EE E6820: Speech & Audio Processing & Recognition

Lecture 1: Introduction & DSP

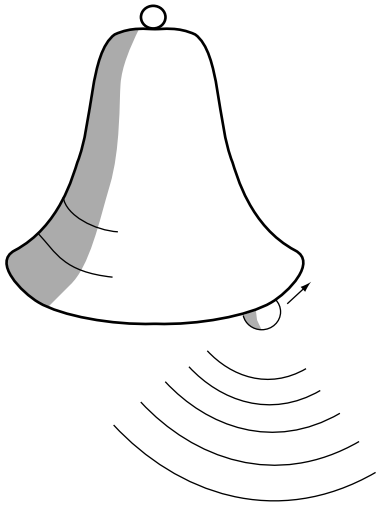
1. **Sound and information**
2. **Course structure**
3. **DSP review: Timescale modification**

Dan Ellis <dpwe@ee.columbia.edu>
http://www.ee.columbia.edu/~dpwe/e6820/

Columbia University Dept. of Electrical Engineering
Spring 2006

1 Sound and information

- **Sound is air pressure variation**

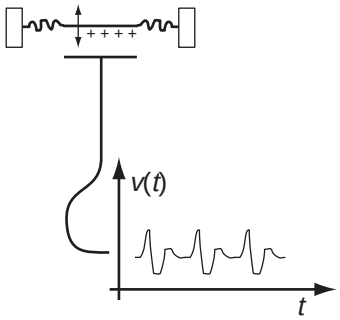


Mechanical vibration

↓

Pressure waves in air

↓



Motion of sensor

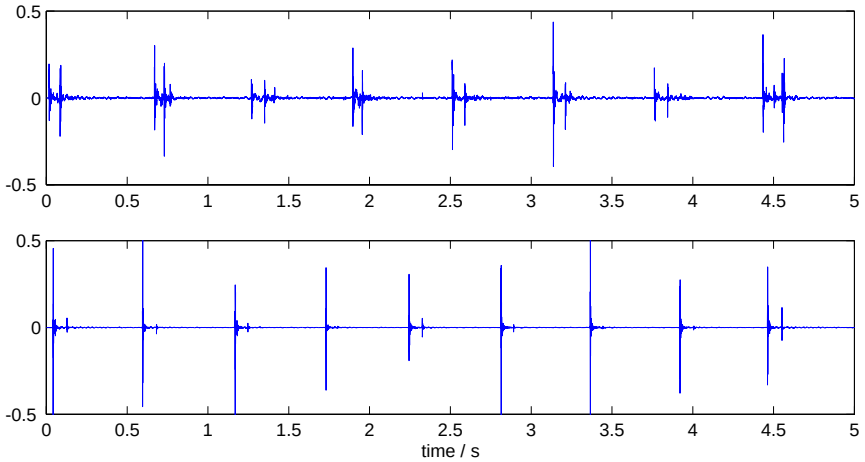
↓

Time-varying voltage

- **Transducers convert air pressure ↔ voltage**

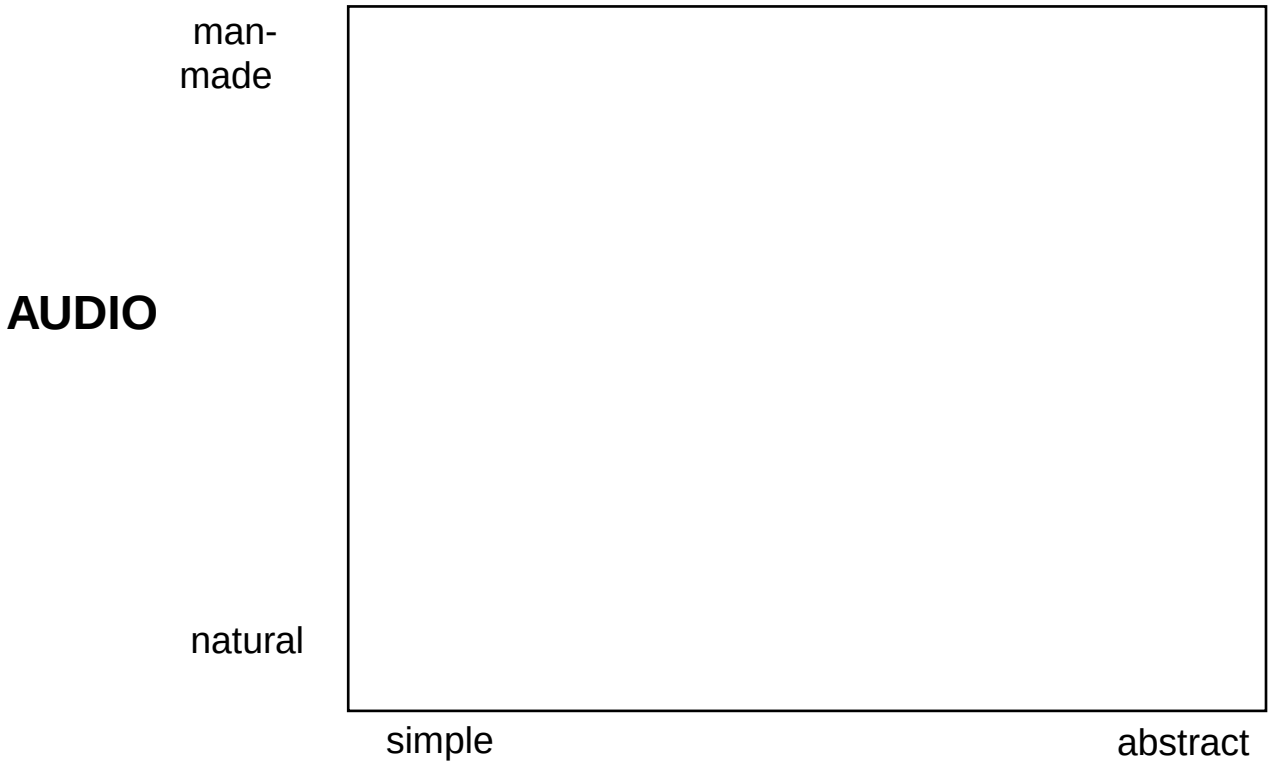
What use is sound?

- **Footsteps examples:**

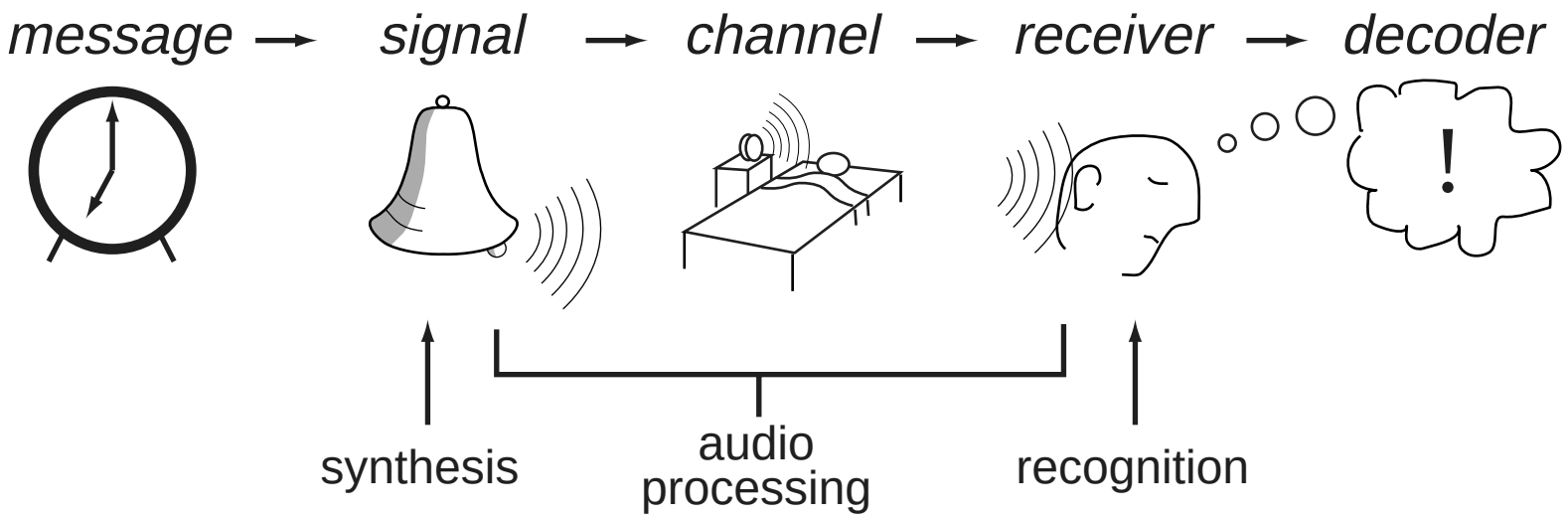


- **Hearing confers an evolutionary advantage**
  - useful information, complements vision
  - ...at a distance, in the dark, around corners
  - listeners are highly adapted to 'natural sounds' (including speech)

The scope of audio processing



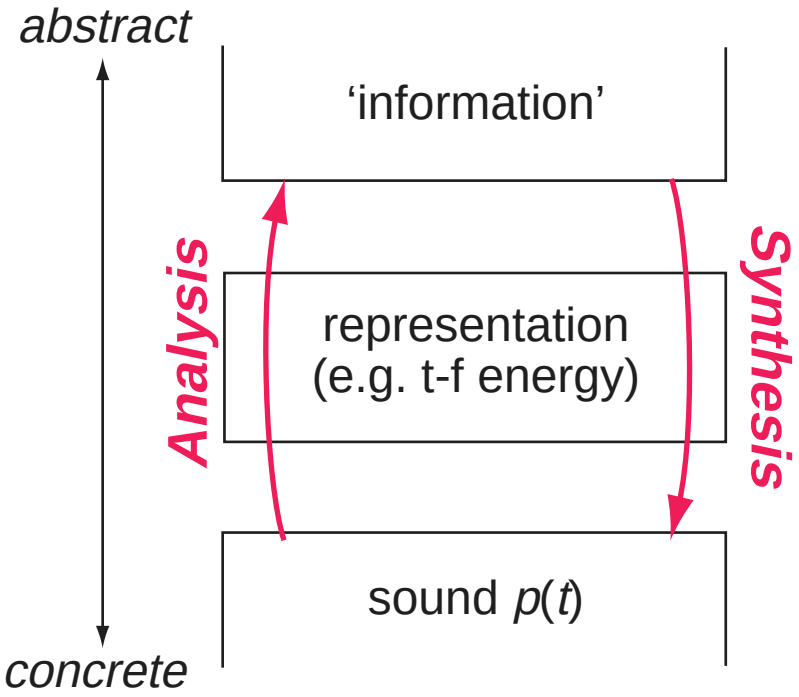
The acoustic communication chain



- **Sound is an information bearer**
- **Received sound reflects source(s) plus effect of environment (channel)**

Levels of abstraction

- **Much processing concerns shifting between levels of abstraction**



- **Different representations serve different tasks**
  - separating aspects, making things explicit ...

2 Course structure

- **Goals:**
  - survey topics in sound analysis & processing
  - develop an intuition for sound signals
  - learn some specific technologies
- **Course structure:**
  - weekly assignments (25%)
  - midterm event (25%)
  - final project (50%)
- **Text:**

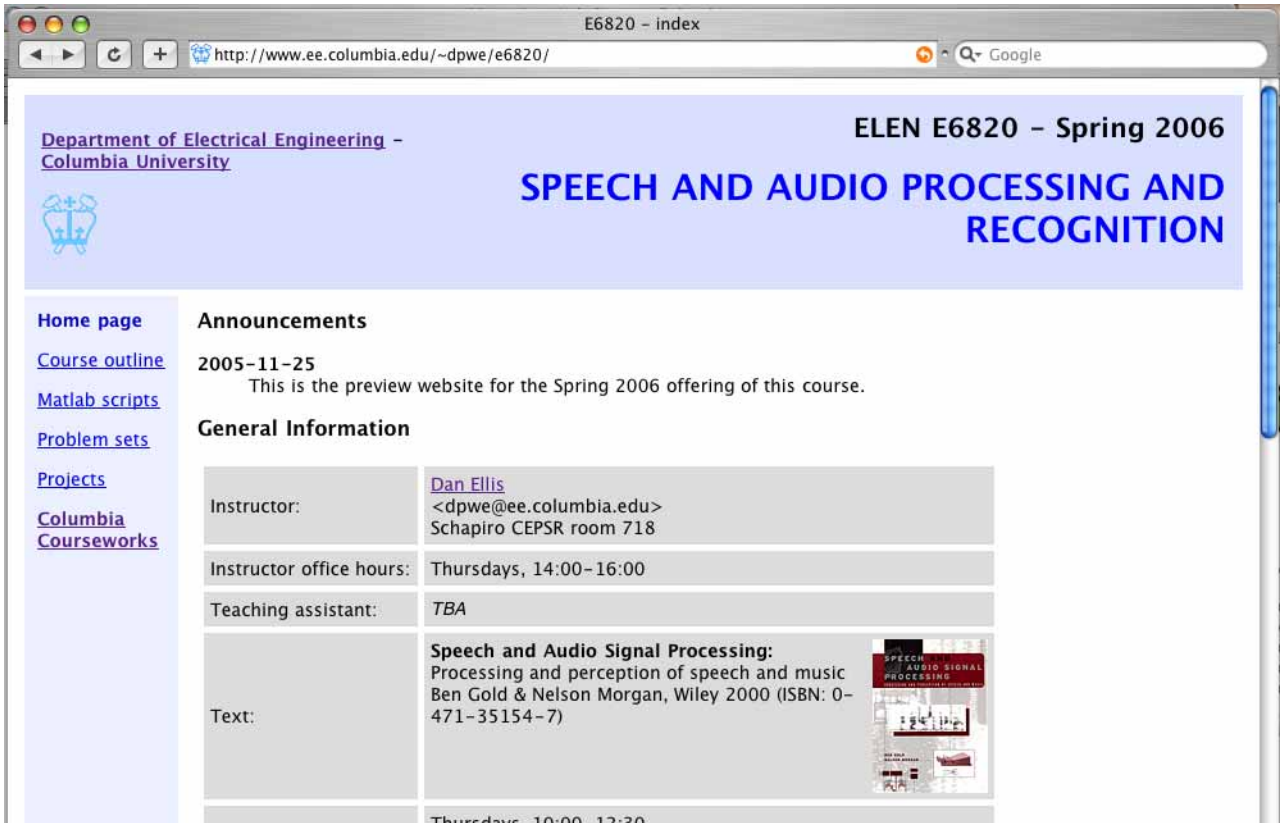
  Speech and Audio Signal Processing

  Ben Gold & Nelson Morgan, Wiley, 2000

  ISBN: 0-471-35154-7

Web-based

- **Course website:**
  http://www.ee.columbia.edu/~dpwe/e6820/
  for lecture notes, problem sets, examples, ...



- **+ student web pages for homework etc.**

Course outline

Fundamentals

L1: **DSP**	L2: **Acoustics**	L3: **Pattern recognition**	L4: **Auditory perception**

Audio processing

L5: **Signal models**	L6: **Music analysis/ synthesis**
L7: **Audio compression**	L8: **Spatial sound & rendering**

Applications

L9: **Speech recognition**	L10: **Music retrieval**
L11: **Signal separation**	L12: **Multimedia indexing**

Weekly Assignments

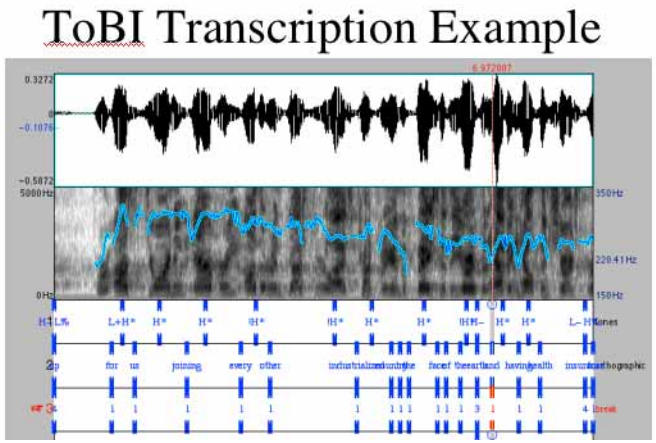
- **Research papers**
  - journal & conference publications
  - summarize & discuss in class
  - written summaries on web page
- **Practical experiments**
  - MATLAB-based (+ Signal Processing Toolbox)
  - direct experience of sound processing
  - skills for project
- **Book sections**

Final Project

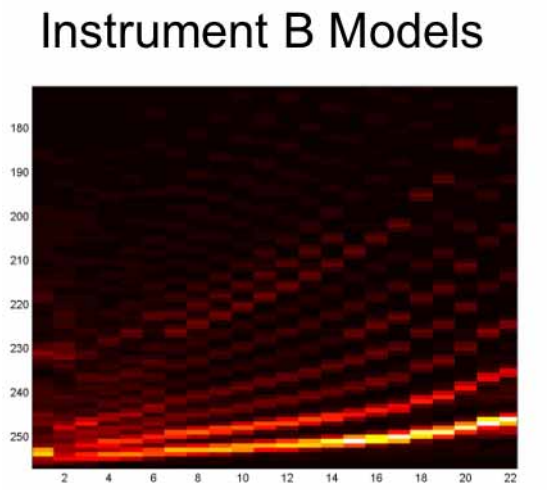
- **Most significant part of course (50% of grade)**
- **Oral proposals mid-semester;**
  Presentations in final class
  + website
- **Scope**
  - practical (Matlab recommended)
  - identify a problem; try some solutions
  - evaluation
- **Topic**
  - few restrictions within world of audio
  - investigate other resources
  - develop in discussion with me
- **Copying**

Examples of past projects

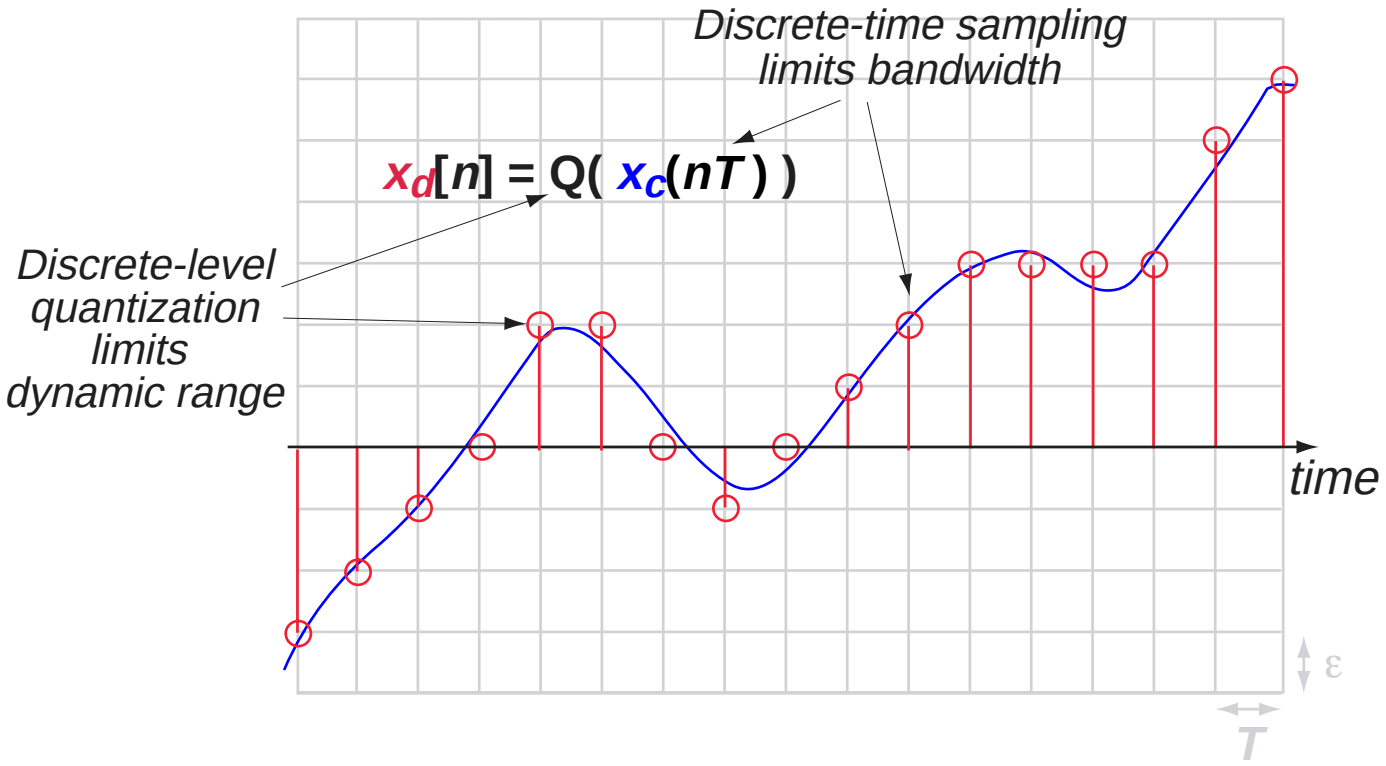
- **Automatic Prosody Classification**



- **Model-based note transcription**



3 DSP review: Digital Signals



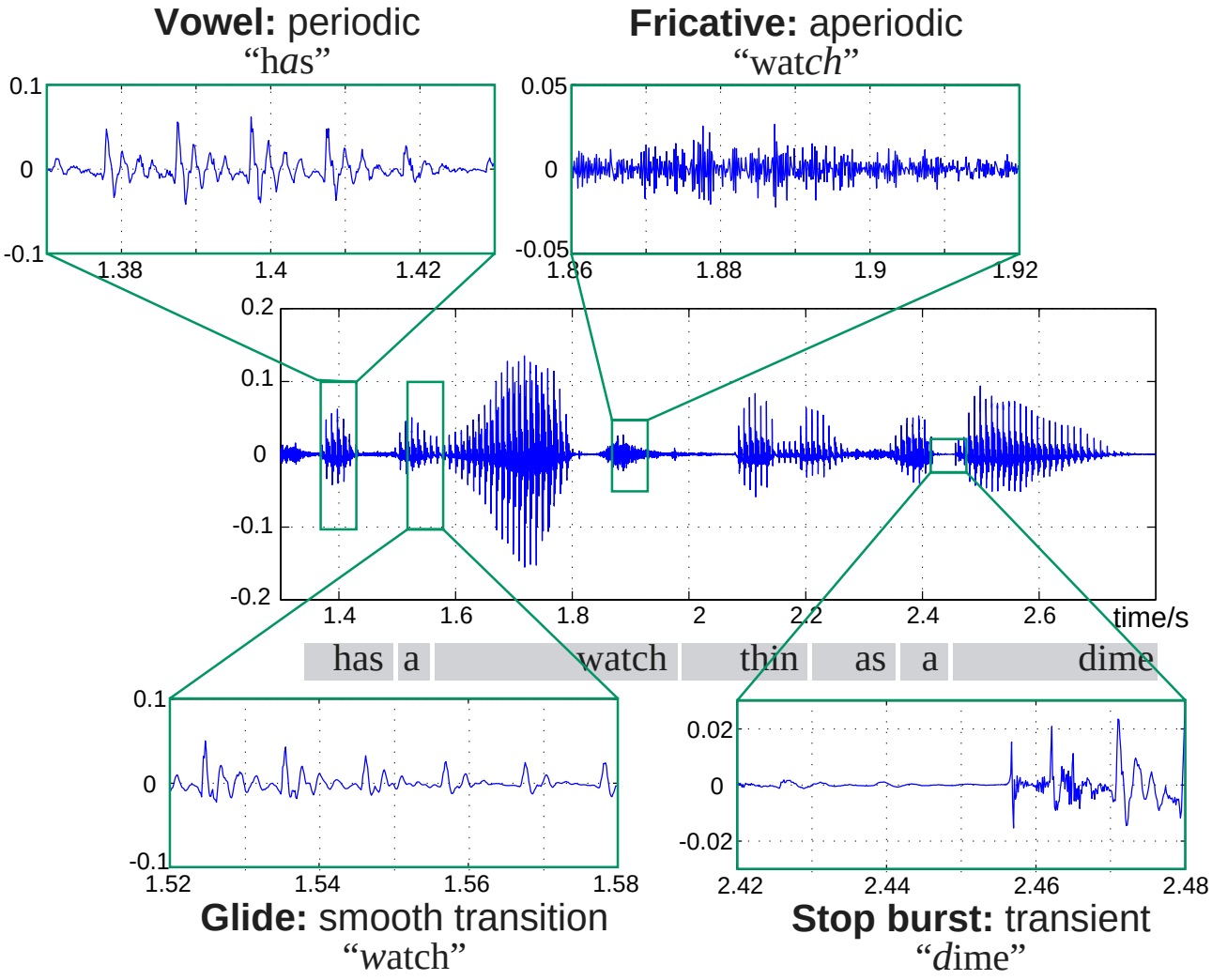
- sampling interval T,

  sampling frequency $\Omega_T = \frac{2\pi}{T}$

- quantizer $Q(y) = \varepsilon \cdot \lfloor y/\varepsilon \rfloor$

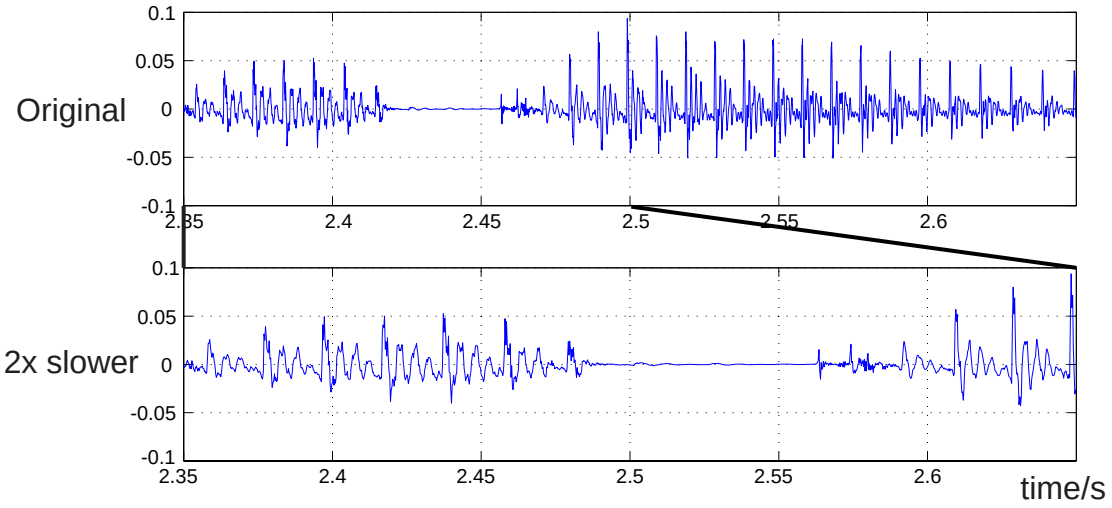
The speech signal: time domain

- **Speech is a sequence of different sound types:**



Timescale modification (TSM)

- **Can we modify a sound to make it 'slower'? i.e. speech pronounced more slowly**
  - e.g. to help comprehension, analysis
  - or more quickly for 'speed listening'?
- **Why not just slow it down?**
  - $x_s(t) = x_o\left(\frac{t}{r}\right)$, r = slowdown factor (>1 → slower)
  - equiv. to playback at a different sampling rate

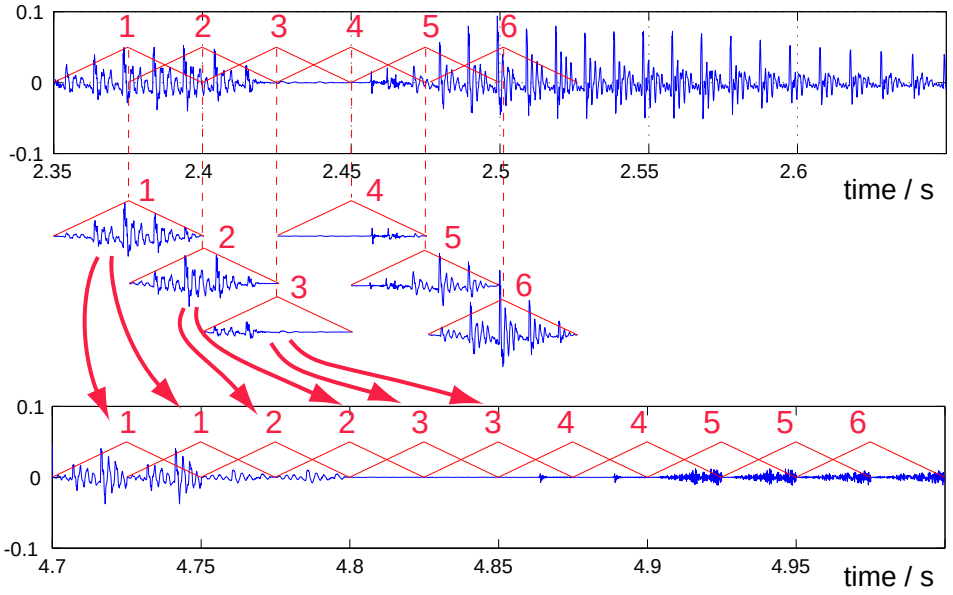


$r = 2$

Time-domain TSM

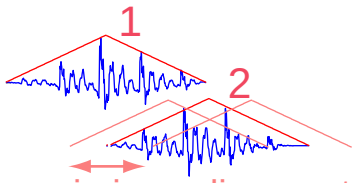
- **Problem: want to preserve local time structure but alter global time structure**
- **Repeat segments**
  - but: artefacts from abrupt edges
- **Cross-fade & overlap**

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



Synchronous Overlap-Add (SOLA)

- **Idea: Allow some leeway in placing window to optimize alignment of waveforms**



K_m maximizes alignment of 1 and 2

- **Hence,**

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

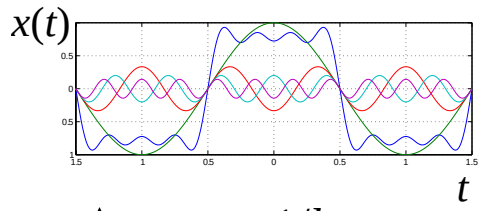
where K_m chosen by *cross-correlation*:

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

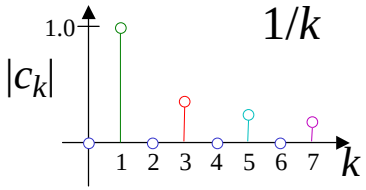
The Fourier domain

Fourier Series (periodic continuous x)

$$x(t) = \sum_k c_k \cdot e^{jk\Omega_0 t} \qquad \Omega_0 = \frac{2\pi}{T}$$



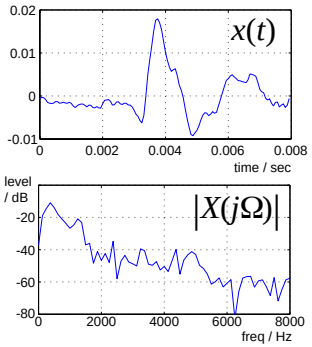
$$c_k = \frac{1}{2\pi T}\int_{-T/2}^{T/2} x(t) \cdot e^{-jk\Omega_0 t} dt$$



Fourier Transform (aperiodic continuous x)

$$x(t) = \frac{1}{2\pi}\int X(j\Omega) \cdot e^{j\Omega t} d\Omega$$

$$X(j\Omega) = \int x(t) \cdot e^{-j\Omega t} dt$$

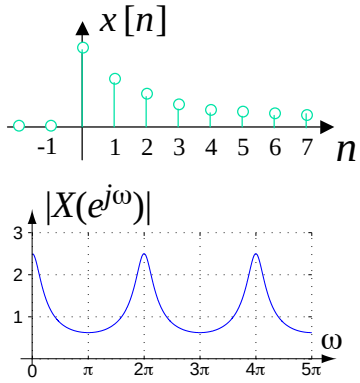


Discrete-time Fourier

DT Fourier Transform (aperiodic sampled x)

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega}) \cdot e^{j\omega n} d\omega$$

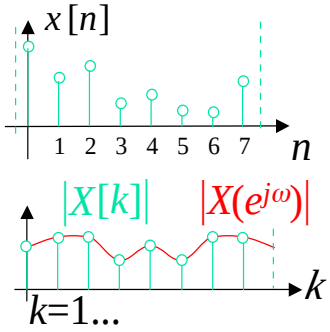
$$X(e^{j\omega}) = \sum x[n] \cdot e^{-j\omega n}$$



Discrete Fourier Transform (N-point x)

$$x[n] = \sum_k X[k] \cdot e^{j\frac{2\pi kn}{N}}$$

$$X[k] = \sum_n x[n] \cdot e^{-j\frac{2\pi kn}{N}}$$

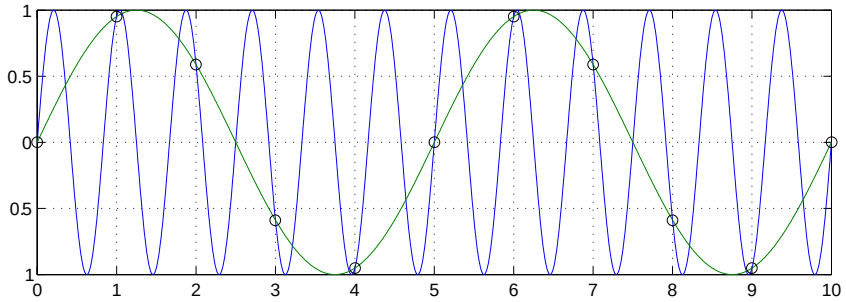


Sampling and aliasing

- **Discrete-time signals equal the continuous time signal at discrete sampling instants:**

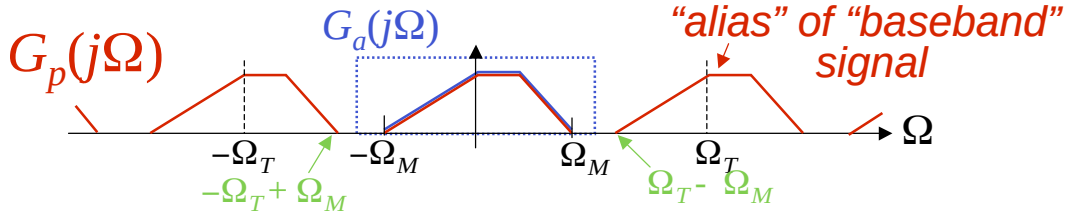
$$x_d[n] = x_c(nT)$$

- **Sampling cannot represent rapid fluctuations**

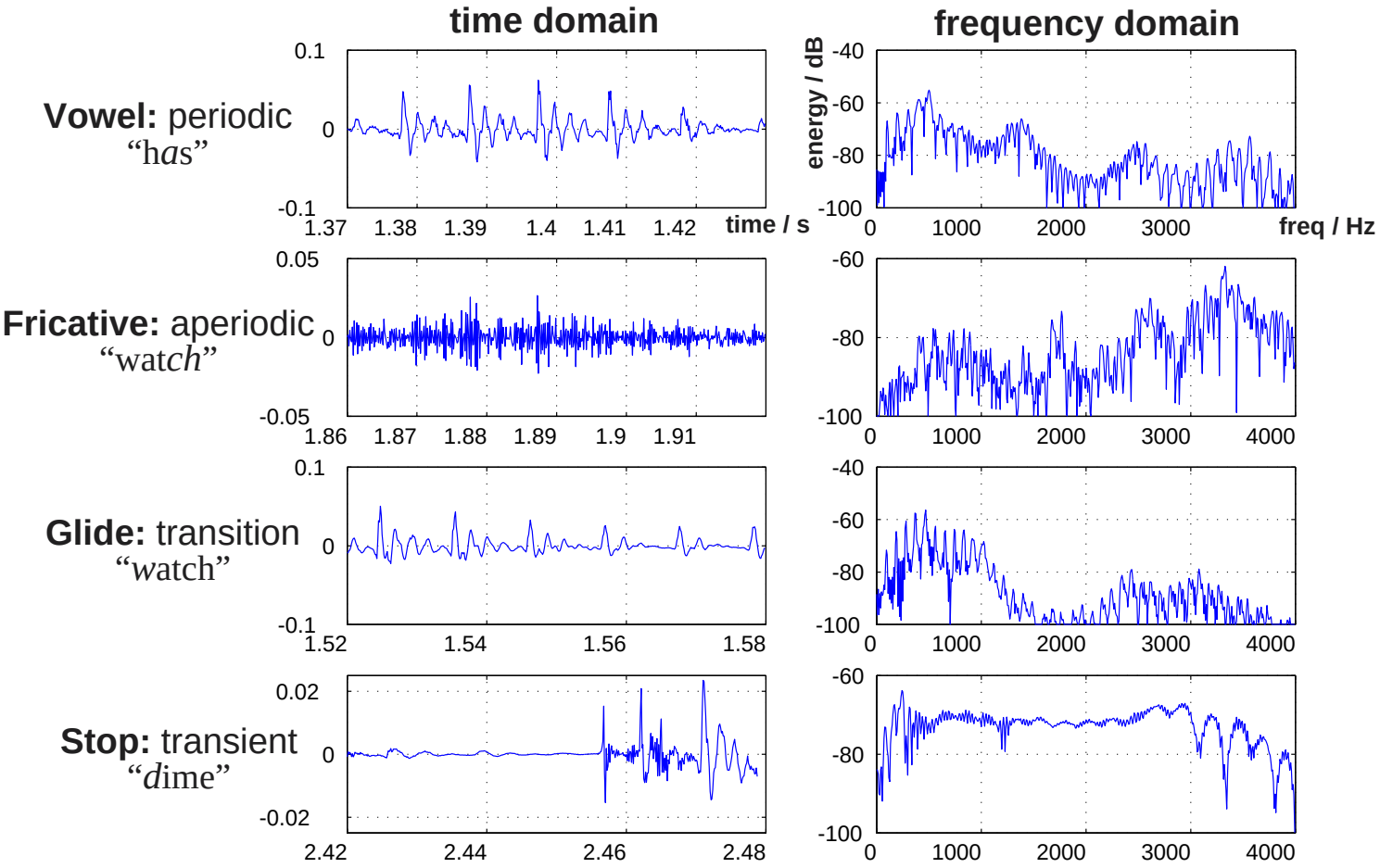


$$\sin\left(\left(\Omega_M + \frac{2\pi}{T}\right)Tn\right) = \sin(\Omega_M Tn)\ \forall n \in I$$

- **Nyquist limit ($\Omega_T/2$) from periodic spectrum:**



Speech sounds in the Fourier domain



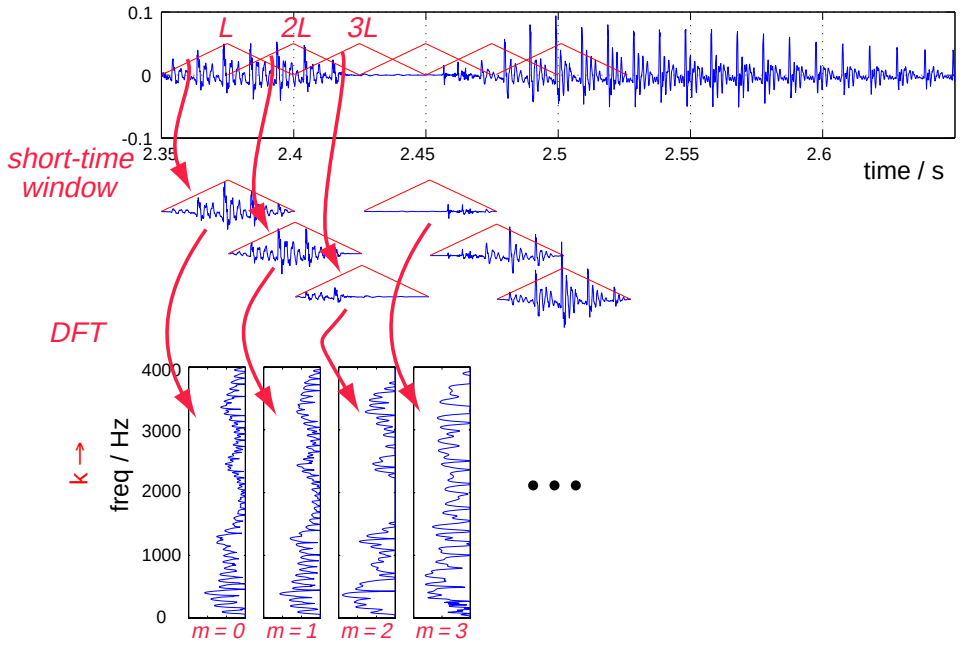
- dB = 20·log10(amplitude) = 10·log10(power)

- **Voiced spectrum has pitch + formants**

Short-time Fourier Transform

- **Want to localize energy in both time and freq**

  →break sound into short-time pieces
  calculate DFT of each one

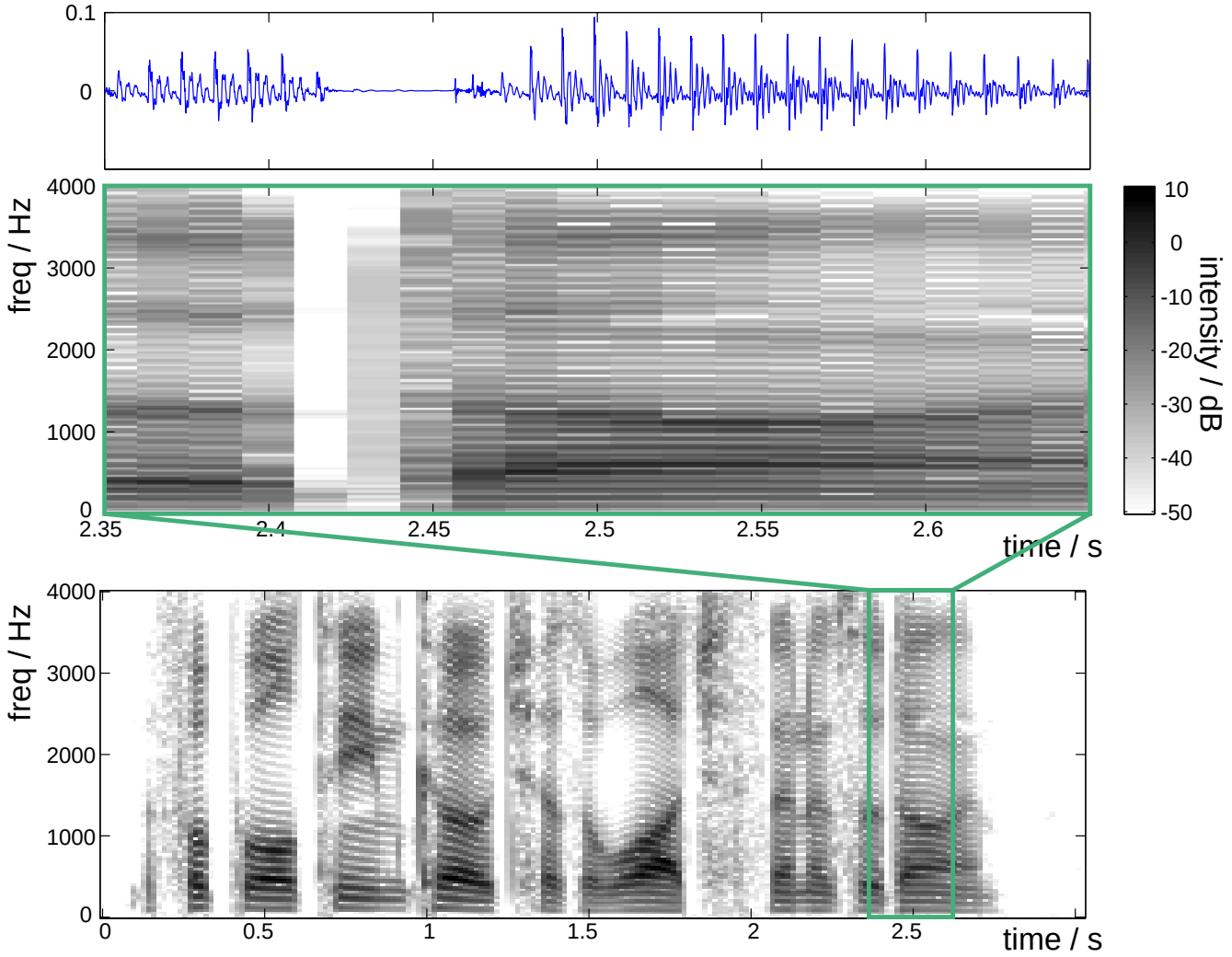


- **Mathematically:**

$$X[k,m] = \sum_{n=0}^{N-1} x[n] \cdot w[n-mL] \cdot \exp -j\left(\frac{2\pi k(n-mL)}{N}\right)$$

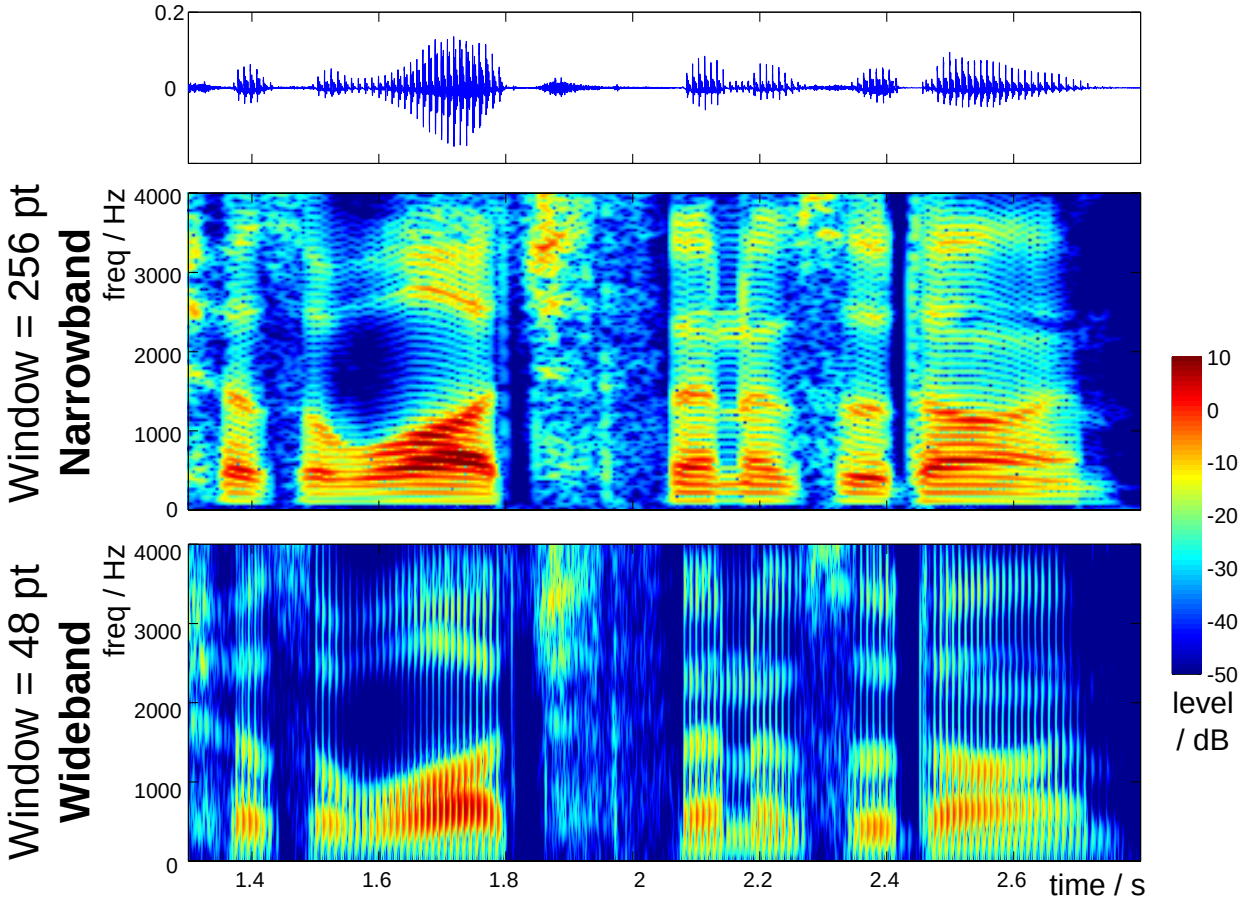
The Spectrogram

- **Plot STFT $X[k, m]$ as a grayscale image:**



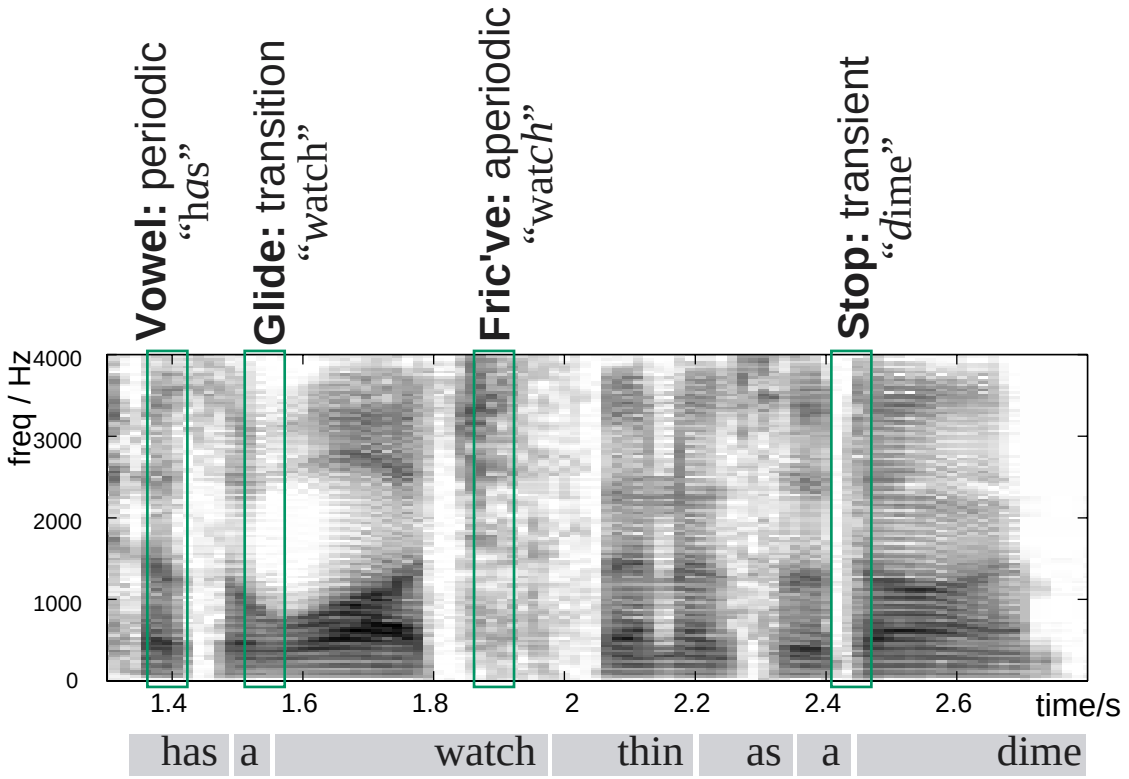
Time-frequency tradeoff

- **Longer window $w[n]$ gains frequency resolution at cost of time resolution**



Speech sounds on the Spectrogram

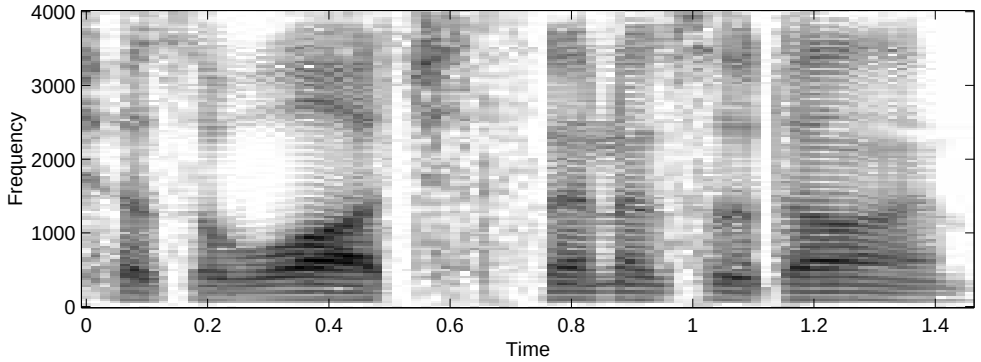
- **Most popular speech visualization**

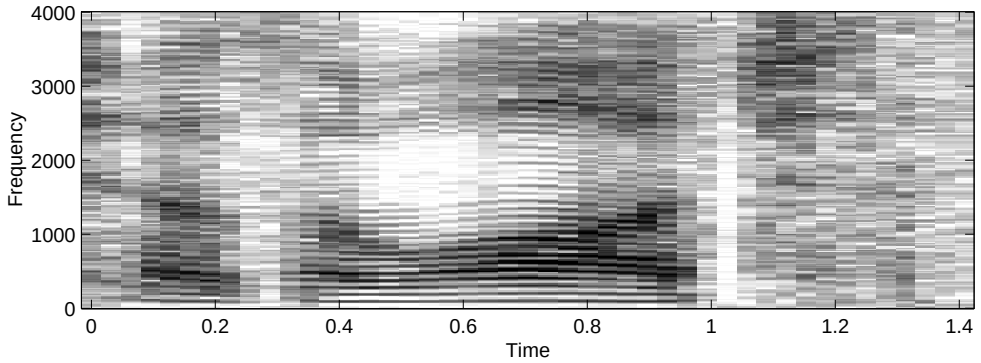


- **Wideband (short window) better than narrowband (long window) to see formants**

TSM with the Spectrogram

- **Just stretch out the spectrogram?**





  - how to resynthesize?
    spectrogram is only $|Y[k, m]|$

The Phase Vocoder

- **Timescale modification in the STFT domain**
- **Magnitude from 'stretched' spectrogram:**

$$|Y[k, m]| = \left|X\left[k, \frac{m}{r}\right]\right|$$

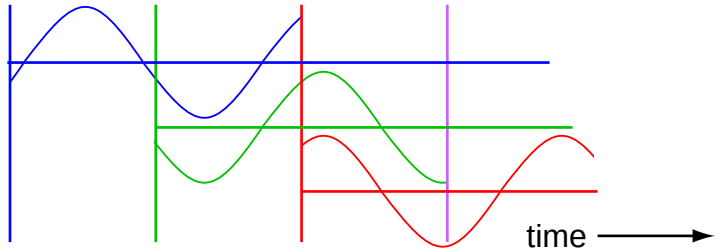
  - e.g. by linear interpolation

- **But preserve phase increment between slices:**

$$\dot{\theta}_Y[k, m] = \dot{\theta}_X\left[k, \frac{m}{r}\right]$$

  - e.g. by discrete differentiator

- **Does right thing for single sinusoid**
  - keeps overlapped parts of sinusoid aligned



General issues in TSM

- **Time window**
  - stretching a narrowband spectrogram
- **Malleability of different sounds**
  - vowels stretch well, stops lose nature
- **Not a well-formed problem?**
  - want to alter time without frequency
    ... but time and frequency are not separate!
  - 'satisfying' result is a subjective judgement

  →solution depends on auditory perception...

Summary

- **Information in sound**
  - lots of it, multiple levels of abstraction
- **Course overview**
  - survey of audio processing topics
  - practicals, readings, project
- **DSP review**
  - digital signals, time domain
  - Fourier domain, STFT
- **Timescale modification**
  - properties of the speech signal
  - time-domain
  - phase vocoder