Clap Detection and Discrimination for Rhythm Therapy

Nathan Lesser & Dan Ellis
Laboratory for Recognition and Organization of Speech and Audio
Dept. Electrical Engineering, Columbia University, NY USA
{nathan,dpwe}@ee.columbia.edu

1. "Rhythm Therapy"
2. Clap Range Estimation
3. Experiments
4. Conclusions

1. "Rhythm Therapy"

- Rhythmic clapping may help neural development
  - sensori-motor planning
  - focus and attention
- "Interactive metronome" devices
  - give feedback on synchrony
  - sensor-based
- Classroom deployment?
  - acoustic-based?
  - for multiple simultaneous users??

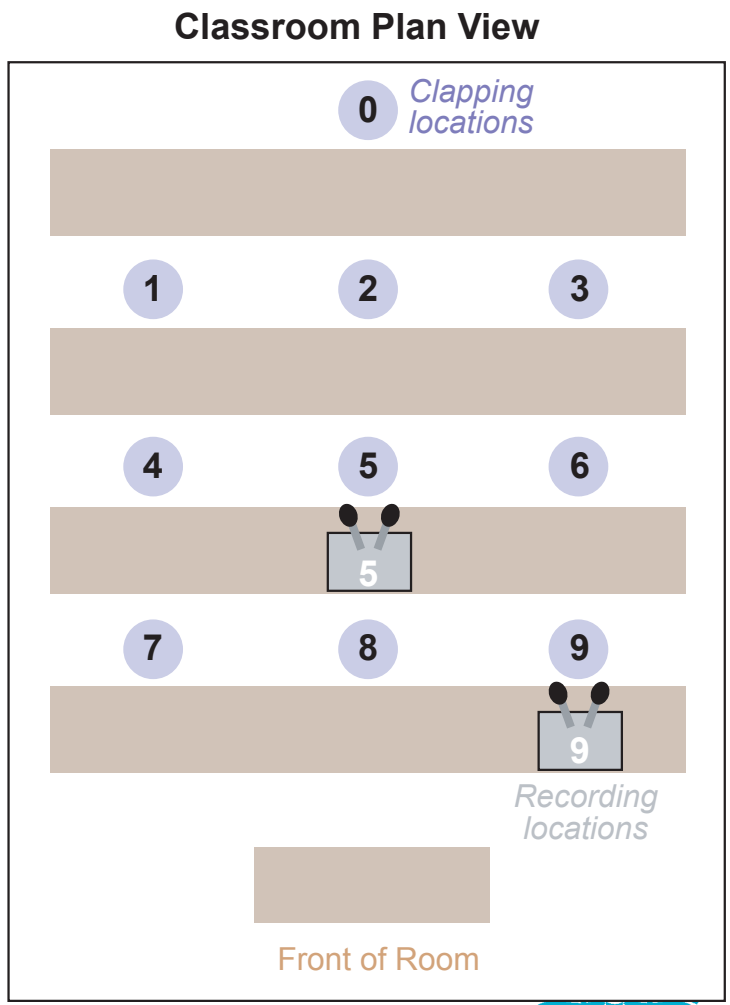
from interactivemetronome.com

Clap Discrimination

- Scenario:
  Many students in same classroom
  each clapping in time to their own laptop
  - students wear headphones (but no sensor)
  - computer hears neighbors
- Goal:
  Discriminate between 'near-field'
  and 'far-field' claps
  - 'near-field' = ~1 meter, on-axis
  - 'far-field' = > 2 meters, maybe off-axis

Data Collection

- Record isolated claps at various locations
  - can superimpose them later...
- Grid of seats:
  - claps from locations 0..9
  - record at locations 5 & 9 only
- Multiple rooms
  - pilot: 1 room,
    2 x 5 claps/location
  - main data: 2 (+2) rooms,
    1 x 50 farfield claps/location
    + 300 nearfield claps/rec.loc.
    = 1500 claps/room

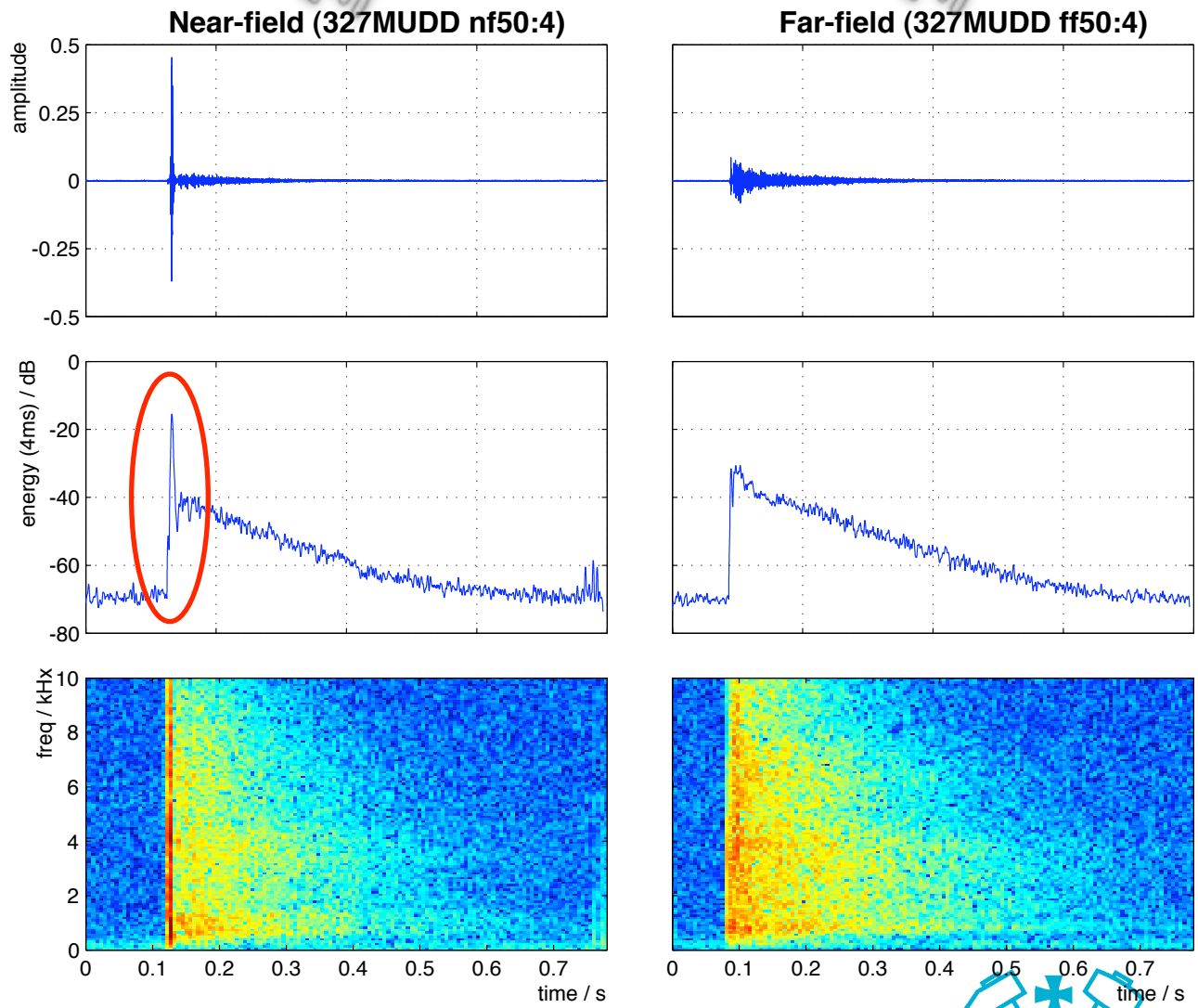


2. Clap Range Estimation

- **Task:**
  Discriminate claps from in front of rig from all others (more distant)
  - main perceptual cue to distance (range): direct-to-reverberant ratio (DRR)
  - how to differentiate direct and reverb?
- **Novel problem: Acoustic range estimation**
  - define correlates of DRR
  - exploit properties of claps (wideband, compact)
  - .. then just feed to classifier

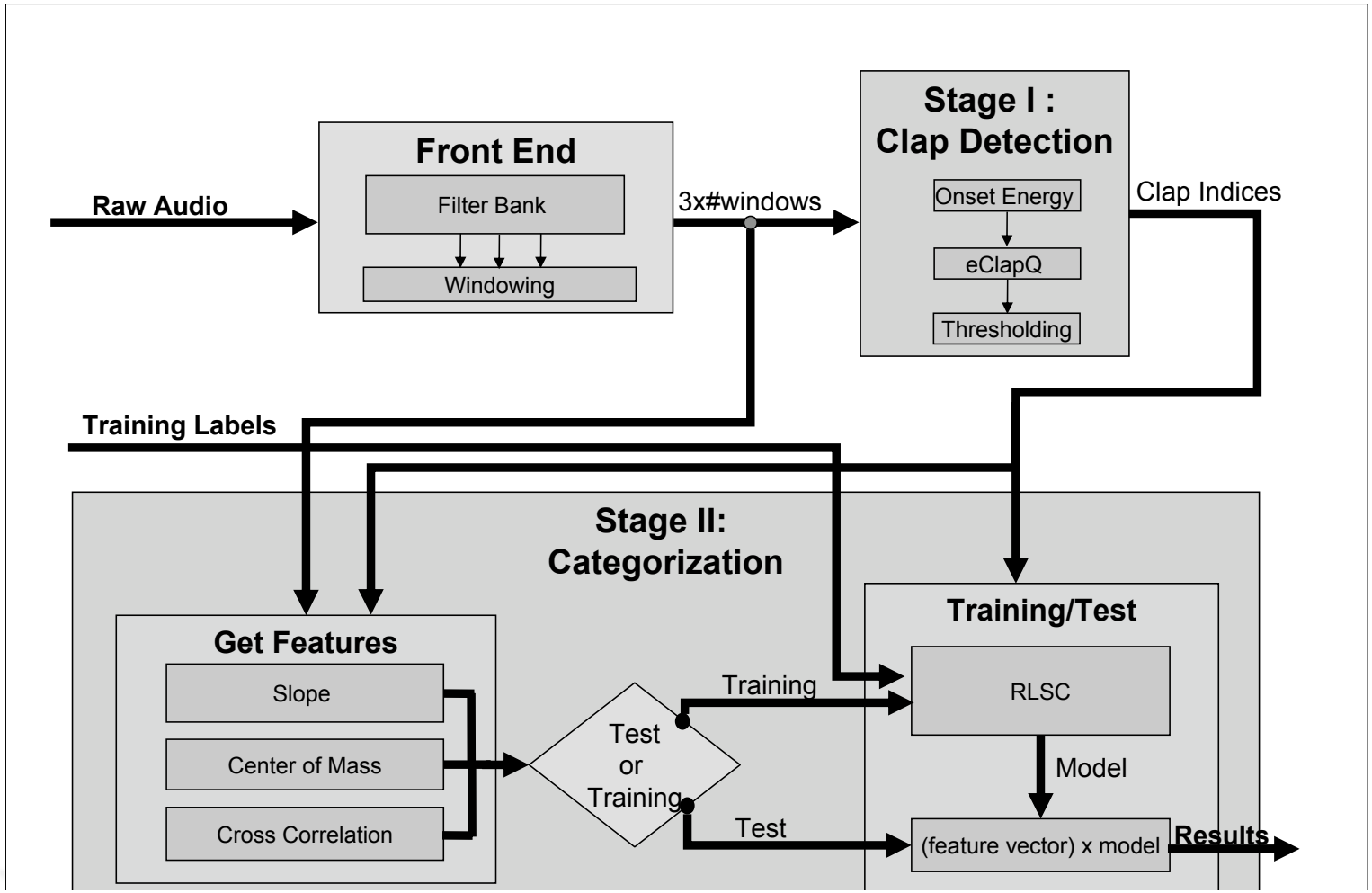
Clap Examples

- Absolute level varies
- Decay slopes ~ same
  - reverberation
  - (RT_{60} ~ 900ms)
- Initial burst for near-field
  - "direct sound"



Processing

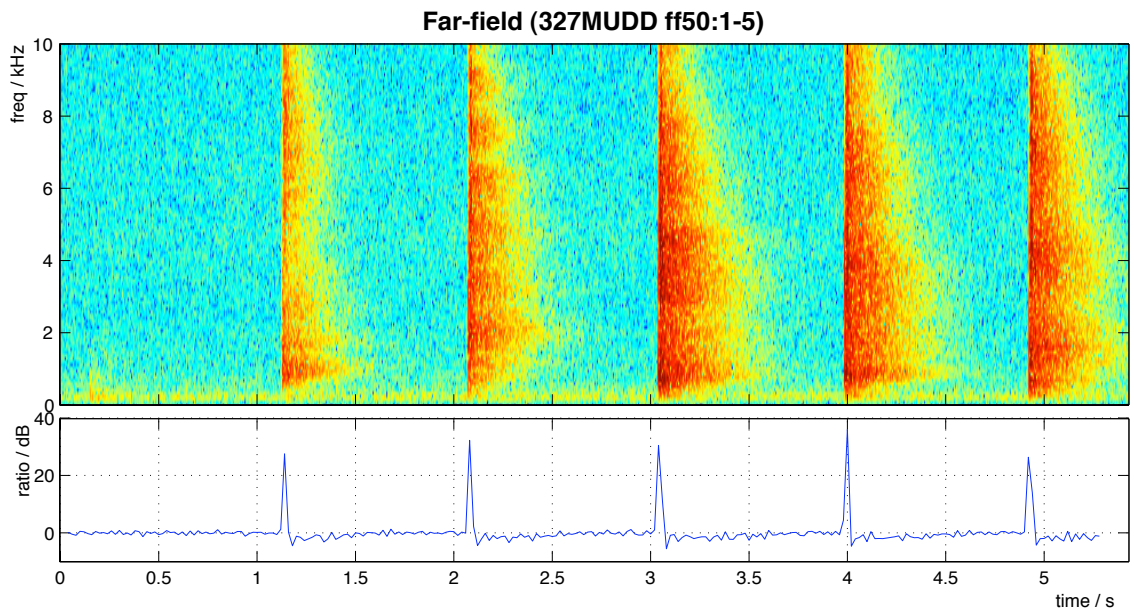
- Detection → Features → Classifier



Clap Detection

- Simple transient detector
  limits feature calculation to 'clap events'

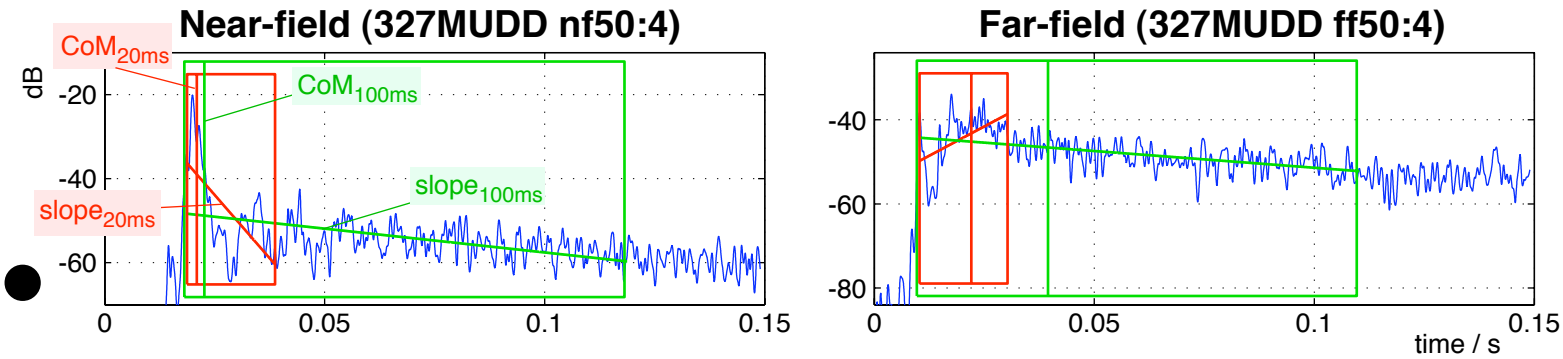
- Adjust threshold on $\Delta(\text{Energy}_{20ms})$ to get desired number of claps
  - known for our data



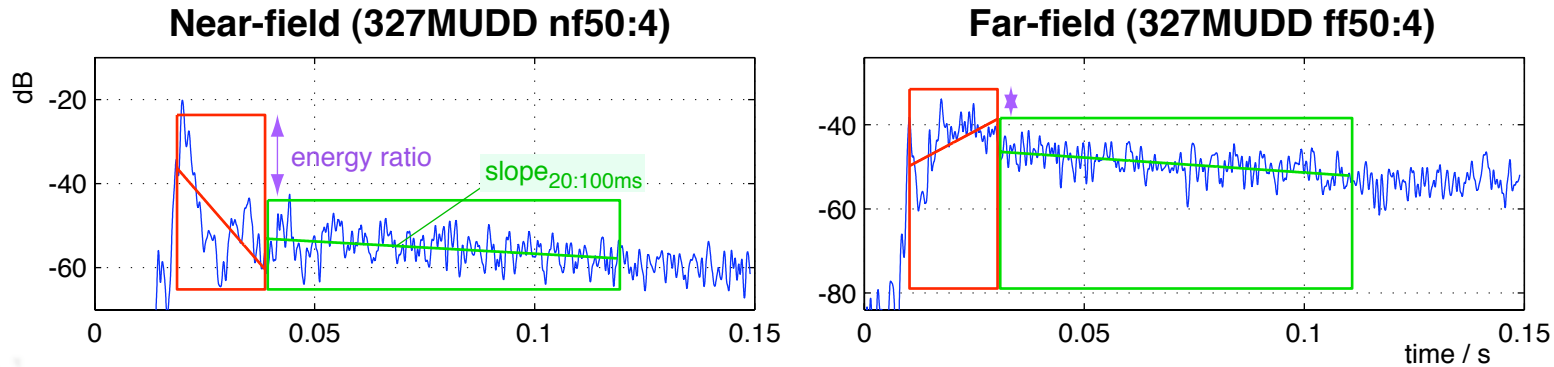
- Backup from maxima to find precise onset
  - Fielded system will need to adapt threshold and reject non-claps

Range Features

- Paper: Ctr. of Mass, Slope in 0..20 , 0..100ms

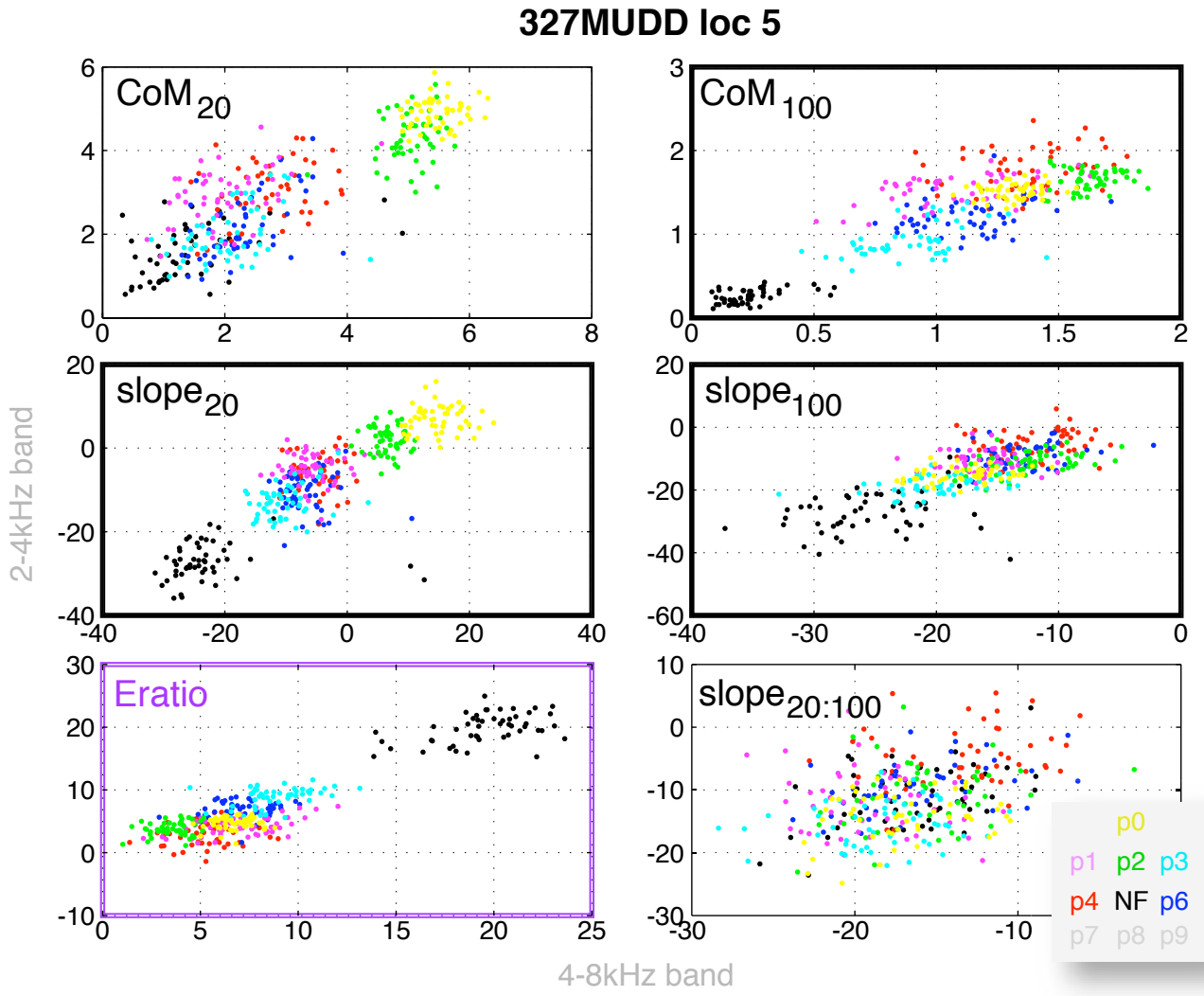


- New: Slope in 0..20ms , 20..100ms
  + Energy Ratio 0..20ms / 20..100ms



Range Feature Behavior

- **Original 4 features**
  - good separation except CoM_{20}
- **New features**
  - Eratio excellent
  - $slope_{20:100}$ useless...
- **Range estimation?**
  - CoM_{20}, $slope_{20}$ show promise



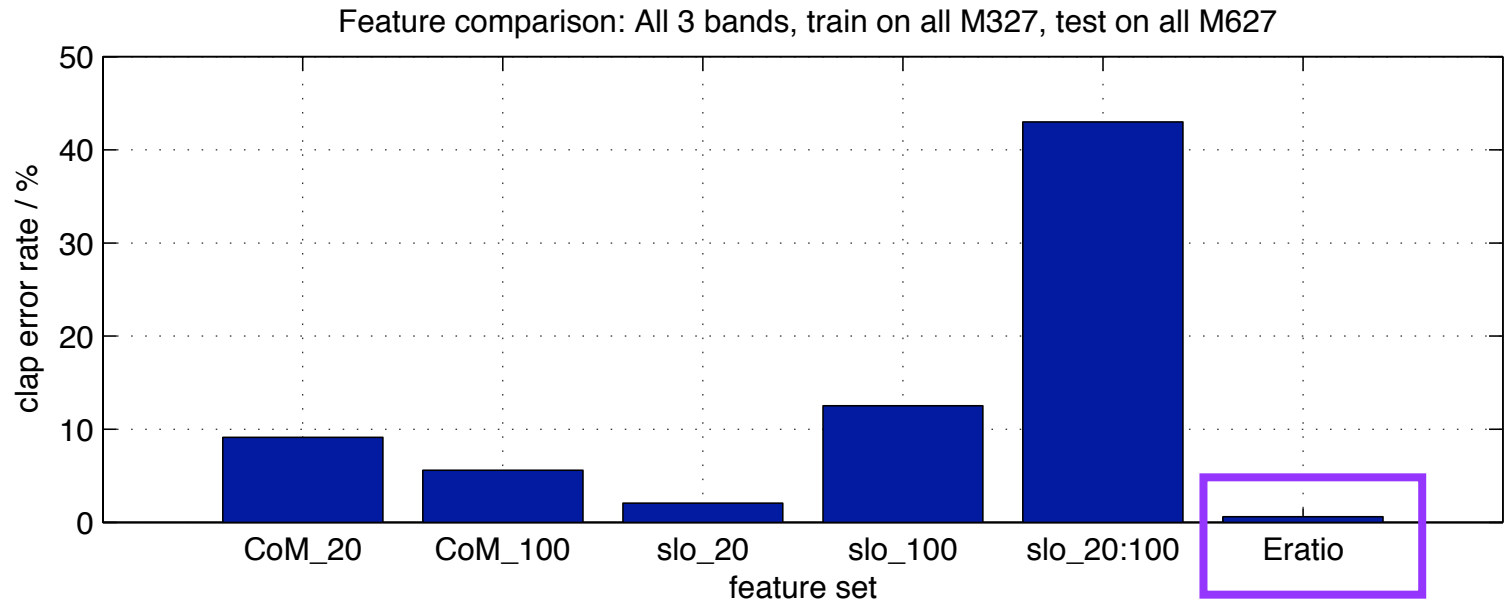
(each plot shows 4-8 kHz band vs. 2-4 kHz band)

3. Experiments

- Build and test actual near/far-field classifier
- Feature experiments
  - quantitative feature comparison
  - best combinations
- Data experiments
  - training data: amount, locations
  - test data: same/different room/location
- Regularized Least-Squares Classifier (RLSC)
  - find a hyperplane in (expanded) feature space
  - ~ simplified Support Vector Machine - no QP

Feature Comparisons

- Train on room 327Mudd; Test on 627Mudd



- Eratio alone (9/1500 = 0.6% errors) beats best combination of rest:
  (CoM_{20}+ CoM_{100}+ slo_{20} = 0.9% errors)

difference of ~0.5% required for signficance

Generalizing Location, Room

- Matrix of 2 rooms x 2 recording locations

CER%		Test			
		M627L5	M627L9	M327L5	M327L9
Train	M627L5	**2.0**	**0.5**	**0.4**	**0.0**
	M627L9	**3.7**	**0.4**	**0.7**	**0.0**
	M327L5	**1.5**	**0.5**	**0.4**	**0.0**
	M327L9	**0.1**	**0.7**	**0.4**	**0.0**

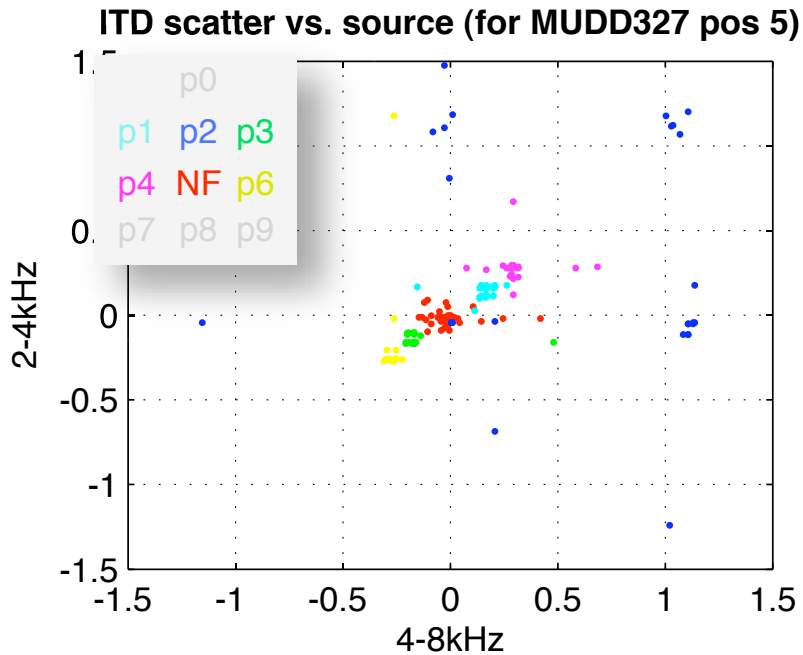
- 627Mudd loc5 is hard data; 327Mudd loc9 is easy!
- Cross-room (shaded) cases generalize better !?
- Plenty of data: 5 claps/loc (20%) just as good

4. Conclusions

- Discriminating isolated near- and far-field claps is feasible (use Eratio 0..20/20..100ms)
- Detection of candidate claps likely to limit accuracy in practice
  - but have 'rhythmic' expectations...
- Applicability to general range estimation?
  - Eratio relies on short-duration direct-sound
  - ..but other sounds have clicks (e.g. speech bursts)
  - CoM_{20}, $slope_{20}$ closer to proportional to range

Azimuth Features

- Cross-correlation of L and R for azimuth:

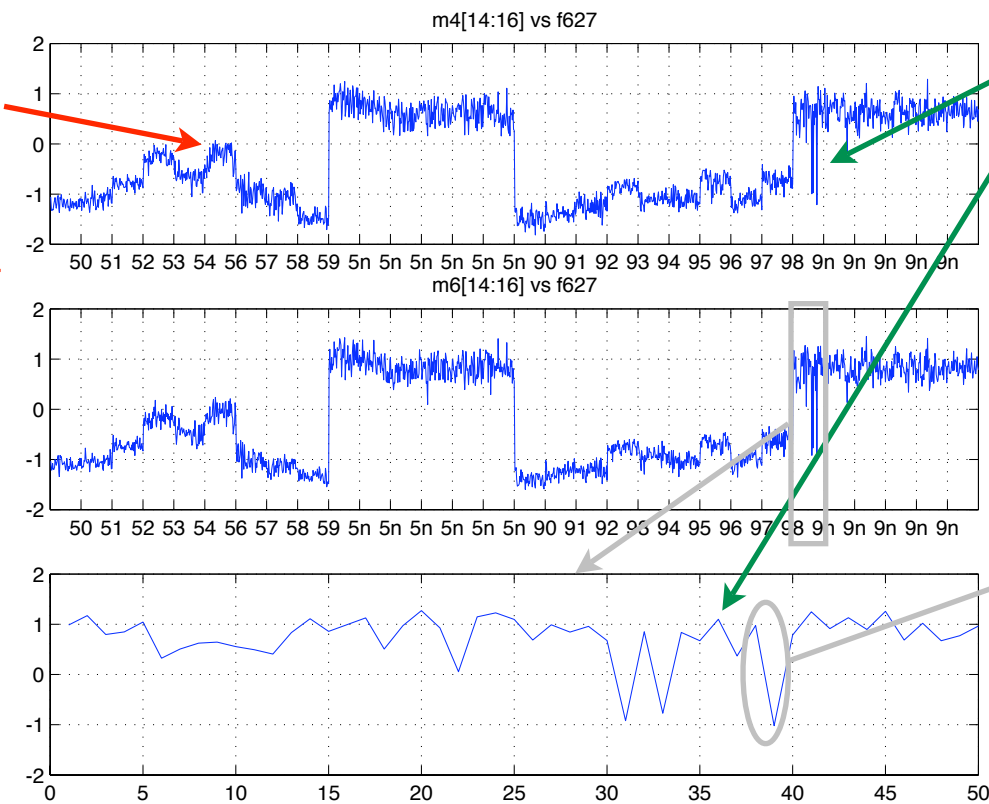


  - nearby locations distinguished - useful
  - distant locations (p2) give random results
  - needs nonlinear feature space expansion!

Error Analysis

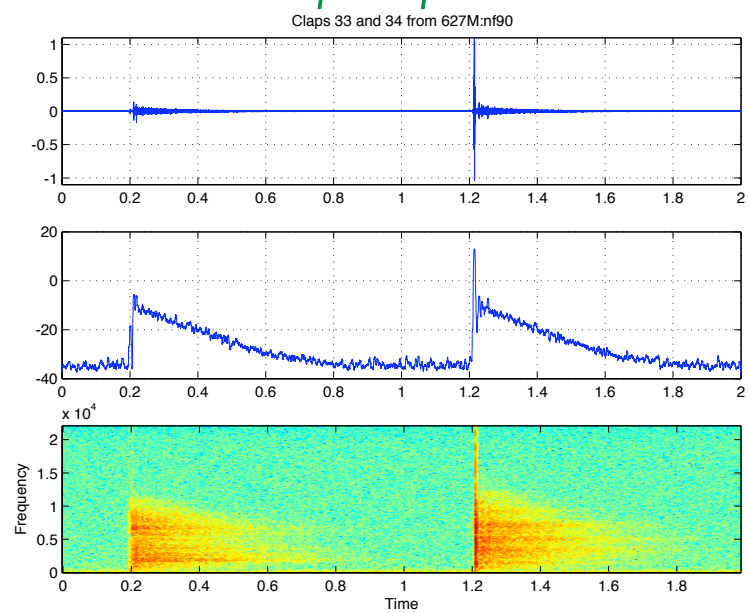
- 627Mudd (record loc 5) is the tough set; look at classifier margins:

false accepts for loc 6 (ambiguous Eratio)



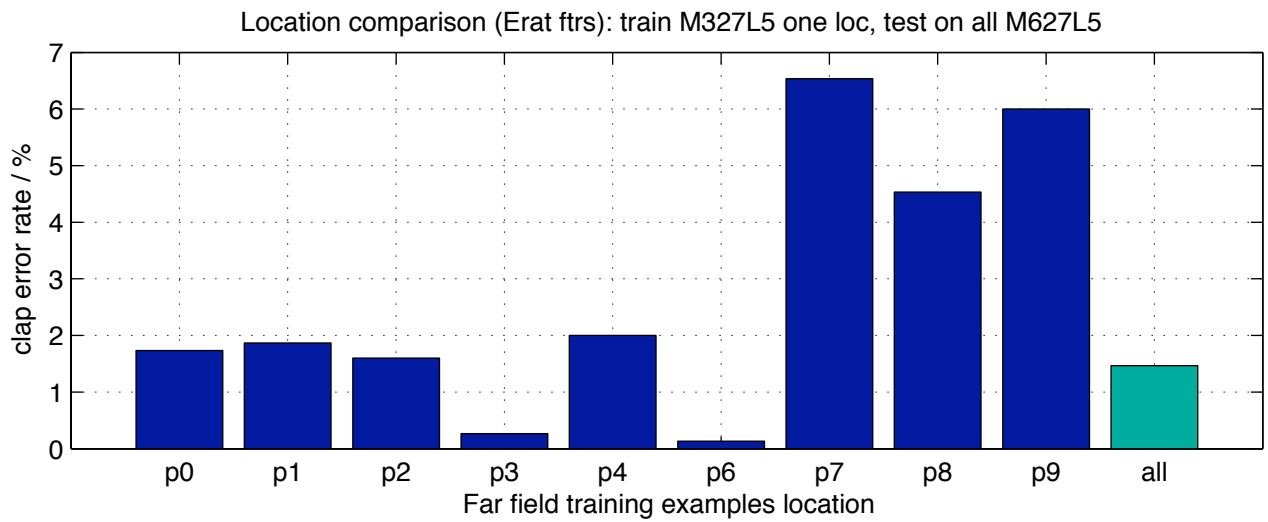
a few solid false rejects...

... really look like far-field???



Usefulness of Each Position

- Train on 50 near-field claps + 50 far-field claps from a single location:



  - all recorded at location 5
  - 'behind' (p7-p9) less useful
  - right-side (p3, p6) most useful !?

p0
p1 p2 p3
p4 p5 p6
p7 p8 p9