Pitch Estimation

Marián Képesi
Signal Processing and Speech Communication Laboratory

13. March 2008

Outline

- Pitch and Formants
- Pitch estimation methods
- Combining pitch and spectrum estimation
- Examples and discussion

Sources

- Klapuri, A. " Signal processing methods for the automatic transcription of music," **Ph.D. thesis**, Tampere University of Technology, Finland
- Pitch Estimation Algorithms, **Chapter 10** in T.Quatieri: Discrete-Time Speech Signal Processing, Prentice Hall, 2002.
- Pitch-related topics on **Wikipedia.org**
- Malcolm Slaney: Auditory Toolbox (for Matlab)
- Matlab code at: http://cvsp.cs.ntua.gr/courses/patrec/OnlineSpeechDemos/speechDemo_2004_Part1.html

Recall: Pitch vs. Formants

Pitch vs. Formants

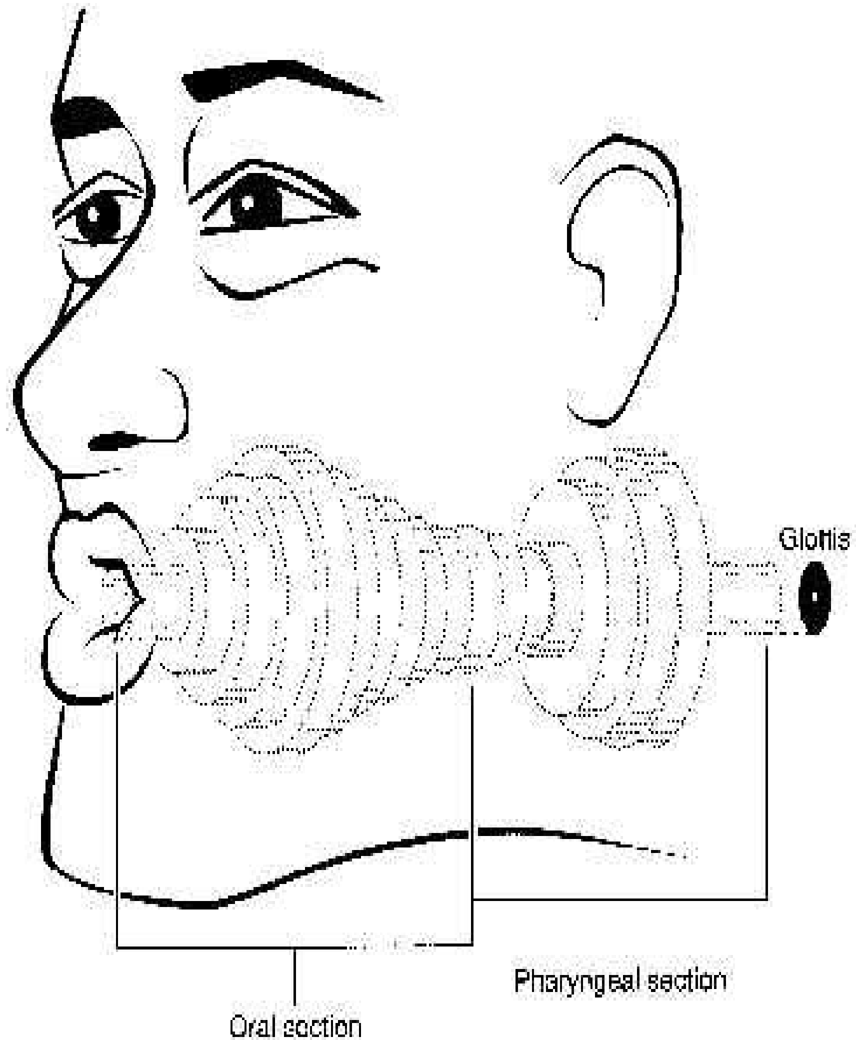
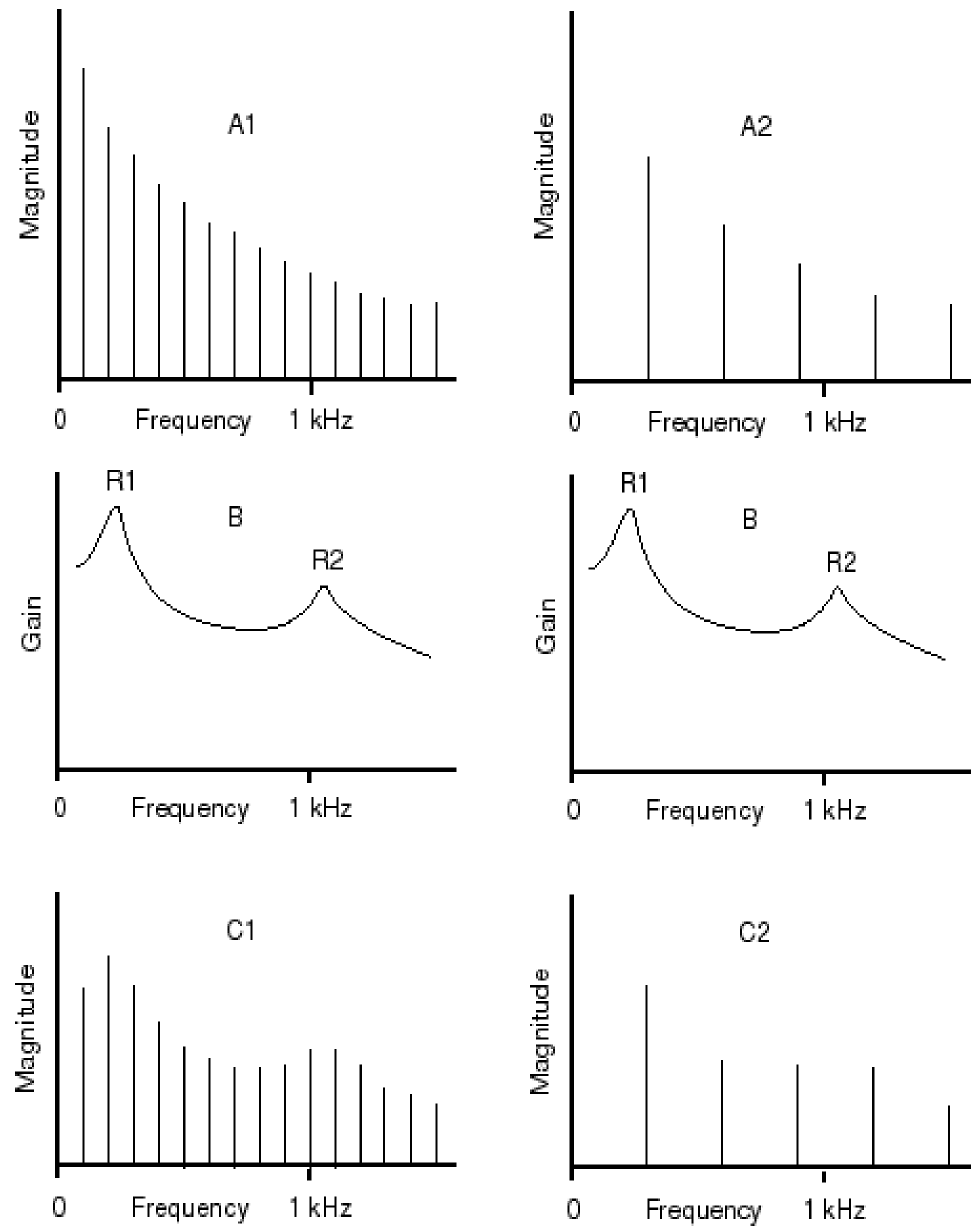
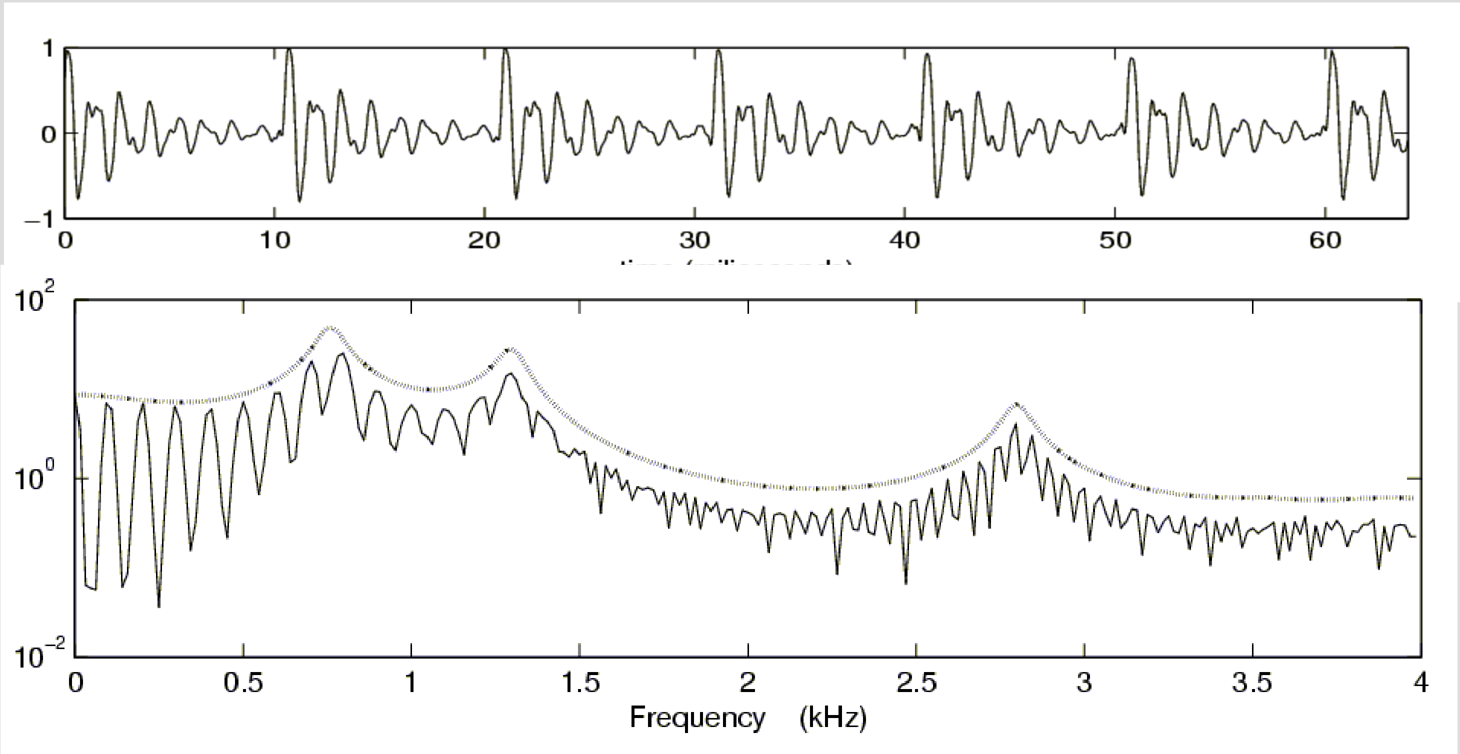


Figure 2.20. Cylindrical-tube approximation of the vocal tract for a simulated /u/ vowel (from Titze, *Principles of Voice Production*, 1994. All rights reserved. Reprinted by permission of Allyn & Bacon).



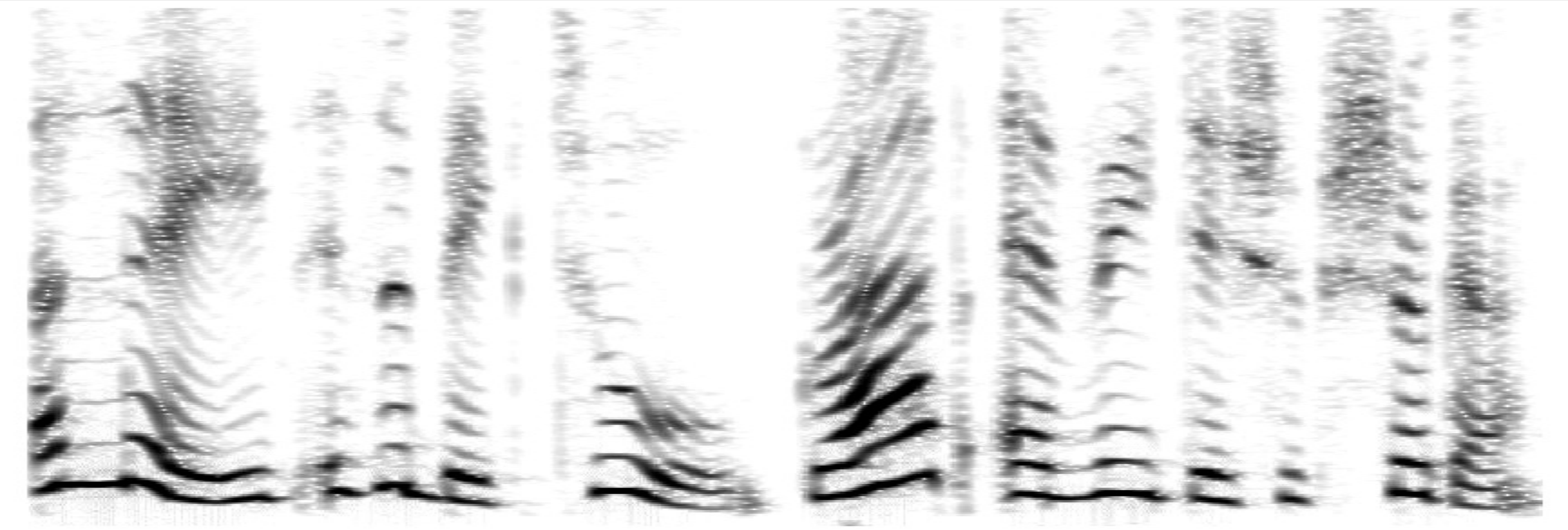
Pitch of Natural Speech

Pitch is always changing, speech is **semi**-periodic!



Reading Pitch from Spectrograms

Note the time-frequency regions where the pitch changes:



The higher the frequency the higher the "distortion" is.

Pitch Estimation Methods

- Zero crossing based
- Autocorrelation based
- Average Magnitude Difference Function (AMDF)
- Cepstral peak - based
- Spectral peak-picking - based
- Auditory Model Based
- etc...

Pitch Estimation - Applications

- **Single-Pitch estimation** for:

- emotion recognition (level of articulation),
- voice coding (mobile phones), voice compression (dictaphones),
- speech analysis (lie detectors, speech disorders, etc.)

-**Multipitch tracking algorithms for**:

- speech segregation (coctail-party problem),
- music transcription, source separation (speech from non-speech).

-**F0 extraction algorithms**:

1) event detection methods
   (peak-picking, zero crossing, etc.),

2) Short-term estimation methods
   (Autocorrelation, AMDF, etc.)

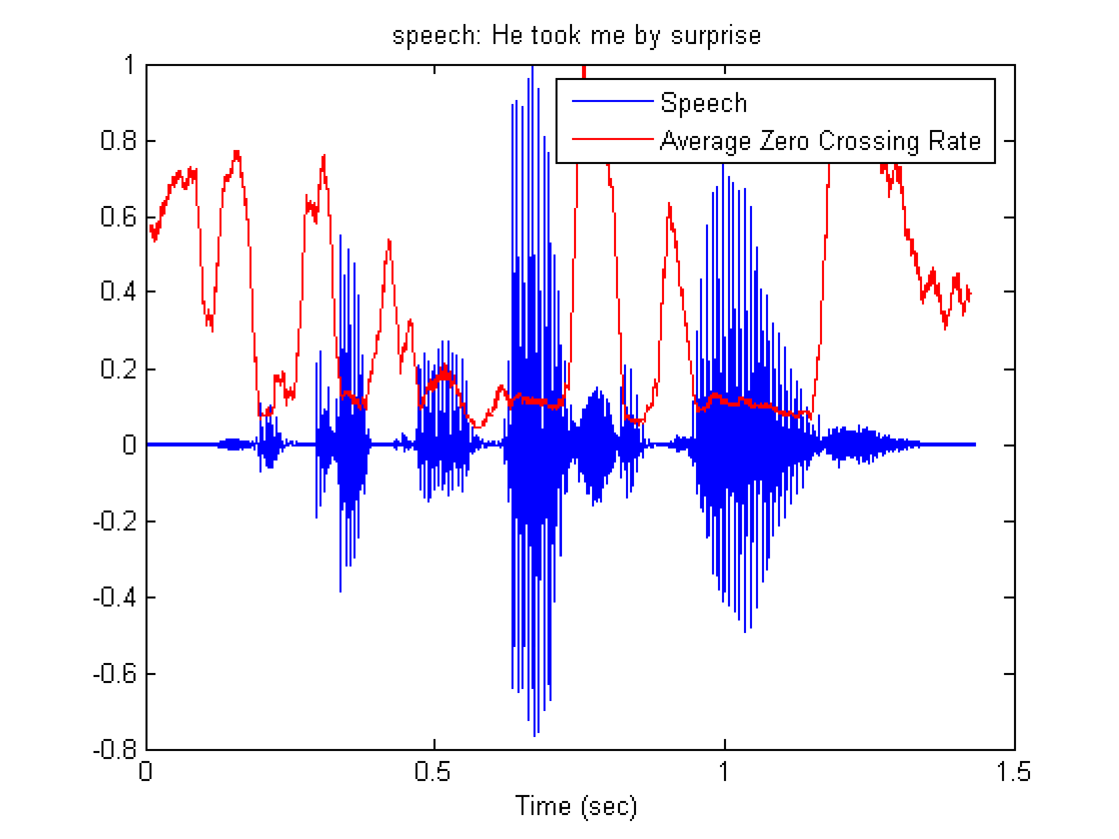
• Zero-crossing Based

$$zcr = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{I}\{s_t s_{t-1} < 0\}$$

$\mathbb{I}\{A\}$ is 1 if its argument A is true and 0 otherwise.

Rate of sign changes
along the signal.
Heavily used both in speech
and musical signal analysis.

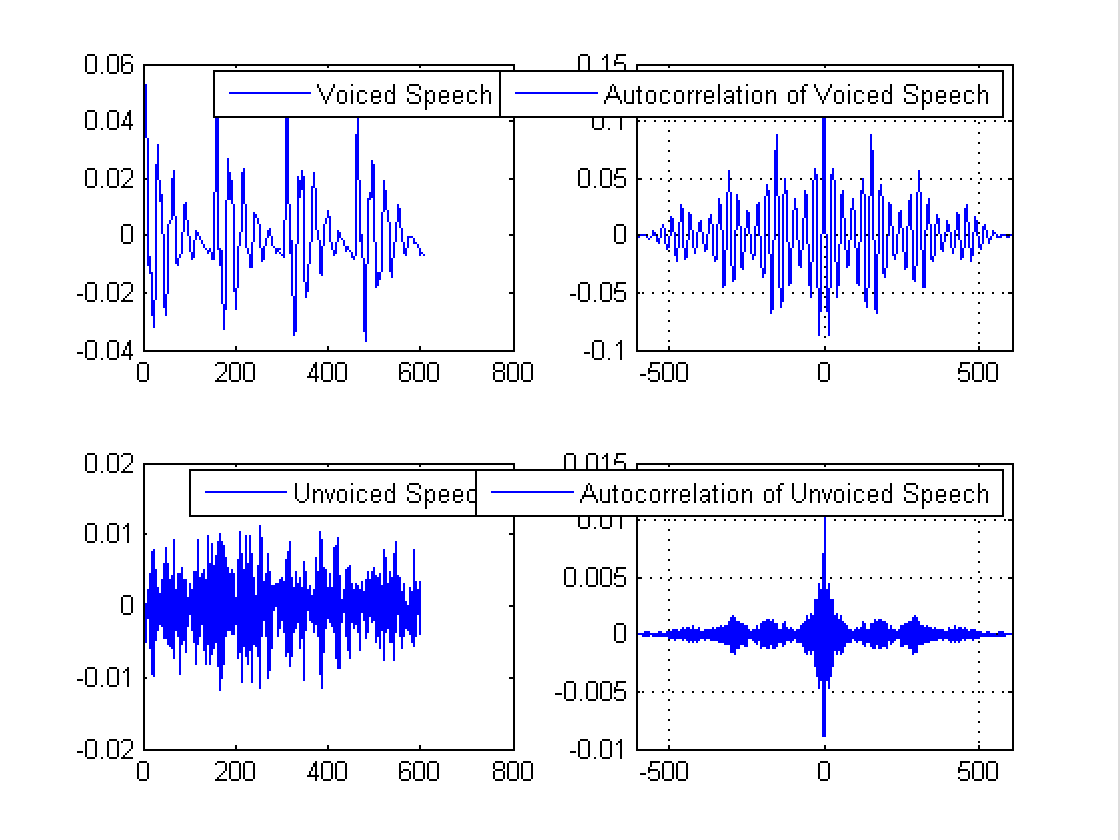
But useful only for
for single- source scenarios



Autocorrelation Based

$$R_{ff}(\tau) = \overline{f}(-\tau) * f(\tau)$$
$$= \int_{-\infty}^{\infty} f(t+\tau)\overline{f}(t)\,dt$$
$$= \int_{-\infty}^{\infty} f(t)\overline{f}(t-\tau)\,dt$$

$$R_{xx}(j) = \sum_n x_n \overline{x}_{n-j}\,.$$

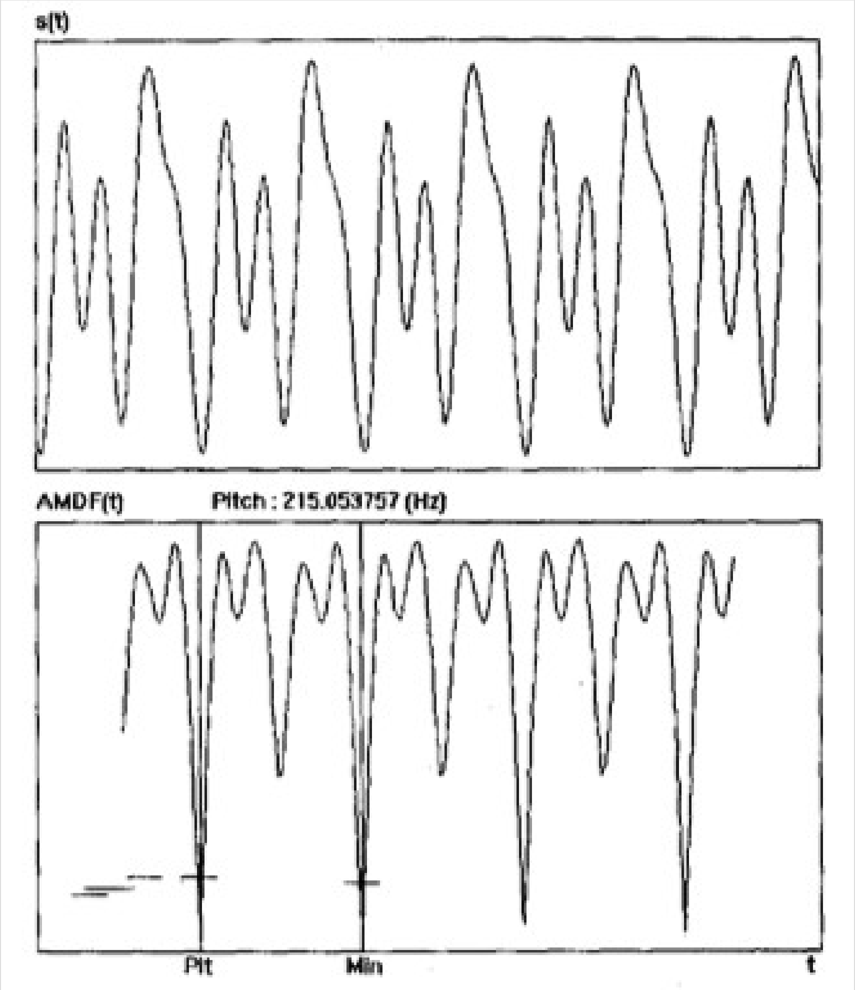


Autocorrelation methods need at least two pitch periods to detect pitch. To detect a fundamental frequency of 40 Hz this means that at least 50 milliseconds (ms) of the speech signal must be analyzed. However, during 50 ms, speech with higher fundamental frequencies may not necessarily have the same fundamental frequency throughout the window

AMDF Function

$$AMDF(t) = \frac{1}{L}\sum_{i=1}^{L}|s(i) - s(i-t)|$$

where s(i) : the samples of input speech
s(i)=[s(1),s(2), ... ,s(L)]
s(i- τ) : the samples time shifted



Average Magnitude Difference Function
- faster than autocorrelation
- is related to autocorrelation
- multiplication replaced either by abs (x-x') or by $(x-x')^2$

Cepstral Peak Picking

signal → FT → abs() → log → phase unwrapping → FT → cepstrum

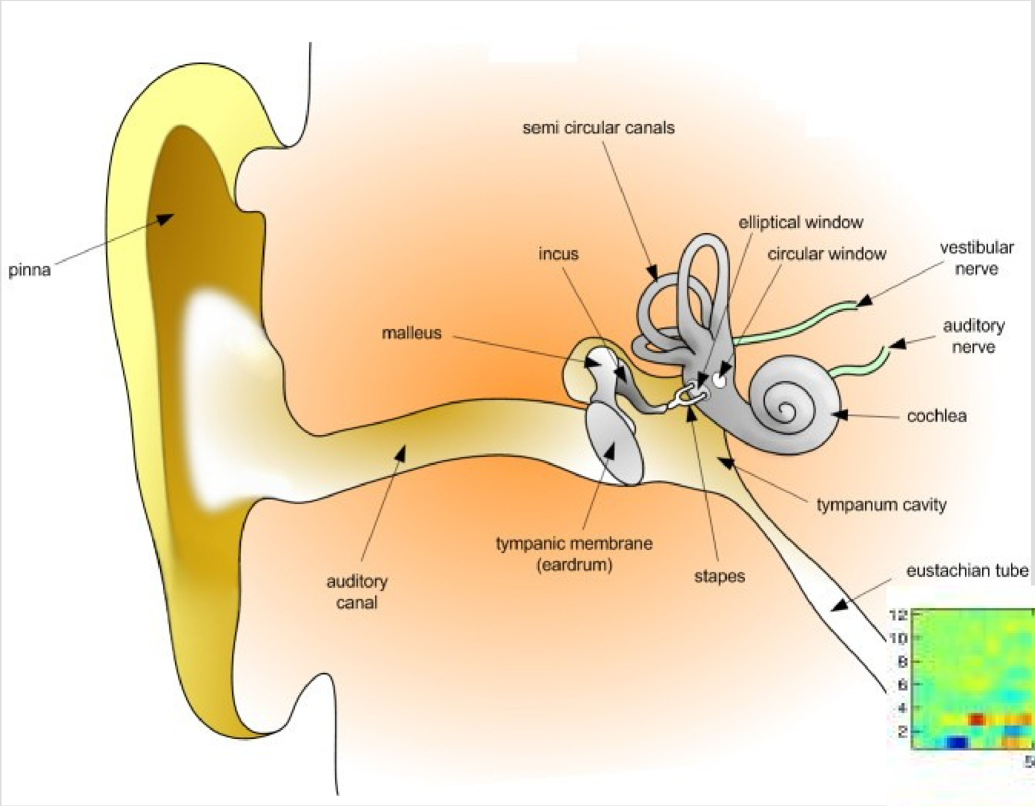
Quefrency Analysis – Cepstral analysis – “Spectrum of Spectrum”

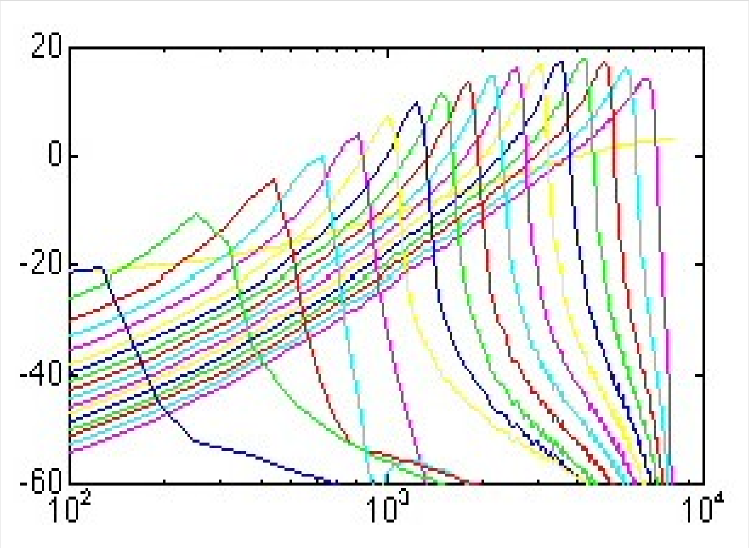
A very important property of the cepstral domain is that the convolution of two signals can be expressed as the addition of their cepstra:

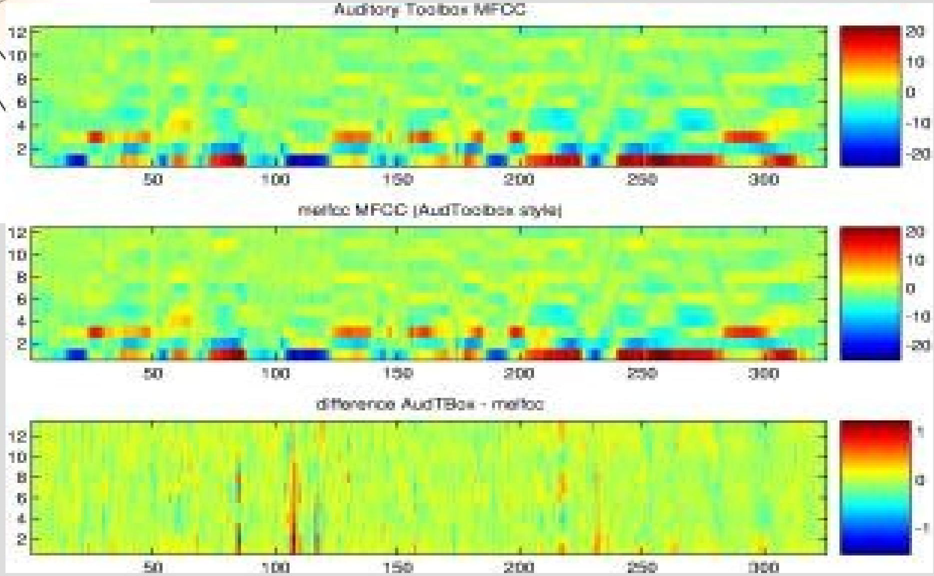
$$x_1 * x_2 \longrightarrow x_1' + x_2'$$

The independent variable of a cepstral graph is called the quefrency. The quefrency is a measure of time, though not in the sense of a signal in the time domain. For example, if the sampling rate of an audio signal is 44100 Hz and there is a large peak in the cepstrum whose quefrency is 100 samples, the peak indicates the presence of a pitch that is 44100/100 = 441 Hz. This peak occurs in the cepstrum because the harmonics in the spectrum are periodic, and the period corresponds to the pitch.

Auditory Model Based



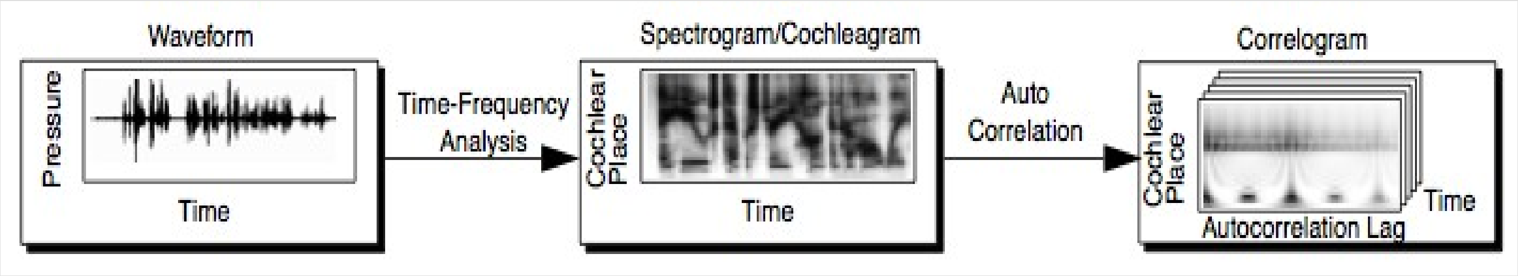




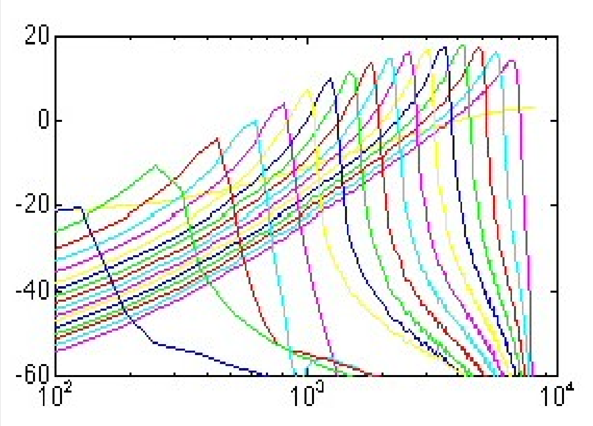
Different models proposed:

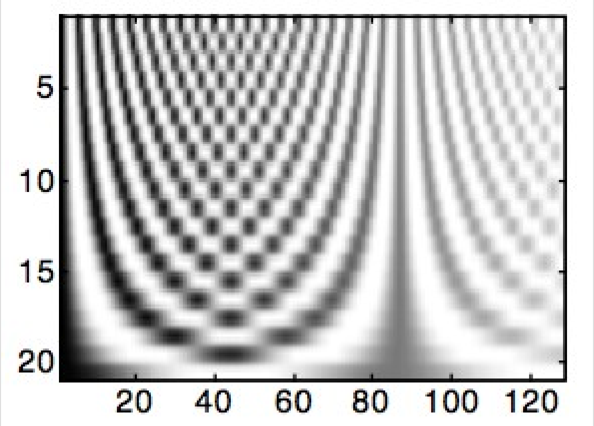
- Lyon Passive Cochlear Model,
- Patterson-Holdsworth Filters
- Seneff Auditory Model, etc.

Auditory Model Based ..



Malcolm Slaney: Auditory Toolbox – for Matlab





Auditory Fiulterbank ,   Correlogram of voiced Signal

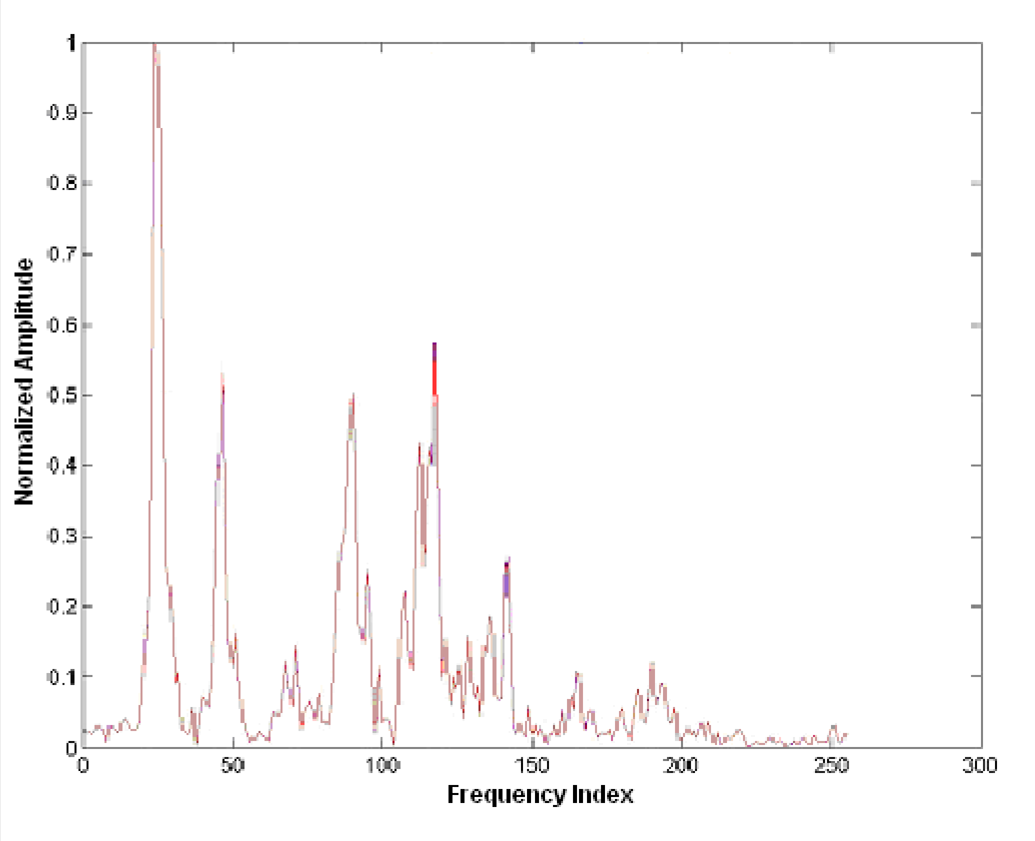
Freq. Domain Methods

In the frequency domain, polyphonic detection is possible, usually utilizing the Fast Fourier Transform (FFT) to convert the signal to a frequency spectrum. This requires more processing power as the desired accuracy increases, although the well-known efficiency of the FFT algorithm makes it suitably efficient for many purposes.

Popular frequency domain algorithms include: the harmonic product spectrum[4]; cepstral analysis and maximum likelihood which attempts to match the frequency domain characteristics to pre-defined frequency maps (useful for detecting pitch of fixed tuning instruments); and the detection of peaks due to harmonic series[5].

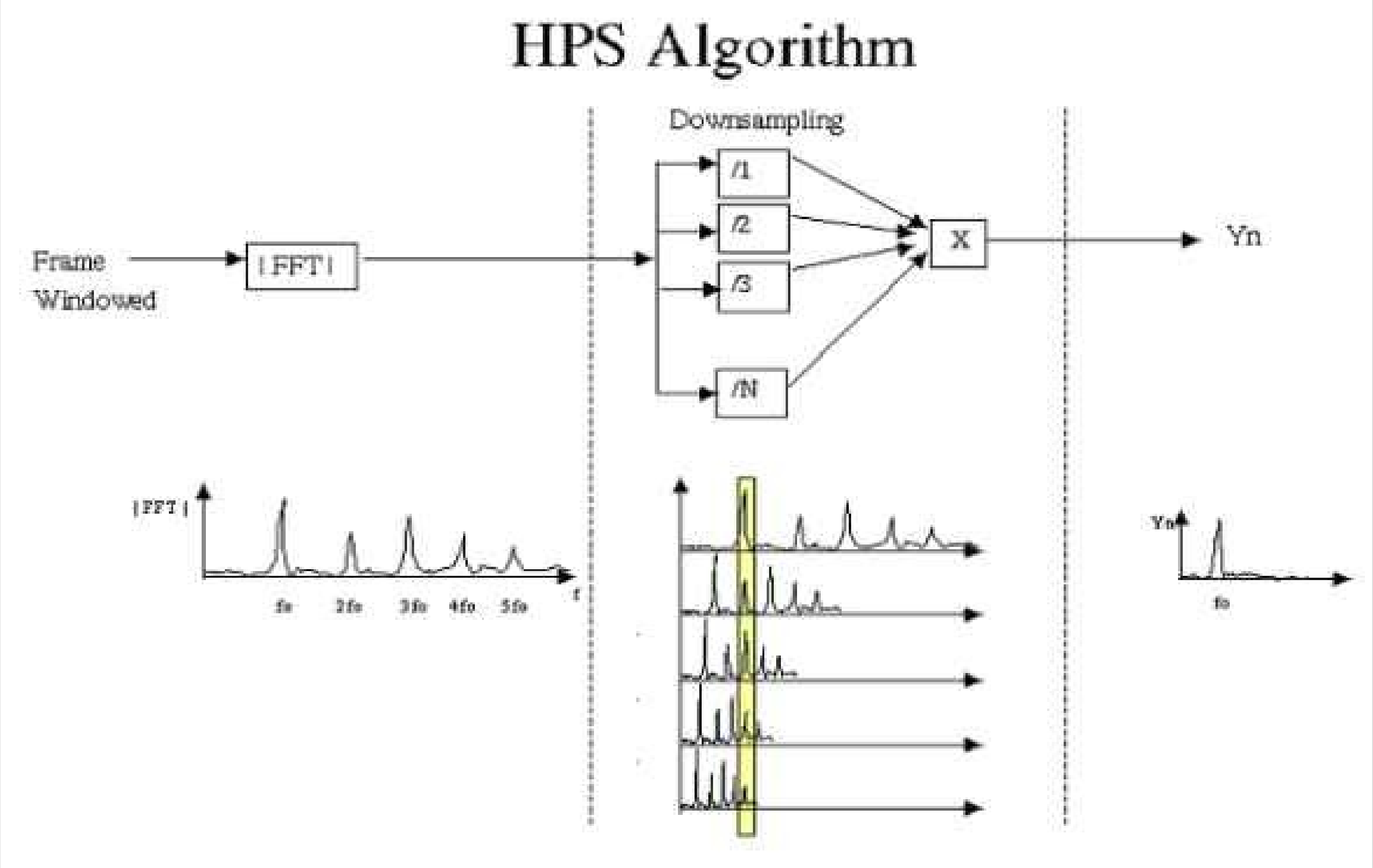
(from wikipedia)

Spectral Peak Picking



Picking up 2-10 harmonics, checking their position, merging the estimated values of Pitch..
Problem: missing harmonics, misplaced harmonics, sources in the background of speech, etc..

Harmonic Product Spectrum



Problem: downsampling reduces frequency resolution!!
remember: finite resolution of STFT (>30Hz/bin)

Thoughts about Pitch Estimation

- **Time-domain representation..**
  - Average Different Magnitude Function (ADMF)
  - Autocorrelation (ACF)
  - Center-clipped autocorr, Zero crossing.

- **Frequency-domain representation..**
  - “four-harmonic” method
  - Harmonic product spectrum (HPS)

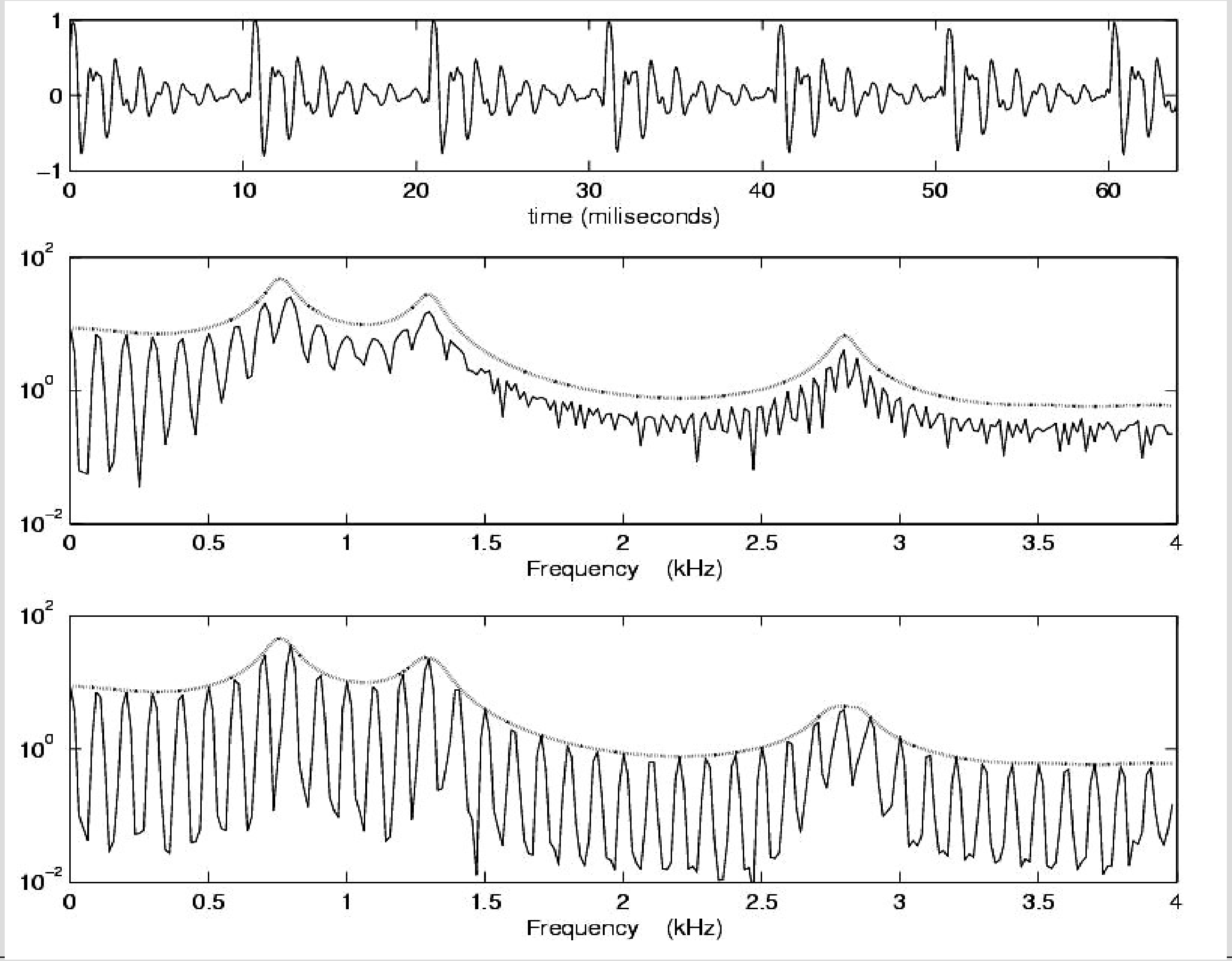
- **Quefrency-domain representation..**
  - Cepstral peak picking..

- !!! Pitch is never stationary!!!

=> which domain?

Recall: Harmonic Chirp Transform

Pitch as controlling feature of the analysis method



s(t)

Fourier transform

Harm.Chirp Trf.

$\alpha=\gamma$

The Short-Time Harmonic Chirp Transform

Discrete-time definition:

$$S[m,k] = \sum_{n=0}^{N-1} s[n+mM]\, w[n]\, \xi_N[n,k,\hat{\alpha}_m]$$

..***w*** is the analysis window, *M* is the time-domain stepsize.

$$\xi_N[n,k,\hat{\alpha}_m] = e^{j\frac{2\pi}{N}k(1+\hat{\alpha}_m(n-N))n}$$

$$k = \left[-K(\hat{\alpha}_m), \ldots, 0, \ldots, K(\hat{\alpha}_m)\right] \qquad K(\hat{\alpha}_m) = \frac{N/2}{1+|\hat{\alpha}_m|N}$$

where can we get $\hat{\alpha}_m$*?*

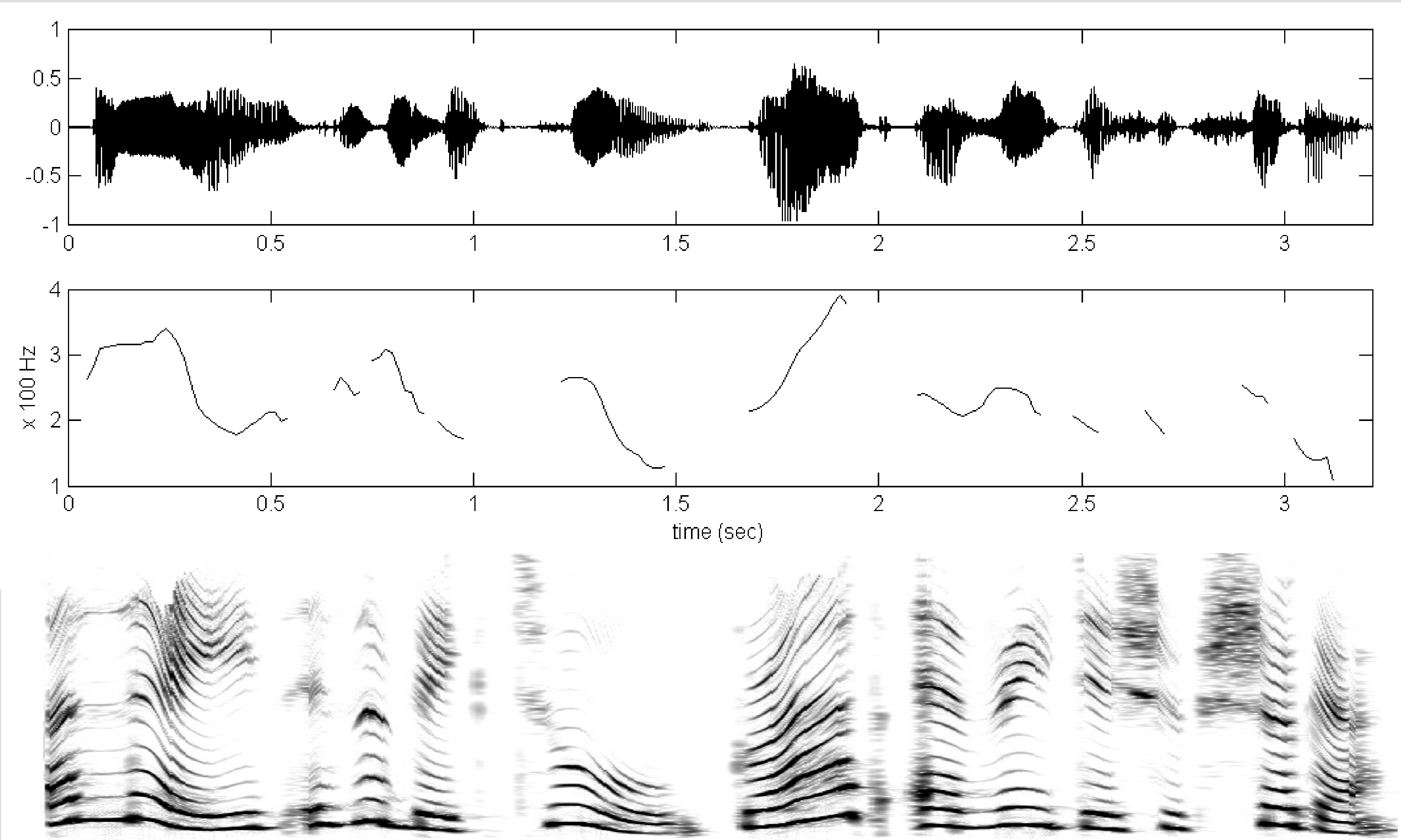
Chirp-rate estimation

- The frequency variation-rate is derived from the pitch trajectory.
- The pitch is estimated from the corresponding spectral frame.
- The chirp rate is finally computed as

$$\hat{\alpha}_m \propto \frac{\Delta f_o[m]}{f_o[m]}$$

Real speech examples

The Short-Time Harmonic Chirp Transform



Gathering the short-time spectrum

- What makes the perception in noisy environment robust?
- We introduce the "spectral gathering"...

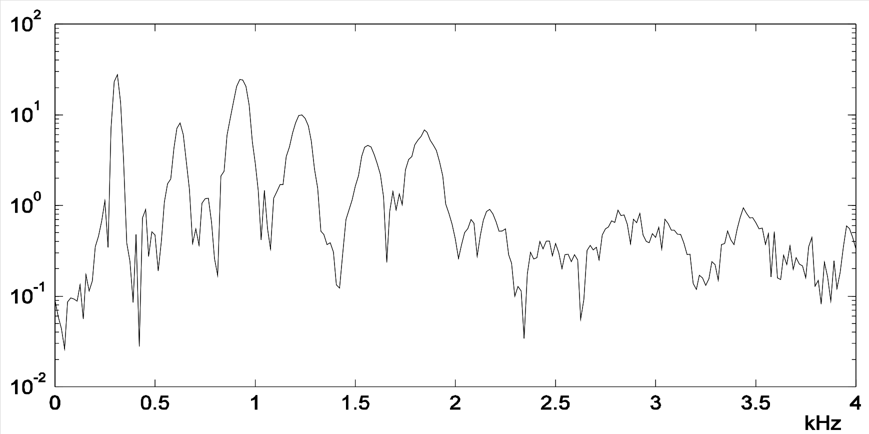
$$\rho_0(f_0) = \frac{1}{H} \sum_{h=1}^{H} \log_{10}[S(hf_0)]$$

Fo .. position of the highest peak:

$$F_0 = \arg\max_{f_0} \rho(f_0)$$

Limits of Frequency Domain Pitch Estimation

- **Problem 1:**
  - finite resolution of STFT (>30Hz/bin)



Applying linear interpolation before the gathering:

$$S(f_0) = (1-d)\; S(\hat{k}_0) + d\; S(\hat{k}_0 + 1)$$

$$d = k_0 - \hat{k}_0$$

Limits of Frequency Domain Pitch Estimation

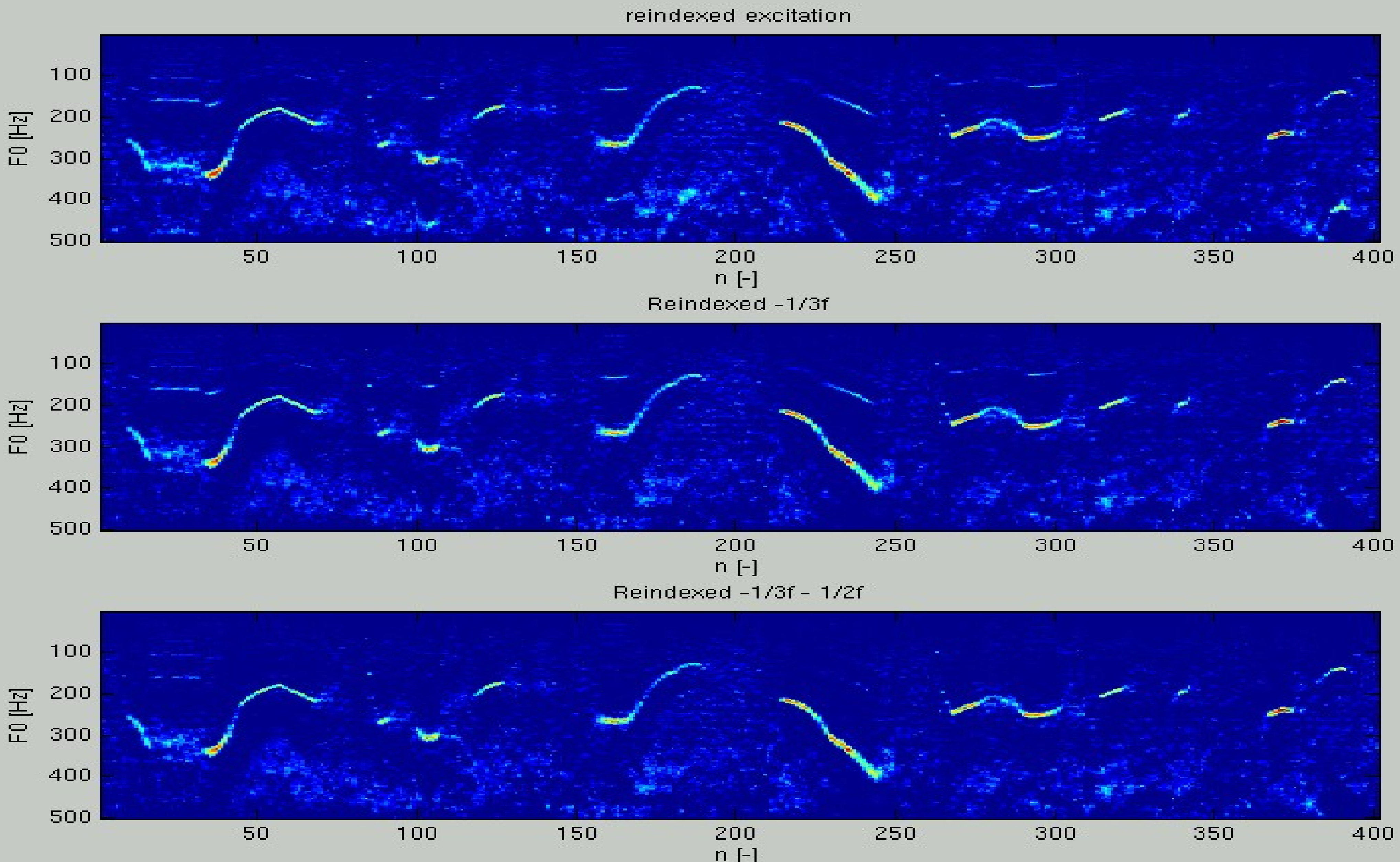
- **Problem 2:**
  - extra peaks present in GlogS.

We recognise that *fo* is NOT the real *Fo* if a GlogS(*f*/*q*) is comparacle to GlogS(f), q=1,2,3,..

Let's define a transformation to remove the anomalies:

$$\rho(f) = \rho_0(f) - \max_q \{ \rho_0(f/q) \}$$

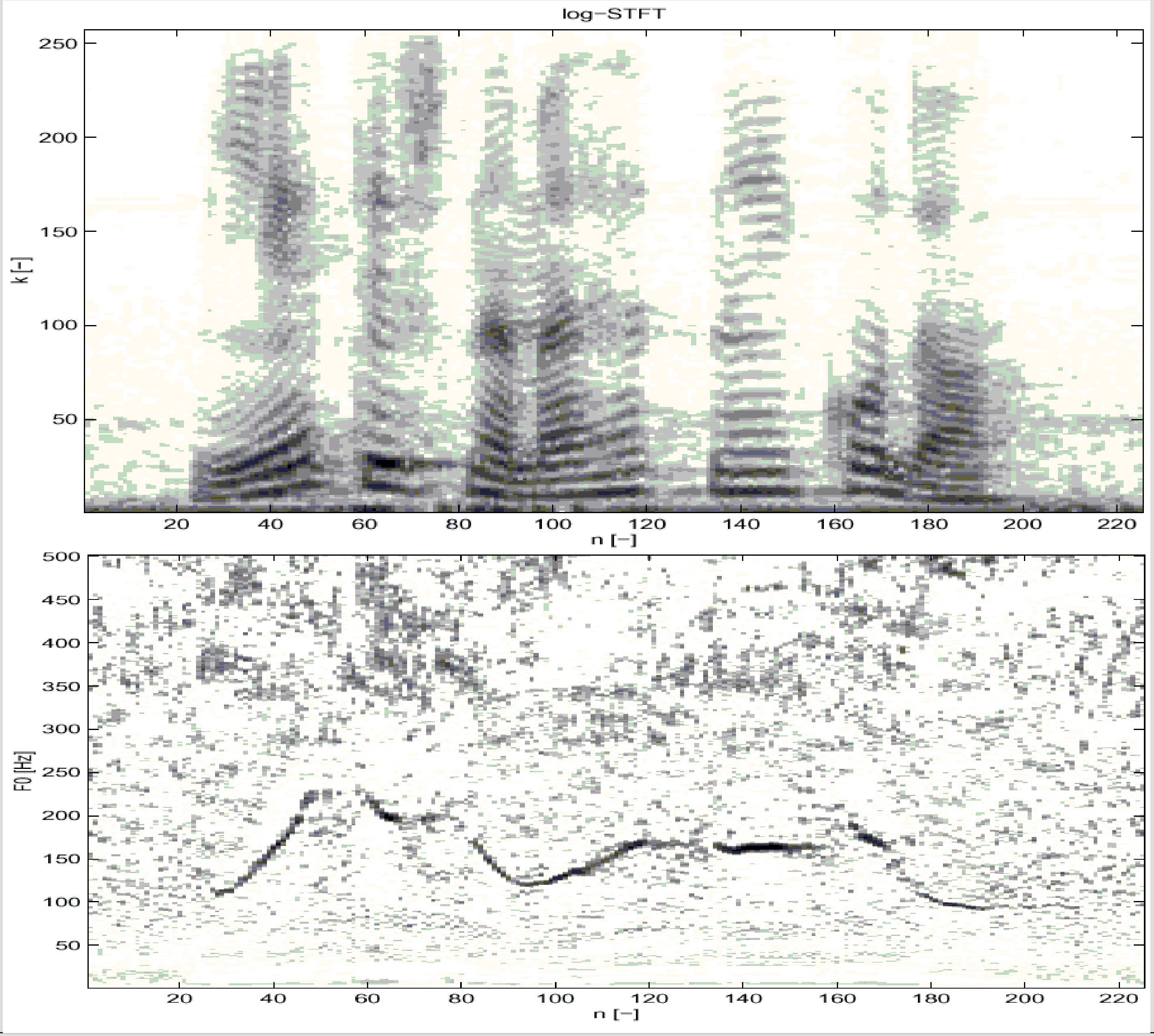
Example of Frequency Domain Pitch Estimation



STFT + GlogS

Real speech examples

STFT + GlogS


log−STFT
k [-]
250
200
150
100
50
20
40
60
80
100
120
140
160
180
200
220
n [−]
F0 [Hz]
500
450
400
350
300
250
200
150
100
50
20
40
60
80
100
120
140
160
180
200
220
n [−]

Boosting Up the Pitch Estimation

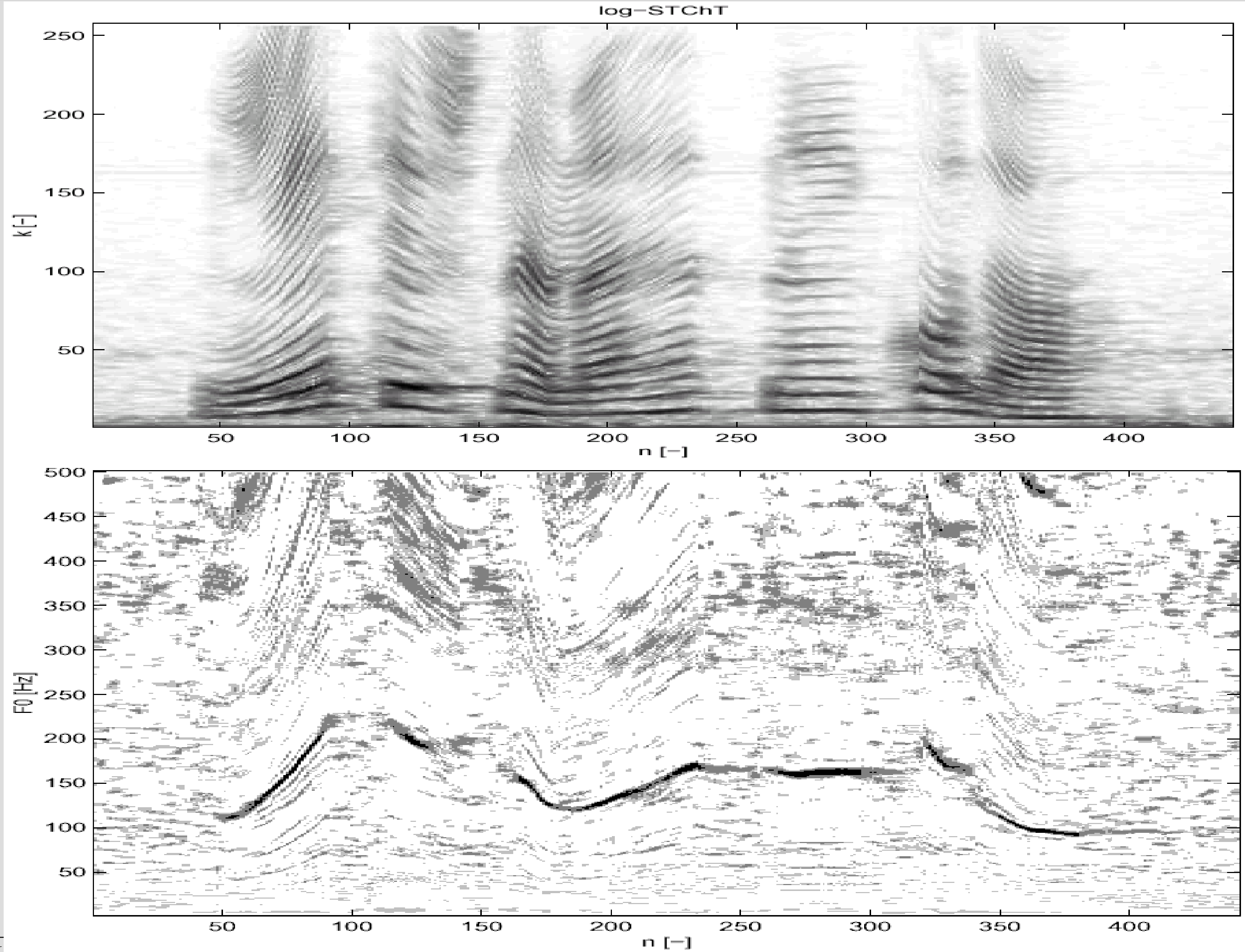
- Application of the Short-time Harmonic Chirp Transform (STHChT) instead of STFT:

$$C_x(n,k,\alpha_n) = \sum_{m=0}^{N-1} x[m+nM]\, w[m]\, \xi(m,k,\alpha_n)^*$$

- The output of the pitch-tracking is fed to the STChT to calculate the "Pitch-change-rate":

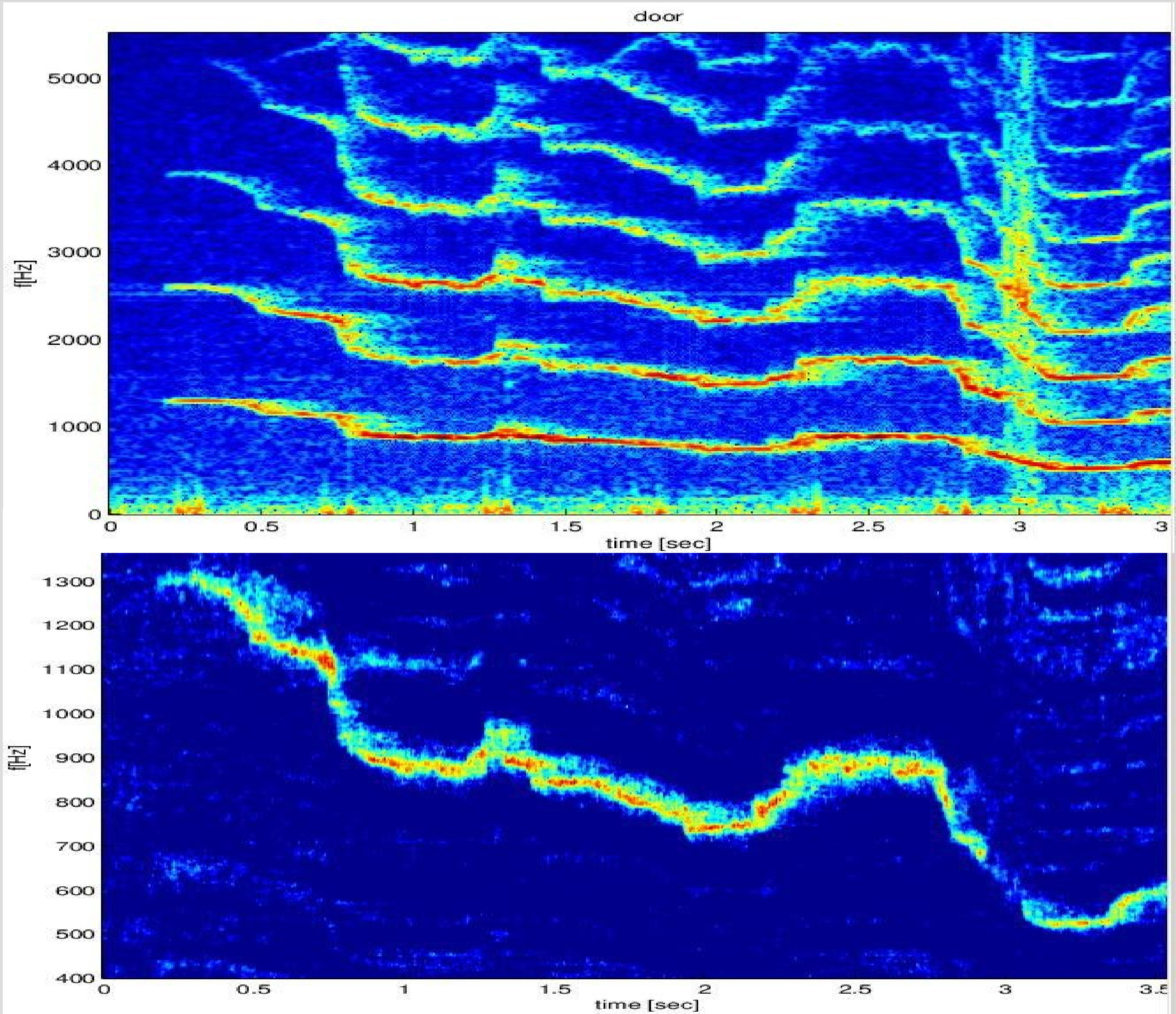
$$\alpha_n = \frac{\triangle F_0}{F_0\, M} = \frac{2(F_n - F_{n-1})}{M(F_n + F_{n-1})}$$

STHChT + GlogS



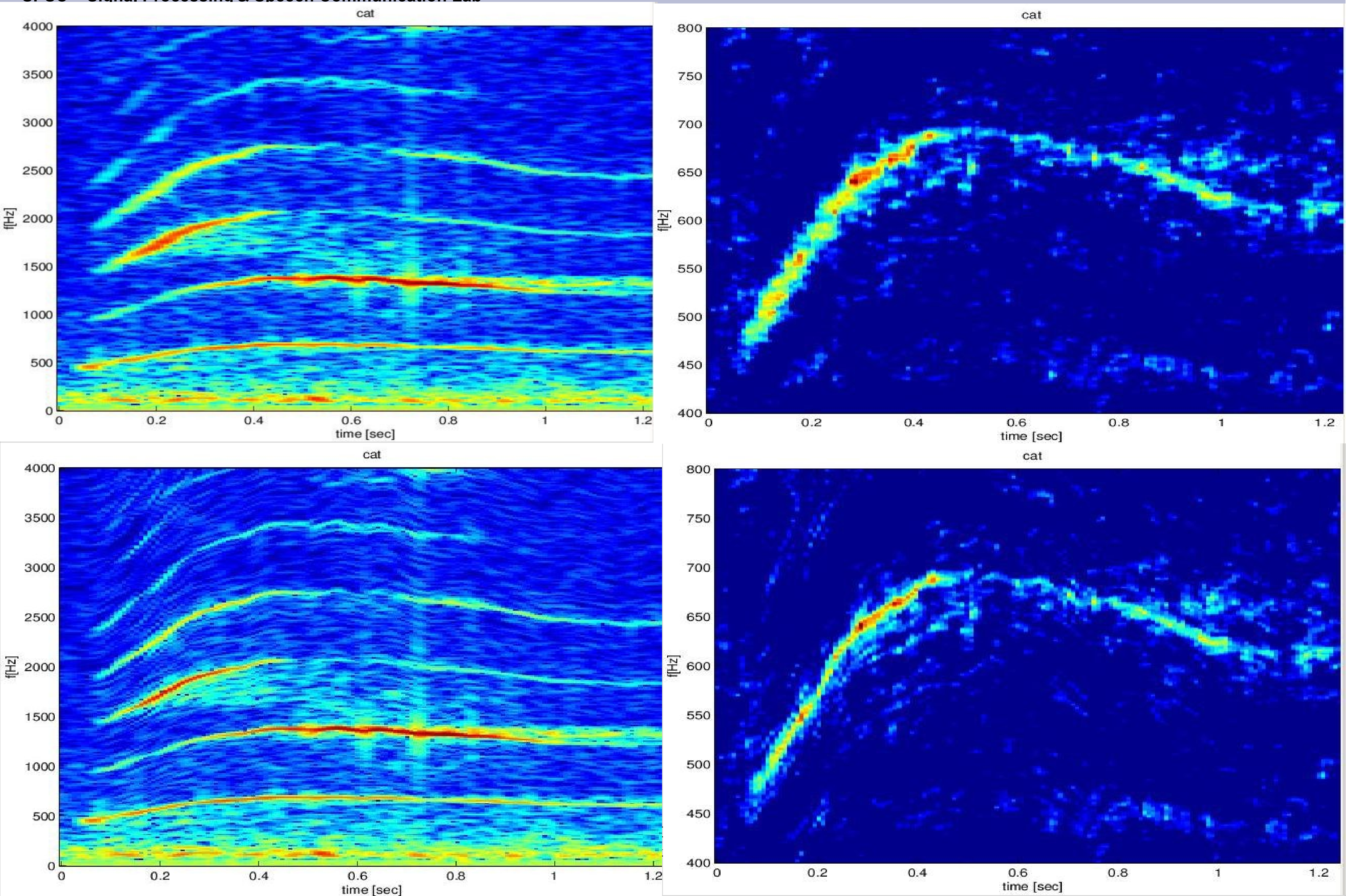
Real world examples

Closing door : STFT + GlogS

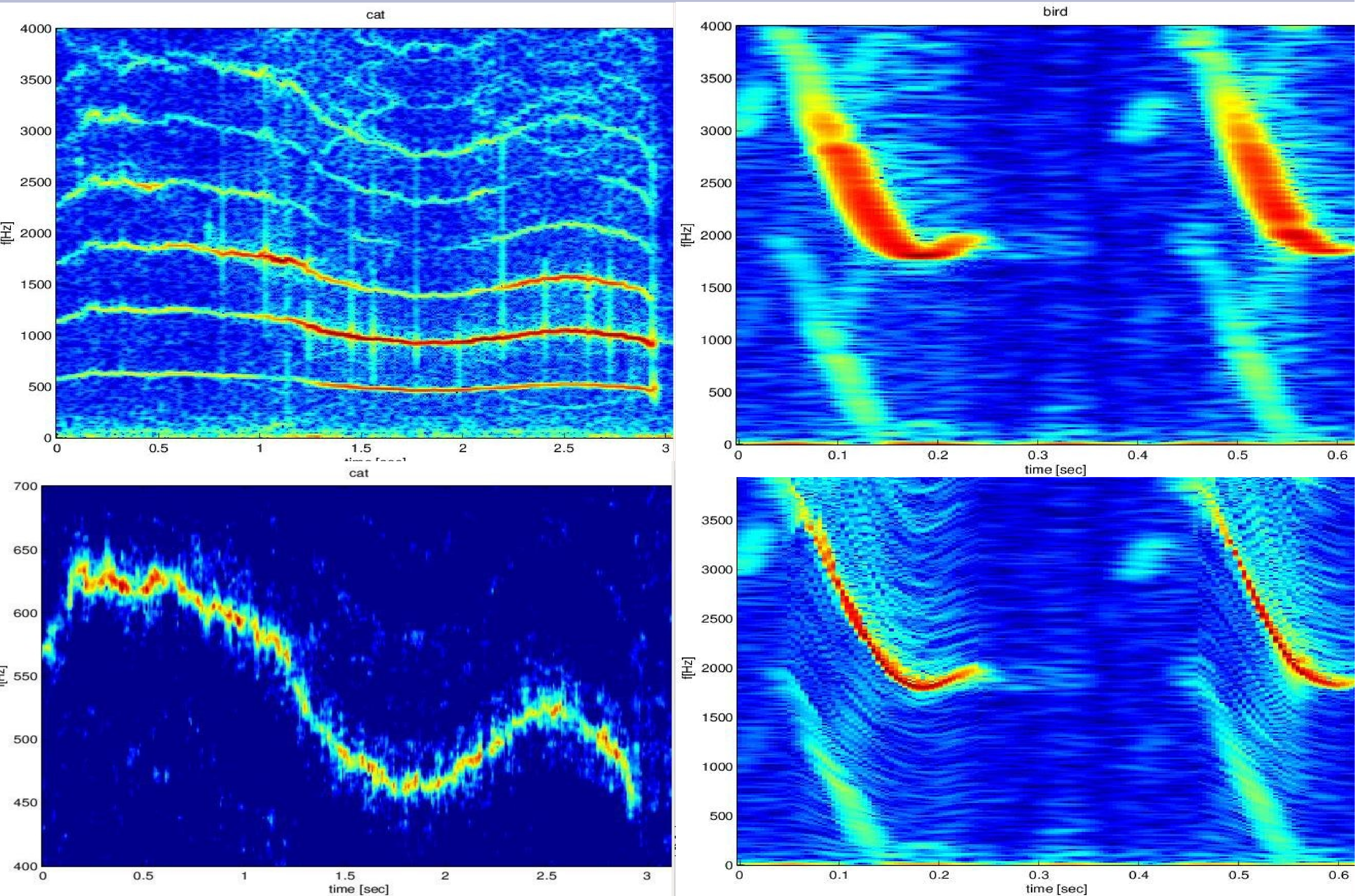


Cat: STFT / STHChT + GlogS

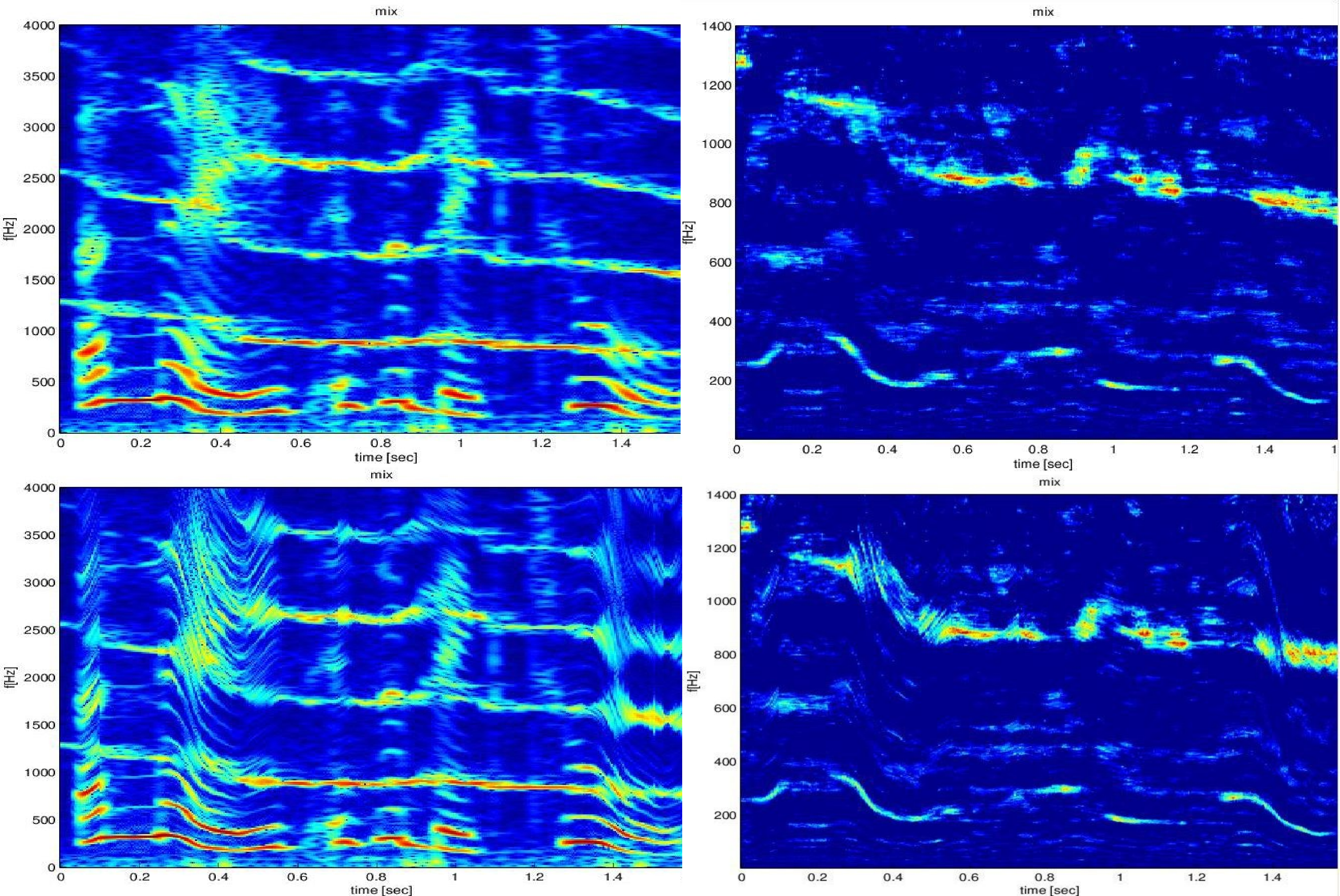
SPSC – Signal Processing & Speech Communication Lab



Cat: STFT + GlogS and Birds: STFT vs. STHChT



STHChT + GlogS



Voice Activity Detection (VAD)

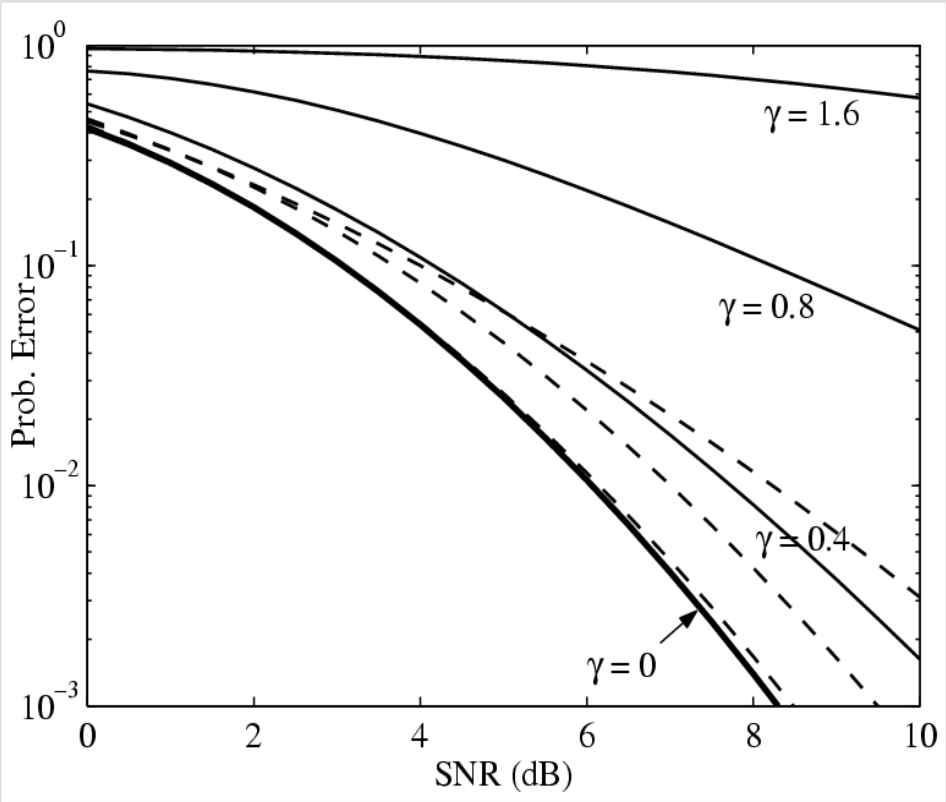
Speech/non-speech classification is a fundamental part in many speech processing algorithms and applications

Applications of VAD:
- Voice transmission (digital telephony)
- Voice control (voice-dialing on mobile phone)
- Disctation systems (ASR)
- Speech enhancement (calling while driving)

Problems:
- High level background noise
- speakers in the background
- low-energy speech signal
- distorted recordings(low SNR => few bits)

Voice Activity Detection (VAD)

Most of the speech utterances are voiced sounds, that is, signals with harmonic spectral structure. Since GlogS is robust against additive noise ...

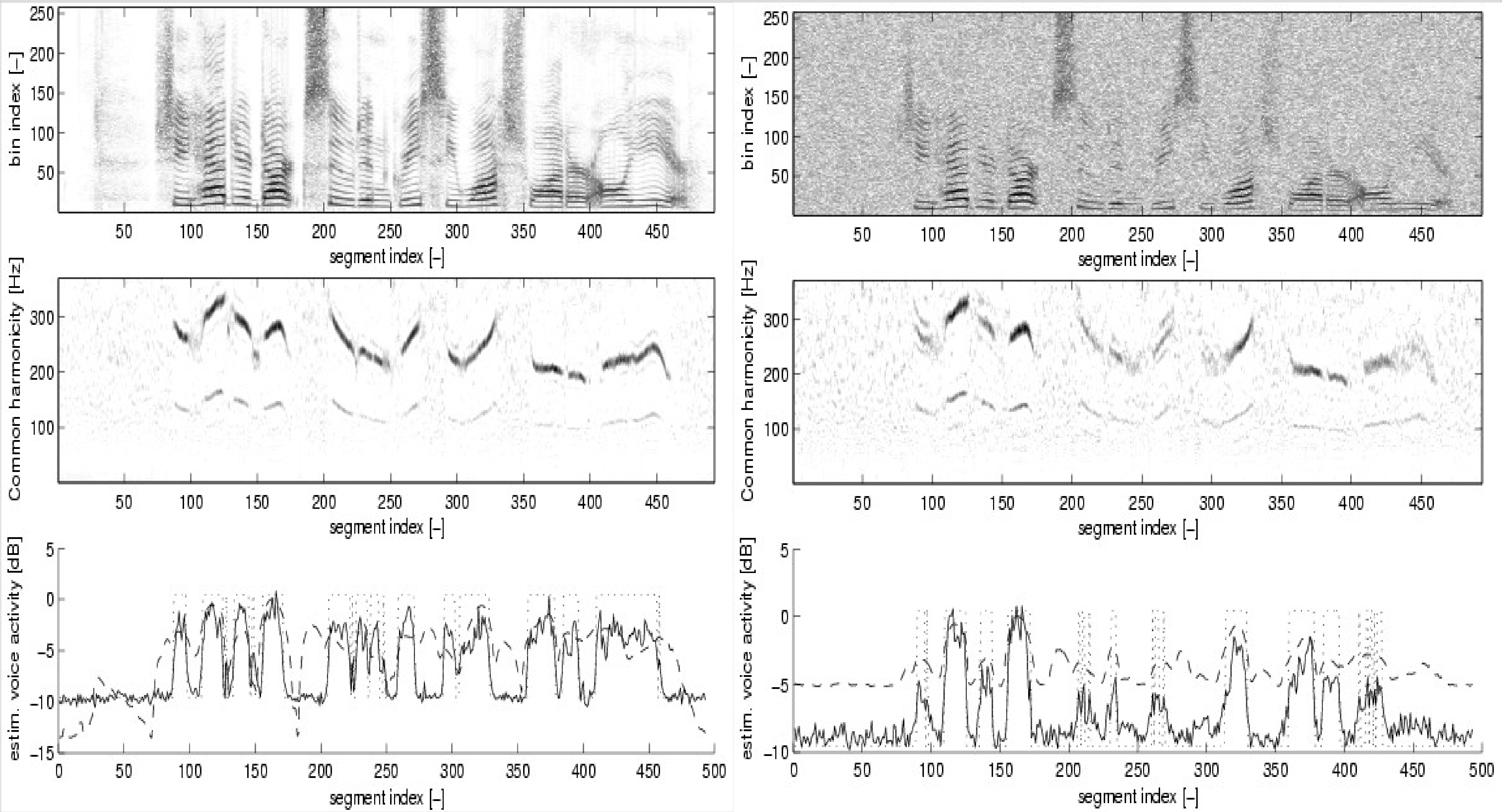


.. a simple thresholding could be applied on the normalized $\rho(f)$. Then, the 15ms/200ms could be used to "bridge" short voice activity regions.

$$\rho_0(f_0) = \frac{1}{H}\sum_{h=1}^{H}\log_{10}[S(hf_0)]$$

$$\rho(f) = \rho_0(f) - \max_q\{\,\rho_0(f/q)\,\}$$

Voice Activity Detection (VAD)



Conclusion

Event detection vs. short-term methods diucussed.
Applications of pitch estimation discussed.
Problems of semi-periodicity and multipitch discussed.
Semi-periodic speech model for pitch estim. introduced.
Harmonic Chirp transform -related pitch estimation algorithms discussed.
Fast implementation discussed.
Real-world recordings used for demonstrations.

Thank you!