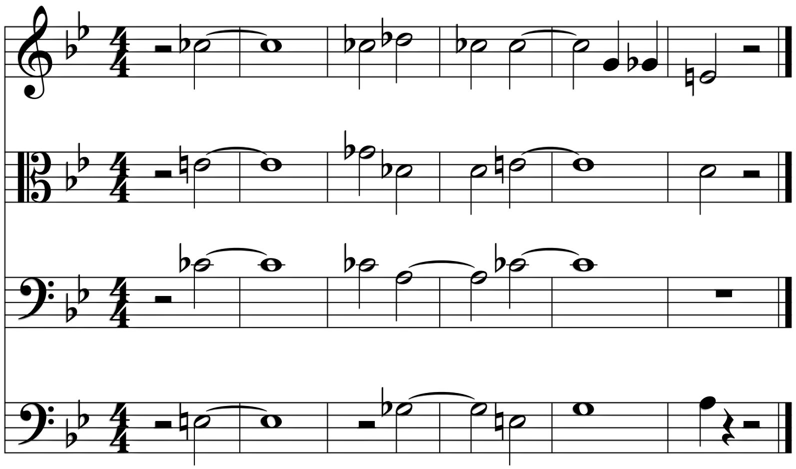
An evaluation of score-informed methods for estimating fundamental frequency and power from polyphonic audio

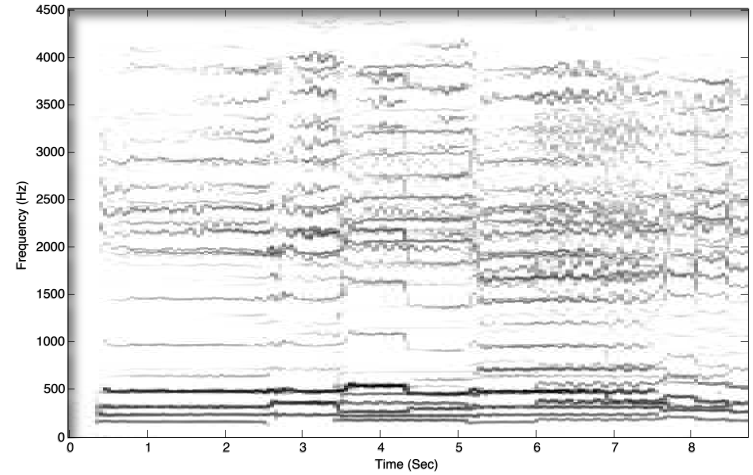
Johanna Devaney (The Ohio State University)
Michael Mandel (Brooklyn College, CUNY)

Motivation

Automatically extracting musical performance data

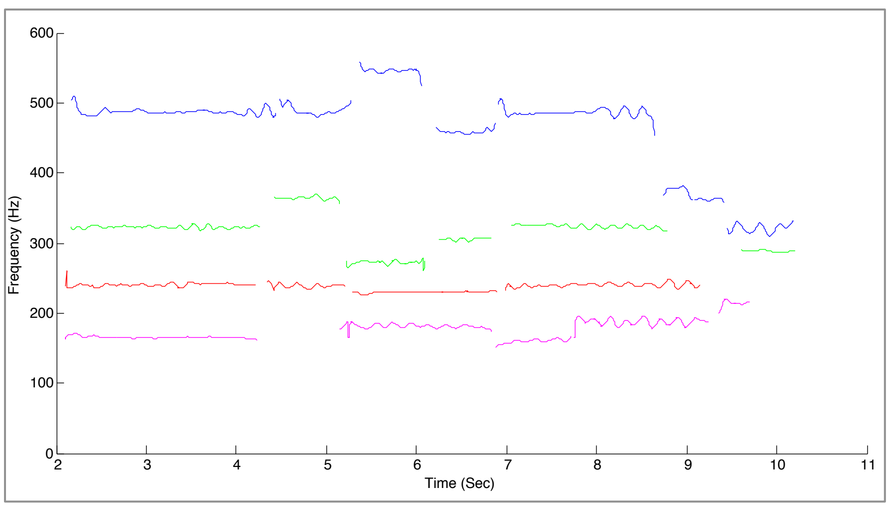
- **Robust extraction of performance data from polyphonic musical performances requires precise frame-level estimation of fundamental frequency (f_0) and power**
    - pitch- and dynamic-deviations in expressive musical performance
- **Currently blind f0 estimation has an ~70% accuracy ceiling (Benetos et. al. 2013)**



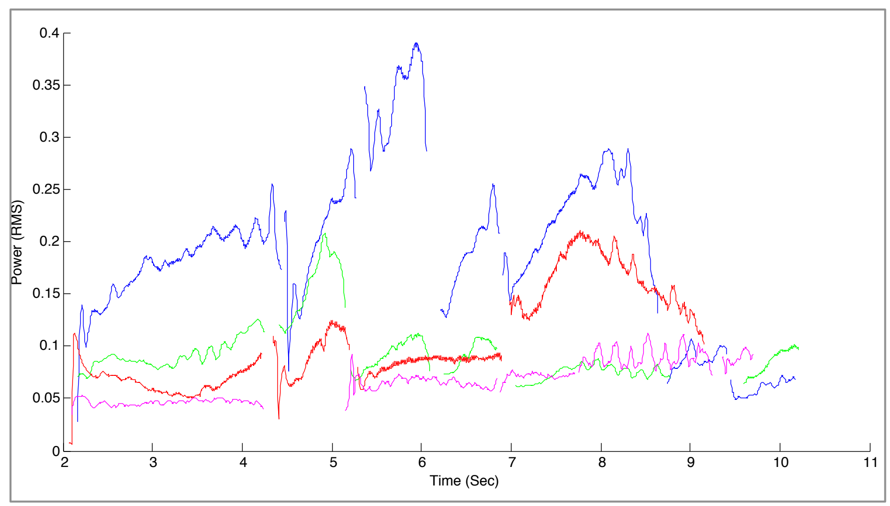


Alignment

Note-wise f_0 estimations



Note-wise power estimations



Evaluated Approaches

General approach

- **Aligned score indicates time-frequency regions of interest**
- **For each region, we calculated frequency and magnitude estimates using one of**
  - discrete Fourier transform (DFT)
  - instantaneous frequency (IF)
  - high resolution spectral analysis (HR)
  - high resolution spectral analysis with comb filtering at harmonics of initial f_0 estimate (HR-C)

Evaluated Approaches

General approach

- **DFT: 64 ms window size and 16 ms hop size**
- **IF: derivative of phase spectrum of DFT**
  - Abe, Kobayashi, and Imai (1995, 1997)
- **HR: estimate mixtures of complex exponentials modulated by polynomials using ESPRIT algorithm**
  - Badeau, Richard, and David (2008)

Evaluated Approaches

General approach

- **f_0 estimates calculated with a weighted sum**

set of frequencies considered

$$\hat{f}_0 = \frac{\sum_n \sum_{\omega_i \in \boxed{\mathcal{N}(nf_0)}} \frac{\omega_i}{n} \boxed{x(\omega_i)}}{\sum_n \sum_{\omega_i \in \boxed{\mathcal{N}(nf_0)}} \boxed{x(\omega_i)}}$$

magnitude measurements of set of frequencies

- **The process is repeated 10 times in order to refine f_0 estimate**

Evaluated Approaches

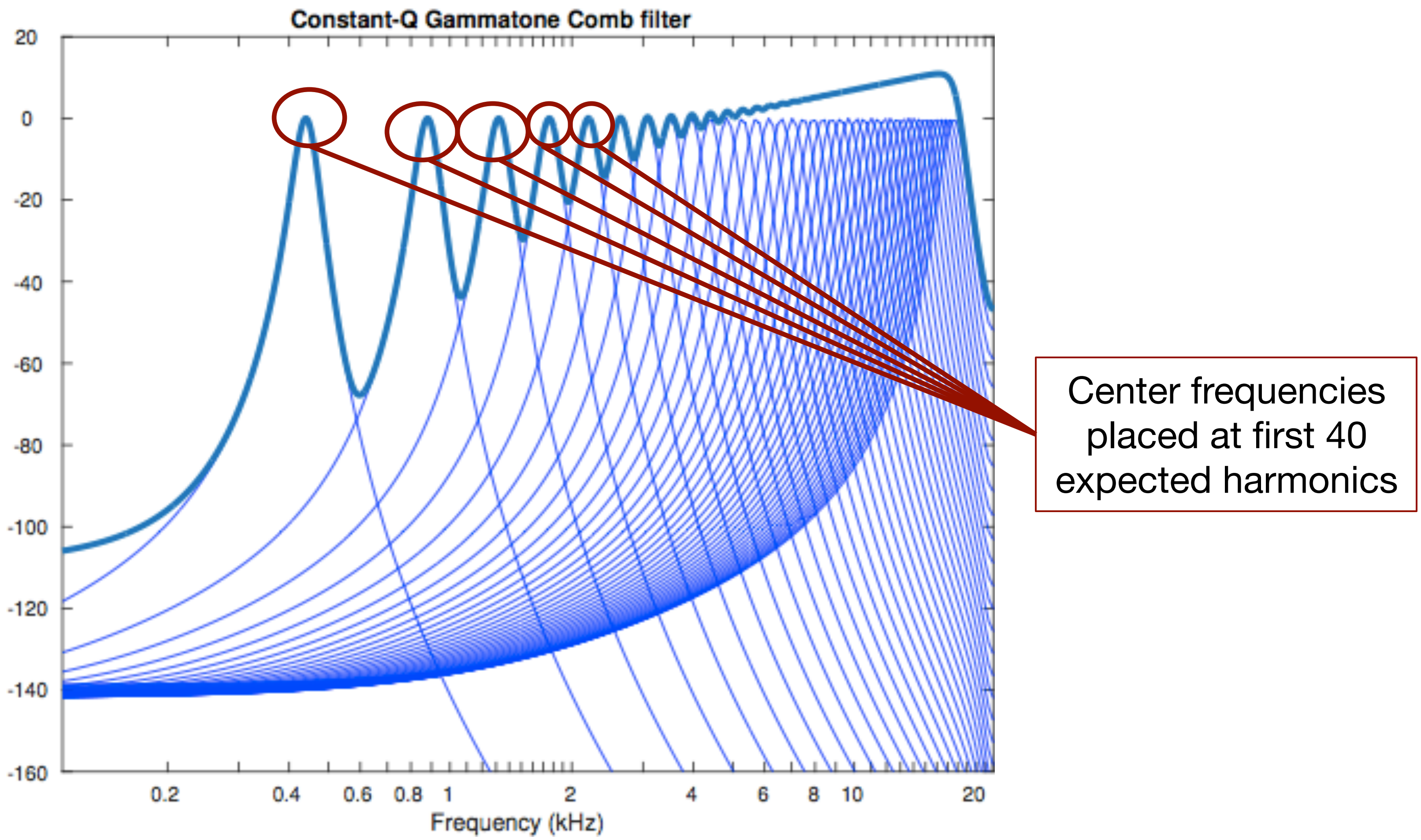
DFT, IF, and HR

$$\hat{f}_0 = \frac{\sum_n \sum_{\omega_i \in \mathcal{N}(nf_0)} \frac{\omega_i}{n} x(\omega_i)}{\sum_n \sum_{\omega_i \in \mathcal{N}(nf_0)} x(\omega_i)}$$

	$x(\omega_i)$	ω_i	$\mathcal{N}(nf_0)$
DFT	cube roots of the DFT magnitudes	uniformly spaced between DC and Nyquist (plus interpolation from weighted sum)	two DFT bins, below and above the predicted frequency (27 Hz)
IF	cube roots of the DFT magnitudes	frequency values estimated from the time derivative of the phase spectrum	estimated frequency within 27 Hz
HR	the cube root of the amplitude of each modulated complex exponential	the frequency of each modulated complex exponential	estimated frequency within 40 Hz

Evaluated Approaches

High Resolution + Comb Filter



Bank of zero-delay, constant-Q, one-zero gammatone filters (Lyon 2010)

Evaluated Approaches

Power estimates

- **The power estimates were derived from the same data as the f_0 estimates**
  - with squared magnitudes used instead of cube root magnitudes
- **For each estimated f_0, the power was estimated as**

$$\hat{p}(f_0) = \sum_{n} \sum_{\omega_i \in \mathcal{N}(n f_0)} \tilde{x}(\omega_i)$$

Introduction
Motivations and Background

1

Evaluated Approaches
f_0 and Power

2

Experiment
Materials and Results

3

Conclusions
Future Directions and Summary

4

Experiment

Materials

- **Two multi-tracked datasets**
  - Bach 10 dataset: violin, clarinet, saxophone and bassoon (Duan, Pardo and Zhang 2010)
  - Machaut “Kyrie” dataset: soprano, alto, tenor, and bass (Devaney and Ellis 2008)
- **Hand annotated note boundaries were used**
  - Further refined by removing any leading or trailing sections that had a pYIN (Mauch and Dixon, 2014) periodicity estimate of less than 95%

Experiment

Materials

- **Signals were convolved with three impulse responses from a room with an RT60 reverberation time of 1.4 sec resulting in four sets of recordings**
  - Original anechoic
  - Reverberant with microphone 3.6 meters away
  - Reverberant with microphone 7.7 meters away
  - Reverberant with microphone 11 meters away

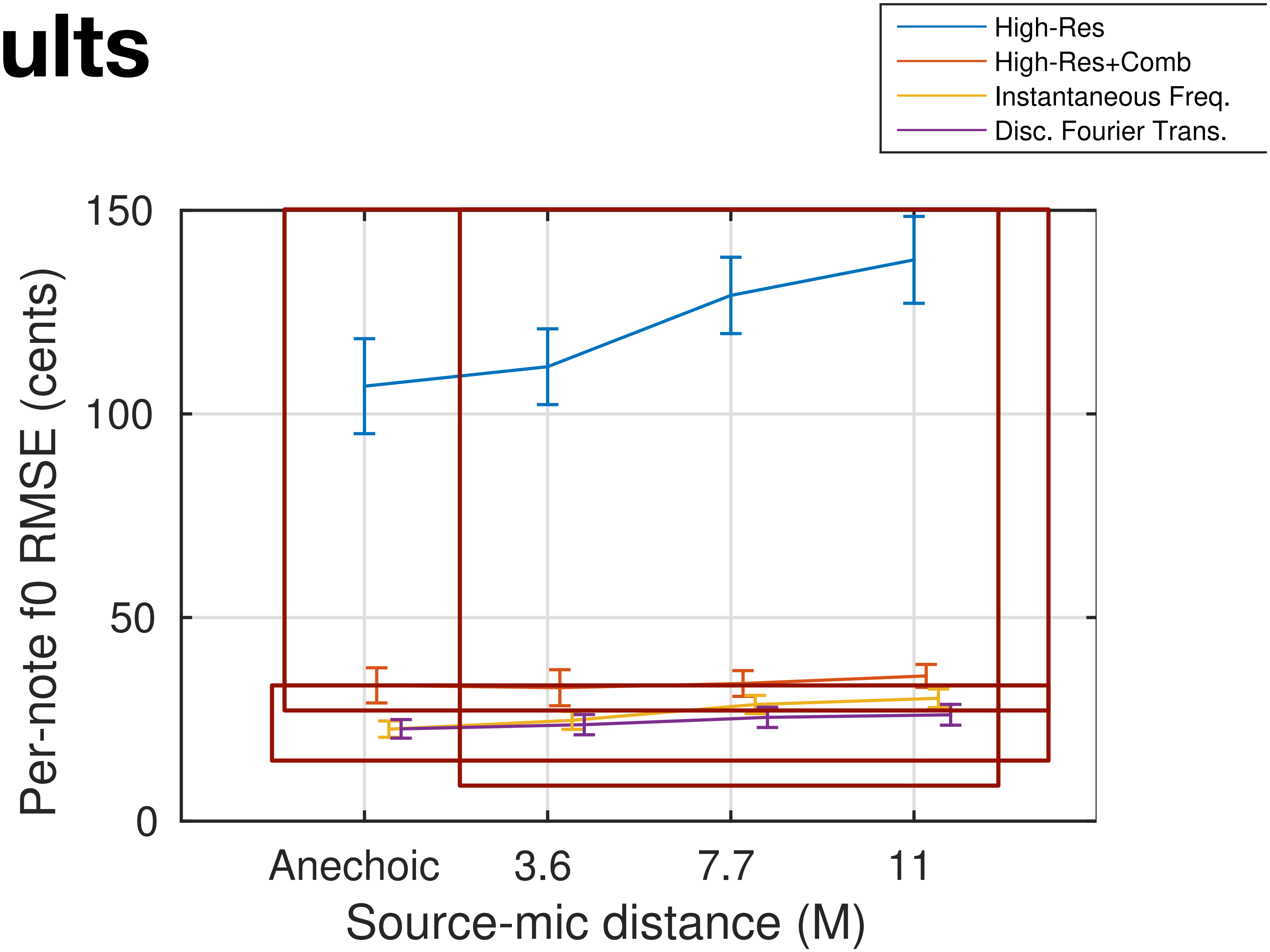
Experiment

Ground Truth

- **Ground truth** was calculated from monophonic tracks using the pYIN algorithm (Mauch and Dixon, 2014)
- **Error** was calculated in **cents** (or **db**) between the estimates and the ground truth across each note

$$E = \sqrt{\sum_{n} \left(\hat{f}_0(n) - f_0(n) \right)^2}.$$

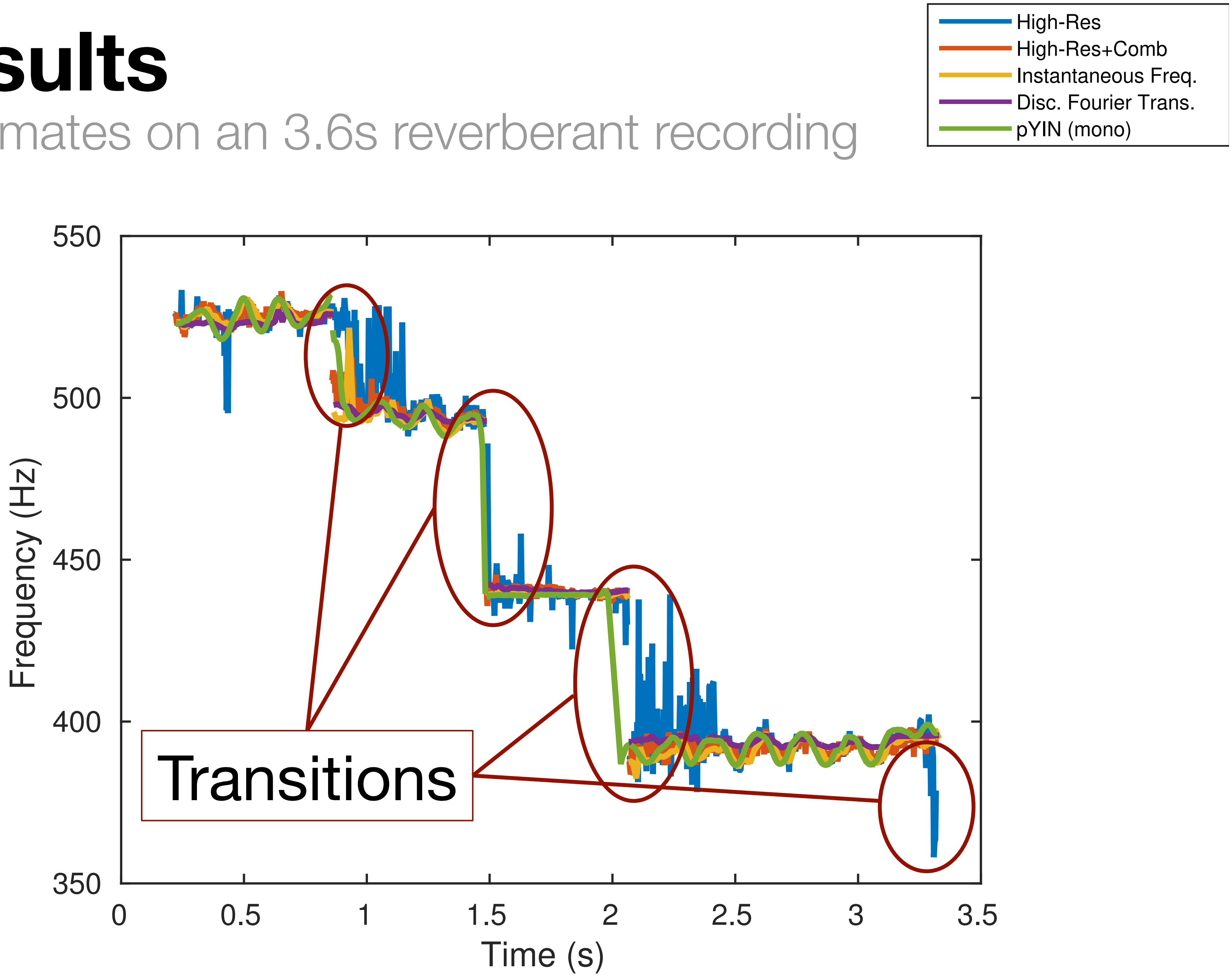
Results



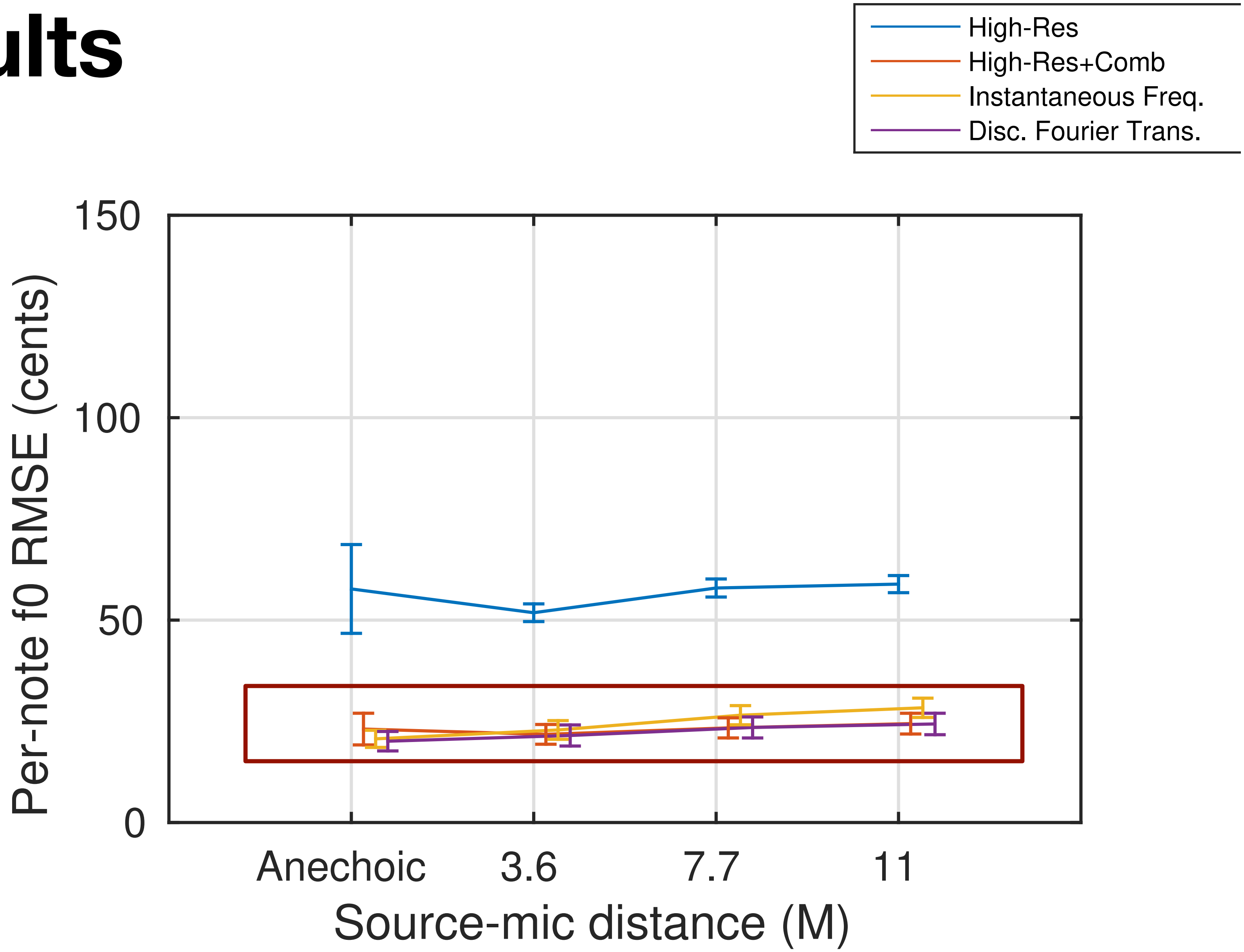
f_0, entire notes

Results

f_0 estimates on an 3.6s reverberant recording



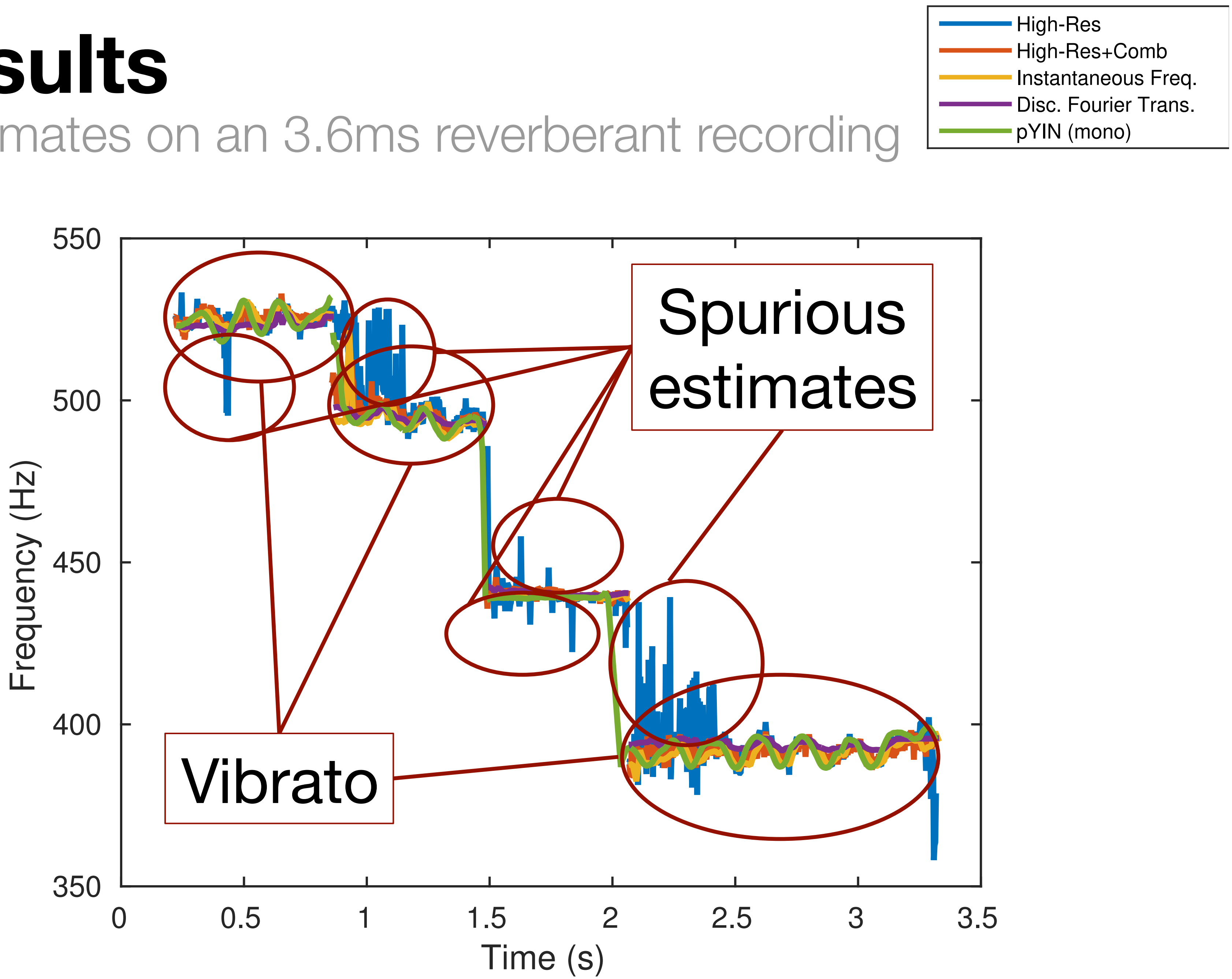
Results



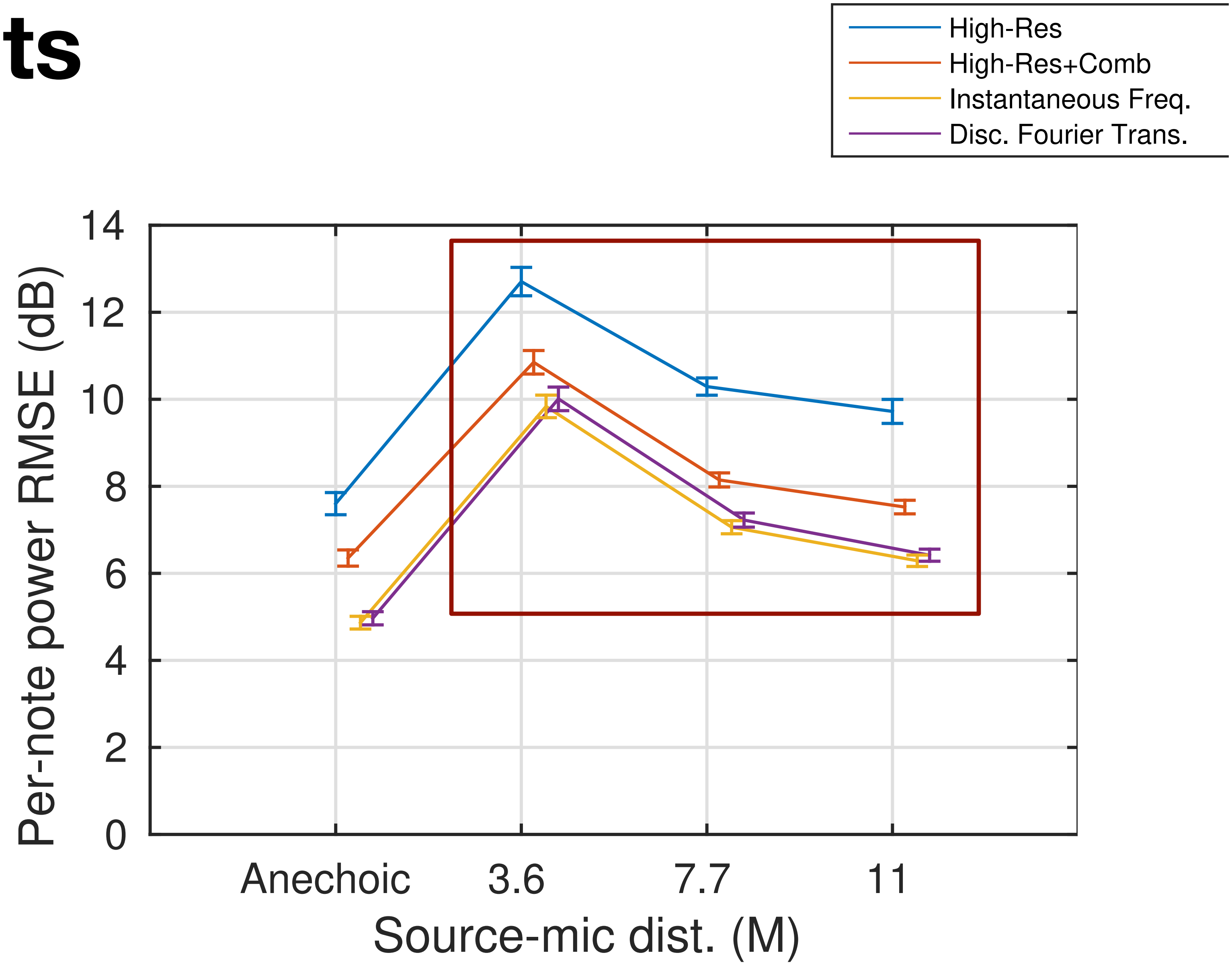
f_0, central 80%

Results

f_0 estimates on an 3.6ms reverberant recording



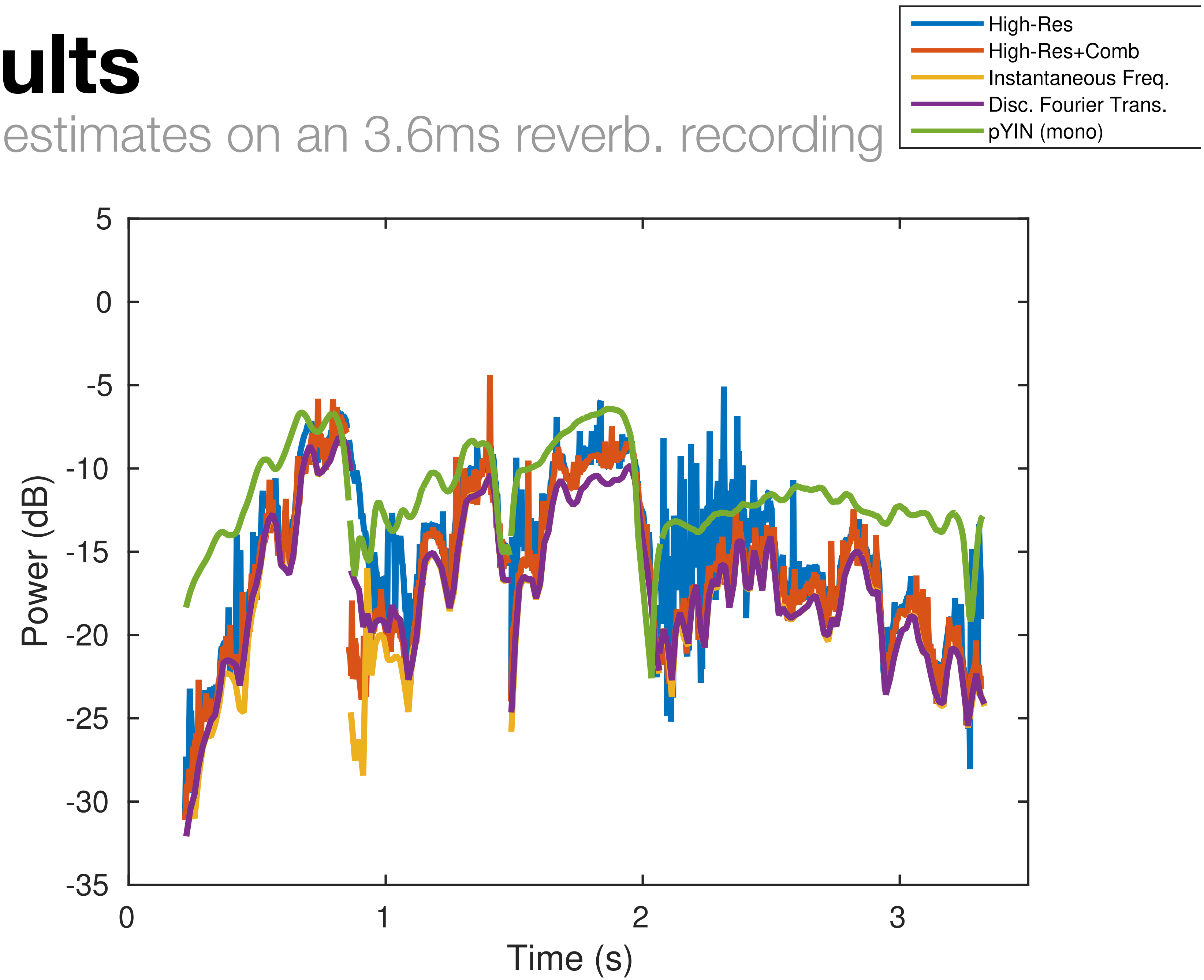
Results



Power, entire notes

Results

Power estimates on an 3.6ms reverb. recording



Introduction
Motivations and Background

1

Evaluated Approaches
f_0 and Power

2

Experiment
Materials and Results

3

Conclusions
Future Directions and Summary

4

Summary

- **Simple** and **accurate** score-guided f0 and power estimation method
- **Instantaneous frequency** features performed best in terms of accuracy, capturing frame-wise variation (e.g., vibrato), and computational cost
- f_0 estimates, in particular, shown to be **robust to reverberation**

Future Directions

- **Improve power estimates**
  - Add a de-reverberation step
- **Move from frame-wise measurements to note-wise estimates informed by perceptual models**
  - relatively straightforward in the case of perceived pitch
  - more complicated for perceived loudness
  - much more so for timbre
- **Investigate alternatives to using pYIN to generate ground truth, such as synthesized tracks using high-quality models**

Acknowledgements

NATIONAL ENDOWMENT FOR THE

Humanities

OFFICE OF DIGITAL HUMANITIES

Thank you