EE679: Speech Processing

A preview

Dept of Electrical Engineering
I.I.T. Bombay
Outline

• Speech production (physiology)

• Classification of sounds: articulatory, acoustic

• Speech analysis (signal processing methods for information extraction)

• Hearing, and speech perception

• Speech technology (speech compression, ASR, TTS)

• Audio/music technology
Speech communication

Diagram:
- **SPEAKER**
  - Sound source (lungs + vocal cords)
  - Acoustic system (vocal tract)
  - Neuro-muscular actions
  - Language code
  - Message formulation

- **LISTENER**
  - Basilar membrane motion
  - Neural transduction
  - Language code
  - Message comprehension

Acoustic wave
Acoustic waves

\[ \text{Speed} = \text{wavelength} \times \text{frequency} \]
Information in speech

- Linguistic (phone->word->sentence->message)
- Paralinguistic:
  --speaker-based (pronunciation, age, sex, etc.),
  --expressive (emotions, mood)

*The speech signal is characterised by an enormous range of perceptually contrasting sounds!*
Generating speech*

Respiration $\rightarrow$ phonation $\rightarrow$ articulation

Vibrating vocal cords create puffs of air giving rise to air pressure variations which reach our ears.

*HyperPhysics, Sound and Hearing, Georgia State University
Vocal tract: Acoustic resonances*

\[ f_1 = \frac{c}{4L} ; \quad f_2 = \frac{3c}{4L} ; \quad f_3 = \frac{5c}{4L} ; \quad \ldots \ldots \]

*HyperPhysics, Sound and Hearing, Georgia State University
(http://hyperphysics.phy-astr.gsu.edu/hbase/sound/)
Speech production (Childers, Speech Overview, 1993)
Articulation: producing the various sounds of speech*

*Securivox tutorial

Department of Electrical Engineering, IIT Bombay
Von Kempelen's talking machine

1791
1875

- Alexander Bell invents the method of, and apparatus for, “transmitting vocal or other sounds telegraphically ... by causing electrical undulations, similar in form to the vibrations of the air accompanying the said vocal or other sound”.

=> Major impetus to modern speech processing.

- 1930s: Electrical synthesis of speech by Dudley’s vocoder
Sound -> electrical form*

*The Physics Classroom: http://www.glenbrook.k12.il.us/gbssci/phys/Class/sound/u11l2a.html
Speech “waveform”
Speech Waveforms from “my speech”

(a) start of “y” vowel

(b) “ee” vowel

(c) “s” consonant
Components of sound

A sound is usually comprised of several frequency components.

Depending on the relationships of the frequency components, the sound can elicit a sensation of pitch.
Speech production

• **Vocal cords** (larynx) modulate the airflow from the lungs by rapid opening-closing; the *rate of vibration* is determined by their mass and tension.
  Pitch frequency ranges:
  male: 80-160 Hz; female:160-320 Hz;
  singers: over 2 octaves.

• **Vocal tract** shapes the vocal cord vibrations into the intricate sounds of speech via *changes in shape* to produce various *acoustic resonances*. 
Vocal tract “filter”*

- The sound spectrum is modified by the shape of the vocal tract.
- The resonant frequencies of the vocal tract cause peaks in the spectrum called formants.

*Childers, Speech Overview
Most important aspects of speech...

• The intelligence in speech is encoded in the *power spectrum* of the acoustic pressure wave.

• Different *articulatory configurations* result in signals with different spectra, esp. different resonance frequencies called formants, which are perceived as different sounds.

• The different spectra make up the *finite alphabet of symbols* (linguistic code) governed by a hierarchy of linguistic rules.
Basic sounds of speech: Phones

• The speech signal can be divided into sound segments with fixed articulation and acoustics over short intervals. i.e. articulatory configuration <=> acoustic properties

Smallest meaningful sound unit: “phone”
(i.e. set of distinctive sounds of a language)

In Indian written scripts, one symbol represents one phone.
Classification of speech sounds

Vowels and Consonants

- **Vowels**: steady sounds specified by position of the articulators (typically, tongue)

- **Consonants**: are (dynamic) sounds classified by place and manner of articulation
Place of articulation
(constriction of vocal tract)
<table>
<thead>
<tr>
<th>MoA</th>
<th>PoA</th>
<th>Velar</th>
<th>Palatal/ Palato Alvelor</th>
<th>Retroflex</th>
<th>Dentals</th>
<th>Bilabials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stop and Affricates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UvUa</td>
<td>क</td>
<td>च</td>
<td>ट</td>
<td>त</td>
<td>प</td>
<td></td>
</tr>
<tr>
<td>UvAs</td>
<td>ख</td>
<td>छ</td>
<td>ठ</td>
<td>थ</td>
<td>फ</td>
<td></td>
</tr>
<tr>
<td>VoUa</td>
<td>ग</td>
<td>ज</td>
<td>ड</td>
<td>द</td>
<td>ब</td>
<td></td>
</tr>
<tr>
<td>VoAs</td>
<td>घ</td>
<td>झ</td>
<td>ढ</td>
<td>ध</td>
<td>भ</td>
<td></td>
</tr>
<tr>
<td>Nasal</td>
<td>ड</td>
<td>ज</td>
<td>ण</td>
<td>न</td>
<td>म</td>
<td></td>
</tr>
</tbody>
</table>

UvUa: Unvoiced Unaspirated  
UvAs: Unvoiced Aspirated  
VoUa: Voiced Unaspirated  
VoAs: Voiced Aspirated  
MoA: Manner of Articulation  
PoA: Place of Articulation  

Table 1: The table shows some consonants used in Hindi language and their classification depending on their phonetic properties.
“Decoding” the speech signal: visible speech

Dark areas of spectrogram show high intensity

– Voiced segments are much louder than unvoiced

– Horizontal dark bands are the formant peaks

– “s” has high frequency content
– Vertical bands are individual larynx closures
– The “y” of “my” is a diphthong: two successive vowels

“my speech”
Figure 1: Speech spectral characteristics
Machli jal ki hai raani  jeevan uska he paani
Indian costumes are quite colourful
Speech perception

Distinct stages of physiological processing in the auditory system:

Peripheral auditory system (Ears)  ← analysis
Auditory nervous system (Brain)    ← synthesis
Audible sound

Diagram showing the range of frequencies and sound levels for audible sounds, including the threshold of pain and the threshold of hearing.
Sound and Sensation

A sound of given frequency components and sound pressure levels leads to perceived sensations that can be distinguished in terms of:

- **loudness**  <-- intensity
- **pitch**  <-- fundamental frequency
- **timbre** ("quality" or "colour")  <-- their spectro-temporal properties
Our auditory apparatus

Cochlea: Ear’s microphone

HyperPhysics, Sound and Hearing, Georgia State University
(http://hyperphysics.phy-astr.gsu.edu/hbase/sound/soucon.html#soucon)
Basilar Membrane

Location-dependent frequency “resonance”

- Thickness and tension vary along its length
- Traveling wave has maximum vibration amplitude at a location depending on its frequency
Basilar Membrane

*Frequency-to-place transformation (Fourier analysis)*

HyperPhysics, Sound and Hearing, Georgia State University
(http://hyperphysics.phy-astr.gsu.edu/hbase/sound/soucon.html#soucon)
Applications

- Automatic speech recognition/ understanding
- Text-to-speech synthesis
- Speaker verification (biometric)
- Digital storage/transmission of speech
- Aids to the handicapped
- Enhancement of quality
Transmission/storage

Waveform coding:

**distortion vs bit rate**

What distortion is “acceptable” depends on the application and on human perception.
Digital audio bit rates: Waveform coding

<table>
<thead>
<tr>
<th>Format</th>
<th>Sample Rate (kHz)</th>
<th>Bits/sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephony</td>
<td>8</td>
<td>12 (=&gt; 96 kbps)</td>
</tr>
<tr>
<td>Wideband audio</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Hi-fidelity audio</td>
<td>44.1</td>
<td>16</td>
</tr>
</tbody>
</table>
Source-filter model parameters

Pitch and vocal tract shape vary slowly in time
Frame-based coding of speech

Feature vector (short-time spectrum) extraction from speech
Automatic speech recognition

• To extract the linguistic code (a *structured sequence of discrete symbols*) from an *analysis of the acoustic speech signal*.

• That is, *continuous, noisy* measurements of a non-stationary function of time only are available.
Automatic speech recognition

- Feature calculation (to a more distinctive domain)
- Pattern classification with respect to previously trained models of phones/words
- Improved transcription based on language model
ASR: block diagram*

*K.Samudravijaya, A Tutorial on Speech and Speaker Recognition*
ASR: Challenges

• Inter- and intra-speaker variations
• Effects of coarticulation in continuous speech
• Background noise and variable channels
Categories of speech recognition tasks

Human to machine:
• Database query/ information retrieval
• Dictation

Human to human:
• Broadcast news
• Lectures
• Voice mail
• Meeting
• Telephone conversation
Speaker recognition
(voice-based biometric)

• The voice signal is considered relatively easy to acquire/collect.
• Speech enables an (indirect) measurement of physiological features (i.e. characteristics of the speaker’s voice production system).

• Applications:
  Commercial (access control, segmentation)
  Military, Forensic
Speech Synthesis

**What:** To convert a text string into a speech waveform

**Why:** For technology to communicate when a display would be inconvenient.
Basic TTS System

Text-to-speech synthesizer

Natural language processing

Phonemes

Prosody

Digital signal processing

Speech

Prosody => A phone is long/short, loud/soft, high/low-pitched
Outline

• Speech production (physiology)

• Classification of sounds: articulatory, acoustic

• Speech analysis (*signal processing methods for information extraction*)

• Hearing, and speech perception

• Speech technology (speech compression, ASR, TTS)

• Audio/music technology
Text / References

• **Douglas O'Shaughnessy**, Speech Communications: Human and Machine, Universities Press (India) Ltd., 2001

• **Rabiner and Schafer**, Digital Processing of Speech Signals

• IITB Moodle for all course-related hand-outs
Recognition: “Vowel triangle”
Speaker variability: due to differences in vocal physiology