My research group: Laboratory for Computational Audition

- We study auditory scene analysis and sound recognition
- Contact with speech technology through assistive devices and machine intelligence
- Funded by McDonnell Foundation and NSF
Recent approach in our lab: train deep convolutional neural networks on speech tasks, compare representations to brain

- So far: word recognition, speaker identification in noise
- CNN performs about as well as humans
- Can use CNN as a hypothesis about neural representation
Ability of shallow vs. deep CNN layers to predict brain responses provides insights into computational complexity:

Primary auditory cortex

Variance explained

Speech-selective cortex

CNN layer

Shallow → deep

Legend:
- Auditory CNN
- Arch. CNN
- Acoustic features
- Spectrum
- Temp. mod.
- Spec. mod.
- Spetemp. mod.
- All feats
Using speech analysis/synthesis to manipulate grouping cues:

- **STRAIGHT** decomposes speech into excitation and filtering.
- Excitation modeled sinusoidally
- Altered to inharmonic, or replaced with noise to simulate whispering:
- Do these manipulations affect ability to segregate speech?

*joint work with Kawahara & Ellis*
Task: “WORD” or “WORD” + “WORD”

Type in all the words you hear.

- Single word recognition similar for all conditions.
- For word pairs, recognition worse for inharmonic than harmonic speech, suggestive of effect on segregation.
- But much larger effect of whispering.
- Potentially suggestive of importance of sparsity.
Reverberation profoundly distorts sound signals: \( s(t) \ast f(t) = r(t) \)

Problem for machine speech recognition:

Reverberation is also a challenge for hearing-impaired listeners.
Characterizing the distribution of real-world reverberation

What is the empirical distribution of environmental impulse responses?

**IR Measurement**

- Broadcast fixed source signal
- Record resulting reverberant signal
- From this, infer environmental IR

**IR Survey**

- 24 text messages/day
- Phone returns GPS coordinates
- Participants reply to text with photo, address
Everyday impulse responses are pretty stereotyped

- Exponential decay
- Faster at high frequencies
- Exaggerated asymmetry in large rooms
- Suggests prior for dereverberation…

271 IRs from 301 surveyed locations

1st quartile
4th quartile

Survey
KEMAR HATS
8m

Mean subband RT60 (s)
Frequency asymmetry (skew of subband RT60)
Challenges to Impacting Technology

• Lack of large high-quality labeled data sets in some domains
  • Emotional speech
  • Environmental sounds

• Cultural divides between neuroscience and engineering
  • Different meetings, departments, jargon, funders
  • Possibly getting worse?
  • Workshops help, particularly if students have access