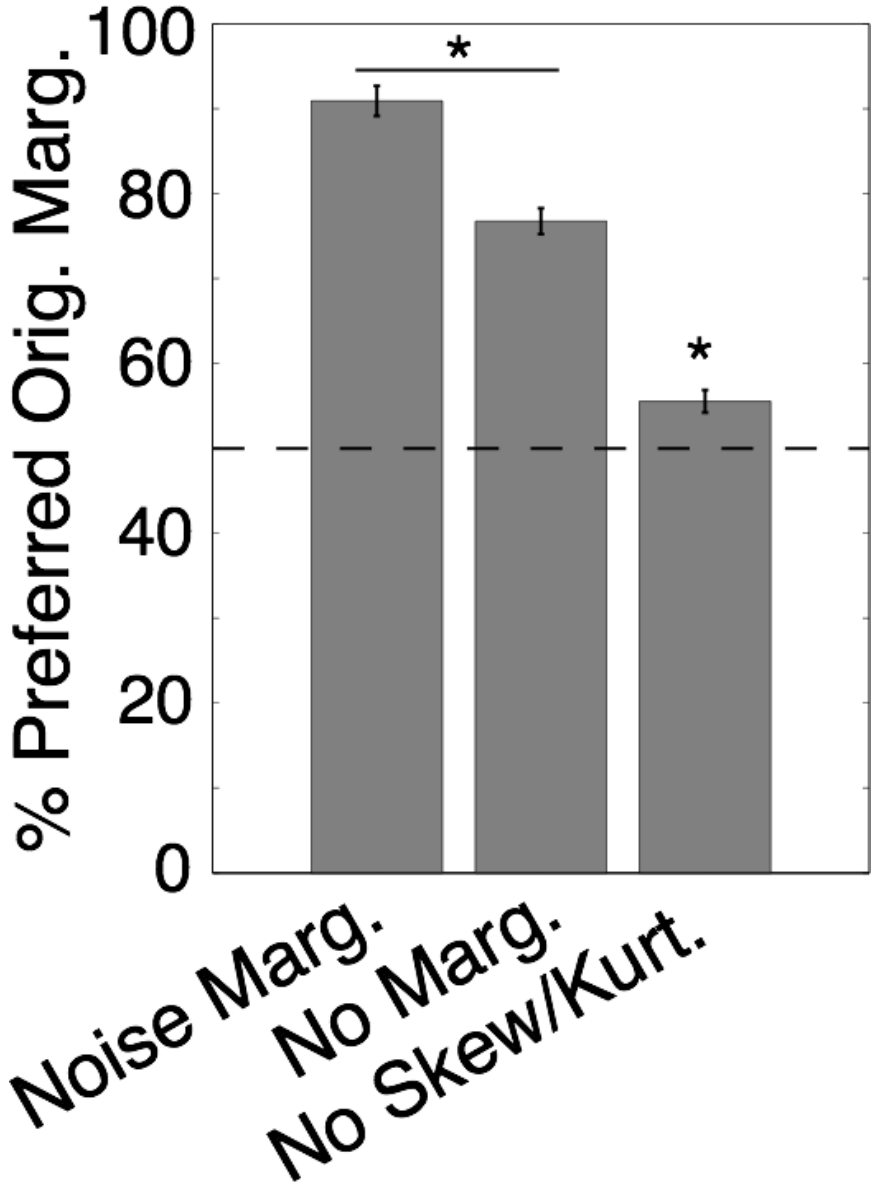
High variance, skew are characteristic of natural sound textures:

Noise marginals seem to be a lower bound on what occurs naturally:

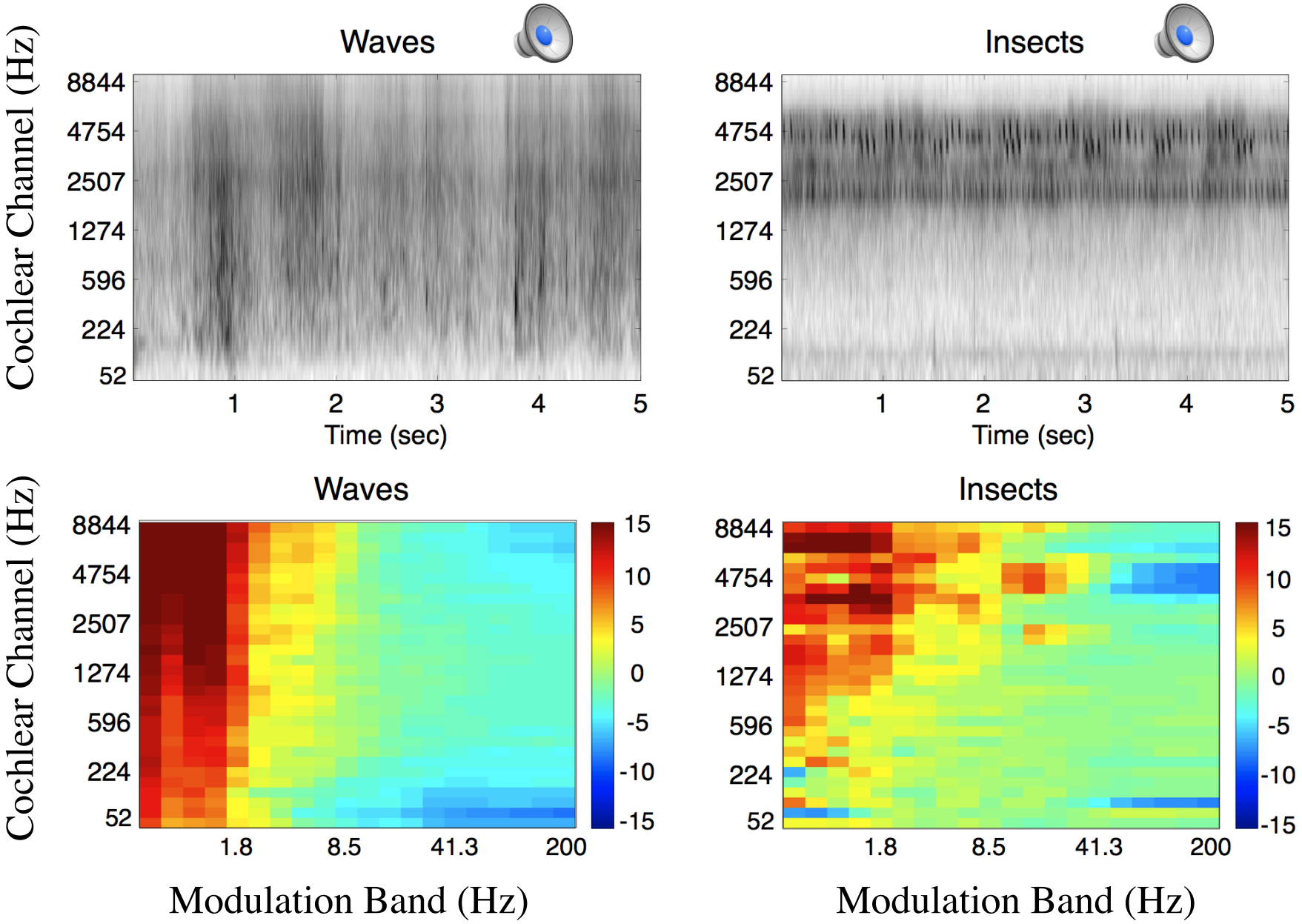




Forcing marginals to values for noise (making signals less sparse) impairs synthesis realism:



Variance (power) in each modulation band conveys temporal structure within channels:



Successful synthesis is nice, but failures are often more informative. They imply the need for new statistics.

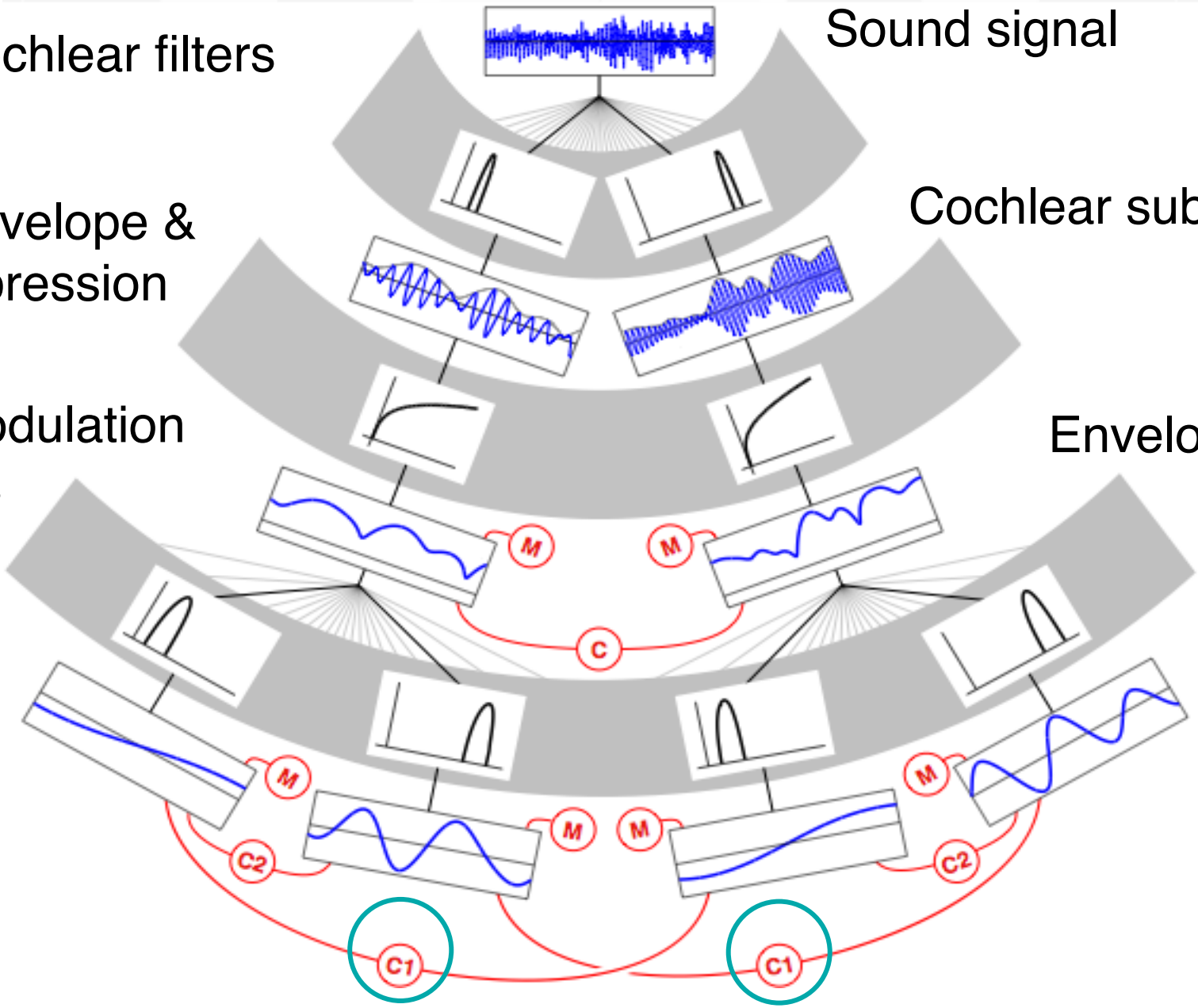
Synthetic waves, some wind examples, sound funny:

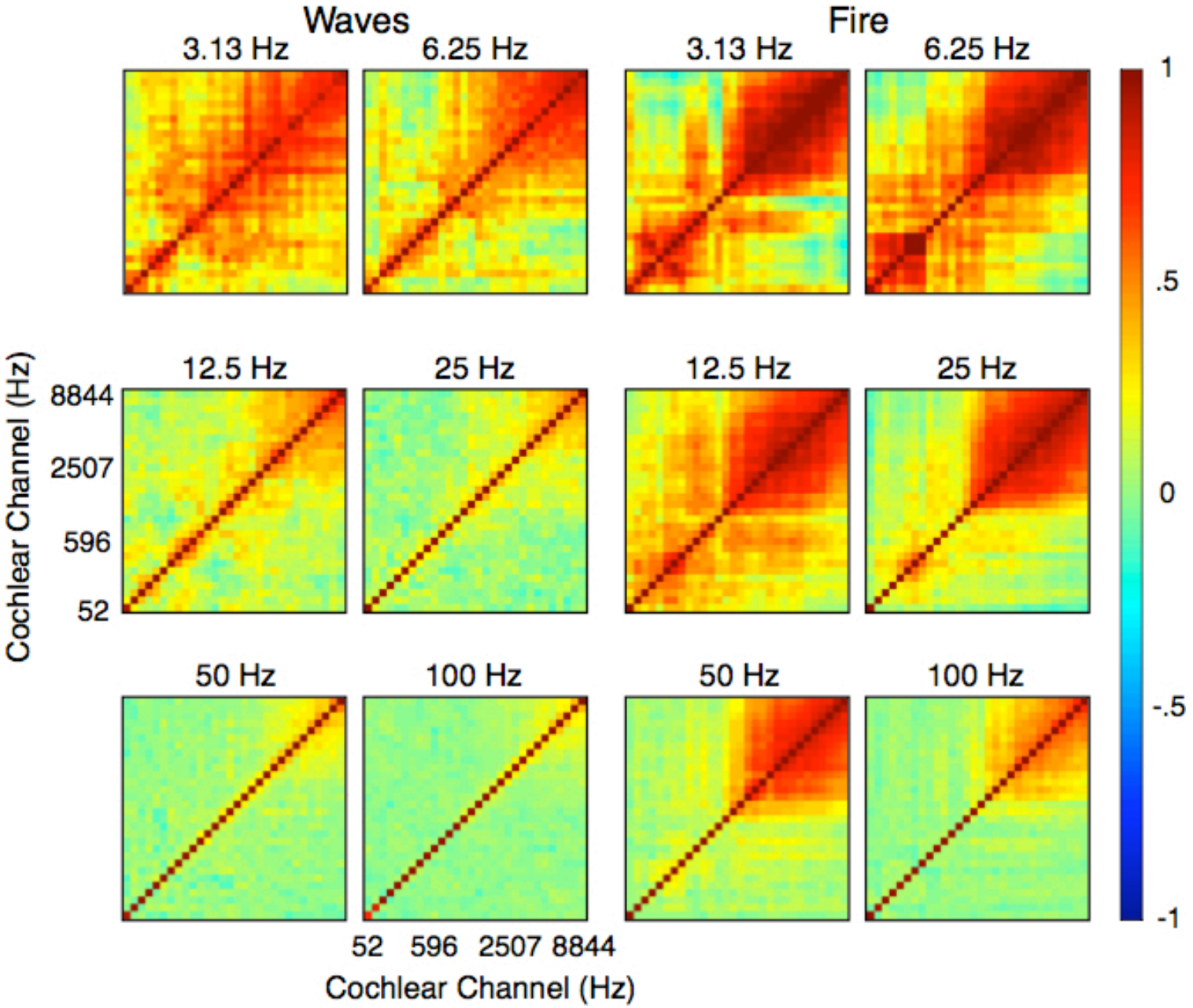
	Real	Marg	CBC+ModPow
Waves			
More waves			
Wind			

1. Cochlear filters

2. Envelope & compression

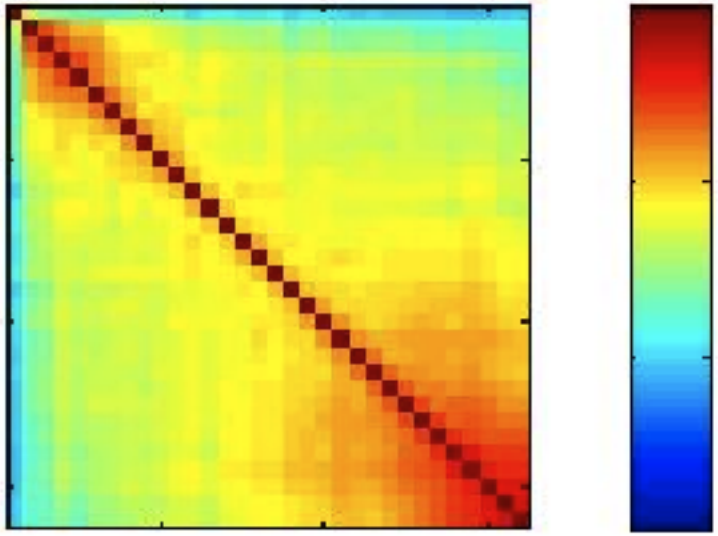
3. Modulation filters



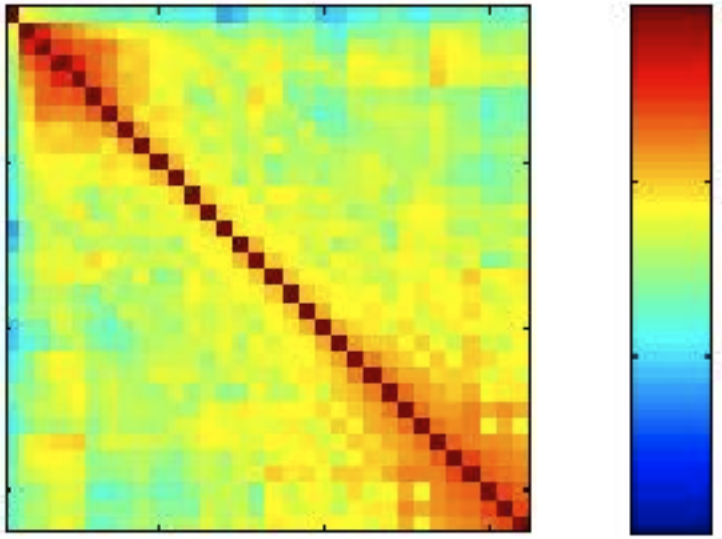

Waves
Fire
3.13 Hz
6.25 Hz
3.13 Hz
6.25 Hz
12.5 Hz
25 Hz
12.5 Hz
25 Hz
50 Hz
100 Hz
50 Hz
100 Hz
Cochlear Channel (Hz)
8844
2507
596
52
52
596
2507
8844
Cochlear Channel (Hz)
1
.5
0
-.5
-1

Cross-band correlation is being imposed correctly, to first order:

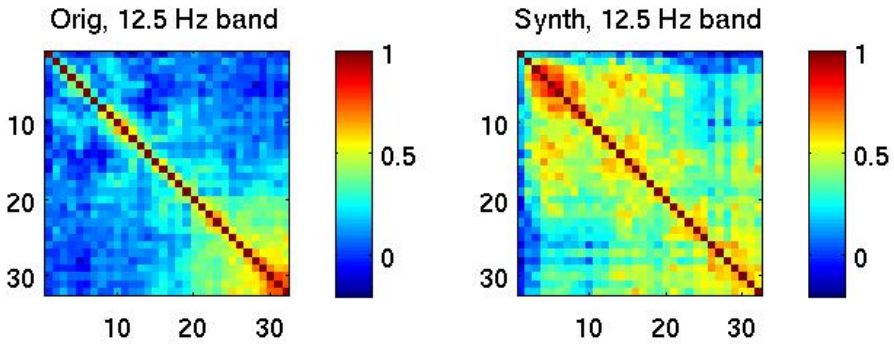
Original Waves



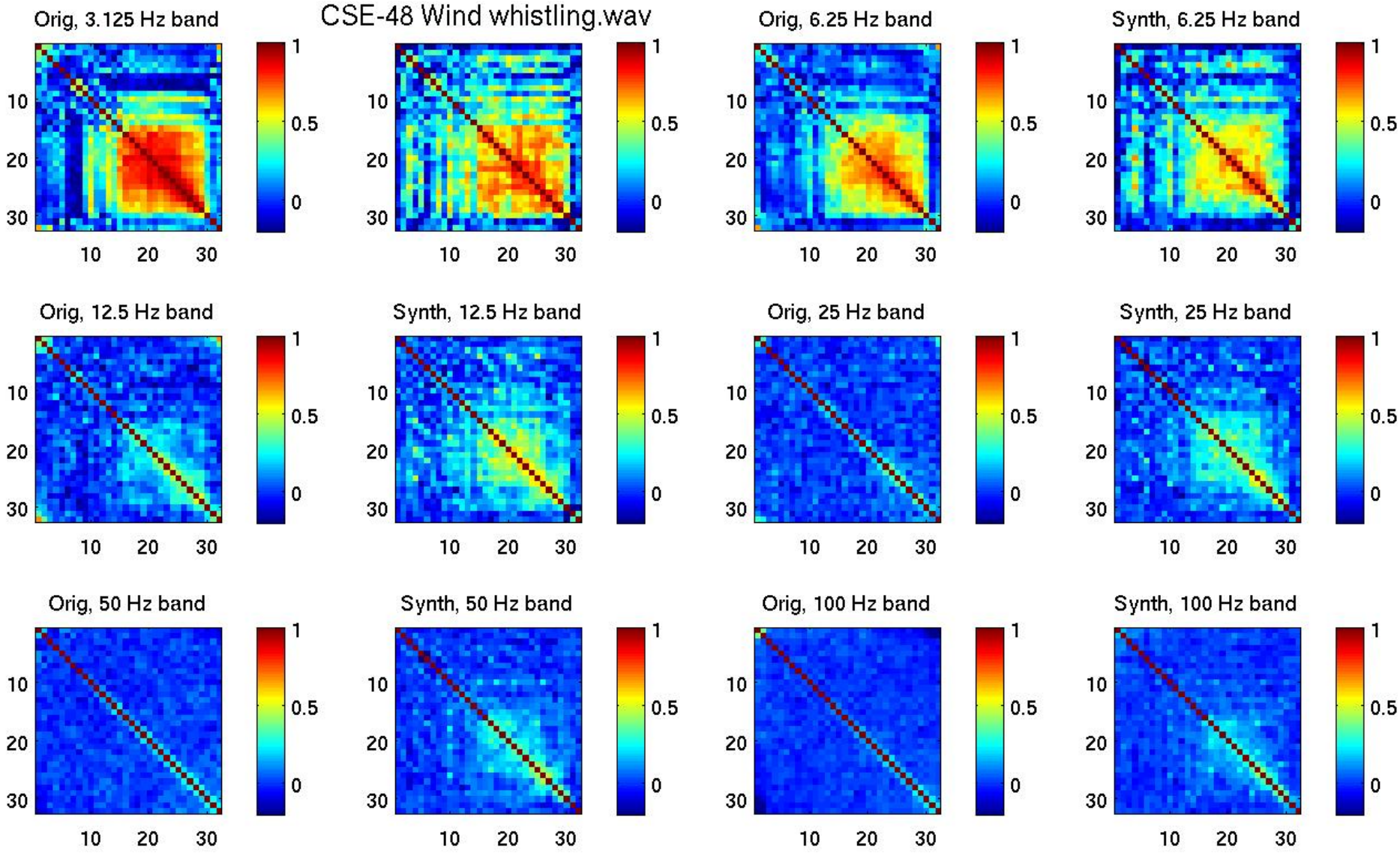
Synthetic Waves



But correlation is applied to wrong modulation frequencies!



Suggests we may be sensitive to correlation at different modulation frequencies…



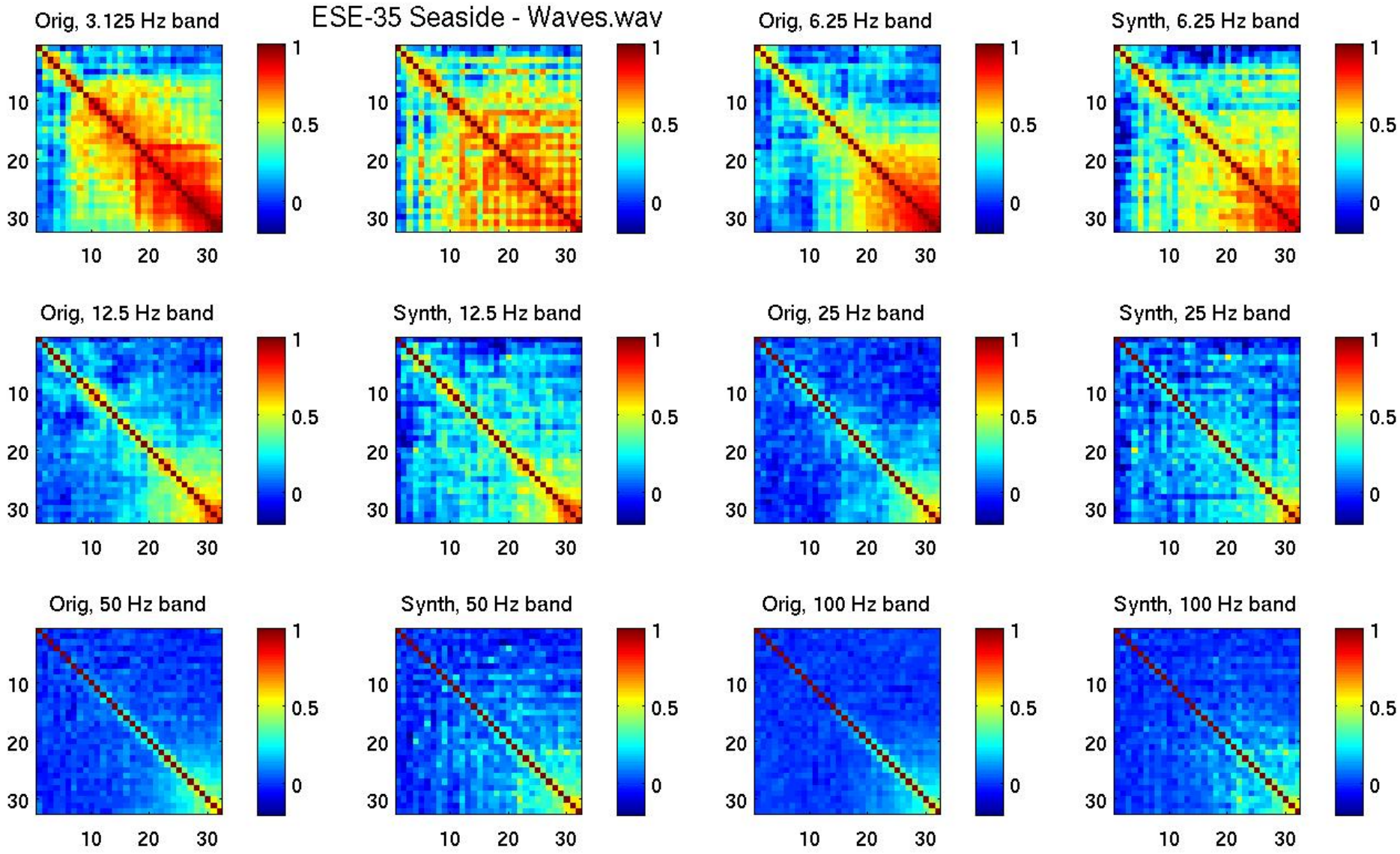
Imposing cross-band correlation between modulation filters, in addition to cochlear filters, is a natural extension.

1-2 offsets is enough

	Real	Marg	CBC+ModPow	w Mod CBC too
Waves				
More waves				
Wind				

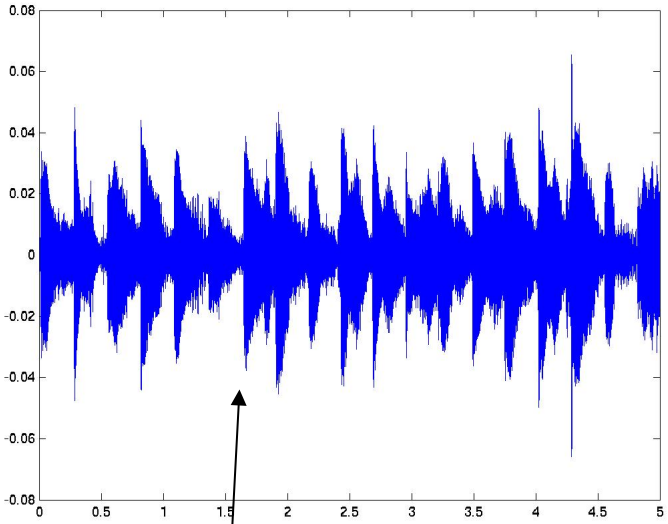
We now have waves!

Not fully imposed, but good enough for big perceptual effect.



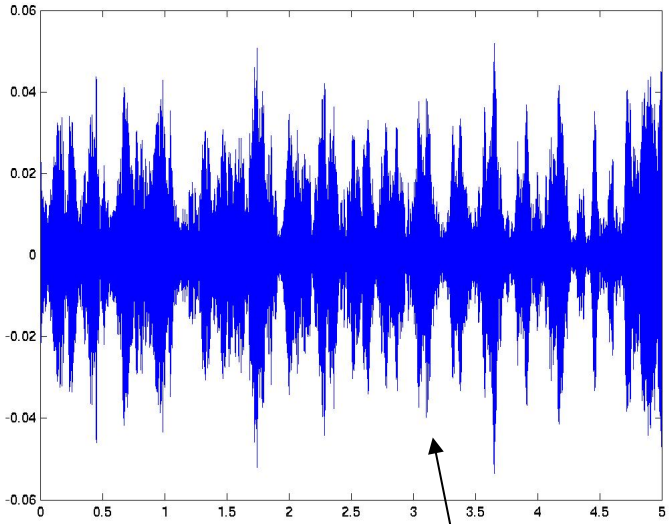
Attack/decay asymmetries are another common mode of failure:

Railroad - original



Note asymmetric envelopes

Railroad - synthetic



Synthetic version lacks asymmetry

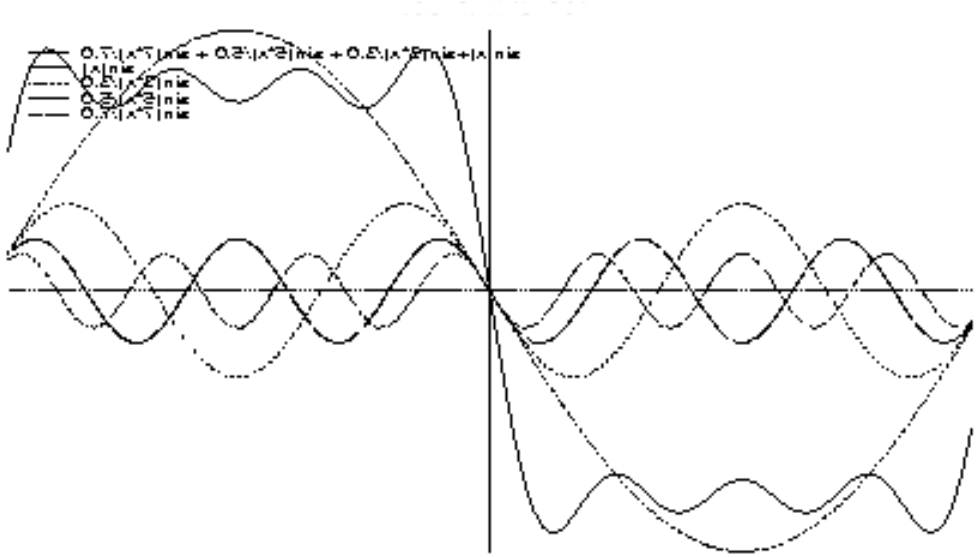
Real sounds often have rapid onsets and slower decays.

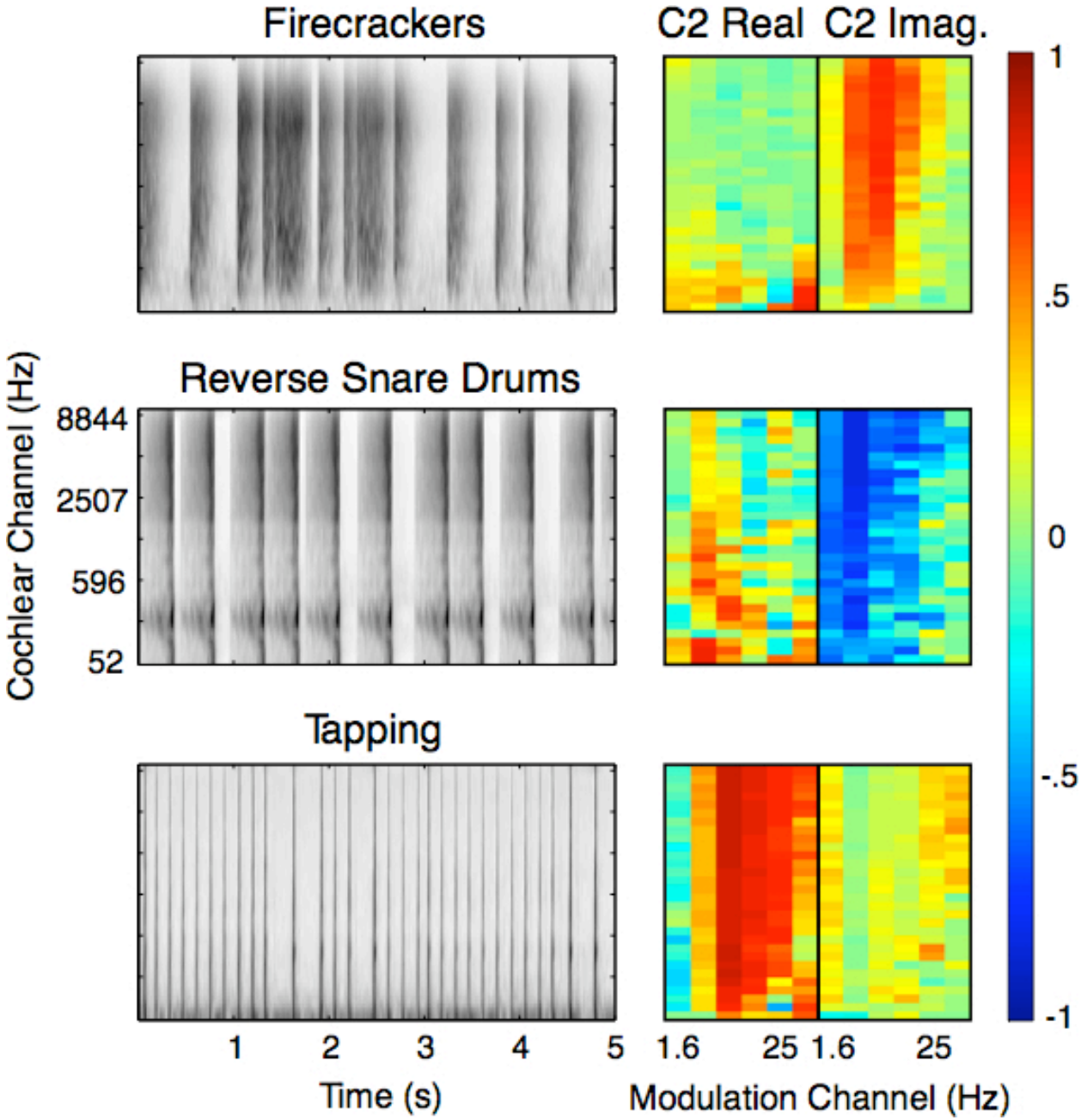
Listeners are sensitive to direction of time, but modulation power is not…

What, then, is the auditory system measuring to capture asymmetry?

Intuition:
relative phase of different modulation frequencies matters.

Step edge is generated by summing bandlimited components in sine phase.




Firecrackers
C2 Real
C2 Imag.
Reverse Snare Drums
Tapping
Cochlear Channel (Hz)
8844
2507
596
52
1
2
3
4
5
Time (s)
1.6
25
1.6
25
Modulation Channel (Hz)
1
.5
0
-.5
-1

Relative phase can be measured between different modulation bands with a form of correlation:

$$\frac{(f_L)^2 \cdot f_H^*}{|f_L|}$$

f_L, f_H are analytic versions of low/ high frequency filter responses

Filters must be octave spaced.

	Real	Without RP	With RP
War			
Typing			
Bubbling Water			
Coins