Data Neutering

Henry Kautz

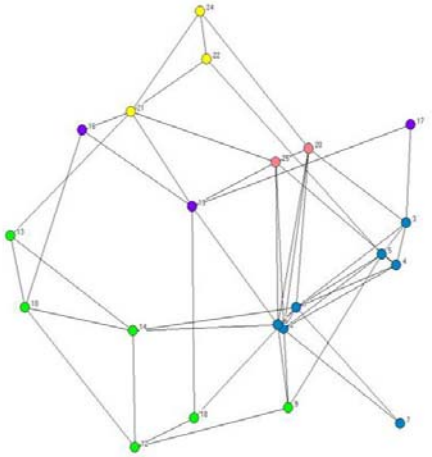
UW → {University of Rochester,
Kodak Laboratories}

Behavioral Research & Data Privacy

- It is now possible to automatically gather lots of interesting data on human behavior for important applications
  - Sociology, Assistive Technology, Epidemiology, Urban Planning, *etc.*
- We try to keep the data private… but complete security is a practical impossibility
- Therefore: whenever possible, replace the raw data with a "less invasive" (neutered) version

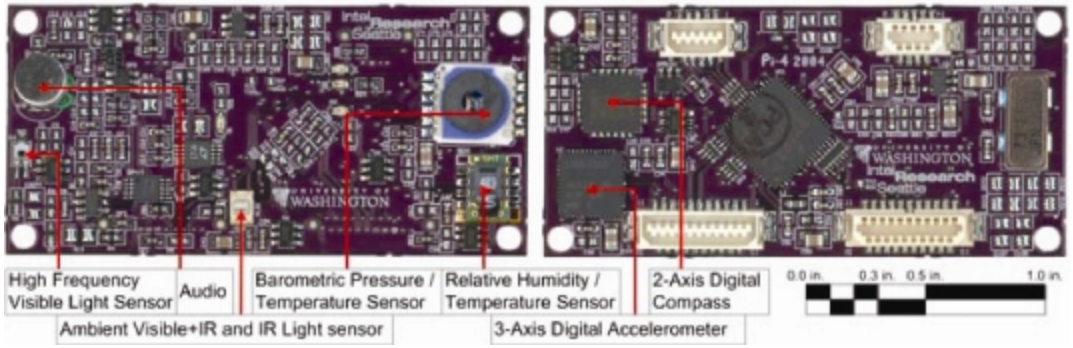
Example: Human Social Dynamics

- Much recent work on social network models of human interaction
- UW Human Social Dynamics Projects: collect data on *face-to-face conversational interaction* among a group of incoming graduate students over the course of year
  - When & where conversations occur
  - Conversational style – informal, formal, …
  - Tanzeem Choudhury (Intel), Dieter Fox, Henry Kautz(UW CSE), James Kitts (UW Sociology)
- How to gather useful data that would not be harmful if it escaped?



Data Collection

Intel Multi-Modal Sensor Board



Coded Database

Real-time audio feature extraction

audio features

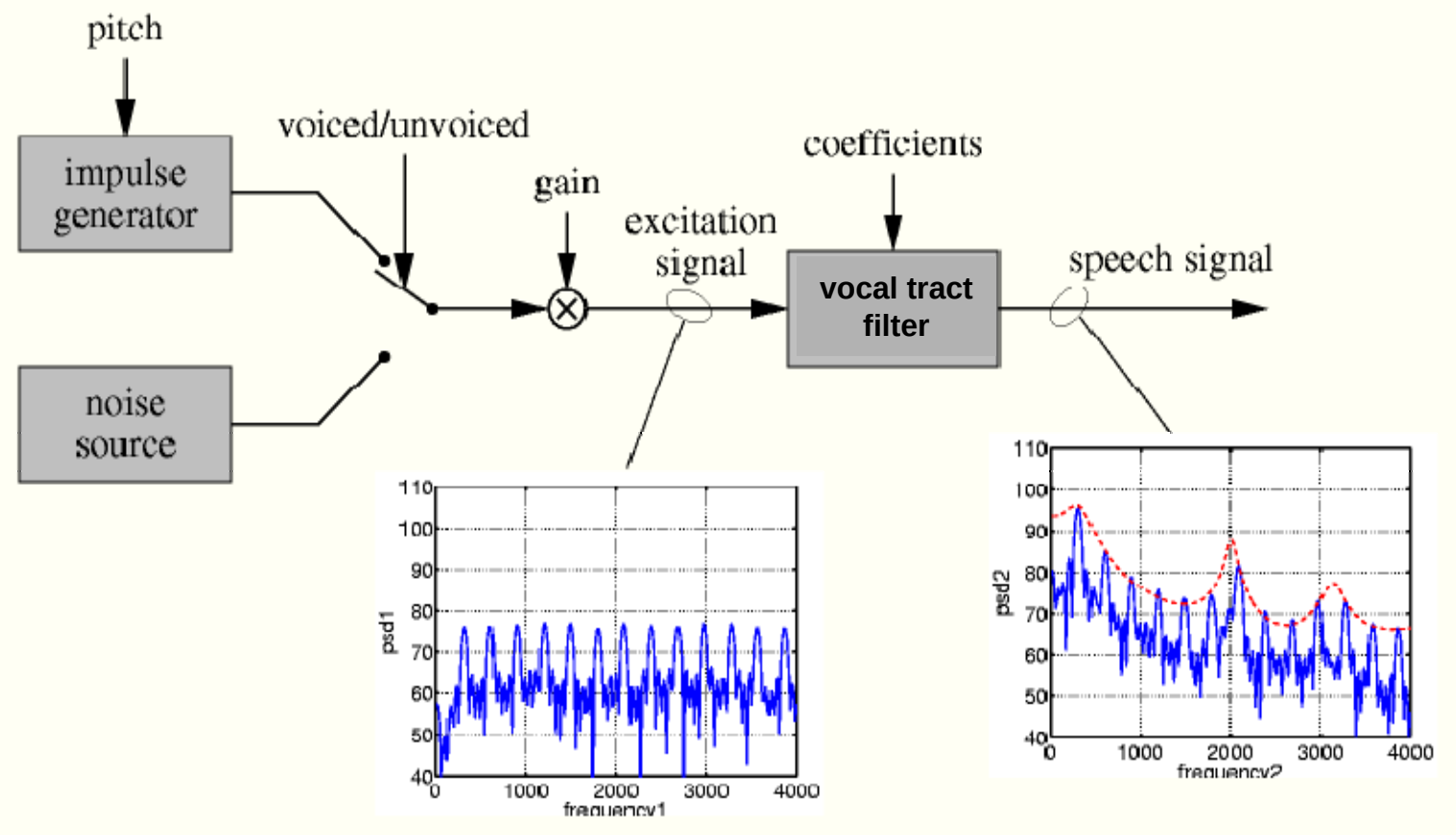
WiFi strength

code identifier

Speech Detection

- From the audio signal, we want to extract features that can be used to determine
  - Speech segments
  - Number of different participants (but not identity of participants)
  - Turn-taking style
  - Rate of conversation (fast versus slow speech)
- But the features must not allow the audio to be reconstructed!
  - UW Human Subjects Division
  - Input from concerned faculty members
  - WA state law on recording conversations

Speech Production (Simplified)



Fundamental frequency encodes voicing & pitch

Formant frequencies encode phonemes

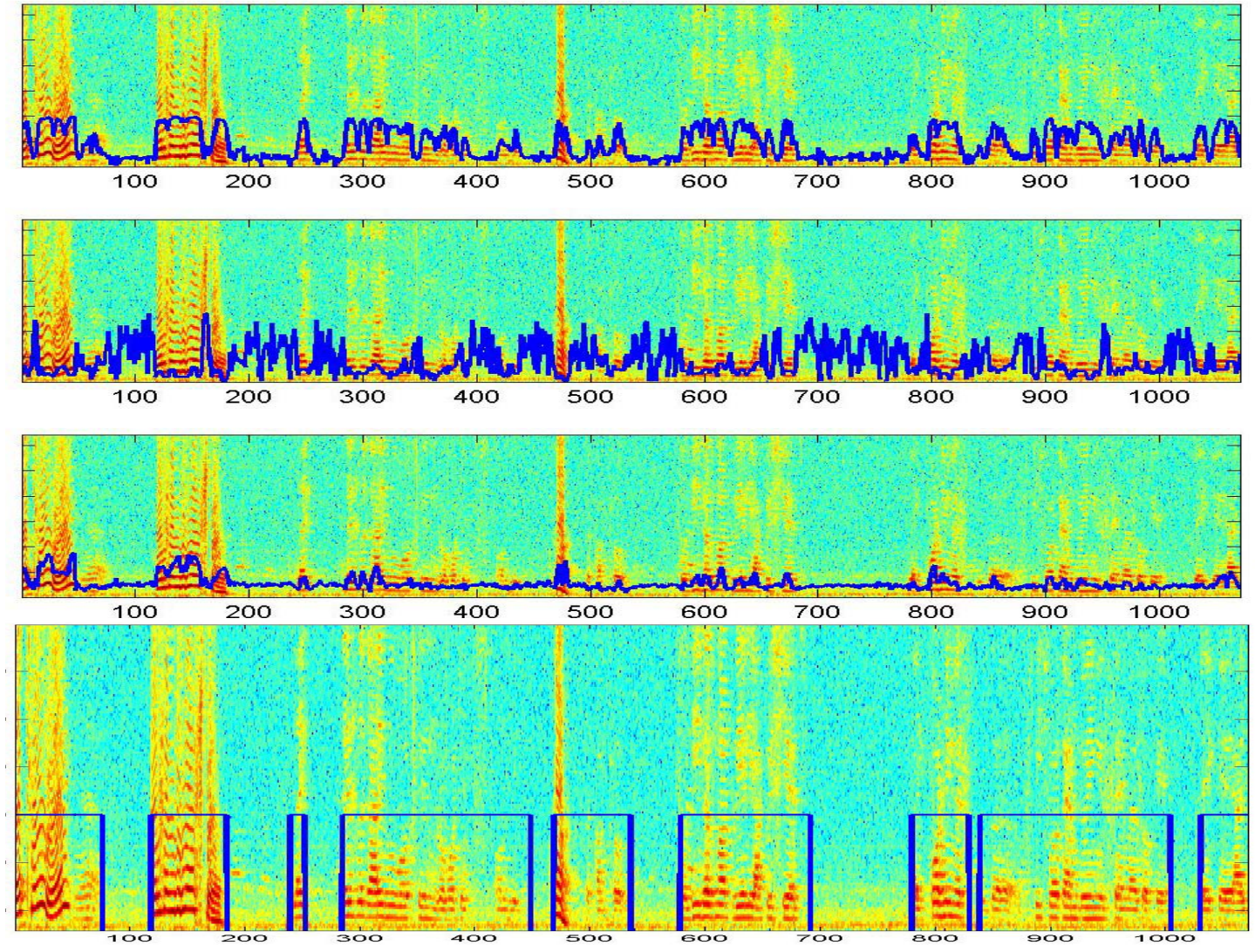
Basic Idea

- On the fly, compute & record audio features that relate to the fundamental – sufficient to determine
  - Speech / non-speech
  - Rate of speech
  - Subject / non-subject as speaker
- Can easily compute features on an iPAQ in real time
  - Full *analysis* of features done off-line
- Do not store information from the formants
  - At least 3 formants needed to reconstruct intelligible speech

Speech Features Computed

1. Spectral entropy
2. Relative spectral entropy
3. Total energy
4. Energy below 2kHz (low frequencies)
5. Autocorrelation peak values and number of peaks
6. High order MEL frequency cepstral coefficients

Segmenting Speech Regions



Data Security

- Standard human subjects data privacy procedures used:
  - Database indexed by random user keys
  - Access to data restricted to approved researchers
- But if these measures fail…
  - Data without user key completely harmless
  - Data + user key could reveal only when subjects spoke to each other, but not what they spoke about!
- Can we do better?
- How can we neuter *location* data?