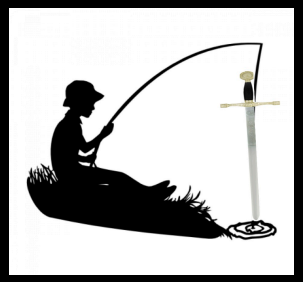
Enhancement of Very Noisy Speech Signals

Dan Ellis & Zhuo Chen
Team Swordfish / ICSI / Columbia

dpwe@ee.columbia.edu http://labrosa.ee.columbia.edu/

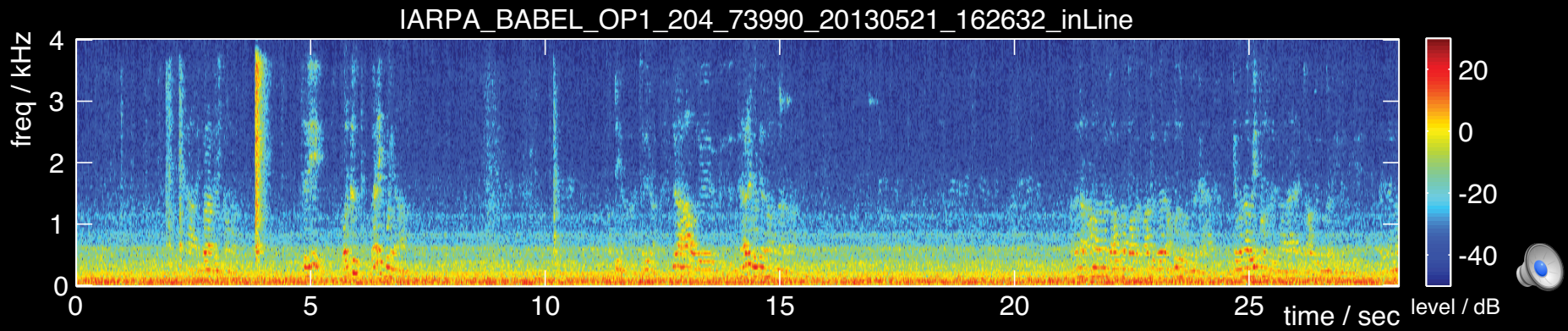
1. Speech Enhancement
2. Flat-Pitch Enhancement
3. Results & Future



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

1. Speech Enhancement

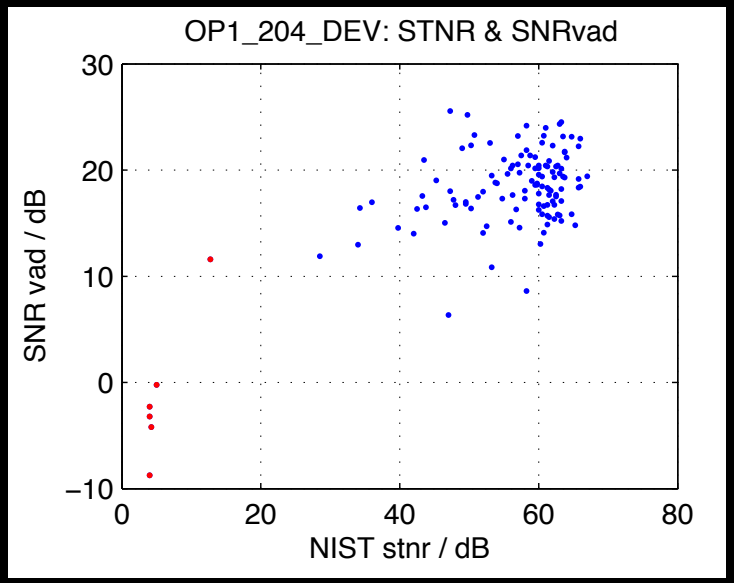
- Noisy speech is a challenge:
  - Surprise channel in surprise language

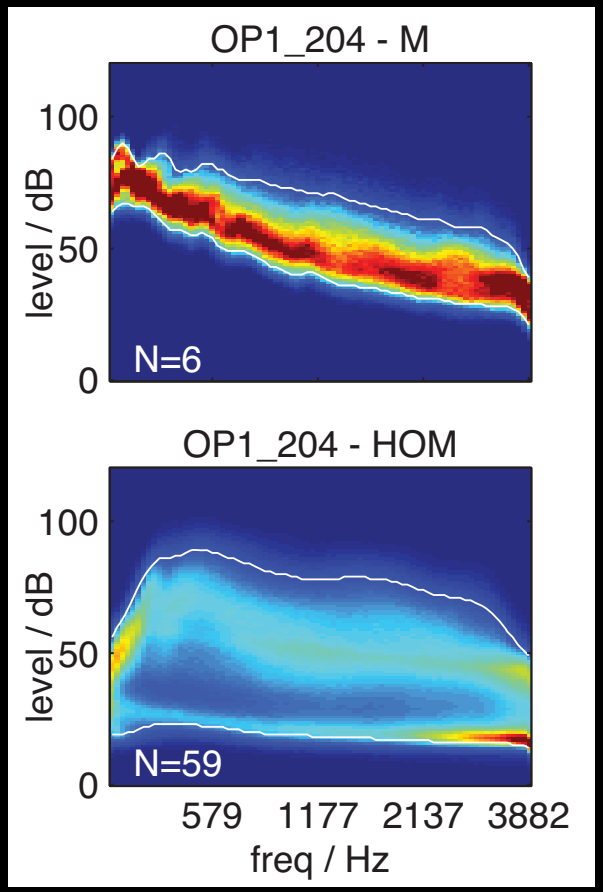


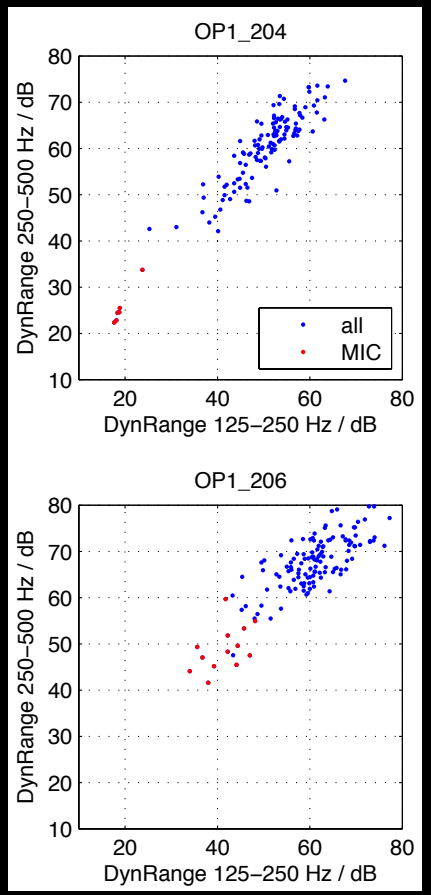
- How to distinguish speech and interference?
  - Energy peaks are speech (spectral subtraction)
  - Energy troughs are noise (Wiener, log-mmse)
  - Speech has a known form (Factorial HMM)
  - Voiced speech is periodic (Pitch-based)

Noisy Channel Detection

- MIC channels are in a different formant
- Even after resampling,
  noisy channels are very distinct:



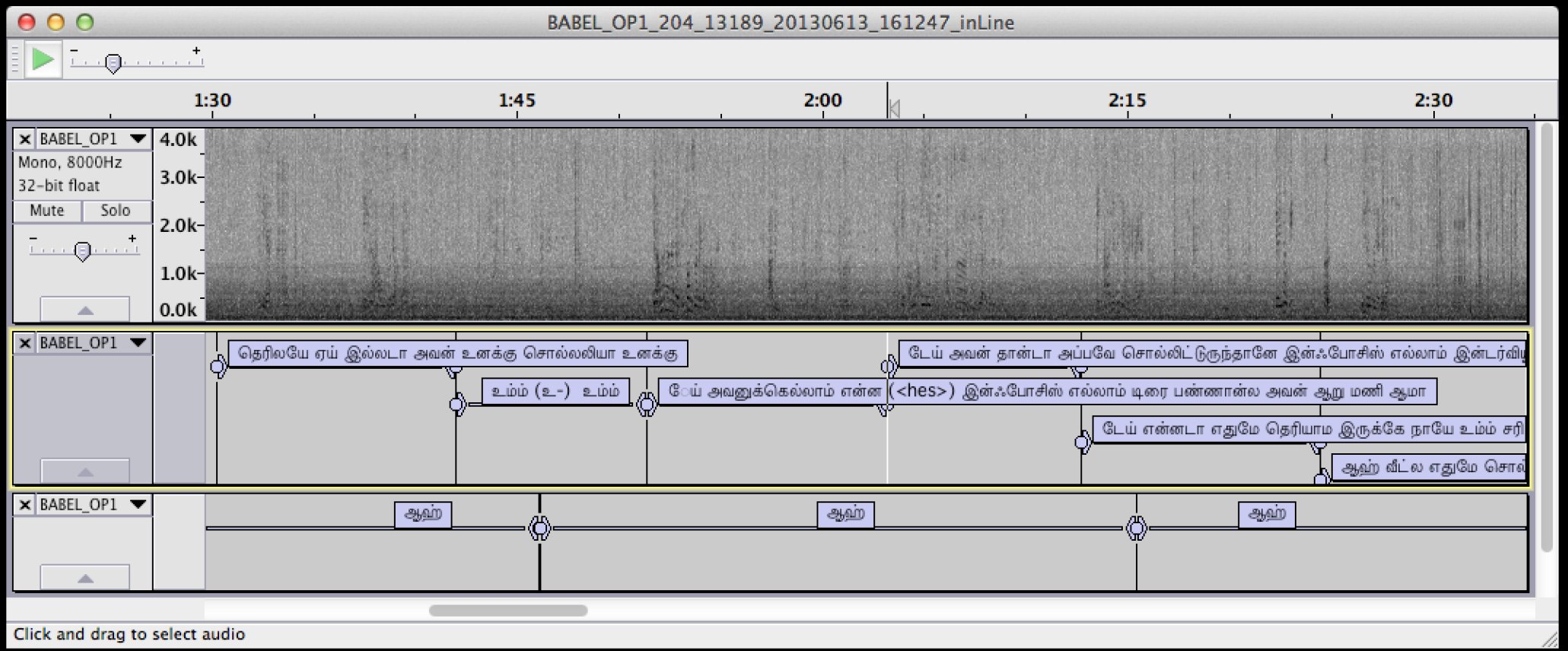




KWS on Noisy Signals

- WER is very poor
  - only nonzero thanks to common word ஆஹ் (“ah”)

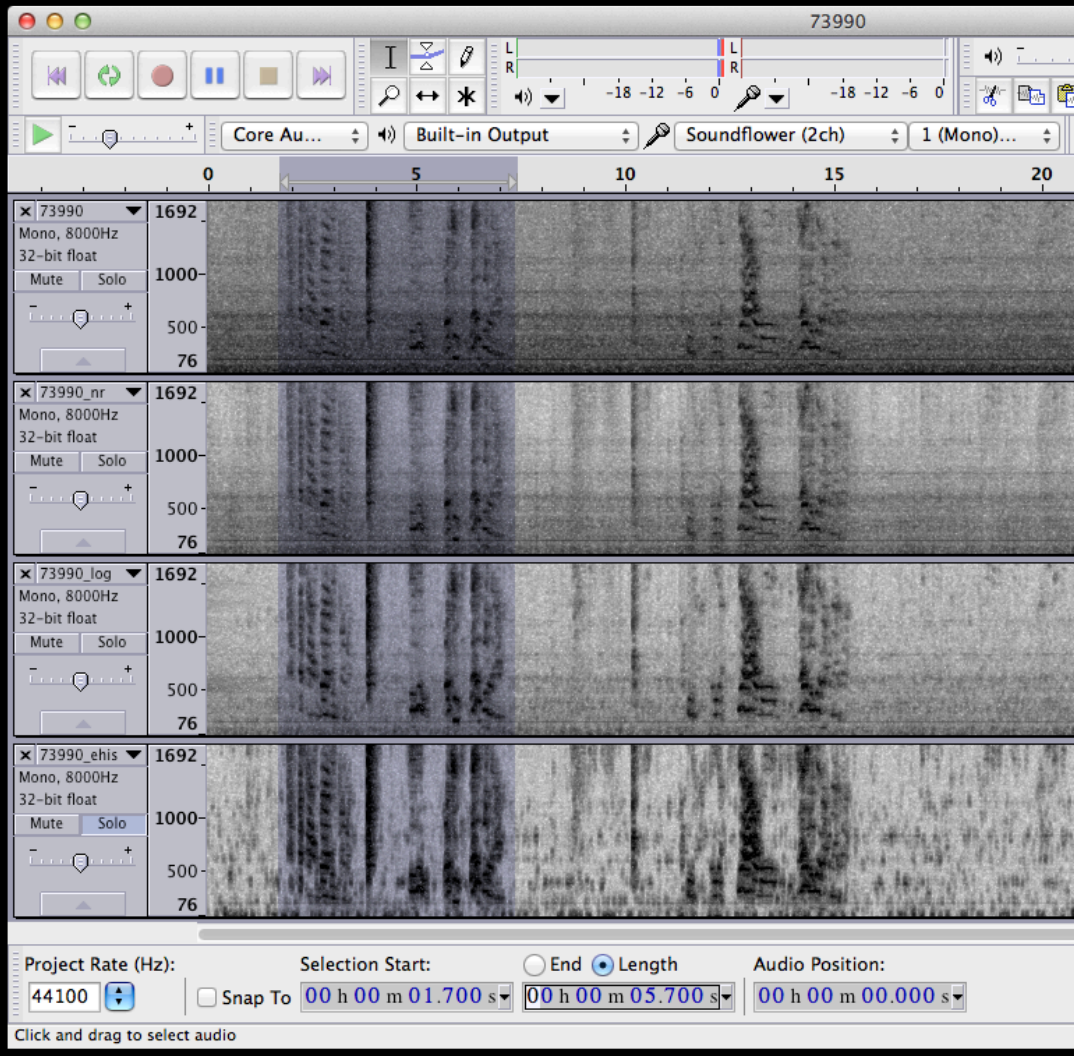
	Corr	Sub	Del	Ins	Err
OP1_204_13189__inLine	5.1	5.1	89.8	1.1	96.0
OP1_204_61440__inLine	7.9	4.8	87.4	0.0	92.1
OP1_204_73990__inLine	5.4	2.6	92.0	1.2	95.9
OP1_204_78161__inLine	15.1	54.4	30.5	6.0	90.8
OP1_204_90937__inLine	3.0	3.0	93.9	0.0	97.0
OP1_204_91808__inLine	2.4	10.0	87.5	1.2	98.8



Spectral-Based Enhancement

- Classic enhancement boosts loudest parts
  - evaluated over 6 MIC utterances of OP1_204_DEV

	Corr	Sub	Del	Ins
Original	6.8	10.6	82.6	0.7
Wiener	8.8	17.5	73.7	1.5
log-MMSE	9.9	31.1	59.0	1.4
E.hist norrm	8.6	23.1	68.4	0.6



RPCA Enhancement

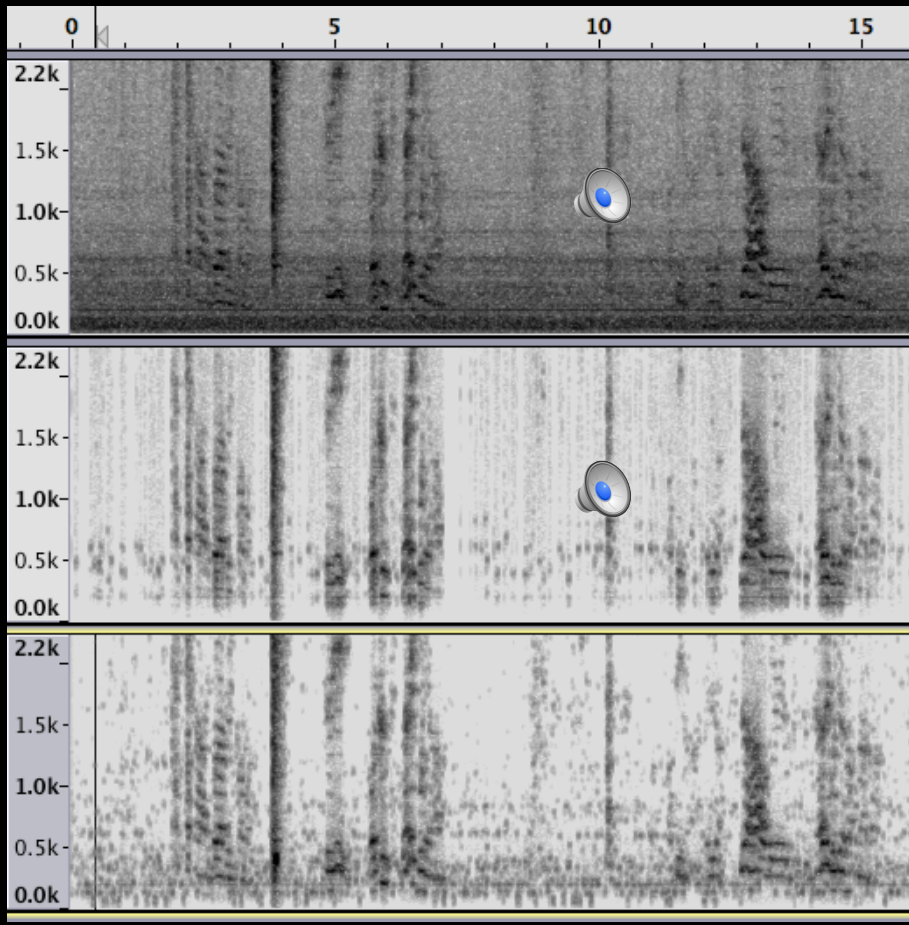
Chen, McFee & Ellis '14

- Decompose spectrogram into sparse + low-rank
- Sparse activation H of dictionary W

$$\min_{H,L,S} \lambda_H \|H\|_1 + \lambda_L \|L\|_* + \lambda_S \|S\|_1 + \mathcal{I}_+(H)$$
$$\text{s.t. } Y = WH + L + S$$

- ASR benefits:

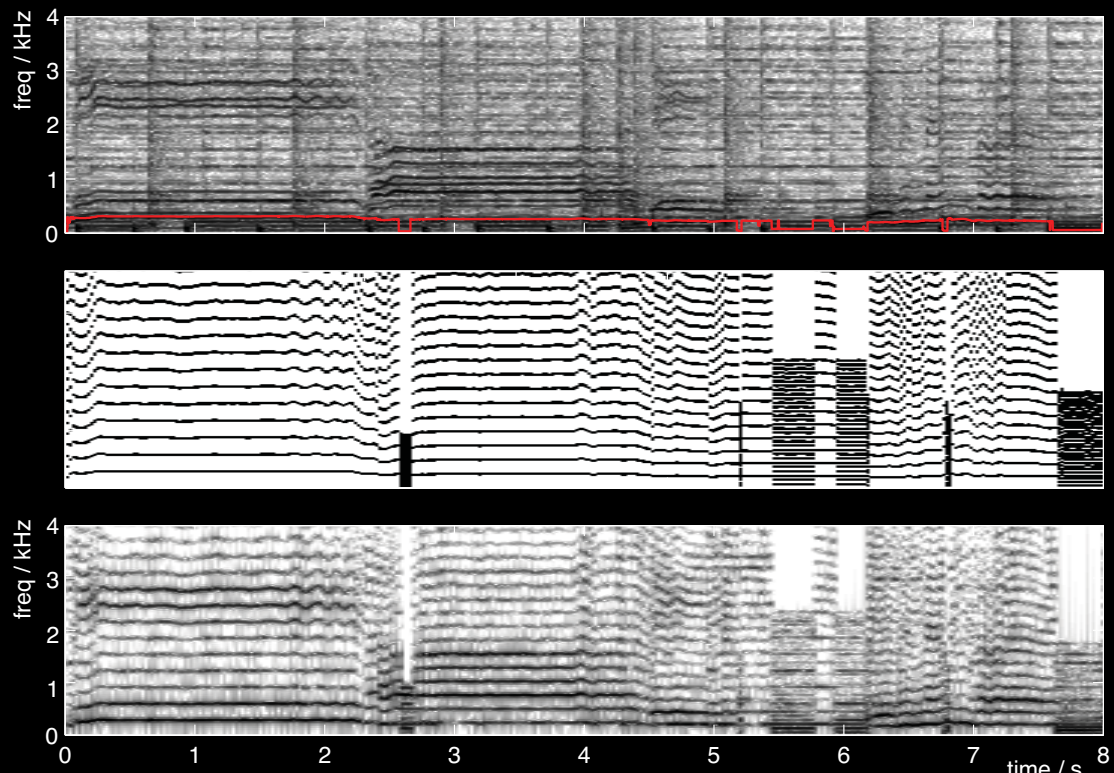
	Corr	Sub	Del	Ins
Original	6.8	10.6	82.6	0.7
log MMSE	9.9	31.1	59.0	1.4
L+WH	15.5	46.1	38.4	1.6



2. Pitch-Based Enhancement

Denbigh & Zhao '92

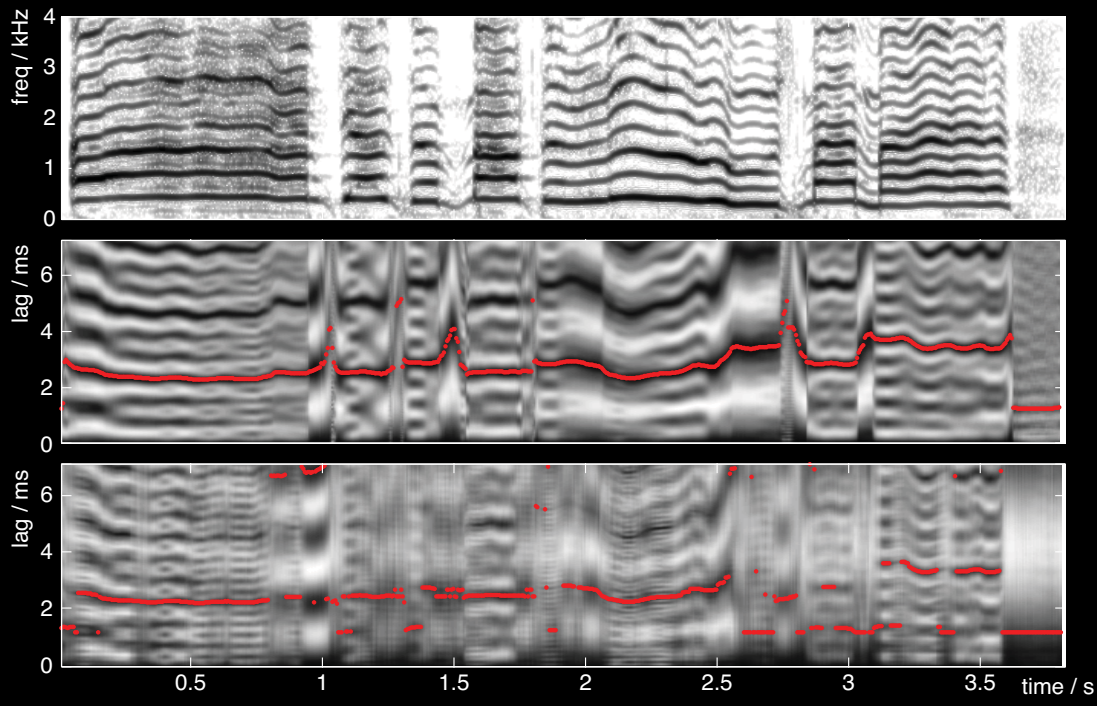
- Voiced Speech has near-periodic waveform
  - Energy concentrated in harmonics
- Given pitch, keep only those harmonics?
  - time-varying filter
  - sinusoidal model
- Problems
  - pitch errors
  - filtering artefacts
  - unvoiced speech, graceful degradation



after Wang '95

Pitch Estimation in Noise

- Conventional pitch trackers are based on periodic structure
  - e.g. finding peaks in autocorrelation
  - not robust to noise

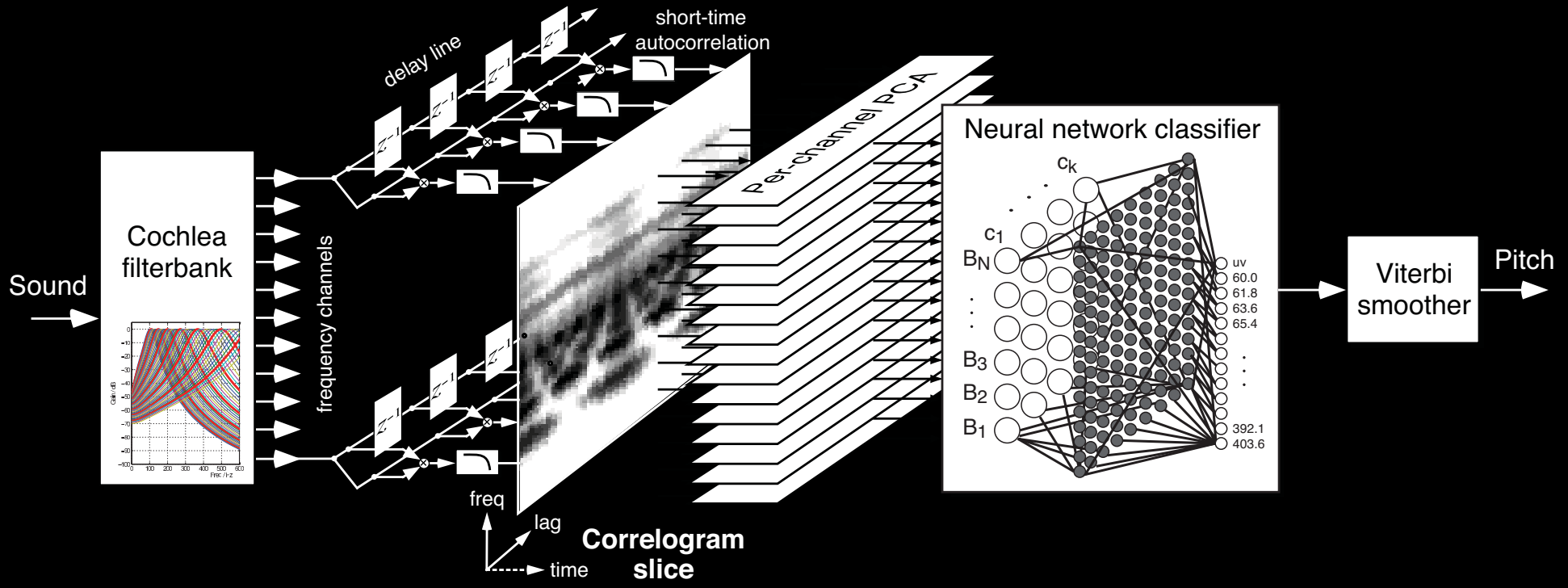


- Classifier-based approach
  - don't predefine nature of pitch
  - let a classifier learn from examples

Classification-based Pitch Tracker

- Subband Autocorrelation Classification (SAcC) Pitch Tracker:
  - Trained on noisy speech with true pitch targets

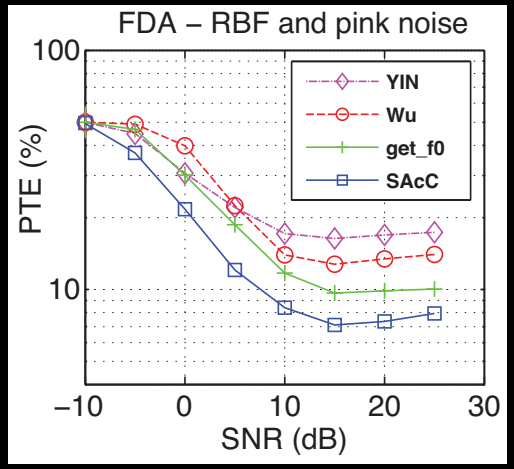
Lee & Ellis '12



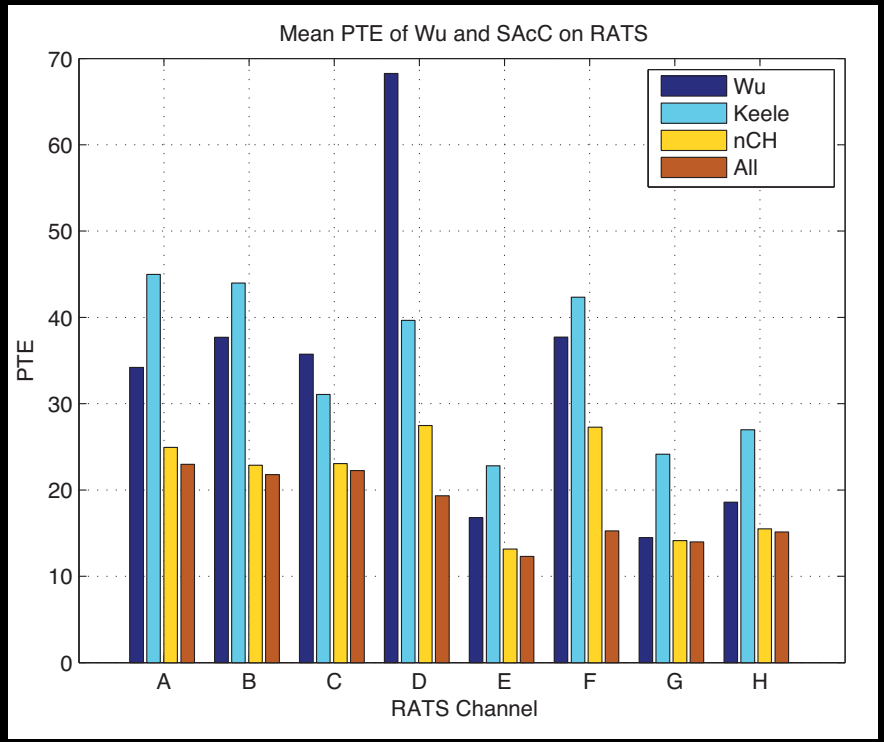
- Subband autocorrelation features
- PCA to reduce dimensions

SAcC Results

- **Excellent in-domain results**
  - at low SNRs
  - errors dominated by V/UV



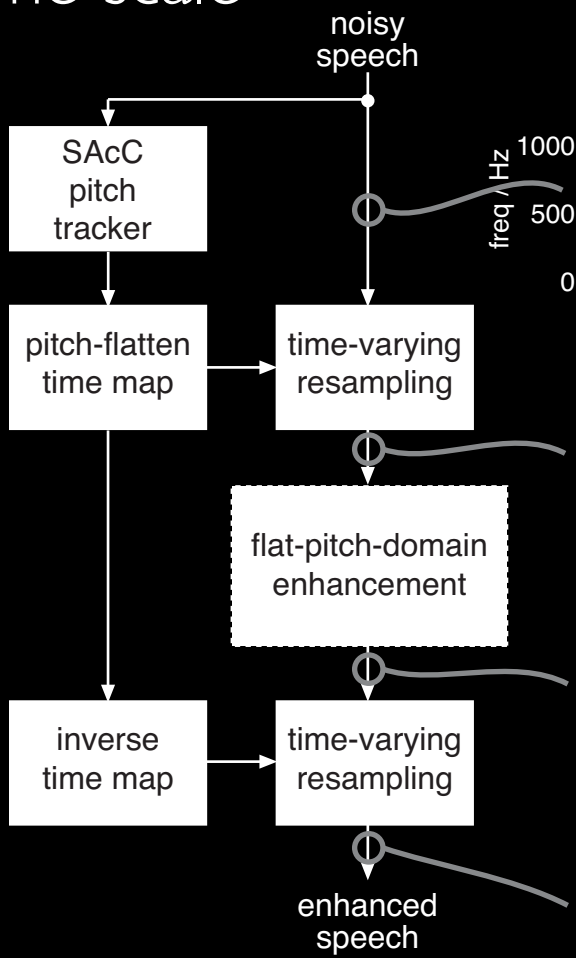
- **Generalization is good**
  - between different RATS channels

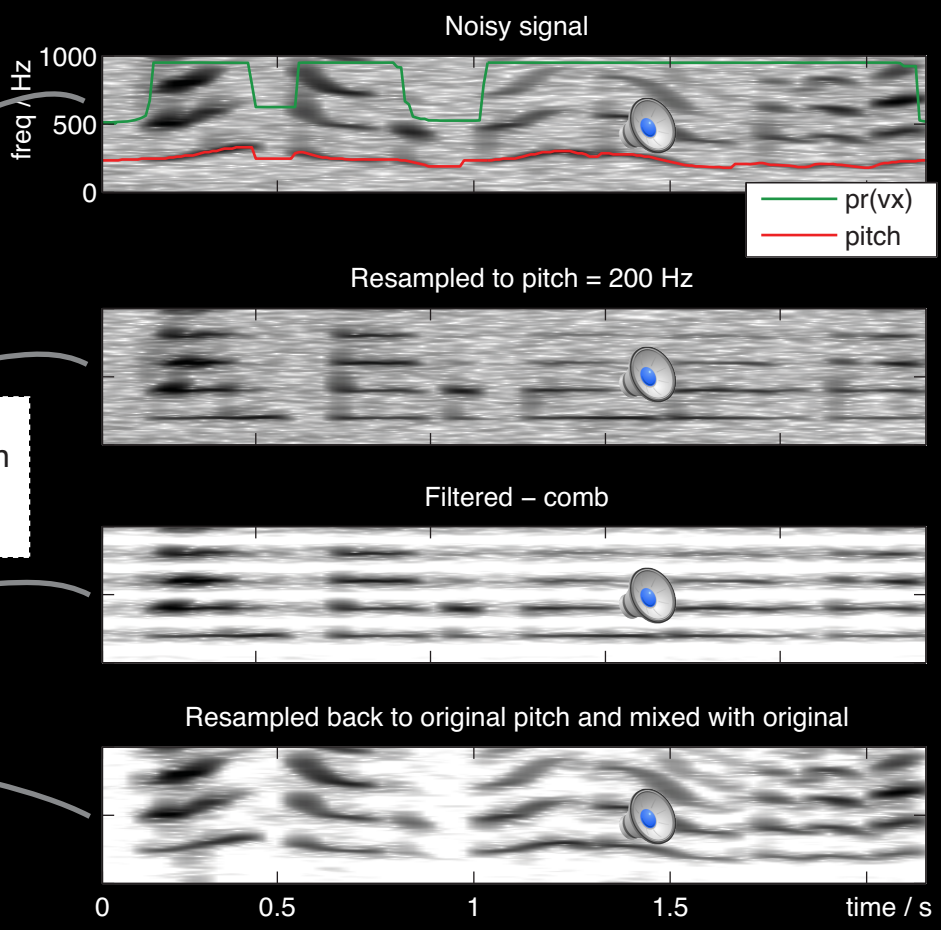


Flat-Pitch Processing

- Time-varying filtering is tricky
  - if pitch variation and filter impulse response are on a similar time-scale

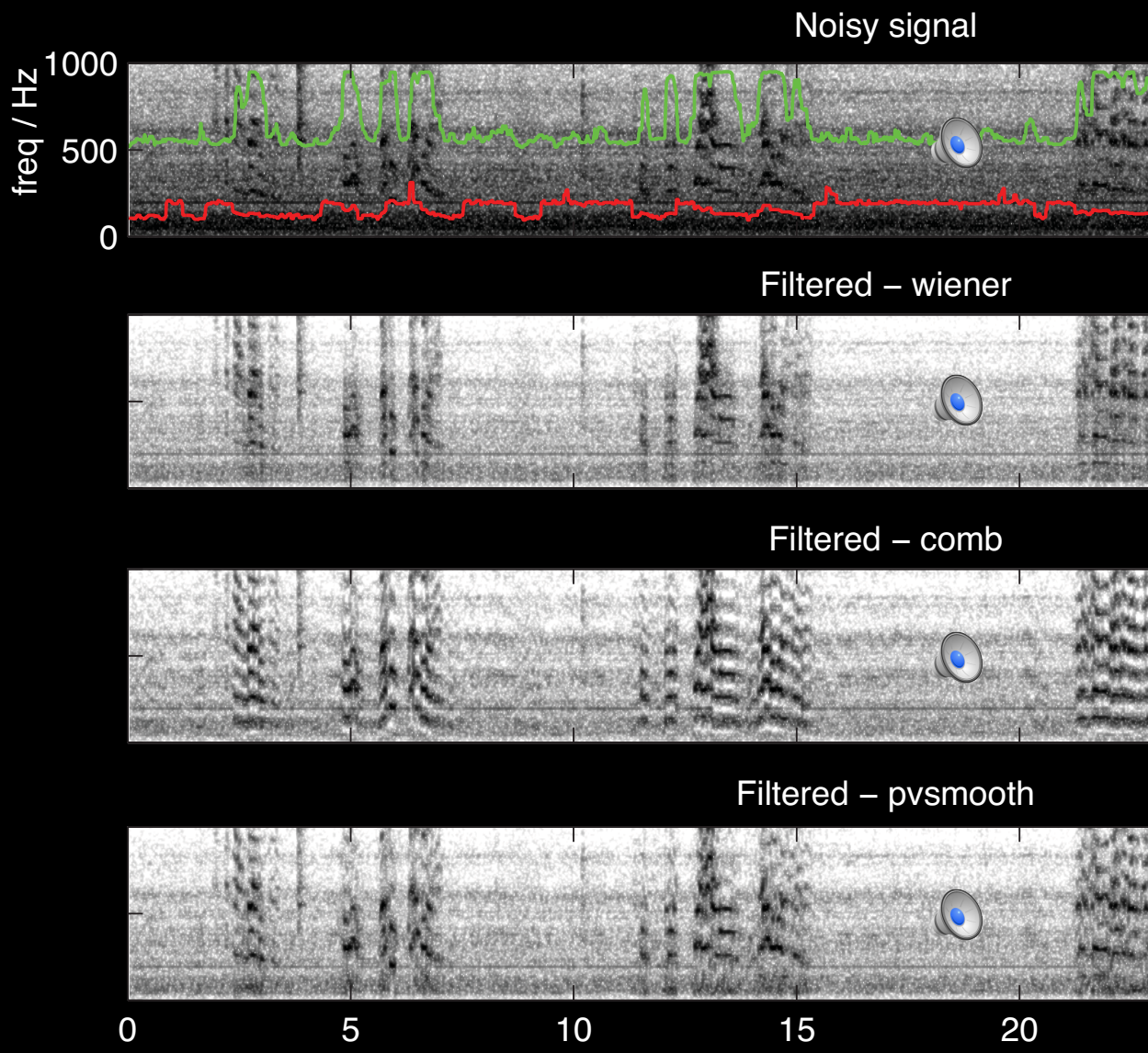
- Solution: Flatten the pitch
  - use local pitch estimate to resample
  - process constant-pitch
  - resampling is (near) invertible





Flat-Pitch Processing

- How to enhance flat pitch?
  - Wiener filtering
  - Comb filtering
  - "Phase Vocoder" emphasis



Flat-Pitch Results

- Over 6 MIC utterances of OP1_204_MIC

	Corr	Sub	Del	Ins
Original	6.8	**10.6**	82.6	**0.7**
log MMSE	9.9	31.1	59.0	1.4
L+WH	**15.5**	46.1	38.4	1.6
flat_pitch-comb	10.0	25.2	64.9	**0.7**
MMSE+f_p-comb	13.6	42.5	43.9	2.8
f_p-comb+L+WH	15.0	52.4	**32.6**	3.1

Summary

- Noisy Speech
  - single distant mic in real-world environments
- Enhancement
  - boosting spectrogram energy that appears to be speech
  - low-rank + sparse dictionary exploits knowledge
- Flat-Pitch enhancement
  - trained noise-robust pitch classifier
  - dynamic resampling to flatten pitch for enhancement