Speech production & perception

Professor Marie Roch

Phonetics & Phonology

- Phoneme – A minimal unit of sound which can be used to distinguish one word for another. i.e. “pet” /pɛt/ vs. “bet” /bɛt/.
- Phone – A sound that corresponds to a phoneme.
Speech Production

Air, driven by our lungs, drives speech production.

The sound, or phoneme produced depends upon voicing & the configuration of our articulators.

Articulators

- Vocal folds (cords) - Responsible for voiced/unvoiced speech
- Velum (soft palate) – Serves as a valve to the nasal cavity.
Articulators

• Tongue – Flexible muscle, shape & position very important to phoneme production.
• Hard palate – Hard part of the roof of your mouth.
• Alveolar ridge

http://www.personal.rdg.ac.uk/~llsroach/phon2/artic-basics.htm

Articulators

• Teeth – Target for the tongue for some consonants, i.e. /dh/ in “then.” (Teeth are actually moved by the jaw.)

• Lips – Rounding can extend the length of the vocal tract. Closure can produce a stop, i.e. the /p/ in “apple.”
Voicing

- Voiced sounds occur when the vocal folds open & close at a regular interval:
  - Subglottal pressure forces open the vocal folds
  - As the pressure differential drops, the folds close.

Huang et al., 2001, p. 26

Voicing “sees”

/s/

/iy/

/z/
Magnified portions of “sees” (times scales are different)

unvoiced s /s/

voiced ee /iy/

voiced s /z/ (constriction contributes to irregular pattern unlike the vowel)

F0 – Fundamental Frequency

• The fundamental frequency, or F0, is the number of times per second that the vocal folds open & close

• Each cycle in the figure to the left is about 8.33 ms.

• As

\[
\text{Frequency} = \frac{\text{cycles}}{\text{sec}}
\]

• F0 is about 120 Hz

\[
1 \text{cycle} \quad 1000 \text{ms.} = \frac{120 \text{Hz}}{8.33 \text{ms.} \quad 1 \text{s.}}
\]
F0 and Harmonics

- F0 (if present), is not the only frequency.
- Harmonics are frequencies which occur at multiples of F0.
- Frequencies from a small portion of ee /iy/

Formants

- For any vocal tract shape, certain frequencies are reinforced.
- Harmonics (multiples) of F0 near resonances are reinforced.
Formants

• These reinforced harmonics are called formants, and can play an important role in recognizing vowels.

• Note that F0 is not a formant!

The Human Ear

• Outer
• Middle
• Inner
The outer ear

- Pinna - protect & filter
- Ear canal & concha - amplify frequencies between 1.5-7kHz.
- tympanic membrane (ear drum)

The middle ear

- Outer ear’s tympanic membrane connected to the inner ear’s oval window by ossicles
  - malleus
  - incus
  - stapes
Middle ear contd.

- Ossicle functioning
  - mechanical transfer of energy
  - compression to prevent overload
  - stapes connected to the inner ear’s oval window

- Eustachian tube
  - Connects to nasal cavity
  - Normally closed
  - When open, permits pressure equalization between outer/middle ear.

The inner ear

- Vestibule
- Semicircular canals
  - sense of balance
- Cochlea
  - coiled ≈ 2 and ¾ turns.
  - mechanical → neural impulses

Yost, 1994
**Cochlea (simplified view)**

- filled with fluid
- scala vestibuli and tympani joined at apex (helicotrema)
- traveling waves vibrate the basilar membrane moving hair cells which fire neurons

---

**Deformation of basilar membrane**

- Point of maximum deformation is frequency dependent
- The cochlea acts as a spectrum analyzer.

finite element model animations from WADA laboratory, Japan
Masking

- Simultaneous tones close in frequency:
  - Louder tone can “hide” the softer ones.
  - Lower frequency tones are better maskers.
- When a short tone follows a sound closely (20-30 ms), the tone may be hidden (forward masking).

Masking Demonstration

Low vs. high frequency masker
- Masker/Test 1200/2000Hz then 2000/1200 Hz.
- Ten repetitions, volume of test tone decreases each time.
- Basilar membrane response
  - Lower pitch masks more effectively than lower pitch tone.
Spectral shape and Timbre

- Spectral shape is the shape of the frequency domain:

- Timbre is our perception of the frequencies, i.e. a sound is “rich” or “tinny.”

Frequency discrimination

- 0-4000 Hz – Good frequency resolution
- > 4000 Hz – Requires greater separation of frequency to distinguish
Mel Scale

- Subjective scale
- $2N$ mel seems twice as high pitched as $N$ mel.

Sundberg, 1991

Classes of phonemes

Note that this is not exactly the same characters as the ARPABET used in our text.

Note that experts sometimes disagree on some of the classifications, e.g. OW.
Vowels

/ARPABET, IPA/
/i/ i feel, elite, /ih/ fill, /ae/ gas, /a/ father,
/a/ cut, /ao/ dog, /ax/ ex comply, /eh, /e/ pet,
/er, /e/ turn, /uh, /u/ good, /uw, /u/ tool

• Phonemes whose phones are characterized by:
  – voicing
  – lack of major constrictions of the air
  – pharyngeal cavity produces F1, oral cavity F2
  – rounding the lips increases the oral cavity length, lowering F2

Diphthongs (vowels)

/ARPABET, IPA/
/ay, /a/ tie, /ey, /e/ ate, /oy, /o/ coin, /aw, /a/ foul, /ow, /ou/ coach, /ow, /ou/ tone

• Articulators start to form one vowel & move into another:

<table>
<thead>
<tr>
<th>diphthong</th>
<th>from</th>
<th>to</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ay/ tie</td>
<td>/aa/ father</td>
<td>/iy/ eve</td>
</tr>
<tr>
<td>/ey/ ate</td>
<td>/eh/ ten</td>
<td>/iy/ eve</td>
</tr>
<tr>
<td>/oy/ coin</td>
<td>/ao/ dog</td>
<td>/iy/ eve</td>
</tr>
<tr>
<td>/aw/ foul</td>
<td>/aa/ father</td>
<td>/uw/ tool</td>
</tr>
<tr>
<td>/ow/ coach</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Ladefoged, 2001, p. 200
Major articulators for vowels

- Tongue height
  - high (i.e. /iy, iː/ eve)
  - versus low (i.e. /æ, æ/ at)
- Tongue position
  - front (i.e. /iy, iː/ eve)
  - back (i.e. /uh, u/ book)
- Lip rounding
  - flat (i.e. /iy, iː/ see)
  - rounded (i.e. /uw, u/ blue)

Vowels

- Vowels can typically be characterized by F1 & F2

/iy, iː/ “we”

- F2~2400
- F1~350
Consonants

• Manner of articulation describes the major distinction between different consonant classes.
• Many consonants come in pairs, where the only difference between them is whether or not they are voiced, i.e. /s/ vs. /z/

Note: Many IPA consonants are the same as for ARPABET. Only one symbol is shown when there is no distinction.

Consonants: Approximants

• Voiced with less obstruction of the vocal tract than normal consonants:
  – Liquids (/l/ edible, /r/ far) are very vowel-like and can even take the place of a vowel in a syllable.
  – Glides (/y, j/, /w/) are shortened & unstressed versions of the vowels /i/ evi & /u/ moo.

• Semivowels & vowels form the category of sonorants.
Consonants: Nasals

- Nasals, /m/, /n/, /ng, η/, are characterized by:
  - Constriction of oral cavity making it difficult for air to pass through it.
  - Lowering of the velum, permitting air to move through the nasal passage.

Consonants: Plosives (Stops)

- Complete blockage of the oral cavity
- Voiced & unvoiced pairs: /b/-/p/, /d/-/t/, /k/-/g/, g/
- Easy to recognize in a spectrogram from the lack of energy right before the plosive.
Consonants: Fricatives

- Nearly complete closure of the vocal tract creates turbulent, noise like sound.
- Can be voiced or unvoiced:
  - /v/-/f/ voiced, free
  - /dh, ð/-/th, θ/ then, math
  - /z/-/s/ mizzen, sigh
  - /zh, ʒ/-/sh, ʃ/ Zsa-Zsa, sheepish

Consonants: Affricates

- Combination: stop followed by a fricative
- voiced: /d/ + /zh, ʒ/ = /jh, dz/ agile
- unvoiced: /t/-/sh, ʃ/ = /ch, tj/ cheese
Distinctions between consonants

• We’ve indicated that many consonants belong to the same classes which are determined by the manner of articulation.

• What makes consonants within a class unique?

Place of articulation

• The distinction is caused by where the manner of articulation occurs.

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Labial</th>
<th>Labio- dental</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Palatal</th>
<th>Velar</th>
<th>Glottal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voice</td>
<td>/p, b</td>
<td>ʔ</td>
<td>/t, d</td>
<td>/t, d</td>
<td>/k, g</td>
<td>/ʔ</td>
<td>/ʔ</td>
</tr>
<tr>
<td>Nasal</td>
<td>/m</td>
<td>/n</td>
<td>/n</td>
<td>/n</td>
<td>/n</td>
<td>/n</td>
<td>/n</td>
</tr>
<tr>
<td>Plosive</td>
<td>/k, g</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
</tr>
<tr>
<td>Aspirate</td>
<td>/kʰ, gʰ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
</tr>
<tr>
<td>Lateral</td>
<td>/l</td>
<td>/l</td>
<td>/l</td>
<td>/l</td>
<td>/l</td>
<td>/l</td>
<td>/l</td>
</tr>
<tr>
<td>Affricate</td>
<td>/k l</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
<td>/ʔ</td>
</tr>
</tbody>
</table>

Huang et al., 2001, p 47
Other languages

- Other subsets of the phonemes - Spanish, French
- Use of pitch as pitch to distinguish words - Mandarin Chinese
- Some languages use of vowel length - Japanese

Allophones & Coarticulation

- Allophone – Phone which is recognizable even though it is atypical.
- Coarticulation
  - Surrounding phonemes affect production.
  - Try “pin” versus “spin” (The plosive /p/ is stronger in pin)
  - As speech rate increases, these effects will be more prominent.
Insertions and Deletions

• We sometimes insert (epenthisis) sounds:
  strength:  [strɛŋkθ]

• Similarly, we can drop sounds
e.g. alveolar stops between consonant pairs
  “last game” becomes [læs ɡeɪm]

Syllables

<table>
<thead>
<tr>
<th>ham</th>
<th>green</th>
<th>eggs</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ</td>
<td>σ</td>
<td>σ</td>
</tr>
<tr>
<td>Onset</td>
<td>Rime</td>
<td>Rime</td>
</tr>
<tr>
<td>h</td>
<td>ae</td>
<td>ɛh</td>
</tr>
<tr>
<td>Nucleus</td>
<td>Coda</td>
<td>Nucleus</td>
</tr>
<tr>
<td>m</td>
<td>iy</td>
<td>ɡz</td>
</tr>
</tbody>
</table>

Jurafsky & Martin 2000, p. 223
Syllables

• Linguist consider *phonotactics*, rules about syllable construction

• In practice, not a serious issue for speech recognition systems as cross syllable boundaries are usually modeled.