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 of sounds in the acoustic waveform
• Location of sounds in spectrograms
• Articulatory properties of speech sounds—place and manner of articulation

Basic Speech Processes

• idea Æ sentences Æ words Æ sounds Æ waveform Æ sounds Æ words Æ sentences Æ idea
  — 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?
  — Sounds: /h/ /æ/-/l/-/d/-/ih/ /d/-/y/
  — Coarticulated Sounds: /h-æ/-/l/-/d/-/ih/-/d/-/y/-/j/-/æ/-/t/-/j/-/æ/-/t/ (hi-al-dia-jejt)
• remarkably, humans can decode these sounds and determine the meaning that was intended—at least at the idea/concept level (perhaps not completely at the word or sound level); often machines can also do the same task
  — speech coding: waveform Æ (model) Æ waveform
  — speech synthesis: words Æ waveform
  — speech recognition: waveform Æ words/sentences
  — speech understanding: waveform Æ idea

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

Human Vocal Apparatus

• vocal tract—dotted lines in figure begin at the glottis (the vocal cords) and end 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
• valvel—a trapper-like mechanism at the back of the mouth cavity, lowers to couple the nasal tract to the vocal tract to produce the nasal sounds like /m/ (mom), /n/ (night), /ng/ (sing)

Mid-sagittal plane X-ray of human vocal apparatus

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

Acoustic Tube Models Demo

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

Artificial Larynx Demo

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

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 ms line, 0.5 sec for utterance
- S-silence-background-no speech
- U-unvoiced, no vocal cord vibration (aspiration, 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
  - /ch/ 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

Speech Sounds Demo—'should', 'every'

Estimate of Pitch Period - I

Estimate of Pitch Period - II

Source-System Model of Speech Production

Making Speech “Visible” in 1947
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

Sound Spectrogram

WavSurfer: www.wavsurfer.com
VoiceBox: www.ee.ic.ac.uk/hp/staff/dmb/voicbox/voicebox.htm
COLEA UI: www.utdallas.edu/~loizou/speech/colea.htm
HMM Toolkit: www.ai.mit.edu/~murphyk/Software/HMM/hmm.html#hmm

Speech Sentence Waveform

Two plus seven is less than ten

Speech Wideband Spectrogram

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

For vowel production:

\[ P_v(t) = S_v(t) T_v(t) R_v(t) \]

\[ S_v(t) = \frac{u_v(t)}{T_v(t)} \]
\[ T_v(t) = \frac{u_v(t)}{u_v(t)/u_v(t)} \]
\[ R_v(t) = P_v(t)/u_v(t) \]
**Sound Source for Voiced Sounds**

Modelled as a volume velocity source at glottis, $f_v(t)$:

$$f_v(t) = \begin{cases} \frac{V}{A} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

<table>
<thead>
<tr>
<th></th>
<th>$f_v$ ave (Hz)</th>
<th>$f_v$ min (Hz)</th>
<th>$f_v$ max (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Men</td>
<td>125</td>
<td>80</td>
<td>200</td>
</tr>
<tr>
<td>Women</td>
<td>225</td>
<td>150</td>
<td>350</td>
</tr>
<tr>
<td>Children</td>
<td>300</td>
<td>200</td>
<td>300</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, $P_0(t)$:

$$P_0(t) = \begin{cases} \frac{V}{D} & \text{if } t > 0 \\ 0 & \text{otherwise} \end{cases}$$

$V$: Velocity at constriction, $D$: Critical dimension $= \sqrt{\frac{4P_0}{\mu}}$

**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**

- Why do I owe you a letter?

**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/ /iy/ (/L/ /AE/ /L/ /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’ /I/ /AY/ /n/ /IY/ /M/ /IH/ /ZI/-/LU/ /AE/ /R/ /IY/
  - ‘How old are you?’ /H/ /AW/-/OW/ /L/ /D/-/AA/ /R/-/Y/ /UW/
  - ‘Speech processing is fun’ /SH/-/I/ /IY/ /CH/-/P/ /R/-/AY/ /S/-/IH/-/NG/-/IH/ /ZI/-/F/-/AH/-/N/
- word ambiguity abounds
  - ‘lives’ /L/ /IH/ /V/ /ZI/ (he lives here) versus /L/ /AY/ /V/ /ZI/ (a cat has nine lives)
  - ‘record’ /R/ /EH/-/K/ /ER/-/D/ (he holds the world record) versus /R/-/IY/-/K/-/AW/-/D/ (please record my favorite show tonight)

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 musical works (opera)
- 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_g_n_f_c_nt _mpr_v_mnt_s n th_c_mp_n_y's
_m_g_. s_p_rv_s_n_d m_n_g_m_n_t.

**Text 2: all consonants deleted**
A_i_u_e_ _o_a_. _a_e_e_e_e a_e_e_i_e_.

(They noted significant improvements in the company's image, supervision and management.)

**Vowels and Consonants**

**Text 1: all vowels deleted**
Th_y_n_t_d s_g_n_f_c_nt _mpr_v_mnt_s n th_c_mp_n_y's
_m_g_. s_p_rv_s_n_d m_n_g_m_n_t.

(They noted significant improvements in the company's image, supervision and management.)

**Text 2: all consonants deleted**
A_i_u_e_ _o_a_. _a_e_e_e_e a_e_e_i_e_.

(Attitudes toward pay stayed essentially the same, with the scores of occupational employees slightly decreasing)

**More Textual Examples**

**Text (all vowels deleted):**
_n th_n_xt f_w_d_c_d_s_. _dv_n_c_s_.

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

**Text (all consonants deleted):**
_e_o_e_o_o_i_o_o_i_e_.

(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

| [i] | [e] | [a] | [u] |

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 18 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
**Canonic Vowel Spectra**

- **IY**
- **AA**
- **UW**

**100 Hz Fundamental**

**300 Hz Fundamental**

**Eliminating Vowels and Consonants**

- Frequency (kHz)
- Time (s)

**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,w,l,r
  - Mixed source—both voicing and unvoiced—j,ch
  - Whispered—h

**Diphthongs**

- Gliding 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 boy
  - /AW/ in down
  - /EY/ in bait
  - /OY/ in boy
  - /OW/ in boat (usually classified as vowel, not diphthong)
  - /Y/ in you (usually classified as glide)

**Places of Articulation**

- Palato-Alveolar
- Alveolar
- Labial
- Dental
- Velar
- Uvular

**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.
Nasal Consonants

- The nasal consonants consist of /M/, /N/, and /NG/
  - nasals produced using glottal excitation => 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 => low frequency zero
  - /N/ is produced with a constriction just behind the teeth => higher frequency zero
  - /NG/ is produced with a constriction just forward of the velum => even higher frequency zero

Manner: nasal
Place: bilabial (m), alveolar (n), velar (ng)

Nasal Sounds

Nasal Spectrograms

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 => vocal tract is separated into two cavities
  - sound radiated from lips – front cavity
  - back cavity traps energy and produces anti-resonances (zeros of transmission)

Manner: fricative
Place: labiodental (f), dental (th), alveolar (s), palatal (sh)
Unvoiced Fricatives

- Sounds /f/, /θ/, /s/, /z/
- Place of constriction: labiodental, dental, alveolar, palatal

Voiced Fricatives

- Sounds /v/, /ð/, /z/, /ʃ/
- Place of constriction same as for unvoiced counterparts

Voiced and Unvoiced Stop Consonants

- Sounds /b/, /d/, /ɡ/ (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
- Unvoiced stops have no vocal cord vibration during period of closure => brief period of frication (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

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Glide</th>
<th>Nasal</th>
<th>Voiced</th>
<th>Unvoiced</th>
<th>Voiced</th>
<th>Unvoiced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Front</td>
<td>m,m</td>
<td></td>
<td>b</td>
<td></td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>Bilabial</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Labiodental</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Middle</td>
<td>v</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dental</td>
<td></td>
<td></td>
<td>a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alveolar</td>
<td></td>
<td>n</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Palatal</td>
<td></td>
<td></td>
<td>d</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Back</td>
<td></td>
<td></td>
<td>t</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Velar</td>
<td></td>
<td></td>
<td>l</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharyngeal</td>
<td></td>
<td></td>
<td>k</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glottal</td>
<td>r</td>
<td></td>
<td>h</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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
Review Exercises

File: simple10k, sampling rate: 10000, starting sample: 1 number of samples 7152

Simple:
S-IH-M-P-(AX-L | EL)

Offset
Enjoy:
EH-N-JH-OY

This is a test (16 kHz sampling rate)

Ultimate Exercise—Identify Words From Spectrogram

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

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

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

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

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