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**

- **Transducers convert air pressure ↔ voltage**

Mechanical vibration

Pressure waves in air

Motion of sensor

Time-varying 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

Audio

Processing

Natural

Simple

Man-made

Abstract
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 ...
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
Web-based

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

  ![Course website screenshot](image)

- **+ student web pages for homework etc.**
# Course outline

## Fundamentals

|--------|--------------|--------------------------|------------------------|

## Audio processing

<table>
<thead>
<tr>
<th>L5: Signal models</th>
<th>L6: Music analysis/synthesis</th>
<th>L7: Audio compression</th>
<th>L8: Spatial sound &amp; rendering</th>
</tr>
</thead>
</table>

## Applications

<table>
<thead>
<tr>
<th>L9: Speech recognition</th>
<th>L10: Music retrieval</th>
</tr>
</thead>
<tbody>
<tr>
<td>L11: Signal separation</td>
<td>L12: Multimedia indexing</td>
</tr>
</tbody>
</table>
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**
  
  **ToBI Transcription Example**

- **Model-based note transcription**
  
  **Instrument B Models**
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:
  - **Vowel:** periodic
    - "has"
  - **Fricative:** aperiodic
    - "watch"
  - **Glide:** smooth transition
    - "watch"
  - **Stop burst:** transient
    - "dime"
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?

\[
x_s(t) = x_o\left(\frac{t}{r}\right), \quad r = \text{slowdown factor}
\]

- equiv. to playback at a different sampling rate

![Graph showing original and 2x slower sound samples](image-url)
**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**
  
  \[ y^m[mL + n] = y^{m-1}[mL + n] + w[n] \cdot x\left[ \frac{m}{r} \right] L + n \]
Synchronous Overlap-Add (SOLA)

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

Hence, where $K_m$ chosen by cross-correlation:

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

where $K_m$ chosen by cross-correlation:

$$K_m = \arg\max_{0 \leq K \leq K_U} \frac{\sum_{n=0}^{N_{ov}} y^{m-1}[mL + n] \cdot x\left[\frac{m}{r}\right]L + n + K}{\sqrt{\sum (y^{m-1}[mL + n])^2 \sum (x\left(\frac{m}{r}\right) L + n + K)^2}}$$
The Fourier domain

Fourier Series (periodic continuous $x$)

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

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

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

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

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

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

\[ G_p(j\Omega) \quad G_d(j\Omega) \quad \text{“alias” of “baseband” signal} \]
Speech sounds in the Fourier domain

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

- Voiced spectrum has **pitch + formants**

![Graphs showing time and frequency domain comparisons for different speech sounds: Vowel (periodic), Fricative (aperiodic), Glide (transition), Stop (transient).]
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:

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

![Diagram showing time-frequency analysis with different window sizes](image)

- Window = 256 pt (Narrowband)
- Window = 48 pt (Wideband)

<table>
<thead>
<tr>
<th>freq / Hz</th>
<th>0</th>
<th>1000</th>
<th>2000</th>
<th>3000</th>
<th>4000</th>
</tr>
</thead>
<tbody>
<tr>
<td>time / s</td>
<td>0</td>
<td>1.4</td>
<td>1.6</td>
<td>1.8</td>
<td>2</td>
</tr>
</tbody>
</table>

- VegaVis colorbar: level / dB

- Level range: -50 to 10 dB
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:
  \[ |Y[k, m]| = \left| X[k, \frac{m}{r}] \right| \]
  - e.g. by linear interpolation
• But preserve phase increment between slices:
  \[ \dot{\theta}_Y[k, m] = \dot{\theta}_X[k, \frac{m}{r}] \]
  - 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