**Overview**

- The Physics of Sound
- The Production of Speech
- The Perception of Speech and Audio
  - Frequency Masking
  - Noise Masking
  - Temporal Masking
- Phonetics and Phonology
- Analog and Digital Signals

---

**The Physics of Sound**

| Speed of Sound | Units       | Temperature (°C) |
|----------------|-------------|------------------|
| Air (at sea level, 20 C) | 343 m/s (1230 km/h) | V=(331+0.6T) m/s |
| Water (at 20 C)            | 1482 m/s (5335 km/h) |
| Steel                      | 5960 m/s (21460 km/h) |
| Tendon                     | 1650 m/s |

**Speed of Sound**
- Proportional to \( \sqrt{\text{elastic modulus/density}} \)
- Second order dependency on the amplitude of the sound \( \Rightarrow \) nonlinear propagation effects

---

**The Production of Speech**

How is SPEECH produced?
- Characteristics of Acoustic Signal

Speech Signal \( \rightarrow \) Speech Recognition "How are you?"
Human Speech Production

- Physiology
  - Schematic and X-ray Sagittal View
  - Vocal Cords at Work
  - Transduction
  - Spectrogram

- Acoustics
  - Acoustic Theory
  - Wave Propagation

---

Sagittal Plane View of the Human Vocal Apparatus

[Image: Sagittal Plane View of the Human Vocal Apparatus]

---

Characterization of English Phonemes

|     | Place | Consonants |
|-----|-------|------------|
| /p/ | Bilabial | p |
| /b/ | Bilabial | b |
| /t/ | Dental | t |
| /d/ | Dental | d |
| /k/ | Palatal | k |
| /g/ | Palatal | g |
| /f/ | Labiodental | f |
| /v/ | Labiodental | v |
| /θ/ | Dental-velar | θ |
| /ð/ | Dental-velar | ð |
| /s/ | Dental-velar | s |
| /z/ | dental-velar | z |

---

Picture by Y.Alhabib from Wikipedia
The Vowel Space

- We can characterize a vowel sound by the locations of the first and second spectral resonances, known as formant frequencies.
- Some voiced sounds, such as diphthongs, are transitional sounds that move from one vowel location to another.

Phonetics

Formant Frequency Ranges
Vocal Tract Workshop

Articulatory Speech Synthesis using TractSyn (P. Birkholz, 2004).

Reading material:
P. Birkholz, D. Jackel, A Three-Dimensional Model of the Vocal Tract for Speech Synthesis. In Proceedings of the 15th International Congress of Phonetic Sciences, pp. 2597-2600, Barcelona, Spain, 2003.

See: http://www.vocaltractlab.de/
The Perception of Speech: The Ear

- Three main sections, outer, middle, and inner ear.
- The outer and middle ears reproduce the analog signal (impedance matching).
- The inner ear transduces the pressure wave into an electrical signal.
- The outer ear consists of the external visible part and the auditory canal. The tube is about 2.5 cm long.
- The middle ear consists of the eardrum and three bones (malleus, incus, and stapes). It converts the sound pressure wave to displacement of the oval window (entrance to the inner ear).

The Perception of Speech: The Ear

- The inner ear primarily consists of a fluid-filled tube (cochlea) which contains the basilar membrane. Fluid movement along the basilar membrane displaces hair cells, which generate electrical signals.
- There are a discrete number of hair cells (30,000). Each hair cell is tuned to a different frequency.
- Place vs. Temporal Theory: firings of hair cells are processed by two types of neurons (onset chopper units for temporal features and transient chopper units for spectral features).
Perception

Psychoacoustics

- Psychoacoustics: a branch of science dealing with hearing, the sensations produced by sounds.
- A basic distinction must be made between the perceptual attributes of a sound vs measurable physical quantities:
  - Many physical quantities are perceived on a logarithmic scale (e.g. loudness). Our perception is often a nonlinear function of the absolute value of the physical quantity being measured (e.g. equal loudness).
- Timbre can be used to describe why musical instruments sound different.
- What factors contribute to speaker identity?

| Physical Quantity     | Perceptual Quality |
|-----------------------|--------------------|
| Intensity             | Loudness           |
| Fundamental Frequency | Pitch              |
| Spectral Shape        | Timbre             |
| Onset/Offset Time     | Timing             |
| Phase Difference      | Location           |

Perception, Non-Linear Frequency Warping: Bark and Mel Scale

- **Critical Bandwidths**: correspond to approximately 1.5 mm spacings along the basilar membrane, suggesting a set of 24 bandpass filters.
- **Critical Band**: can be related to a bandpass filter whose frequency response corresponds to the tuning curves of auditory neurons. A frequency range over which two sounds will sound like they are fusing into one.

- **Bark Scale**: \( \text{Bark} = 13 \times \text{atan} \left( \frac{6.76f}{1000} \right) + 3.5 \times \text{atan} \left( \frac{f}{7500} \right) \)
- **Mel Scale**: \( \text{mel frequency} = 2595 \log_{10} \left( 1 + \frac{f}{700} \right) \)

Perception

Equal Loudness

- Just Noticeable Difference (JND): The acoustic value at which 75% of responses judge stimuli to be different (limen)
- The perceptual loudness of a sound is specified via its relative intensity above the threshold. A sound’s loudness is often defined in terms of how intense a reference 1 kHz tone must be heard to sound as loud.

Perception

Bark and Mel Scale

- The Bark scale implies a nonlinear frequency mapping
Perception: Bark Scale and Mel Scale

- Filter Banks used in ASR:
- The Bark scale implies a nonlinear frequency mapping

---

Perception: Tone-Masking Noise

- **Frequency masking**: one sound cannot be perceived if another sound close in frequency has a high enough level. The first sound masks the second.
- **Tone-masking noise**: noise with energy $EN$ (dB) at Bark frequency $g$ masks a tone at Bark frequency $b$ if the tone’s energy is below the threshold:
  \[ TT(b) = EN - 6.025 - 0.275g + Sm(b-g) \] (dB SPL)
  where the spread-of-masking function $Sm(b)$ is given by:
  \[ Sm(b) = 15.81 + 7.5(b+0.474) - 17.5 \times \sqrt{1 + (b+0.474)^2} \] (dB)
- **Temporal masking**: onsets of sounds are masked in the time domain through a similar masking process.
  - Thresholds are frequency and energy dependent.
  - Thresholds depend on the nature of the sound as well.

---

Comparison of Bark and Mel Space Scales

Perception: Noise-Masking Tone

- **Noise-masking tone**: a tone at Bark frequency $g$ with energy $ET$ (dB) masks noise at Bark frequency $b$ if the noise energy is below the threshold:
  \[ TN(b) = ET - 2.025 - 0.17g + Sm(b-g) \] (dB SPL)
  - Masking thresholds are commonly referred to as Bark scale functions of just noticeable differences (JND).
  - Thresholds are not symmetric.
  - Thresholds depend on the nature of the noise and the sound.
Perceptual Noise Weighting

Simple Z-Transform interpretation:
- can be implemented by evaluating the Z-Transform around a contour closer to the origin in the z-plane:
  \[ H_{nw}(z) = H(az) \]
- Used in many speech compression systems (Code Excited Linear Prediction).
- Analysis performed on bandwidth-broadened speech; synthesis performed using normal speech.

Effectively shapes noise to fall under the formants.

Perception: Echo and Delay

- Humans are used to hearing their voice while they speak - real-time feedback (side tone).
- When we place headphones over our ears, which dampens this feedback, we tend to speak louder.
- Lombard Effect: Humans speak louder in the presence of ambient noise.
- When this side-tone is delayed, it interrupts our cognitive processes, and degrades our speech.
- This effect begins at delays of approximately 250 ms.
- Modern telephony systems have been designed to maintain delays lower than this value (long distance phone calls routed over satellites).
- Digital speech processing systems can introduce large amounts of delay due to non-real-time processing.
Perception: Adaptation

- **Adaptation** refers to changing sensitivity in response to a continued stimulus, and is likely a feature of the mechano-electrical transformation in the cochlea.
- Neurons tuned to a frequency where energy is present do not change their firing rate drastically for the next sound.
- Additive broadband noise does not significantly change the firing rate for a neuron in the region of a formant.

Visual Adaptation

- The **McGurk Effect** is an auditory illusion which results from combining a face pronouncing a certain syllable with the sound of a different syllable. The illusion is stronger for some combinations than for others. For example, an auditory ‘ba’ combined with a visual ‘ga’ is perceived by some percentage of people as ‘da’. A larger proportion will perceive an auditory ‘ma’ with a visual ‘ka’ as ‘na’. Some researchers have measured evoked electrical signals matching the ‘perceived’ sound.

Perception: Timing

- Temporal resolution of the ear is crucial.
- Two clicks are perceived mono-aurally as one unless they are separated by at least 2 ms.
- 17 ms of separation is required before we can reliably determine the order of the clicks. (~58bps or ~3530bpm)
- Sounds with onsets faster than 20 ms are perceived as "plucks" rather than "bows".
- Short sounds near the threshold of hearing must exceed a certain intensity-time product to be perceived.
- Humans do not perceive individual “phonemes” in fluent speech - they are simply too short. We somehow integrate the effect over intervals of approximately 100 ms.
- Humans are very sensitive to long-term periodicity (ultralow frequency) - this has implications for random noise generation.