ELEN E6820:

Speech and Audio Processing and Recognition:

Selected highlights

Columbia University Dept. of Electrical Engineering
Spring 2001

Professor: Dan Ellis  <dpwe@ee.columbia.edu>

Web site:
http://www.ee.columbia.edu/~dpwe/courses/e6820-2001-01
Some aspects of the speech signal

- **Speech is highly redundant**
  - intelligible despite large distortions
  - multiple cues for each phoneme

- **Speech is very variable**
  - redundancy leaves room for variability
  - speakers can use different subsets of cues
Psychoacoustic-based audio compression

- Exemplified by MPEG-Audio layer 3 (‘MP3’)

→ From CD rate (1.4 Mbps) to 128 kbps or less (< 1.5 bits/sample)
Psychoacoustic masking model

- Based on extensive experiments

- Put noise where it can’t be heard
Automatic Speech Recognition (ASR)

- Observations $X = \{X_1..X_N\} \rightarrow$ States $S = \{S_1..S_N\}$

  $$S^* = \arg\max_S P(S|X)$$

  $$= \arg\max_S \frac{P(S, X)}{P(X)}$$

  $$= \arg\max_S \prod_i P(X_i|S_i) \cdot P(S_i|S_{i-1})$$

  Markov assumption

  - State sequence $\{S_i\}$ (e.g. phones) define words

- Training (on large datasets) is the key
  - EM iteration for acoustic & transition probs.

Feature calculation $\rightarrow$ Acoustic classifier $\rightarrow$ HMM decoder $\rightarrow$ words

- sound $\rightarrow$ feature vectors $\rightarrow$ phone likelihoods $\rightarrow$ words

Word models

<table>
<thead>
<tr>
<th>Language model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(&quot;sat&quot;,&quot;the&quot;,&quot;cat&quot;)$</td>
</tr>
<tr>
<td>$p(&quot;saw&quot;,&quot;the&quot;,&quot;cat&quot;)$</td>
</tr>
</tbody>
</table>
ASR visualization

- Standard speech recognition structure:

- ‘State of the art’ word-error rates (WERs):
  - 2% (dictation) - 30% (telephone conversations)
Element-based audio indexing

- **Search for nonspeech audio databases**
  - e.g. Muscle Fish ‘SoundFisher’ for SFX libraries

- **Segment-level features**
  - well-performing features: spectral centroid, dynamics, tonality ...

- **Each segment is an object**
  - not applicable to continuous recordings
Audio index features
(Musclefish)

- Basic features
  - duration
  - pitch
  - loudness

- ‘Timbre’ features
  - brightness
  - cepstra
  - deltas

- Music-oriented features
  - rhythm
  - (note) events
  - instruments