Human Speech Production

Introduction

- Overview of (human) sound generation
- Basic set of sounds, phonemes
- Textual, waveform, and spectral representations of sounds
- Articulatory properties of sounds

Speech Signals

- The basic meaningful unit in a speech signal is the sound. That is, a speech signal contains a sequence of sounds.
- One needs to classify the sounds in the speech to understand speech.
  - The classification of sounds is called **phonetics**.
- A sequence of sounds is then interpreted to be words, syntax, semantics, etc., which is the subject of **linguistics**.

A Heuristic Model of Speech

- From the speaker (generation)
  - Beginning with a thought or intent to communicate
  - Requiring proper muscular movement of the vocal tract
- Soundwaves in the air (transmission)
- To the listener (analysis)
  - Receiving sounds with the auditory system
  - Converting to neurological signals for the brain to interpret

Vocal Tract and Nasal Tract

- vocal tract
  - from the vocal cords (*glottis*) to the lips
  - consists of the *pharynx*, and *oral cavity*
  - length: $17cm$
  - the cross-sectional function depends on the positions of the *tongue, lips, jaw*, and *velum*, $0 - 20cm^2$
- nasal tract
  - from the velum to the nostrils
  - consists of the and *nasal cavity*

Speech Production Mechanism

- Air enters the lungs via breathing;
- Air is expelled from the lungs, through the trachea, and cause to vocal cords to vibrate;
- Air flow is chopped up into quasi-periodic pulses;
- The pulses are frequency-shaped by the oral cavity and the nasal cavity;
- Body parts involved in speech production: lungs, trachea, vocal cords within the larynx, velum (soft palate), hard palate, tongue, teeth, lips, nasal tract;

Speech Properties

- Voiced, Unvoiced, and Silence
  - should we chase?
- Pitch periods
  - male: $5 - 12.5$ ms
  - female: $3 - 7$ ms
  - child: $2 - 5$ ms
- Certain properties can be seen from the waveform representation of speech.

Simplified Speech Production Model

$$s(t) = e(t) * v(t) \quad \Rightarrow \quad S(j\Omega) = E(j\Omega) \cdot V(j\Omega) \tag{1}$$

- Pitch period is related to the spacing in the spectrum

$$\delta f = \frac{1}{\tau_p} \tag{2}$$

- Spectral shape is related to the vocal tract frequency response;
- The mode of exicitaion, pitch period, and the vocal tract frequency response all vary with time during speech production. We need **short-time analysis**.

Short-time Fourier Representation

Short-time Fourier transform

$$X_c(\hat{t}, j\Omega) = \int_t w_c(\hat{t} - t) x_c(t) e^{-j\Omega t} dt \tag{3}$$

$w(t)$ is a window function, and $w_c(\hat{t} - t)$ is an analysis window around $\hat{t}$.
The discrete version is

$$X_{\hat{n}}(e^{j\omega}) = \sum_m w[\hat{n} - m] x[m] e^{-j\omega m} \tag{4}$$

Wideband and Narrowband Spectrogram

- wideband: short analysis window
- narrow-band: long analysis window

Formants

- The dark strips in spectrogram
- The resonance frequencies of the vocal tract

Acoustic Phonetics

Phonetic Alphabets

- Most languages can be described by a set of distinctive **phonemes**.
- Phonemes can be represented by phonetic alphbets. For example, the phonemes of the American English can be represented by ARPAbet of IPA (international phonetic alphabet).

American English

- 48 phonemes
  - 14 vowels and 4 diphthongs
  - 4 vowel-like consonants (semi-vowels)
  - 21 consonants
  - 4 syllabic sounds
  - 1 glottal stop
- A different set of phonemes can be used to describe the same American English.
- continuant vs. non-continuant sounds
  - continuant: fixed vocal tract configuration
  - non-continuant: time-varying vocal tract configuration, including diphthongs, semivowels, stops, and affricates

The phoneme representation of a sentence is only *approximate*.

Vowels

- well-defined of the sounds in a language
- long durations
- carry very little linguistic information compared to the consonants
- the vowel triangle

Consonants

Distinctive Features

- Place of articulation: maximum constriction point along the vocal tract, related to $V(j\Omega)$ in (1)
- Manner of articulation: the characteristics of the excitation signal, related to $E(j\Omega)$, or the manner that articulators move

Place of Articulation

- labial/bilabial: /p/, /b/, /m/, /w/
- labiodental: /f/, /v/
- dental: /θ/
- alveolar (at the front of palate): /t/, /d/, /s/, /z/, /n/, /l/
- palatal (middle of palate): /š/, /ž/, /r/
- velar (at the velum): /k/, /g/, /ŋ/
- pharyngeal: whispered /h/

Manner of Articulation

- glide (moving articulators): /w/, /l/, /r/, /y/
- nasal (lowered velum): /m/, /n/, /ŋ/
- stop (air flow blocked): /p/, /t/, /k/, /g/, /b/, /d/
- fricative (non-vibrating cords): /f/, /θ/, etc.
- voicing (vibrating cords): /b/, /g/, /d/, etc.

Semivowels

- /w/ is close to /u/
- /l/ is close to /o/
- /y/ is close to /i/
- /r/ is close to /ɝ/

Nasals

- use the nasal tract
- /m/: constriction at the lips
- /n/: constriction at the back of the teeth
- /ŋ/: constriction at the front of the velum
- introduce low formant and low-energy region in the spectrogram

Fricatives

- turbulent air flow
- /f/: constriction at the lips
- /θ/: constriction near the teeth
- /s/: constriction near the middle of the vocal tract
- /š/: constriction near the back of the vocal tract
- noise spectrum at high frequency

Stops

- complete block of air flow followed by sudden release
- /b/, /p/: constriction at the lips
- /d/, /t/: constriction near the teeth
- /g/, /k/: constriction near the velum
- highly influenced by the following vowel (context-dependent)

Affricates and Whisper

- combining a stop and a fricative
- /č/: /t š/, unvoiced
- /j/: /d ž/, voiced
- /h/: exciting the vocal tract by a steady air flow (no constriction, but turbulence at the glottis)

Distinctive Feature Chart

- Figure 3.39

Coarticulation

- **Coarticulation**: The articulation of a phoneme changes due to its neighboring phones.
- Very common in conversational speech
- For example,

Did you hit it to Tom?

  - palatalization (*did you*)
  - reduction of unstressed sound (*you*)
  - flapping (*hit it*)
  - reduction of schwa (*to*)
  - reduction of germinate (double consonant) (*it to*)

Syllable

- **Cohesive units** formed by phonemes
- Generally centered around a **vowel**
  - syllable = **[onset] nucleus [coda]**
  - **rime** = nucleus [coda]
- For an example of very lengthy syllable, *strengths*
  - onset: /s t r/
  - nucleus: /ɛ/
  - coda: /ŋ θ s/

Words

- *Acoustically*, word is the unit just larger than syllable.
- *Grammatically*, words have **part-of-speech (POS)** tags.
- *Semantically*, word bears *meaning* (**morpheme**).
- *Computationally*, word is often the basic unit for **language models**, and the recognition unit in the **acoustic models**.

Syntax

- A **sentence** is a string of words. However, not all strings of words are sentences.
- A **syntax** is a set of rules for “good” sentences.
- A grammatical rule is specified by a set of symbols, including words, part-of-speech tags, and phrases.
  - Each phrase symbol stands for a **constituent**.
- The *structure* of a sentence can be represented by a **parse tree**