EE E6820: Speech & Audio Processing & Recognition

Lecture 1: Introduction & DSP

Dan Ellis <dpwe@ee.columbia.edu>
Mike Mandel <mim@ee.columbia.edu>

Columbia University Dept. of Electrical Engineering
http://www.ee.columbia.edu/∼dpwe/e6820

January 22, 2009

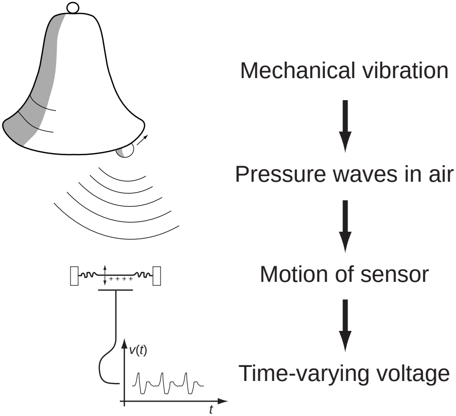
1. Sound and information
2. Course Structure
3. DSP review: Timescale modification

Outline

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

Sound and information

Sound is air pressure variation



Mechanical vibration

Pressure waves in air

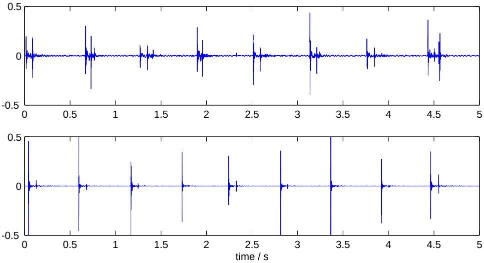
Motion of sensor

Time-varying voltage

Transducers convert air pressure ↔ voltage

What use is sound?

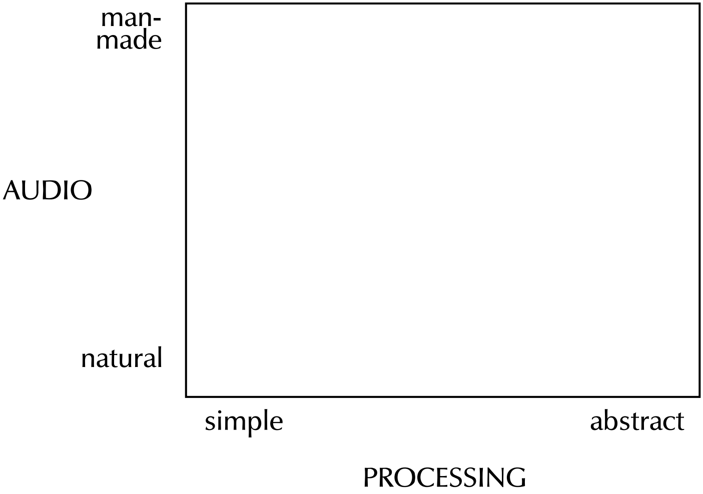
Footsteps examples:



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



man-made

AUDIO

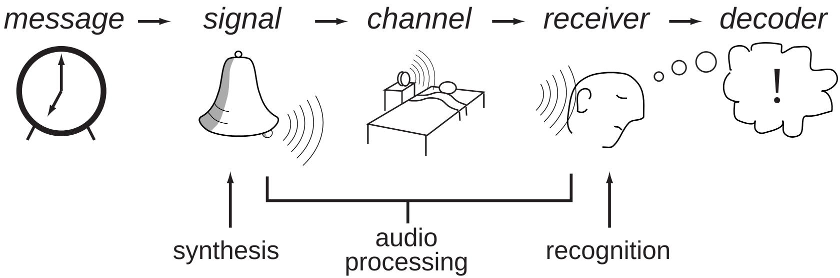
natural

simple

abstract

PROCESSING

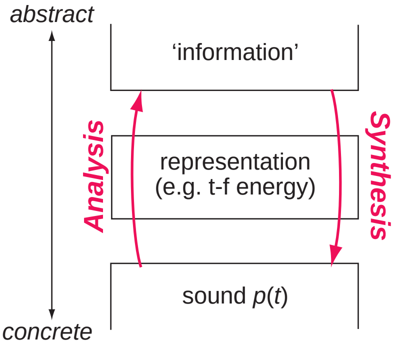
The acoustic communication chain



- 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, . . .

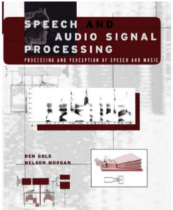
Outline

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

Source structure

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

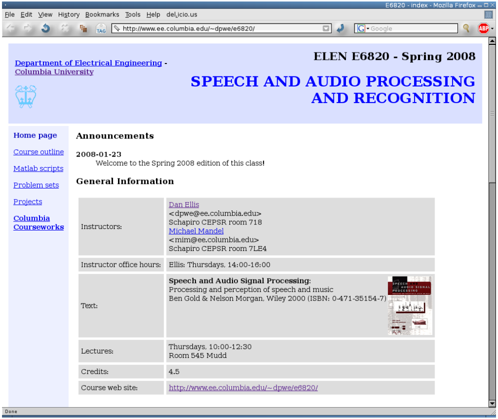
  Speech and Audio Signal Processing
  Ben Gold & Nelson Morgan
  Wiley, 2000
  ISBN: 0-471-35154-7



Web-based

Course website:

- http://www.ee.columbia.edu/∼dpwe/e6820/
- for lecture notes, problem sets, examples, . . .
- + student web pages for homework, etc.


E6820 - index - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.ee.columbia.edu/~dpwe/e6820/
Google
ABP
Done

Department of Electrical Engineering - Columbia University

ELEN E6820 - Spring 2008

SPEECH AND AUDIO PROCESSING AND RECOGNITION

Home page

Course outline

Matlab scripts

Problem sets

Projects

Columbia Courseworks

Announcements

2008-01-23
Welcome to the Spring 2008 edition of this class!

General Information

Instructors:	Dan Ellis <dpwe@ee.columbia.edu> Schapiro CEPSR room 718 Michael Mandel <mim@ee.columbia.edu> Schapiro CEPSR room 7LE4
Instructor office hours:	Ellis: Thursdays, 14:00-16:00
Text:	**Speech and Audio Signal Processing:** Processing and perception of speech and music Ben Gold & Nelson Morgan, Wiley 2000 (ISBN: 0-471-35154-7)
Lectures:	Thursdays, 10:00-12:30 Room 545 Mudd
Credits:	4.5
Course web site:	http://www.ee.columbia.edu/~dpwe/e6820/

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**

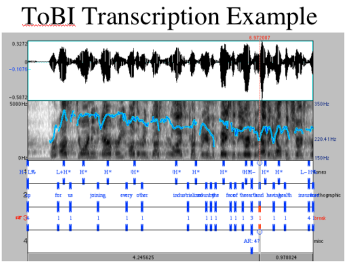
Weekly assignments

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

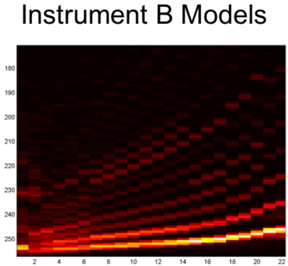
Final project

- 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
- Citation & plagiarism

Examples of past projects



Automatic prosody classification



Model-based note transcription

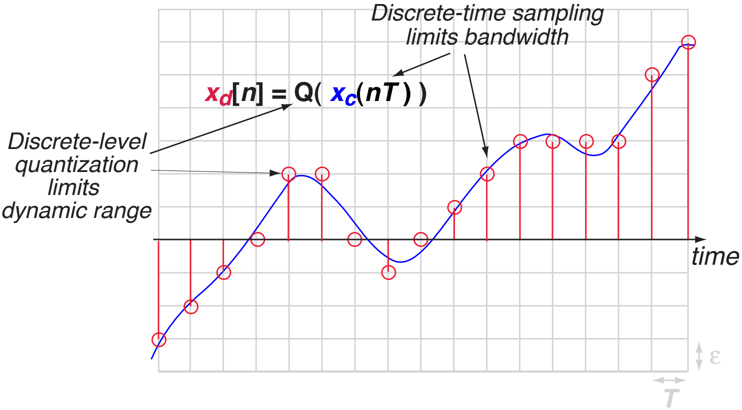
Outline

1. Sound and information

2. Course Structure

3. DSP review: Timescale modification

DSP review: digital signals

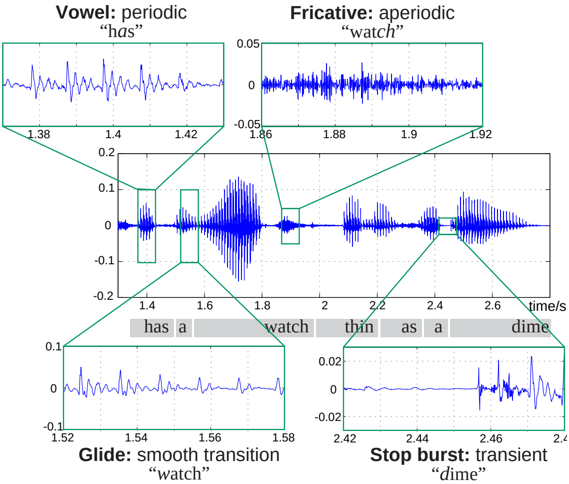


Discrete-time sampling limits bandwidth
$x_d[n] = Q(\ x_c(nT)\ )$
Discrete-level quantization limits dynamic range
time
ε
T

- sampling interval T
- sampling frequency $\Omega_T = \frac{2\pi}{T}$
- quantizer $Q(y) = \epsilon \left\lfloor \frac{y}{\epsilon} \right\rfloor$

The speech signal: time domain

Speech is a sequence of different sound types



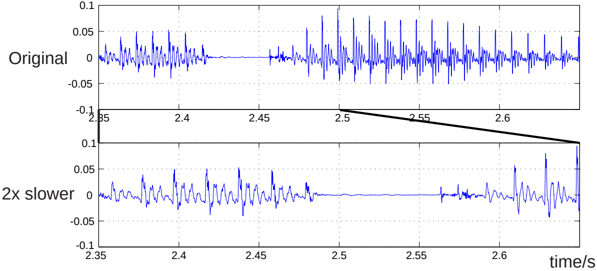
Timescale modification (TSM)

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?

- $x_s(t) = x_o(\frac{t}{r})$, r = slowdown factor ($> 1 \rightarrow$ slower)
- equivalent to playback at a different sampling rate

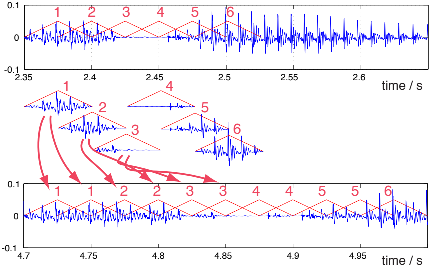


$r = 2$

Time-domain TSM

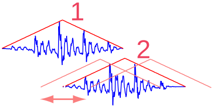
- Problem: want to preserve local time structure but alter global time structure
- Repeat segments
  - but: artifacts from abrupt edges
- Cross-fade & overlap

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



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:

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

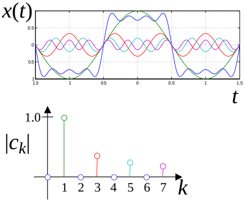
The Fourier domain

Fourier Series (periodic continuous x)

$$\Omega_0 = \frac{2\pi}{T}$$

$$x(t) = \sum_k c_k e^{jk\Omega_0 t}$$

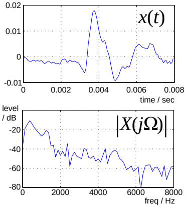
$$c_k = \frac{1}{2\pi T}\int_{-T/2}^{T/2} x(t)e^{-jk\Omega_0 t}dt$$



Fourier Transform (aperiodic continuous x)

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

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

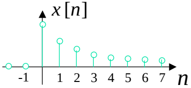


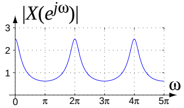
Discrete-time Fourier

DT Fourier Transform (aperiodic sampled x)

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

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

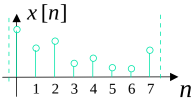


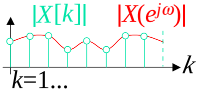


Discrete Fourier Transform (N-point x)

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

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



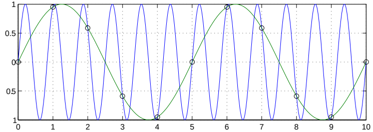


Sampling and aliasing

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

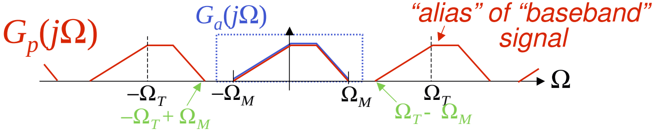
$$x_d[n] = x_c(nT)$$

Sampling cannot represent rapid fluctuations

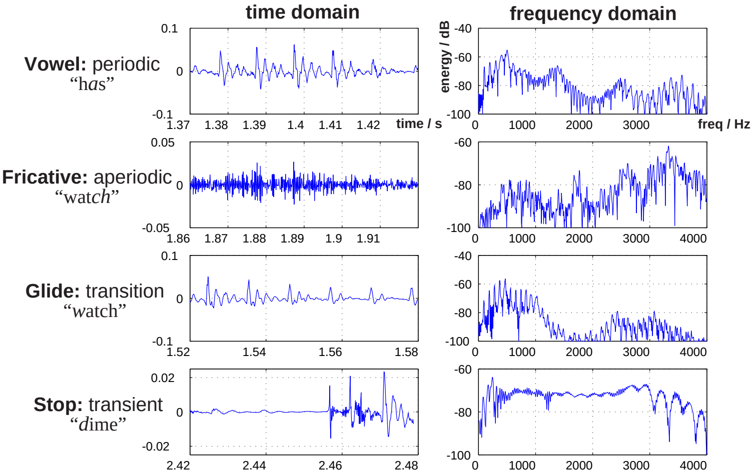


$$\sin\left(\left(\Omega_M + \frac{2\pi}{T}\right) Tn\right) = \sin(\Omega_M Tn) \quad \forall n \in \mathbb{Z}$$

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



Speech sounds in the Fourier domain



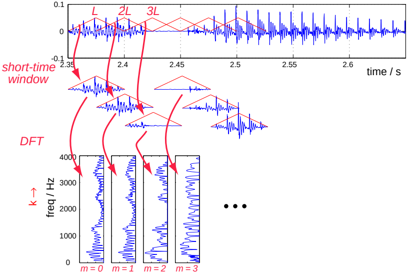
$$\text{dB} = 20\log_{10}(\text{amplitude}) = 10\log_{10}(\text{power})$$

Voiced spectrum has pitch + formants

Short-time Fourier Transform

Want to localize energy in time and frequency

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

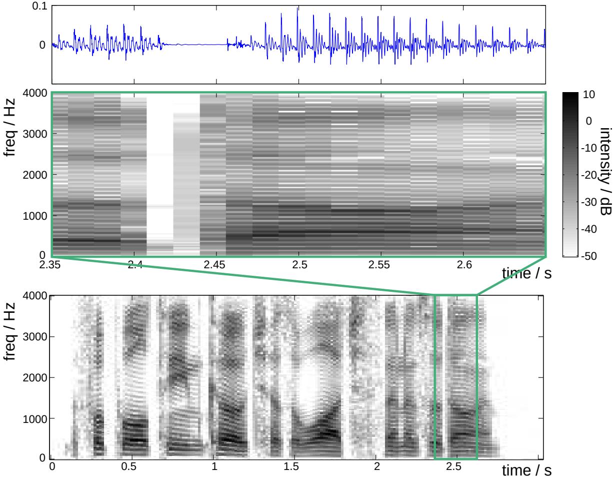


Mathematically,

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

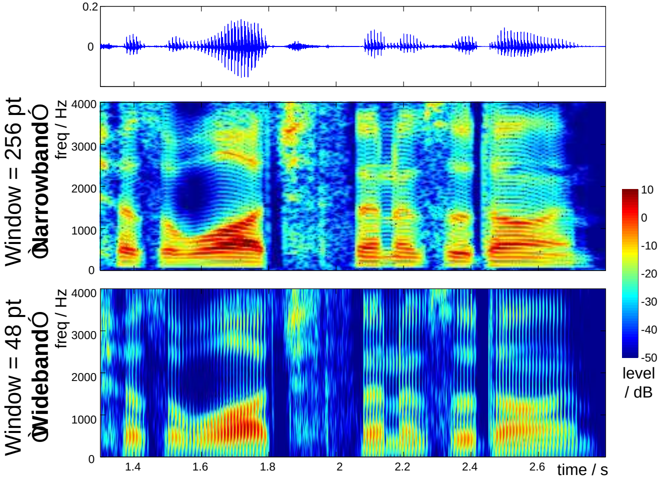
The Spectrogram

Plot STFT $X[k,m]$ as a gray-scale image



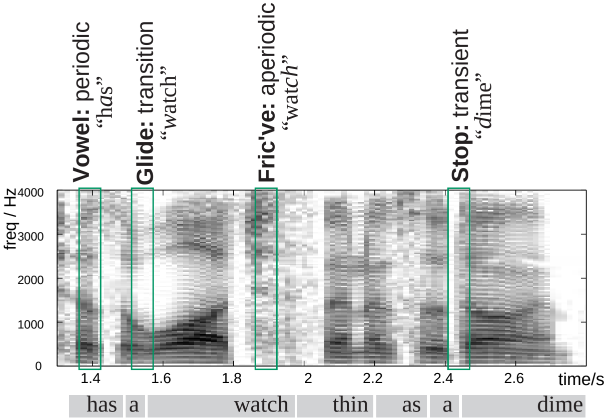
Time-frequency tradeoff

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



Speech sounds on the Spectrogram

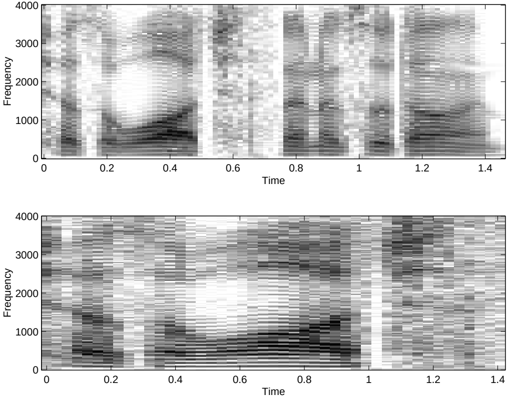
Most popular speech visualization



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

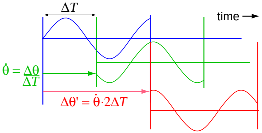
- Timescale modification in the STFT domain
- Magnitude from ‘stretched’ spectrogram:

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

  - *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


ΔT
time →
$\dot{\theta} = \frac{\Delta\theta}{\Delta T}$
$\Delta\theta' = \dot{\theta}\cdot 2\Delta T$

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 judgment

  ⇒ 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

References

J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell System Technical Journal*, pages 1493–1509, 1966.

M. Dolson. The Phase Vocoder: A Tutorial. *Computer Music Journal*, 10(4):14–27, 1986.

M. Puckette. Phase-locked vocoder. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 222–225, 1995.

A. T. Cemgil and S. J. Godsill. Probabilistic Phase Vocoder and its application to Interpolation of Missing Values in Audio Signals. In *13th European Signal Processing Conference*, Antalya, Turkey, 2005.