Digital Speech Processing—Lecture 3

Acoustic Theory of Speech Production

Topics to be Covered

- Sound production mechanisms of the human vocal tract
- Sounds of language => phonemes
- Conversion of text to sounds via letter-to-sound rules and dictionary lookup
- Location/properties of sounds in the acoustic waveform
- Location/properties of sounds in spectrograms
- Articulatory properties of speech sounds—place and manner of articulation

Basic Speech Processes

- Idea: it's getting late, I should go to lunch, I should call Al and see if he wants to join me for lunch today
- Words: Hi Al, did you eat yet?
- Coarticulated Sounds: h- a-/r/-l-d-h- a-/r/-l-d- y-1-e-1 (hi-al-di-jay)

Basics

- speech is composed of a sequence of sounds
- sounds (and transitions between them) serve as a symbolic representation of information to be shared between humans (or humans and machines)
- arrangement of sounds is governed by rules of language (constraints on sound sequences, word sequences, etc)—/spl/ exists, /sbk/ doesn’t exist
- linguistics is the study of the rules of language
- phonetics is the study of the sounds of speech

can exploit knowledge about the structure of sounds and language—and how it is encoded in the signal—to do speech analysis, speech coding, speech synthesis, speech recognition, speaker recognition, etc.

Human Vocal Apparatus

- vocal tract—dotted lines in figure: begins at the glottis (the vocal cords) and ends at the lips
- consists of the pharynx (the connection from the esophagus to the mouth) and the mouth itself (the oral cavity)
- average male vocal tract length is 17.5 cm
- cross sectional area, determined by positions of the tongue, lips, jaw and velum, varies from zero (complete closure) to 20 sq cm
- nasal tract—begins at the velum and ends at the nostrils
- velum—a trapdoor-like mechanism at the back of the mouth cavity, closure to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night), /ng/ (sing)

Vocal Tract MRI Sequences

Mid-sagittal plane X-ray of human vocal apparatus
Schematic View of Vocal Tract

Speech Production Mechanism:

• Air enters the lungs via normal breathing and no speech is produced (generally) on in-take

• As air is expelled from the lungs, via the trachea or windpipe, the tensed vocal cords within the larynx are caused to vibrate (Bernoulli oscillation) by the air flow

• Air is chopped up into quasi-periodic pulses which are modulated in frequency (spectrally shaped) in passing through the pharynx (the throat cavity), the mouth cavity, and possibly the nasal cavity; the positions of the various articulators (jaw, tongue, velum, lips, mouth) determine the sound that is produced

Vocal Tract Tube Models

• Tube areas and lengths variable
  ➢ Each sound has a range of tube areas and tube lengths
  ➢ Tubes simulate neutral vowel (/uh/ sound), /iy/ sound, and /aa/ sound
Vocal Cords

The vocal cords (folds) form a relaxation oscillator. Air pressure builds up and blows them apart. Air flows through the orifice and pressure drops allowing the vocal cords to close. Then the cycle is repeated.

Vocal Cord Views and Operation

Bernoulli Oscillation

Tensed Vocal Cords –
Ready to Vibrate

Lax Vocal Cords –
Open for Breathing

Glottal Flow

Glottal volume velocity and resulting sound pressure at the mouth for the first 30 msec of a voiced sound
- 15 msec buildup to periodicity => pitch detection issues at beginning and end of voicing; also voiced-unvoiced uncertainty for 15 msec

Artificial Larynx

Schematic Production Mechanism

• lungs and associated muscles act as the source of air for exciting the vocal mechanism
• muscle force pushes air out of the lungs (like a piston pushing air up within a cylinder) through bronchi and trachea
• if vocal cords are tensed, air flow causes them to vibrate, producing voiced or quasi-periodic speech sounds (musical notes)
• if vocal cords are relaxed, air flow continues through vocal tract until it hits a constriction in the tract, causing it to become turbulent, thereby producing unvoiced sounds (like /s/, /sh/), or it hits a point of total closure in the vocal tract, building up pressure until the closure is opened and the pressure is suddenly and abruptly released, causing a brief transient sound, like at the beginning of /p/, /t/, or /k/

Abstractions of Physical Model

excitation

voiced
unvoiced
mixed

Time-Varying Filter

speech
The Speech Signal

- speech is a **sequence** of ever changing sounds
- sound properties are highly dependent on **context** (i.e., the sounds which occur before and after the current sound)
- the state of the vocal cords, the positions, shapes and sizes of the various articulators—all change **slowly** over time, thereby producing the desired speech sounds

=> need to determine the physical properties of speech by observing and measuring the speech waveform (as well as signals derived from the speech waveform—e.g., the signal spectrum)

Speech Waveforms and Spectra

- **100 msec/line; 0.5 sec for utterance**
- **S-silence-background-no speech**
- **U-unvoiced, no vocal cord vibration (expiration, unvoiced sounds)**
- **V-voiced-quasi-periodic speech**
- speech is a slowly time varying signal over 5-100 msec intervals
- over longer intervals (100 msec-5 sec), the speech characteristics change as rapidly as 10-20 times/second
- => no well-defined or exact regions where individuals sounds begin and end

Speech Sounds

- “Should we chase”
  - /sh/ sound
  - /ould/ sounds
  - /we/ sounds
  - /ich/ sound
  - /a/ sound
  - /s/ sound

- hard to distinguish weak sounds from silence
- hard to segment with high precision => don’t do it when it can be avoided

COOL EDIT demo—‘should’, ‘every’

Estimate of Pitch Period - I

Estimate of Pitch Period - II

Source-System Model of Speech Production
Making Speech “Visible” in 1947

Visible Speech

Spectrogram Properties

Speech Spectrogram — sound intensity versus time and frequency

- **wideband spectrogram** — spectral analysis on ~15 msec sections of waveform using a broad (125 Hz) bandwidth analysis filter, with new analyzes every 1 msec
  - spectral intensity resolves individual periods of the speech and shows vertical striations during voiced regions
- **narrowband spectrogram** — spectral analysis on ~50 msec sections of waveform using a narrow (40 Hz) bandwidth analysis filter, with new analyzes every 1 msec
  - narrowband spectrogram resolves individual pitch harmonics and shows horizontal striations during voiced regions

Wideband and Narrowband Spectrograms

Speech Sentence Waveform

Speech Wideband Spectrogram

Sound Spectrogram

**Wav Surfer:**
www.wavsurfer.com

**VoiceBox:**
www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.htm

**COLEA UI:**
www.utdallas.edu/~loizou/speech/colea.htm

**HMM Toolkit:**
www.ai.mit.edu/~murphyk/Software/HMM/hmm.html#hmm

**COLEA demo** — ’should’, ’every’

**voicebox demo** — ’k5’, ’k5_synthetic’

**WavSurfer demo** — ’s5’, ’s5_synthetic’

Two plus seven is less than ten
Acoustic Theory of Speech Production

- The acoustic characteristics of speech are usually modelled as a sequence of source, vocal tract filter, and radiation characteristics.

\[ P(z) = S(z)T(z)R(z) \]

- For vowel production:

\[ S(z) = \frac{L(z)}{U(z)} \]
\[ T(z) = \frac{L(z)}{U(z)} \]
\[ R(z) = \frac{P(z)}{L(z)} \]

Sound Source for Voiced Sounds

Modelled as a volume velocity source at glottis, \( U(z) \).

<table>
<thead>
<tr>
<th>Gender</th>
<th>( F_0 ) min (Hz)</th>
<th>( F_0 ) max (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>125</td>
<td>200</td>
</tr>
<tr>
<td>Women</td>
<td>225</td>
<td>350</td>
</tr>
<tr>
<td>Children</td>
<td>300</td>
<td>500</td>
</tr>
</tbody>
</table>

Sound Source for Unvoiced Sounds

- Turbulence noise is produced at a constriction in the vocal tract.
- Aspiration noise is produced at glottis.
- Friction noise is produced above the glottis.
- Modelled as series pressure source at constriction, \( P(z) \).

\[ V = \text{Velocity at constriction} \]
\[ D = \text{Critical dimension} = \sqrt{\frac{4A}{\pi}} \]

Parametrization of Spectra

- Human vocal tract is essentially a tube of varying cross sectional area, or can be approximated as a concatenation of tubes of varying cross sectional areas.

- Acoustic theory shows that the transfer function of energy from the excitation source to the output can be described in terms of the natural frequencies or resonances of the tube.
- Resonances known as formants or formant frequencies for speech and they represent the frequencies that pass the most acoustic energy from the source to the output.
- Typically there are 3 significant formants below about 3500 Hz.
- Formants are a highly efficient, compact representation of speech.

Spectrogram and Formants

Key Issue: reliability in estimating formants from spectral data.

Waveform and Spectrogram
Acoustic Theory Summary

• basic speech processes — from ideas to speech (production), from speech to ideas (perception)
• basic vocal production mechanisms — vocal tract, nasal tract, velum
• source of sound flow at the glottis; output of sound flow at the lips and nose
• speech waveforms and properties — voiced, unvoiced, silence, pitch
• speech spectrograms and properties — wideband spectrograms, narrowband spectrograms, formants

English Speech Sounds

ARPABET representation
- 48 sounds
  - 18 vowels/diphthongs
  - 4 vowel-like consonants
  - 21 standard consonants
  - 4 syllabic sounds
  - 1 glottal stop

Phonemes—Link Between Orthography and Speech

Orthography → sequence of sounds
• Larry → /l/ /æ/ /l/ /eɪ/ /r/ /iː/ (/L/ /AE/ /R/ /IY/)

Speech Waveform → sequence of sounds
• based on acoustic properties (temporal) of phonemes

Spectrogram → sequence of sounds
• based on acoustic properties (spectral) of phonemes

The bottom line is that we use a phonetic code as an intermediate representation of language, from either orthography or from waveforms or spectrograms; now we have to learn how to recognize sounds within speech utterances

Phonetic Transcriptions

• based on ideal (dictionary-based) pronunciations of all words in sentence
  - ‘My name is Larry’ /M/ /AE/ /N/ /IH/ /Z/-/L/ /AE/-/R/ /IY/
  - ‘How old are you’ /H/ /AW/-/OW/-/L/ /D/-/AA/-/R/-/Y/-/UW/
  - Speech processing is fun’ /S/ /P/ /IY/-/CH/-/P/ /R/ /AH/-/S/ /EH/-/S/-/IH/-/NG/-/IH/-/Z/-/F/ /AH/-/N/
• word ambiguity abounds
  - ‘lives’ /L/ /IH/ /V/ /Z/ (he lives here) versus /L/ /AY/-/V/-/Z/ (a cat has nine lives)
  - ‘record’ /R/ /EH/-/K/ /ER/-/D/ (he holds the world record) versus /R/ /IY/-/K/ /ER/-/D/ (please record my favorite show tonight)

She had your dark suit in…

“Wideband” Spectrogram
Reduced Set of English Sounds

- 39 sounds
  - 11 vowels (front, mid, back) classification based on tongue hump position
  - 4 diphthongs (vowel-like combinations)
  - 4 semi-vowels (liquids and glides)
  - 3 nasal consonants
  - 6 voiced and unvoiced stop consonants
  - 8 voiced and unvoiced fricative consonants
  - 2 affricate consonants
  - 1 whispered sound
- Look at each class of sounds to characterize their acoustic and spectral properties

Phoneme Classification Chart

Vowels

- Longest duration sounds – least context sensitive
- Can be held indefinitely in singing and other music productions
- Carry very little linguistic information (some languages don’t display vowels in text—Hebrew, Arabic)

Text 1: all vowels deleted
Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nt_s _n th_c_mp_ny’s _m_g_, s_p_r_v_s__n _nd m_n_g_m_nt.

Text 2: all consonants deleted
A__i_u_e_ _o_a__ _a_ __a_e_ e__e__ia___ __e _a_e, _i__ __e __i_e_ o_ o__u_a_io_a_e__e_e__i___ e__ea_i__.

Vowels and Consonants

Text 1: all vowels deleted
Th_y n_t_d s_gn_f_c_nt _mpr_v_m_nt_s _n th_c_mp_ny’s _m_g_, s_p_r_v_s__n _nd m_n_g_m_nt.
(They noted significant improvements in the company’s image, supervision and management.)

Text 2: all consonants deleted
A__i_u_e_ _o_a__ _a_ __a_e_ e__e__ia___ __e _a_e, _i__ __e __i_e_ o_ o__u_a_io_a_e__e_e__i___ e__ea_i__.
(Attitudes toward pay stayed essentially the same, with the scores of occupational employees slightly decreasing)

More Textual Examples

Text (all vowels deleted):
In the next few decades, advances in communications will radically change the way we live and work.

Text (all consonants deleted):
The concept of going to work will change from commuting…
Vowels

- produced using **fixed vocal tract shape**
- **sustained** sounds
- **vocal cords are vibrating** => voiced sounds
- **cross-sectional area** of vocal tract determines vowel resonance frequencies and vowel sound quality
- **tongue position** (height, forward/back position) most important in determining vowel sound
- usually relatively **long in duration** (can be held during singing) and are spectrally well formed

Vowel Production

- No significant constriction in the vocal tract
- Usually produced with periodic excitation
- Acoustic characteristics depend on the position of the jaw, tongue, and lips

Vowel Articulatory Shapes

- tongue hump position (front, mid, back)
- tongue hump height (high, mid, low)
- /IY/, /IH/, /AE/, /EH/ => front => high resonances
- /AA/, /AH/, /AO/ => mid => energy balance
- /UH/, /UW/, /OW/ => back => low frequency resonances

Vowel Waveforms & Spectrograms

- Synthetic versions of the 10 vowels

Vowel Formants

- Clear pattern of variability of vowel pronunciation among men, women and children
- Strong overlap for different vowel sounds by different talkers ⇒ no unique identification of vowel strictly from resonances ⇒ need context to define vowel sound

The Vowel Triangle

Centroids of common vowels form clear triangular pattern in F1-F2 space
Canonic Vowel Spectra

100 Hz Fundamental

10 Hz

33 Hz

100 Hz

100 Hz Fundamental

300 Hz

500 Hz

200 Hz

400 Hz

600 Hz

Eliminating Vowels and Consonants

Diphthongs

• Gilding speech sound that starts at or near the articulatory position for one vowel and moves to or toward the position for another vowel
  – /AY/ in buy
  – /AW/ in down
  – /EY/ in bait
  – /OY/ in boy

Distinctive Features

Classify non-vowel/non-diphthong sounds in terms of distinctive features

– place of articulation
  • Bilabial (lips)—p,b,m,w
  • Labiodental (between lips and front of teeth)—f,v
  • Dental (teeth)—th,dh
  • Alveolar (front of palate)—t,d,s,z,n,l
  • Palatal (middle of palate)—sh,zh,r
  • Velar (at velum)—k,g,ng
  • Pharyngeal (at end of pharynx)—h

– manner of articulation
  • Glide—smooth motion—w,l,r,y
  • Nasal—lowered velum—m,n,ng
  • Stop—constricted vocal tract—p,t,k,b,d,g
  • Fricative—turbulent source—f,th,s,sh,v,dh,z,zh,h
  • Voicing—voiced source—b,d,g,v,dh,z,zh,m,n,ng,w,l,r
  • Mixed source—both voicing and unvoiced—j,ch
  • Whispered—h

Places of Articulation
### Semivowels (Liquids and Glides)

- Vowel-like in nature (called semivowels for this reason)
- Voiced sounds (w-l-r-y)
- Acoustic characteristics of these sounds are strongly influenced by context—unlike most vowel sounds which are much less influenced by context

\[
\text{Manner: glides} \\
\text{Place: bilabial (w), alveolar (l), palatal (r)}
\]

\[
\text{uh-}\{w,l,r,y\}-a
\]

### Nasal Consonants

- The nasal consonants consist of /M/, /N/, and /NG/
  - Nasals produced using glottal excitation \( \Rightarrow \) voiced sounds
  - Vocal tract totally constricted at some point along the tract
  - Velum lowered so sound is radiated at nostrils
  - Constricted oral cavity serves as a resonant cavity that traps acoustic energy at certain natural frequencies (anti-resonances or zeros of transmission)
  - /M/ is produced with a constriction at the lips \( \Rightarrow \) low frequency zero
  - /N/ is produced with a constriction just behind the teeth \( \Rightarrow \) higher frequency zero
  - /NG/ is produced with a constriction just forward of the velum \( \Rightarrow \) even higher frequency zero

\[
\text{Manner: nasal} \\
\text{Place: bilabial (m), alveolar (n), velar (ng)}
\]

\[
\text{uh-}\{m,n,ng\}-a
\]

### Nasal Production

- Velum lowering results in airflow through nasal cavity
- Consonants produced with closure in oral cavity
- Nasal murmurs have similar spectral characteristics

\[
[m] \\
[n] \\
[b]
\]

### Nasal Sounds

\[
\text{UH M AA} \\
\text{UH N AA}
\]

### Nasal Spectrograms

\[
\text{simmer} /\text{sim}ər/ \\
\text{sinner} /\text{sɪnər/} \\
\text{singer} /\text{sɪŋər/
\]

### Unvoiced Fricatives

- Consonant sounds /F/, /TH/, /S/, /SH/
  - Produced by exciting vocal tract by steady air flow which becomes turbulent in region of a constriction in the vocal tract
  - /F/ constriction near the lips
  - /TH/ constriction near the teeth
  - /S/ constriction near the middle of the vocal tract
  - /SH/ constriction near the back of the vocal tract
  - Noise source at constriction \( \Rightarrow \) vocal tract is separated into two cavities
  - Sound radiated from lips \( - \) front cavity
  - Back cavity traps energy and produces anti-resonances (zeros of transmission)

\[
\text{Manner: fricative} \\
\text{Place: labiodental (f), dental (th), alveolar (s), palatal (sh)}
\]

\[
\text{uh-}\{f,th,s,sh\}-a
\]
Unvoiced Fricative Production

Unvoiced Fricatives

- Sounds /f/, /θ/, /s/, /z/, /ʃ/, /ʒ/
  - place of constriction same as for unvoiced counterparts
  - two sources of excitation; vocal cords vibrating producing semi-periodic puffs of air to excite the tract; the resulting air flow becomes turbulent at the constriction giving a noise-like component in addition to the voiced-like component

Unvoiced Fricative Spectrograms

Voiced Fricatives

- Sounds /v/, /ð/, /z/, /ʒ/
  - place of constriction same as for unvoiced counterparts
  - two sources of excitation; vocal cords vibrating producing semi-periodic puffs of air to excite the tract; the resulting air flow becomes turbulent at the constriction giving a noise-like component in addition to the voiced-like component

Voiced and Unvoiced Stop Consonants

- Sounds /b/, /d/, /g/ (voiced stop consonants) and /p/, /t/, /k/ (unvoiced stop consonants)
  - voiced stops are transient sounds produced by building up pressure behind a total constriction in the oral tract and then suddenly releasing the pressure, resulting in a pop-like sound
  - /b/ constriction at lips
  - /d/ constriction at back of teeth
  - /g/ constriction at velum
  - no sound is radiated from the lips during constriction => sometimes sound is radiated from the throat during constriction (leakage through tract walls) allowing vocal cords to vibrate in spite of total constriction
  - stop sounds strongly influenced by surrounding sounds
  - unvoiced stops have no vocal cord vibration during period of closure => brief period of friction (due to sudden turbulence of escaping air) and aspiration (steady air flow from the glottis) before voiced excitation begins
Stop Consonant Production

- Complete closure in the vocal tract, pressure build up
- Sudden release of the constriction, turbulence noise
- Can have periodic excitation during closure

Voiced Stop Consonant

Unvoiced Stop Consonants

Stop Gap

Stop Consonant Waveforms and Spectrograms

Distinctive Phoneme Features

- the brain recognizes sounds by doing a distinctive feature analysis from the information going to the brain
- the distinctive features are somewhat insensitive to noise, background, reverberation => they are robust and reliable

Distinctive Features

- place and manner of articulation completely define the consonant sounds, making speech perception robust to a range of external factors
Review Exercises

Write the transcription of the sentence “Good friends are hard to find”

G-UH-D F-R-EH-N-D-Z AA-R HH-AA-R-D T-UH (UW) F-AY-N-D
Ultimate Exercise—Identify Words From Spectrogram

Word Choices:
that, and, was, by, people, little, simple, between, very, enjoy, only, other, company, those

was/ — this word can be identified by the voiced initial portion with very low first and second formants (sounds like UW or W), followed by the AA sound and ending with the Z (S) sound.

Ultimate Exercise—Identify Words From Spectrogram

Word Choices:
that, and, was, by, people, little, simple, between, very, enjoy, only, other, company, those

enjoy/ — this word can be identified by the two-syllable nature, with the nasal sound N at the end of the first syllable, and the fricative JH at the beginning of the second syllable, with the characteristic OY diphthong at the end of the word.

Ultimate Exercise—Identify Words From Spectrogram

Word Choices:
that, and, was, by, people, little, simple, between, very, enjoy, only, other, company, those

company/ — this word can be identified by the three syllable nature, with the initial stop consonant K, the first syllable ending in the nasal M, followed by the stop P, and with the second syllable ending with the nasal N followed by an IY vowel-like sound.

Ultimate Exercise—Identify Words From Spectrogram

Word Choices:
that, and, was, by, people, little, simple, between, very, enjoy, only, other, company, those

simple/ — this word can be identified by the two-syllable nature, with a strong initial fricative S beginning the first syllable and the nasal M ending the first syllable, and with the stop consonant P beginning the second syllable.

Summary

- sounds of the English language—phonemes, syllables, words
- phonetic transcriptions of words and sentences — coarticulation across word boundaries
- vowels and consonents — their roles, articulatory shapes, waveforms, spectrograms, formants
- distinctive feature representations of speech