Hearing, Auditory Models, and Speech Perception

Introduction

- The *receiver* side of speech communication, namely human speech perception and understanding
- With a good understanding of how humans hear sounds and perceive speech, we are better able to design and implement robust and efficient speech processing systems
- Black-box approach for complex mechanism not fully understood, such as the processing of neural signals in the brain (psychoacoustic experiments)

Key Findings by Psychoacoustic Experiments

- Frequency is perceived as pitch on a non-linear scale
- Intensity is perceived as loudness on a compressive scale
- Syllable perception is based on a long-term spectral integration process
- Auditory masking effect for robustness

The Speech Chain

- Figure 4.1
- From production to perception
- 3 levels
  - linguistic level: where the words (sounds) are chosen and perceived
  - physiological level: where the articulators in the vocal tract are set in motion to produce the soundwave, and the soundwave is converted by the ear to neural signals
  - acoustic level: soundwave

Auditory System

- Basic function: Converting acoustic signal to perceived sounds, Figure 4.2
- The ear is only a front-end of the auditory system for acoustic to neural conversion. It also includes the neural transduction and neural processing.
- Black-box behavioral model: Using psychophysical (psychoacoustical) observation, Figure 4.3

Psychoacoustic Experiments

- Instead of the more complicated signal of human speech, the simpler signal of *tone* and *noise* are used

- These signals have well-defined and *controllable* parameters, such as the intensity, frequency, and bandwidth
- Note that even with such controlled setting, the relationship between the variables and observables are still complicated

Ear

- outer ear: consisting of the pinna and the external canal; conducting the pressure wave to the *tympanic membrane* or eardrum
- middle ear: including three small connected bones *malleus* (hammer), *incus* (anvil), and *stapes* (stirrup); converting the pressure wave to mechanical wave
- inner ear: consisting of the *cochlea*, the *basilar membrane*, and the neural connections (auditory nerves) to the brain

Auditory Gain

- The outer ear increases the hearing sensitivity by a factor of $2 \sim 3$
- The middle ear is a mechanical transducer with a gain of $3 \sim 15$ from the eardrum to the stirrup; the muscles around the small bones also protect the inner ear against the damage of very loud sounds

Cochlea

- The cochlea in the inner ear is a “microphone” that converts mechanical wave to electrical signal
- $\sim 2\frac{1}{2}$-turn snail shaped, ~ 3 centi-meter long, fluid-filled chamber partitioned longitudinally by the basilar membrane
- Mechanical vibration at the stapes creates (through an oval window) standing waves (the fluid as media), causing the basilar membrane to vibrate accordingly
- $\sim 30,000$ *inner hair cells* (IHC) are lined with the basilar membrane. ~ 10 nerve fibers per IHC. IHCs are set in motion by the vibration of the BM, and fire neural spikes at rates depending on the amplitudes of the vibration.

Basilar Membrane

- Different parts of BM are tuned to different frequencies
- A non-uniform bank of filters, with constant Q (Figure 4.8)
- Performs an effective spectral-temporal analysis via a spatio-temporal representation of information

Critical Bands

- The critical bandwidths of the basilar membrane filters have been defined and measured
- The effective bandwidths are constant at about 100Hz for center frequencies below 500Hz, and 20% of the center frequencies above 500Hz.

- An equation for this relationship is

$$\Delta f = 25 + 75[1 + 1.4(f_c/1000)^2]^{0.69} \tag{1}$$

- For non-overlapping *idealized* critical bands, approximately 25 filters span the frequency range from 0 to $20,000$Hz.

Bark Scale

- The index of the idealized critical-band filter
- That is, each Bark scale corresponds to a filter
- May refer to the correspondence between the index and the high cut-off frequency

$$\Omega = 6\log_e\left\{\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right\} \tag{2}$$

- Table 4.2
- The Bark-scale spectrum

The Perception of Sound

Intensity of Sound

- *Acoustic intensity* (I): the average flow of energy per unit time per unit area (in $\frac{\text{Watt}}{\text{m}^2}$)
- The range of audible intensity is $10^{-12} \sim 10 \quad \frac{\text{Watt}}{\text{m}^2}$ (for a 1000Hz tone)
- The intensity at the threshold of hearing is defined to be

$$I_0 = 10^{-12} \quad \frac{\text{Watt}}{\text{m}^2} \tag{3}$$

- *Intensity level* (IL): the acoustic intensity in dB (decibel), defined as

$$IL = 10\log_{10}\frac{I}{I_0}\ \text{dB}. \tag{4}$$

- *Sound pressure level* (SPL): an alternative way to quantify the intensity level, defined as

$$SPL = 20\log_{10}\frac{P}{P_0}\ \text{dB}, \tag{5}$$

where

$$P_0 = 2 \times 10^{-5} \quad \frac{\text{Newton}}{\text{m}^2} \tag{6}$$

is the pressure corresponding to I_0 defined in (3).

Range of Human Hearing

- Anechoic chamber at the Bell Labs
- From 0 dB to 130 dB
- Sound levels in decibels, Figure 4.12 and Table 4.3

Loudness Level

- *Loudness level* (LL) of a tone: the IL or SPL of a 1000-Hz tone that sounds as loud as the tone

- *phon*: the unit for the loudness level;
- The loudness level of a tone is x phons if it is as loud as a 1000-Hz tone of x-dB SPL
- The equal loudness-level curves (Figure 4.14). The dips at 3500 Hz and 13 kHz corresponds to resonance frequencies of the ear canal.

Loudness

- subjective measure
- A 60-phon tone is not perceived as being twice as loud as a 30-phon tone.
- Above 40 phon, the loudness L doubles with 10-dB increase of the loudness level

$$L = 2^{(LL-40)/10} \quad \Rightarrow \quad \log_{10} L \approx 0.03(LL - 40) = 0.03LL - 1.2 \tag{7}$$

- For a tone of 1000 Hz, $SPL = IL = LL$, so

$$LL = IL = 10\log_{10}(I/10^{-12}) = 10\log_{10} I + 120 \tag{8}$$

- Combining (7) and (8), we have

$$\log_{10} L \approx 0.03(10\log_{10} I + 120) - 1.2 = 0.3\log_{10} I + 2.4 \quad \Rightarrow \quad L = 251 I^{0.3} \tag{9}$$

(9) is also known as the cubic law of loudness.

Pitch

- related to frequency
- the unit of pitch is *mel* (from melody)
- By definition, a tone of 1000 Hz has a pitch of 1000 mels
- According to psychoacoustic experiments, we have

$$\text{pitch in mels} = 1127\log_e(1 + \frac{f}{700}) \tag{10}$$

Masking

- Masking is the phenomenon that some sounds are “covered” by the other sounds. For example, we have to shout in a rock-n-roll concert.
- Masker sound and masked sound
- Masking is complicated. We often simplify matters by using tones (or simple noises) for the maskers and the masked sound.

Masking by Tones

- Masking curves of two masker tones, Figure 4.17 (a) (b)
  - more effective at frequencies above the masker tone than below
  - more effective at higher ILs
  - small notches due to beats
- The presence of a tone alters the hearing threshold, Figure 4.18

Masking by Noise

- With a flat-spectrum masker, the degree of masking of a tone increases with the bandwidth until it reaches a limit

- That actually is one way to measure the critical bandwidths
- Figure 4.19 shows the critical banwidth by masking experiments

Temporal Masking

The phenomenon that a transient sound causes sound adjacent in time to become inaudible. This is illustrated in Figure 4.20.

Auditory Models

Auditory models are constructed to interprete (or make use of) the properties of the human auditory system, Such as

- non-linear frequency scale
- spectral amplitude compression
- loudness compression
- decreased sensitivity at lower and higher frequencies
- temporal features, long spectral integration intervals
- auditory masking effects

Perceptual Linear Prediction

- critical-band spectral analysis using Bark frequency scale
- asymmetric auditory filters with 25 dB/Bark at the high frequency cutoff and 10 dB/Bark at the low frequency
- equal-loudness contour to approximate unequal sensitivity at different frequencies
- cubic root compression for intensity to loudness
- integration of critical bands via all-pole modeling for spectral smoothing

Human Speech Perception Experiments

Sound Perception in Noise

Figure 4.29 is the *confusion matrix* for human consonant recognition with a signal-to-noise ratio (SNR) of 12 dB.

Figure 4.30 is for an SNR of -6 dB

Speech Perception in Noise

Contextual information plays an important role in human speech perception in noise.

Figure 4.31 shows the human speech recognition accuracy of three dfferent tasks, digits, words in sentences, and nonsense syllables, as a function of SNR

Figure 4.32 shows the task for mono-syllabic words

Measurement of Speech Quality and Intelligibility

- Objective vs. Subjective
- Matching a speech signal vs. a speech model

Mean Opinion Score (MOS)

- `excellent`: 5
- `good`: 4
- `fair`: 3
- `poor`: 2
- `bad`: 1

MOS for speech coders in 3 decades, Figure 4.33

Signal to Noise Ratio (SNR)

- objective
- low-cost
- rarely used for speech any more

Perceptual Evaluation of Speech Quality (PESQ)

- an objective measure
- tuned to be highly correlated to the subjective MOS $0.79 \sim 0.98$
- need the *reference* signals

Perceptual Evaluation of Audio Quality (PEAQ)

- extension of PESQ for the wide-band audio coders