Lecture 9: Time & Pitch Scaling

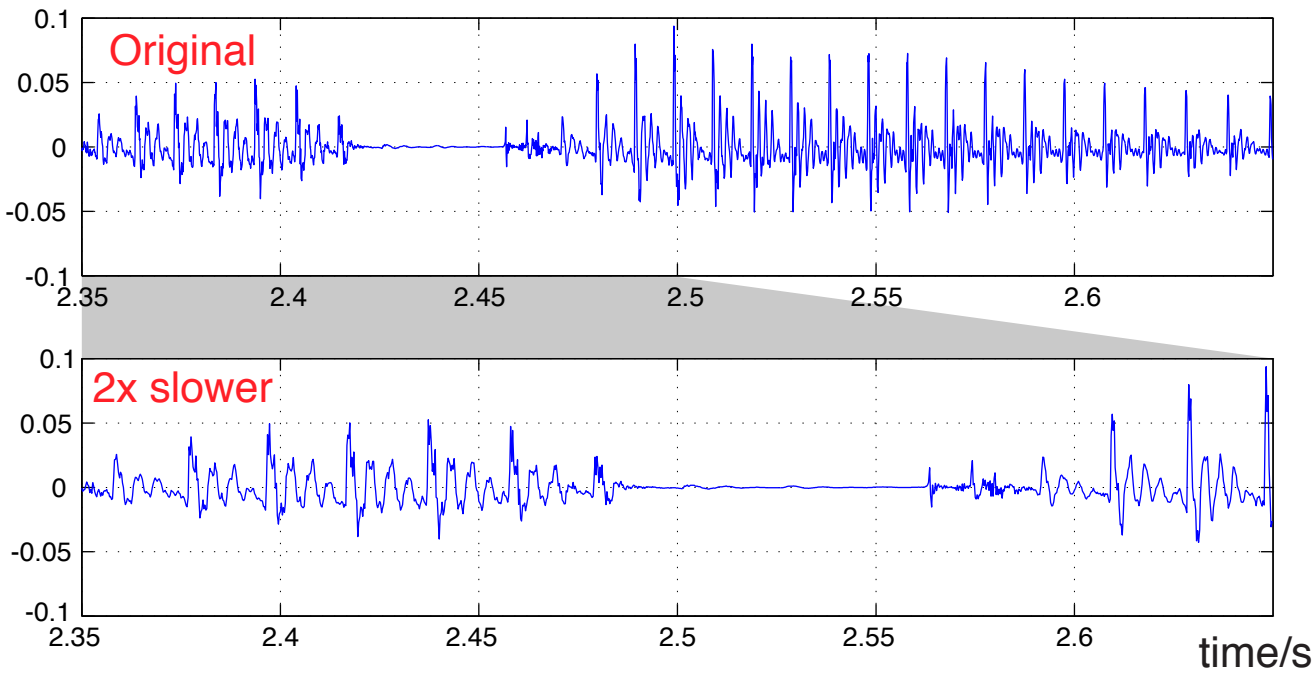
1. Time Scale Modification (TSM)
2. Time-Domain Approaches
3. The Phase Vocoder
4. Sinusoidal Approach

Dan Ellis
Dept. Electrical Engineering, Columbia University
dpwe@ee.columbia.edu http://www.ee.columbia.edu/~dpwe/e4896/

1. Time Scale Modification (TSM)

- The basic problem in time scaling:
  Modify a sound to make it "quicker/slower"?
  - to examine "detail" in speech/performance
  - to synchronize tracks

- Why not just change the sampling rate?
  - "slowing the tape"
  - e.g. for r times longer (slower), $x_s(t) = x(t/r)$

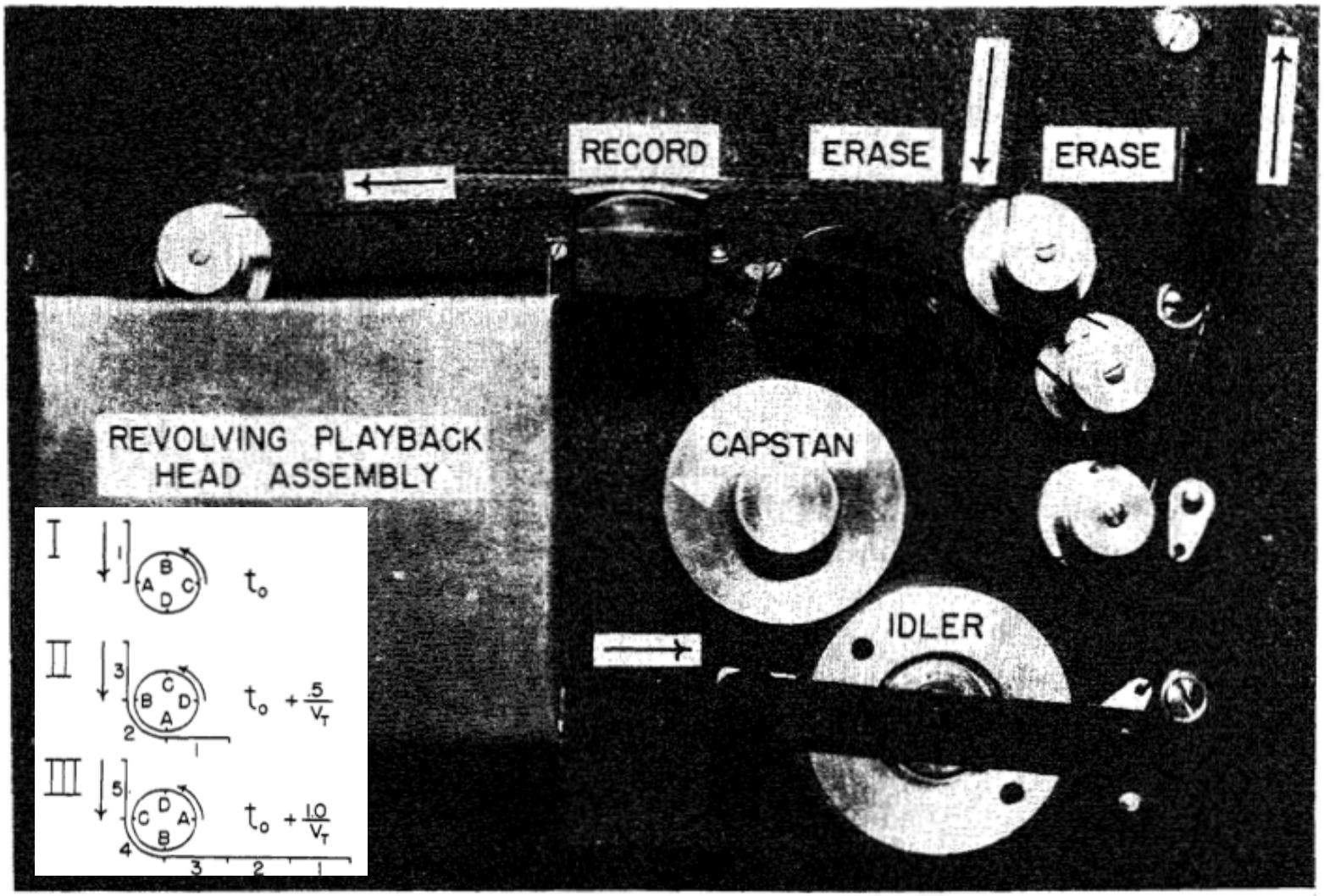


Time & Pitch

- Changing speed alters time AND pitch
  - how do we change just one?
- Preserve pitch, change time
  - preserve pitch = keep local time structure
    but changing global time course?
- Preserve timing, change pitch
  - analogous problem
  - change time scale, then change sampling rate

2. Time-Domain TSM

- Pre-digital time/pitch scaling:
  Revolving tape heads

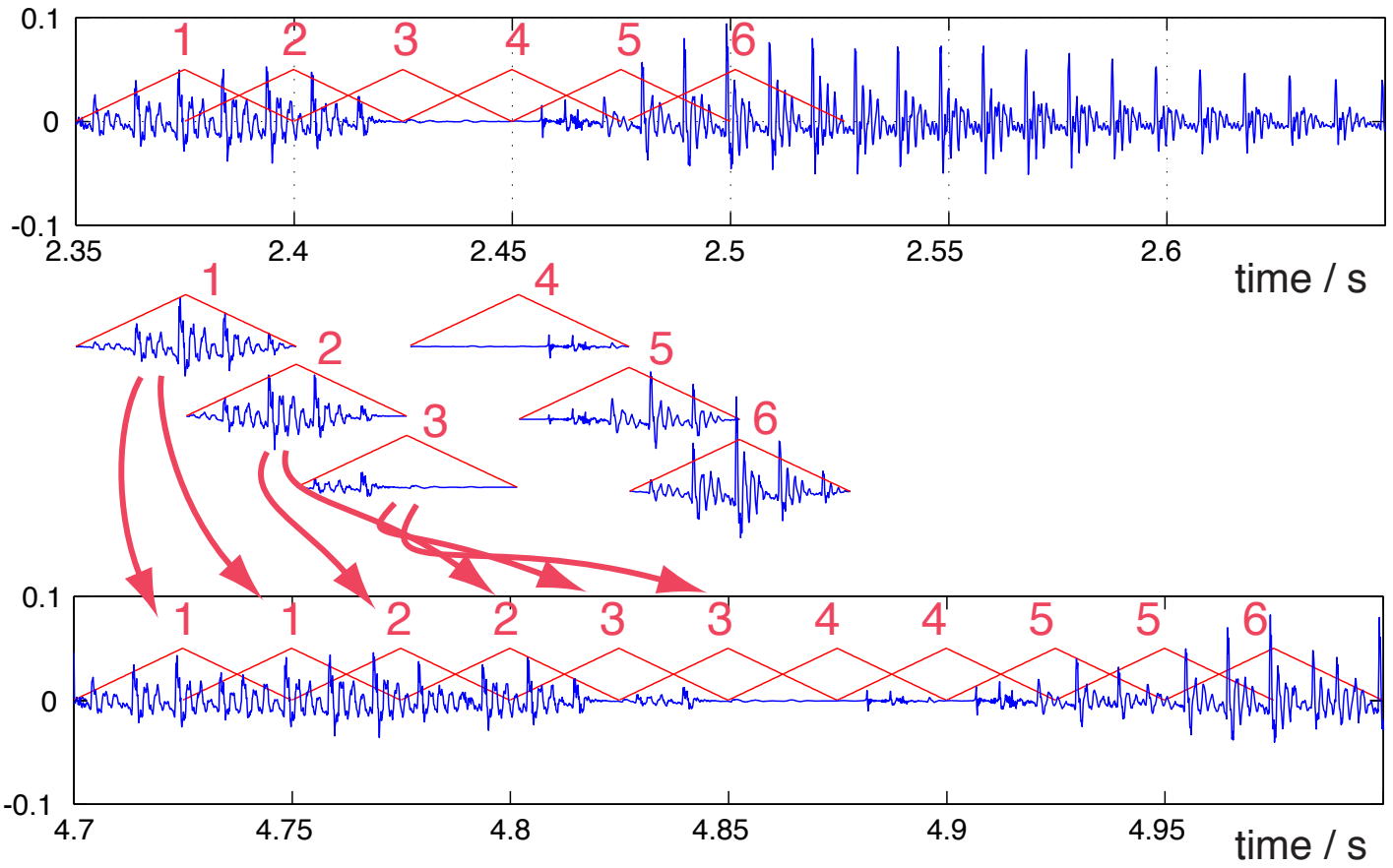


Fairbanks et al., 1954

Simple OLA TSM

- The digital equivalent of revolving tape heads

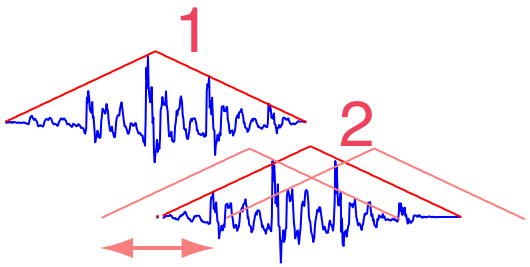
$$y^m[mL+n] = y^{m-1}[mL+n] + w[n] \cdot x\left[\left\lfloor \frac{m}{r} \right\rfloor L + n\right]$$



SOLA / SOLAFS

Roucos & Wilgus 1985
Hejna & Musicus 1991

- Simple OLA leads to random phase interactions during overlap
  - so.. search for optimal alignment K_m via small offset



K_m maximizes alignment of 1 and 2

$$y^m[mL+n] = y^{m-1}[mL+n] + w[n] \cdot x\left[\left\lfloor \frac{m}{r} \right\rfloor L + n + K_m\right]$$

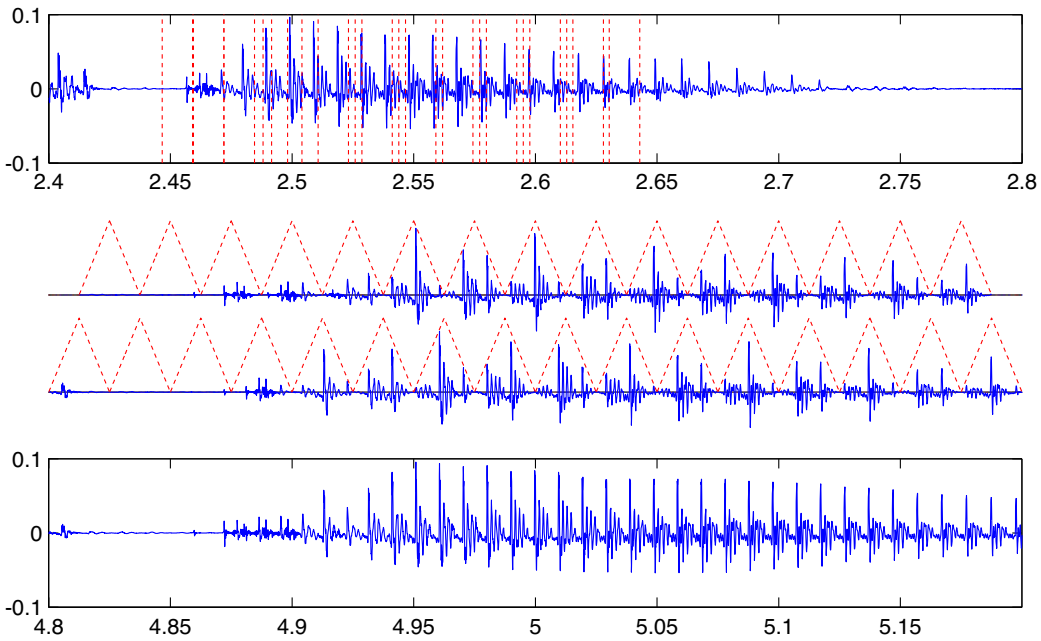
  - find best K_m with cross-correlation:

$$K_m = \arg \max_{0 \le K \le K_u} \frac{\sum_{n=0}^{N_{ov}} y^{m-1}[mL+n] \cdot x\left[\left\lfloor \frac{m}{r} \right\rfloor L + n + K\right]}{\sqrt{\sum (y^{m-1}[mL+n])^2 \sum (x\left[\left\lfloor \frac{m}{r} \right\rfloor L + n + K\right])^2}}$$

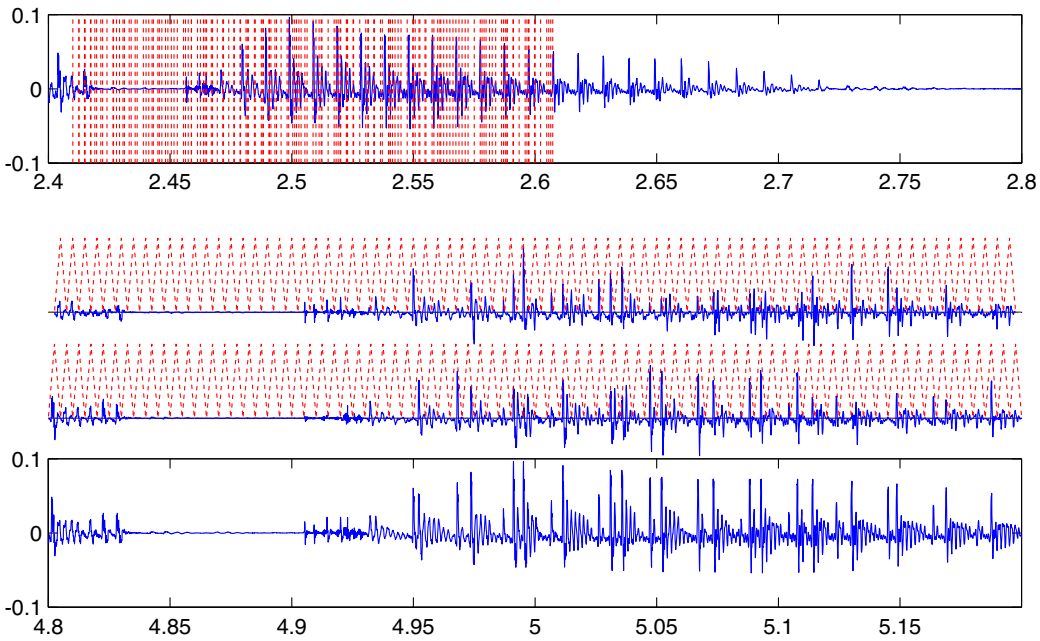
The Importance of Time Window

- Preserved "local structure"
  = structure within each window
  - shorter windows → less structure preserved

$t_w = 25ms$



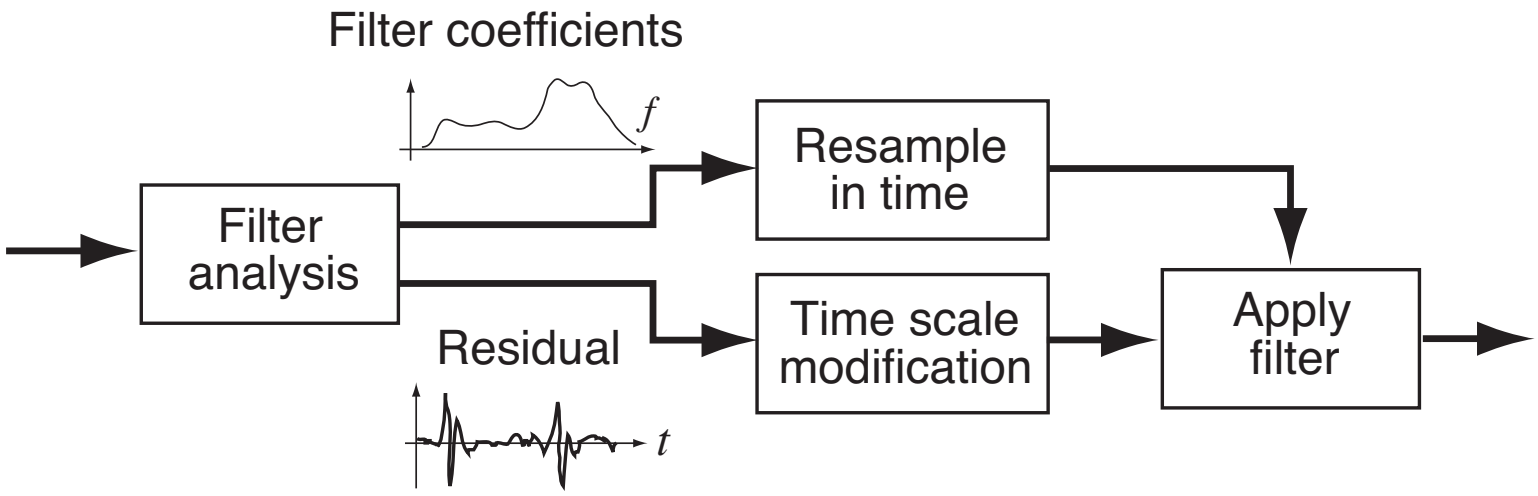
$t_w = 5ms$



Source-Filter TSM

Kawahara 1997

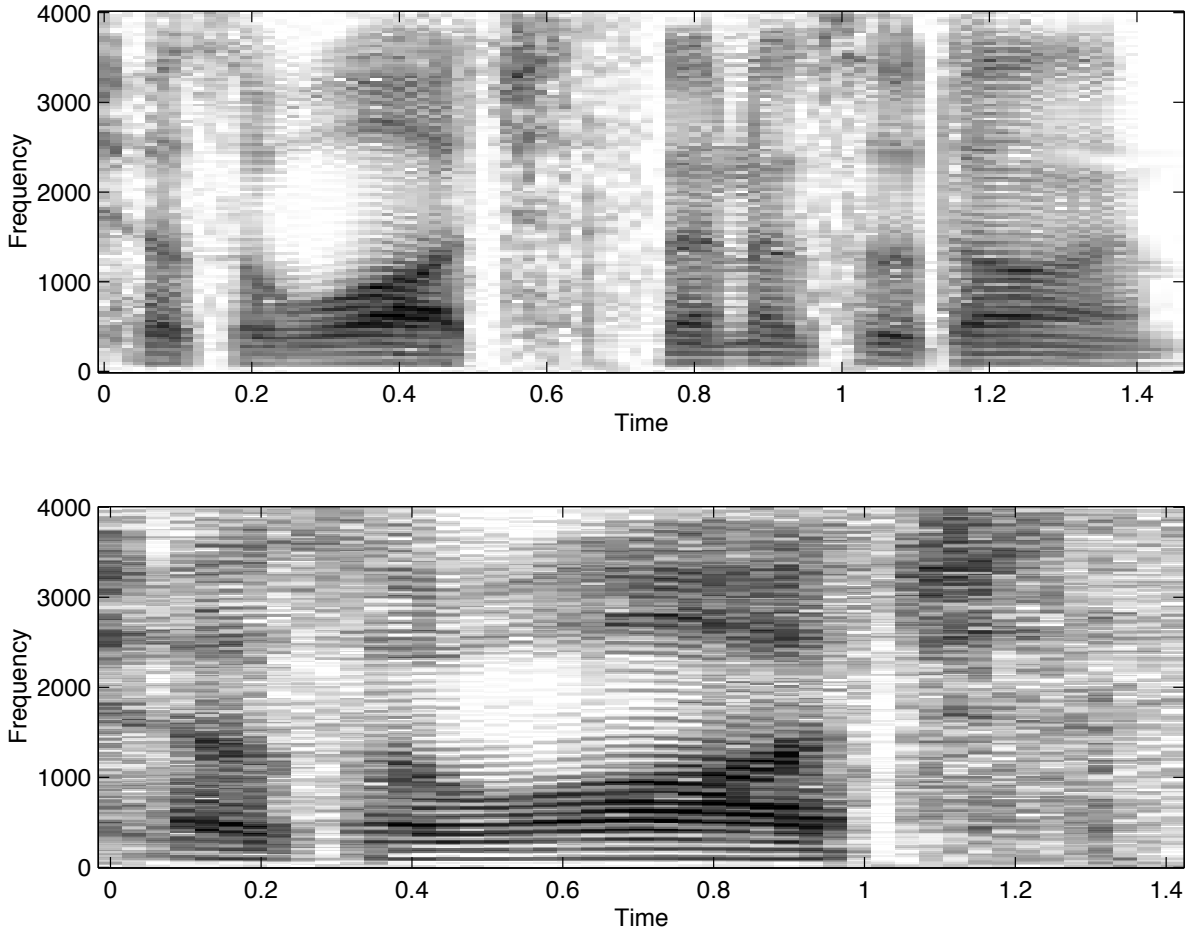
- Decompose into excitation + filter before TSM
  - filter (e.g. LPC) easy to scale - change frame rate
  - excitation scaled e.g. by SOLAFS
  - .. then reapplying filter can reduce artifacts



3. The Phase Vocoder

Flanagan & Golden 1966
Dolson 1986

- TSM based on the spectrogram



- just stretching it out it time gives "what we want":
- but: STFT magnitude isn't quite enough

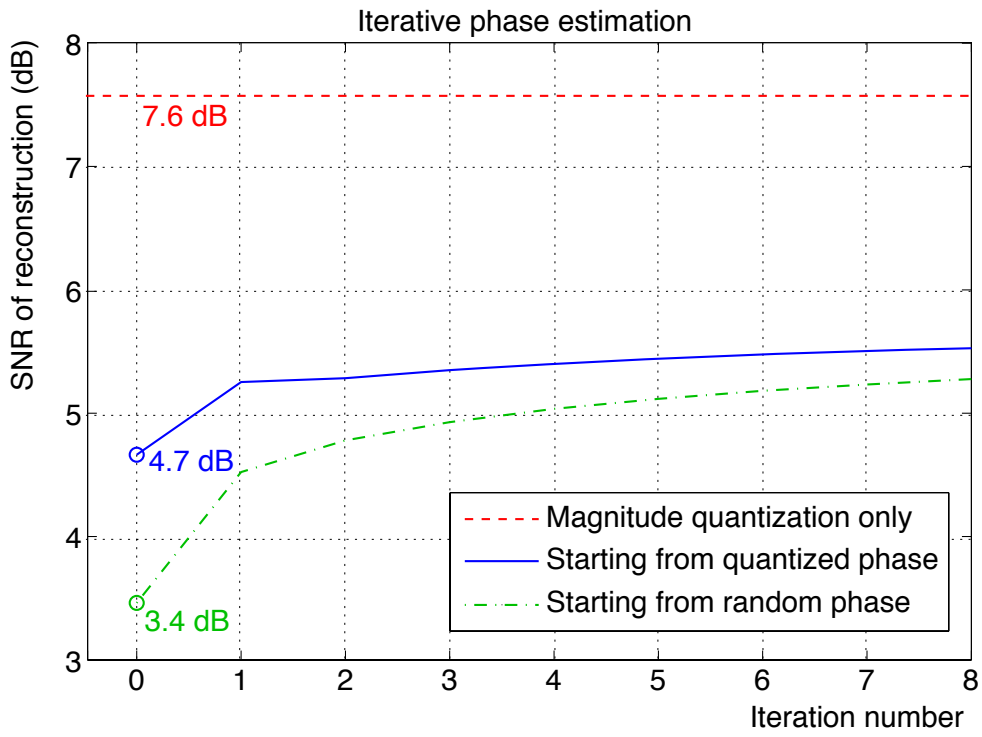
Magnitude-only reconstruction

Griffin & Lim 1984

- How to recover $STFT^{-1}\{|Y(e^{j\omega},t)|\}$?
  - need $\angle\{Y(e^{j\omega},t)\}$ ...
  - iterate

$$y^n(t) = STFT^{-1}\{|Y(e^{j\omega},t)|\phi^n(\omega,t)\}$$
$$\phi^{n+1}(\omega,t) = \angle\{STFT\{y^n(t)\}\}$$

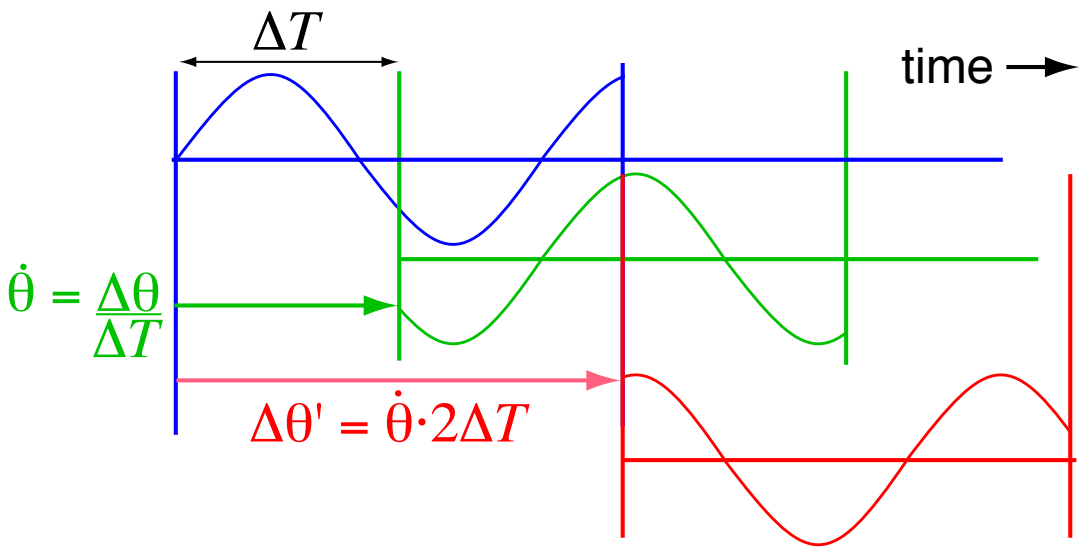
  - converges to a solution, but SLOWLY



Phase Correction

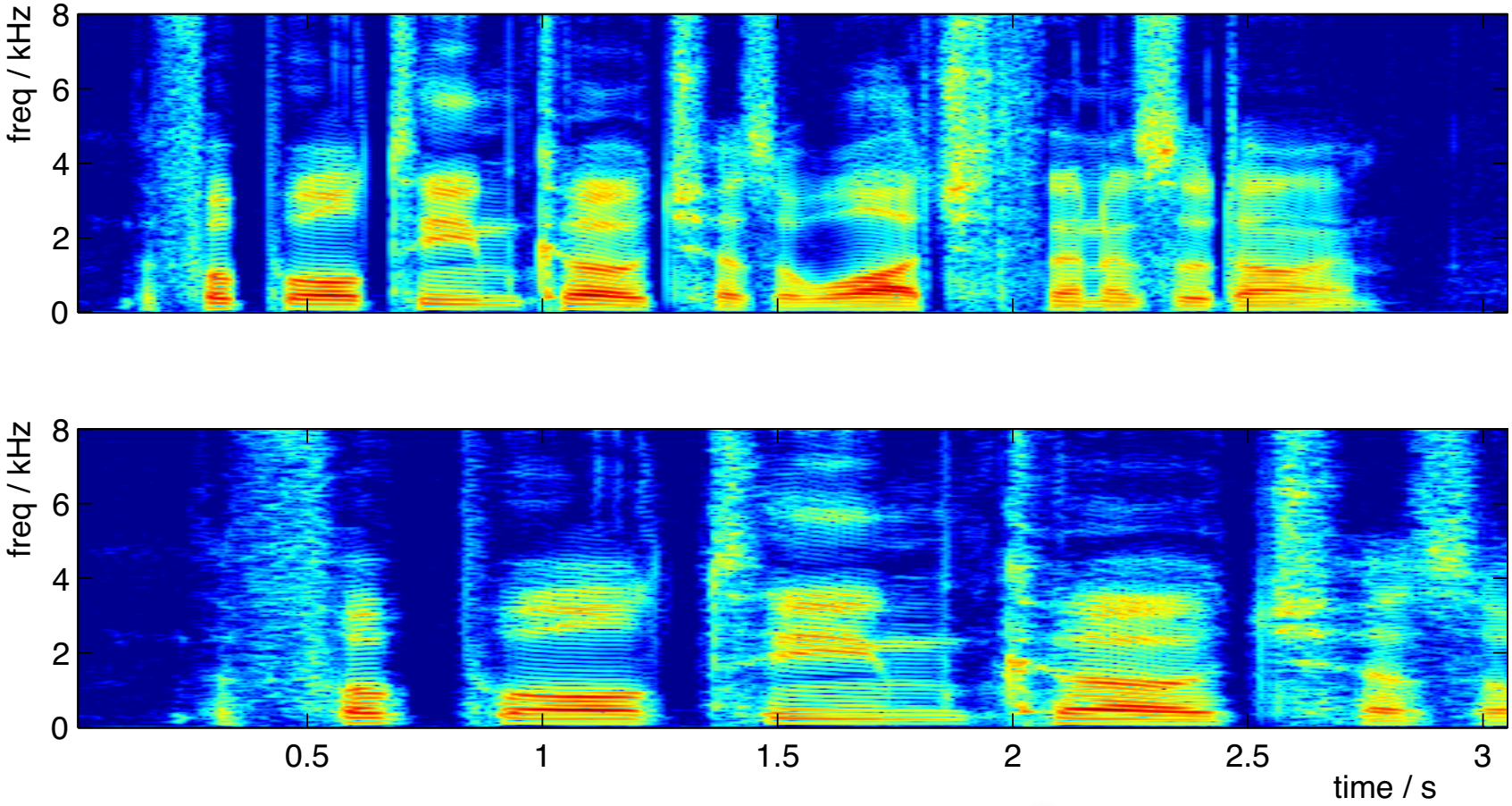
- Essential idea of the phase vocoder
  - phase advance will stretch by time scaling factor
  - ... preserve rate of phase change between slices

    = instantaneous frequency $\dot{\phi}(\omega, t) = \frac{d}{dt}\angle\{Y(e^{j\omega}, t)\}$
  - (from differencing adjacent frames, or directly...)
- Makes sense for a single sinusoid...



Phase Vocoder Results

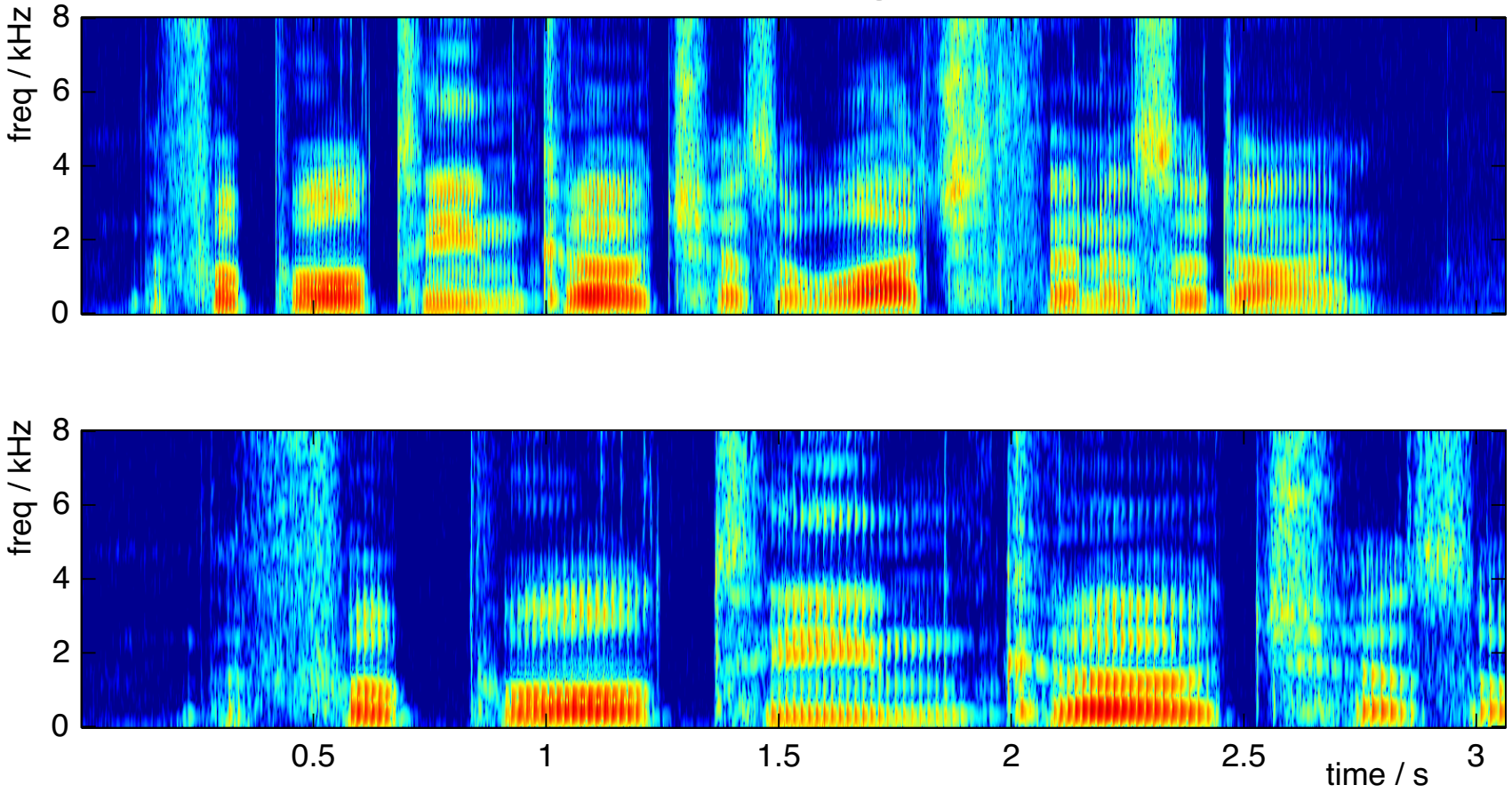
- Reconstructed phase gives smooth evolution



- "metallic" extended noise...

Time Window (again)

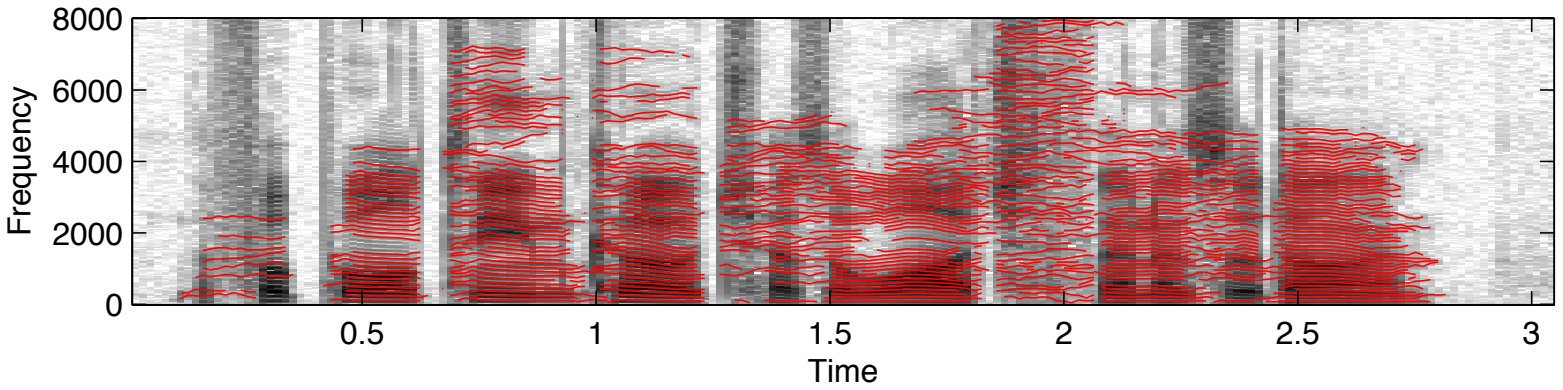
- STFT time-frequency tradeoff determines what "scaling" means



- is pitch structure within DFT window?

4. Sinusoidal TSM

- **Phase Vocoder essentially treats each STFT bin as a sinusoid**
  - with magnitude $|Y(e^{j\omega}, t)|$ and frequency $\dot{\phi}(\omega, t)$
  - time scaling simply projects that sinusoid forward
- **But: bins around peak don't behave that way**
  - need phase alignment to achieve interference
- **Model as sinusoids explicitly**
  - peak-picking, magnitude & frequency estimation
  - separate noise signal?



Summary

- Time / Pitch scale modification
  Intuitive but difficult to define

- Time domain methods
  Preserve local structure

- Spectral methods
  Elongate features in spectrogram

References

- M. Covell, M. Withgott, M., Slaney, "Mach1: Nonuniform Time-Scale Modification of Speech," *Proc. ICASSP*, vol 1, pp. 349-352, 1998.
- Mark Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14-27, 1986.
- G. Fairbanks, W. Everitt, and R. Jaeger, "Method for time of frequency compression-expansion of speech," *Transactions of the IRE Professional Group on Audio*, vol.2, no.1, pp. 7-12, Jan 1954.
- J. L. Flanagan, R. M. Golden, "Phase Vocoder," *Bell System Technical Journal*, pp. 1493-1509, November 1966.
- D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Tr. Acous,, Speech, and Sig. Proc.,* vol. 32, pp. 236–242, 1984.
- Don Hejna and Bruce R. Musicus, "The SOLAFS Time-Scale Modification Algorithm," *BBN Technical Report*, July 1991.
- Hideki Kawahara, "Speech Representation and Transformation using Adaptive Interpolation of Weighted Spectrum: VOCODER Revisited," *Proc. ICASSP*, vol.2, pp.1303-1306, 1997.
- Jean Laroche, "Time and Pitch Scale Modification of Audio Signals." chapter 7 in *Applications of Digitial Signal Processing to Audio and Acoustics*, ed. M. Kahrs & K. Brandenburg, Kluwer, 1998.