

CSE5026: Foundation and Frontiers of Cognitive Science

Auditory and Speech Systems

Yang Xu (徐炆)

Department of Computer Science and Engineering
Southern University of Science and Technology

Overview

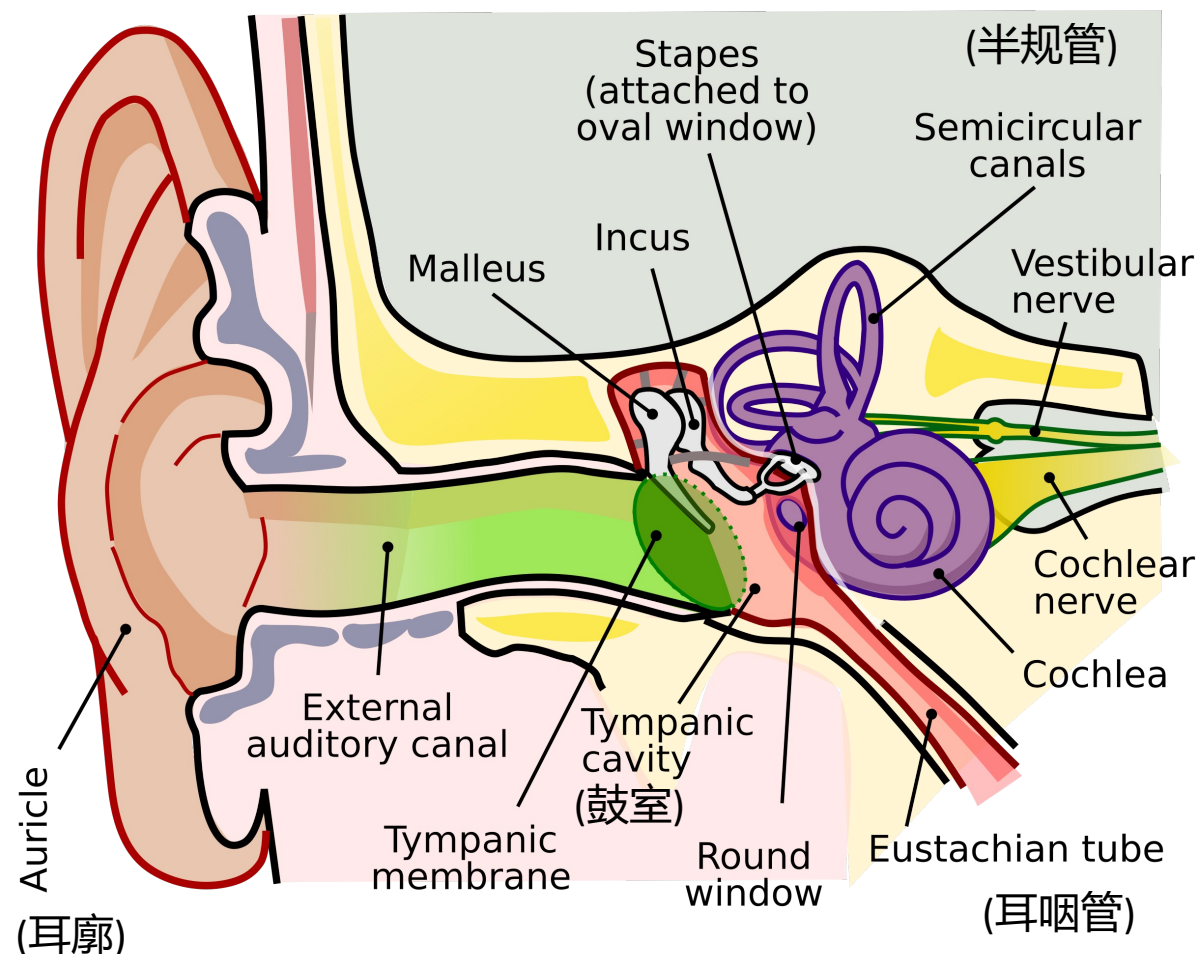
- Auditory Sensory System (听觉系统)
- Auditory Pathways (听觉神经通路)
- Speech Perception (言语感知)
- Speech Production (言语/语音生成)

Auditory Perception (听觉感知)

- Sound waves excite sensory cells (感受细胞), triggering impulses in the auditory nerve (听神经). Auditory centers (听觉中枢) process these signals to produce hearing.
(机械)
- Hearing involves mechanical motion, electrical changes, chemical transmission, neural impulses, and central processing.
 - 1. Sound collection by the outer ear and movement of the basilar membrane (基底膜) are mechanical.
 - 2. Hair cells (毛细胞) convert motion into electrical signals, releasing chemical transmitters and triggering neural impulses.
 - 3. The central nervous system then processes the incoming information.

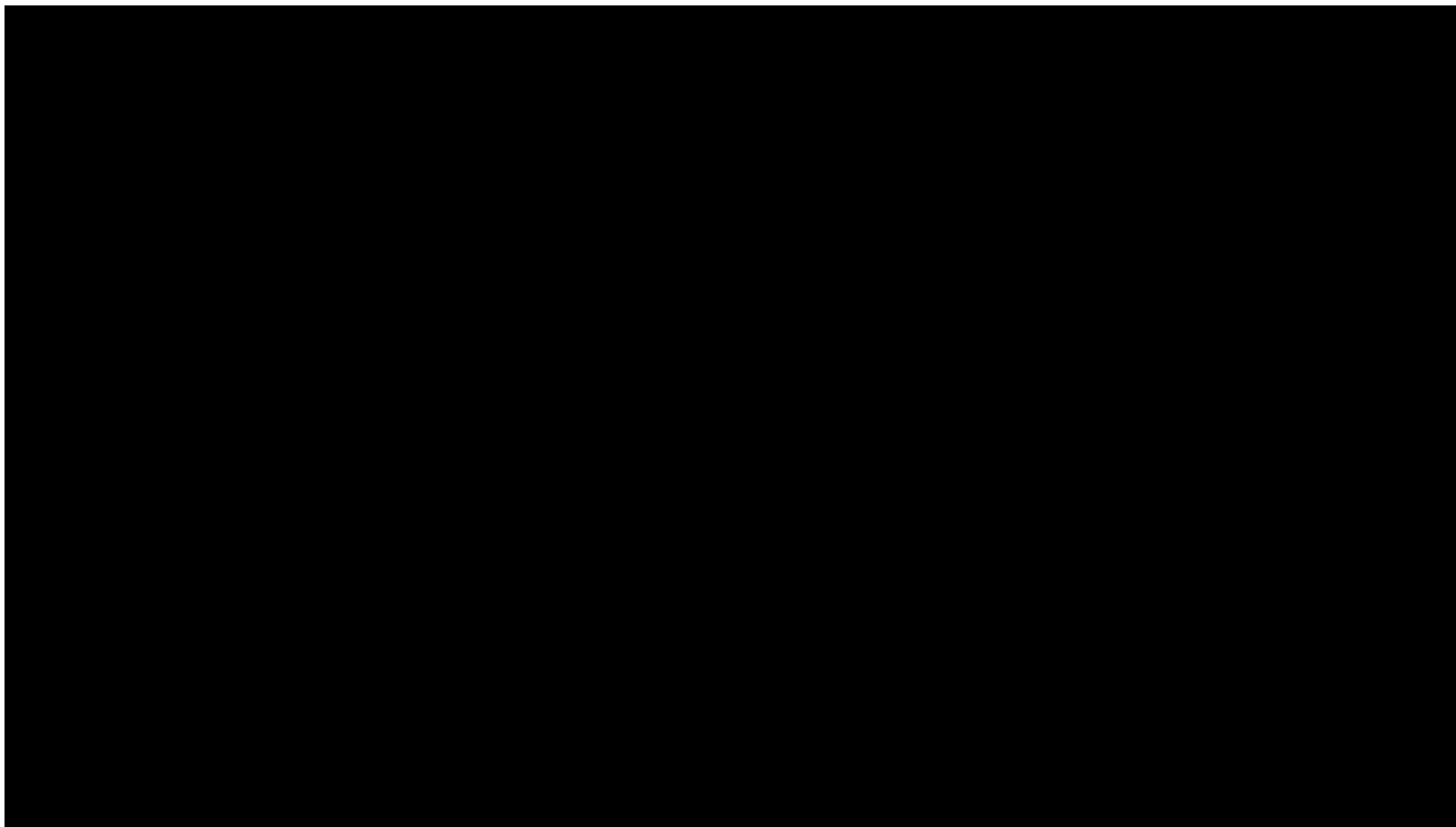
The Auditory Process (听觉形成过程)

- *Sound* $\Rightarrow$ outer ear (外耳)
 $\Rightarrow$ middle ear (中耳)
 $\Rightarrow$ cochlear receptors (耳蜗感受器)
 $\Rightarrow$ auditory nerve (听神经)
- $\Rightarrow$ Auditory pathways (听觉传导通路)
 $\Rightarrow$ auditory cortex (听觉皮层), where auditory perception develops.
- **Tympanic membrane** (鼓膜): $8-10 \text{ mm}^2$
- **Cochlea** (耳蜗)
- **vestibular** nerve (前庭神经) and cochlear nerve (耳蜗神经)



Source: https://en.wikipedia.org/wiki/Auditory_system

The Auditory Process (听觉形成过程)



- Malleus (锤骨)
- Incus (砧骨)
- Stapes (镫骨)

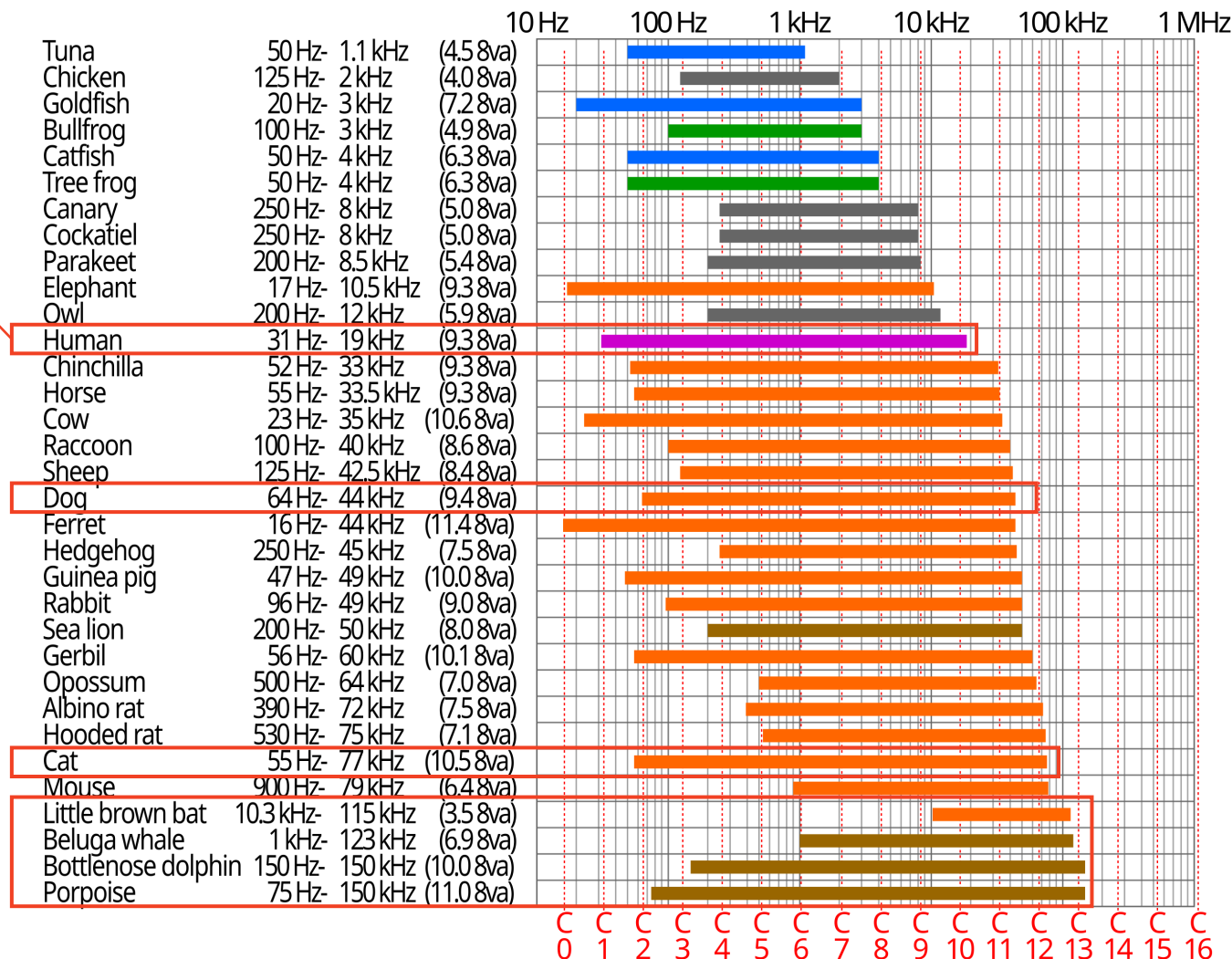
- Cochlea (耳蜗):
filled with fluid.
- Sound vibrations
create ripples in the
fluid.

- Stereocilia (静纤毛)
- Hair cell (毛细胞)

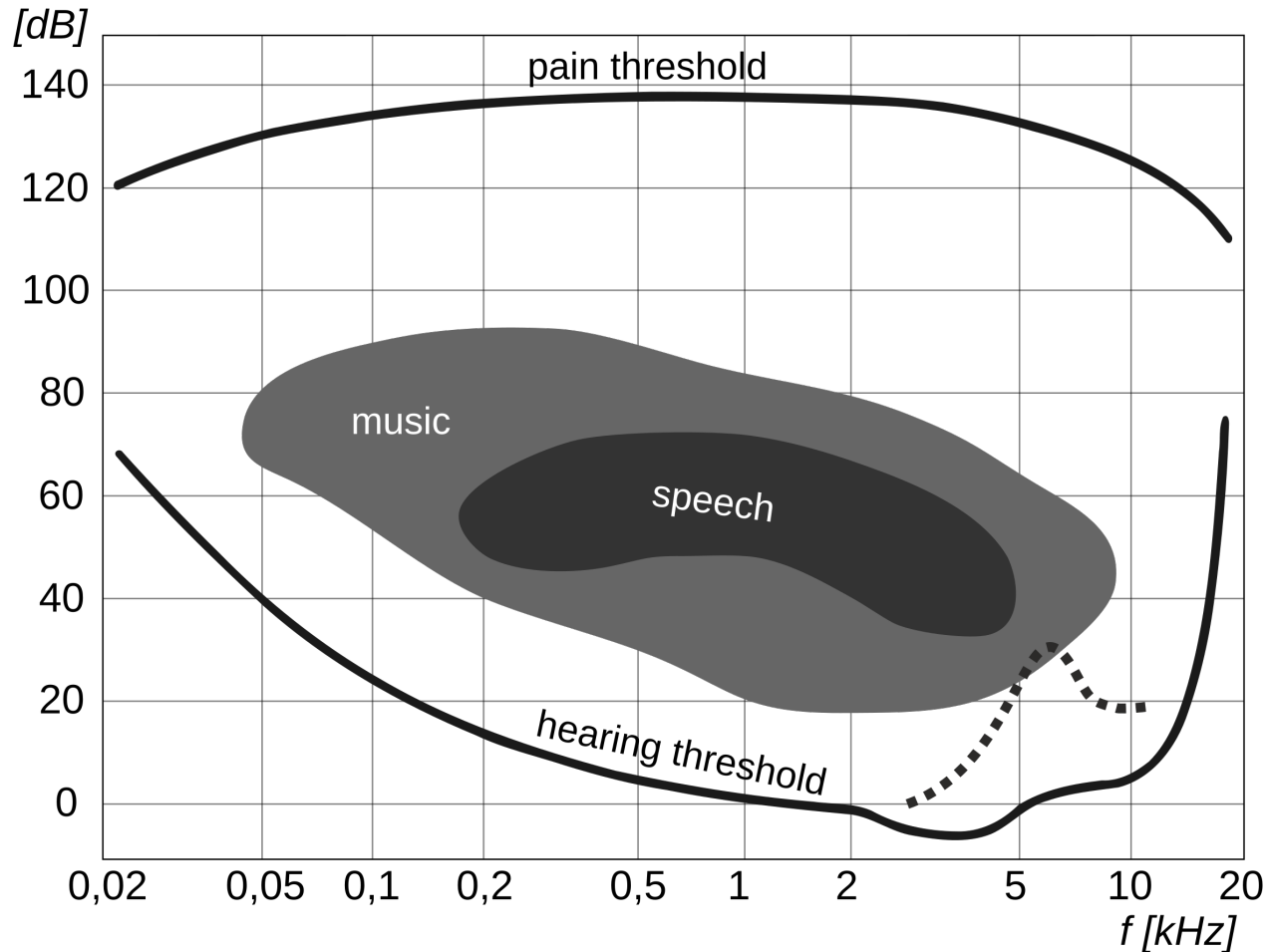
Source: https://en.wikipedia.org/wiki/Hearing_range

Hearing Frequency Range (听觉频率范围)

- Humans: 20–20,000 Hz
- Dogs/cats: similar lower limits, higher upper limits (~ 65,000 Hz)
- Bats: upper limit about 200 kHz
- (白鲸) Beluga whales: 1–123 kHz
- Bottlenose dolphins: 150 Hz–150 kHz
kHz (宽吻海豚)



Intensity–Frequency Range (强度–频率范围)



Measures of sound intensity

Objective (客观) measures:

- **Sound Pressure Level** (声压级)

Subjective (主观) measure:

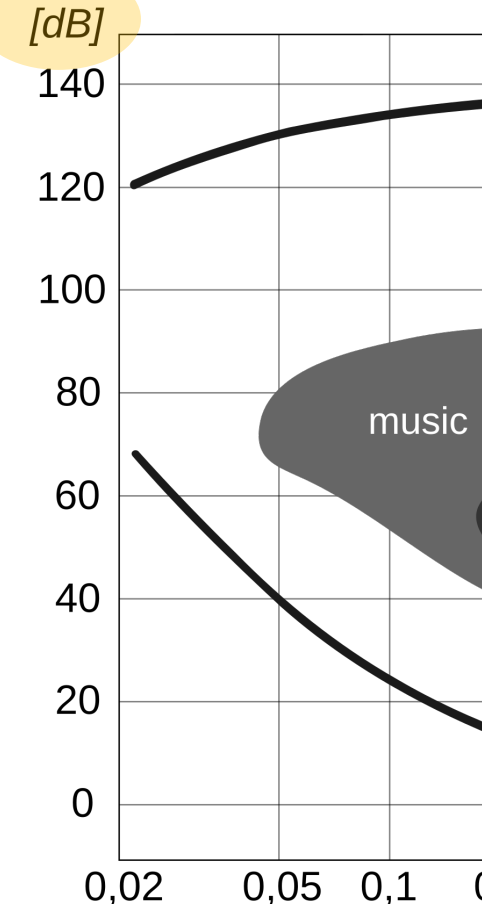
- **Loudness** (响度)

Sound Pressure Level (声压级)

- SPL expresses the effective sound pressure on a logarithmic scale relative to a reference.

$$L_p = \ln\left(\frac{p}{p_0}\right) \text{ Np} = 2 \log_{10}\left(\frac{p}{p_0}\right) \text{ B} = 20 \log_{10}\left(\frac{p}{p_0}\right) \text{ dB}$$

- The standard reference sound pressure in air is $p_0 = 20 \mu \text{ Pa}$.
 - It approximates the human hearing threshold, roughly the sound of a mosquito flying 3 m away.



Source: https://en.wikipedia.org/wiki/Sound_pressure

Examples of Sound Pressure Levels

Sound source	Distance	Pa	dB
Jet engine	1 m	632	150
Loudest human voice	1 inch	110	135
Threshold of pain (疼痛閾限)	At the ear	20–100	120–134
Trumpet	0.5 m	63.2	130
Baby crying	–	–	80–130
Normal conversation	1 m	2×10^{-3} –0.02	40–60
Very quiet room		2.00×10^{-4} ... 6.32×10^{-4}	20–30

Review: Weber's Law (韦伯定律)

- Absolute threshold (绝对阈限): the minimum stimulus intensity needed to detect a single stimulus.
- Difference threshold (差别阈限): the smallest detectable difference between stimuli.
- Weber's law (韦伯定律): within a certain range, the difference threshold is approximately a constant fraction of stimulus intensity.

$$\frac{\Delta I}{I} = k$$

ΔI : difference threshold; I : initial intensity
 k is the constant Weber fraction (韦伯分数)

Weber's Law (韦伯定律)

- Typical Weber fractions k
Initial stimulus intensities appear in parentheses.

$$\frac{\Delta I}{I} = k$$

ΔI : difference threshold; I : initial intensity
 k is the constant Weber fraction (韦伯分数)

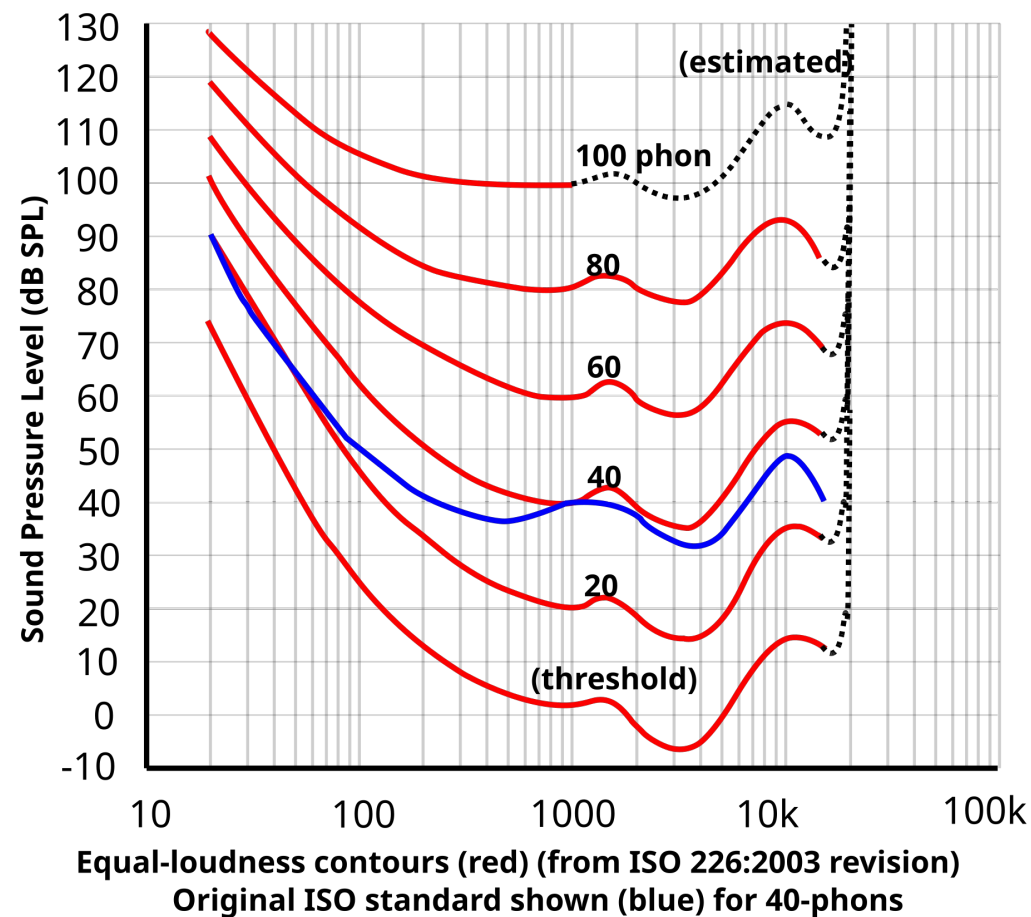
Stimulus (initial intensity)	Weber fraction
Pitch (2,000 Hz)	1/333
Deep pressure (400 g)	1/77
Visual brightness (1,000 photons)	1/62
Lifted weight (300 g)	1/53
Loudness (100 phon, 1,000 Hz)	1/11
Rubber odor (200 odor units)	1/10
Skin pressure (5 g/mm ²)	1/7
Salt taste (3 gram-molecules/kg)	1/5

Loudness (主观响度)

- Loudness (响度) is the subjective perception of sound pressure. An audiometer can measure it using loudness balancing.
- Sensitivity to loudness varies with frequency. Equal-loudness contours (等响度曲线) show this relationship.

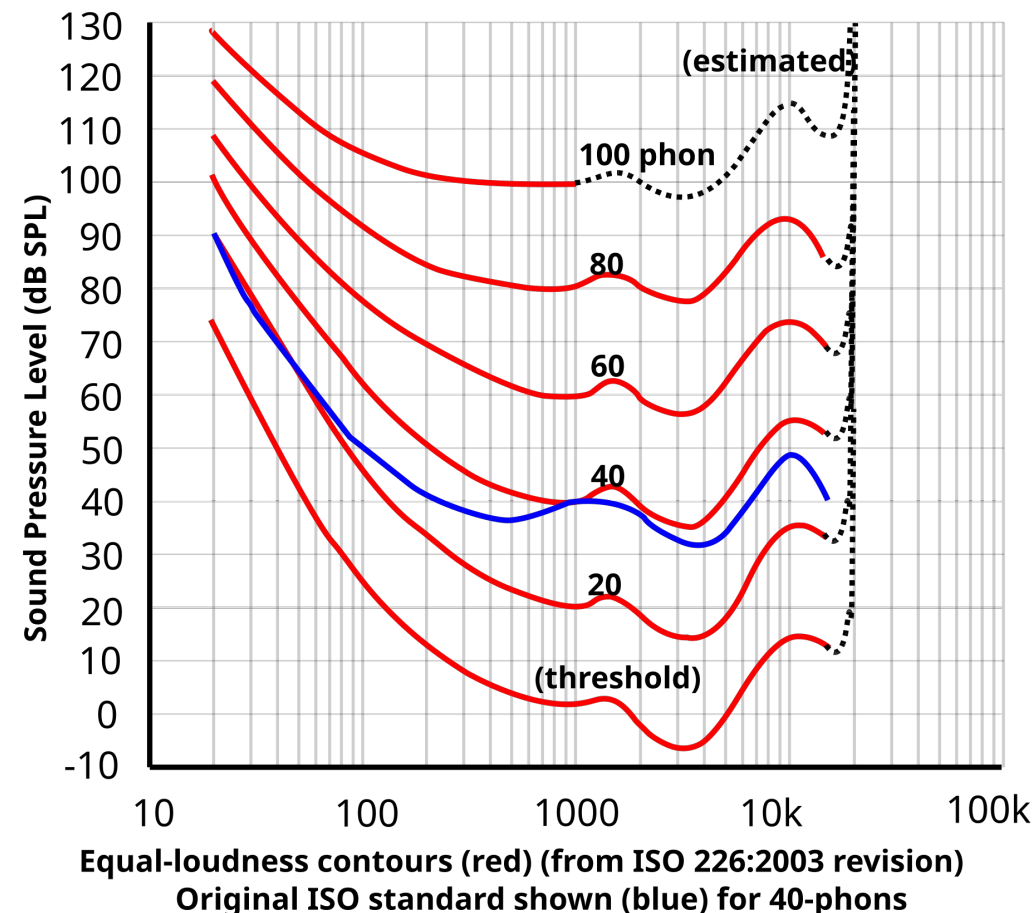


Source: <https://en.wikipedia.org/wiki/Loudness>



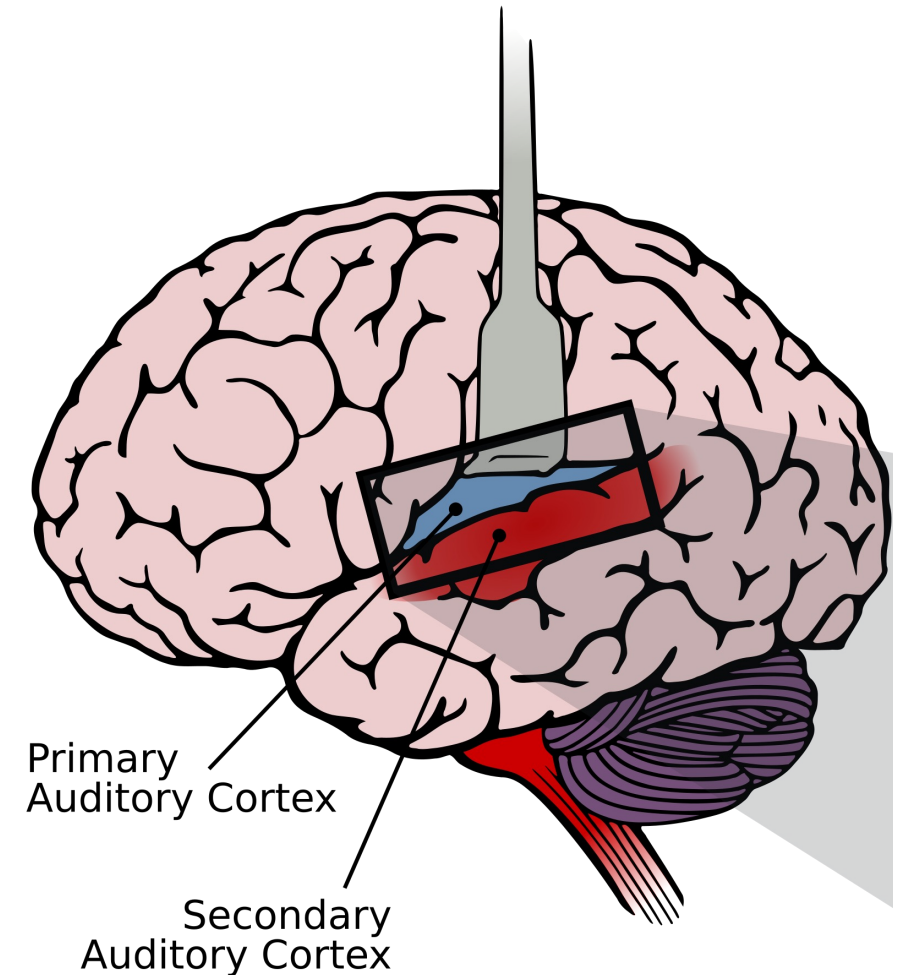
Loudness Level (响度级)

- **Phon** is the unit of **loudness level**.
- A sound has a loudness level of **x phon** if, under specified listening conditions, it *sounds as loud as* a **1,000 Hz reference pure tone (参考纯音)** at **x dB SPL**.
- The labels on the **equal-loudness contours (等响度曲线)** use the same phon scale.



Auditory Cortex (听觉皮层)

- The auditory cortex (听觉皮层) lies in the **temporal lobe (颞叶)**.
- It includes primary auditory cortex (初级听觉皮层, A1), secondary auditory cortex (次级听觉皮层, A2), and association areas (联合区).
- It analyzes the timing and frequency of signals from the cochlea (耳蜗).
- Neurons (神经元) are organized by their preferred frequencies: low at one end and high at the other.
- Functional specialization: A1 encodes complex, abstract features of auditory stimuli, beyond raw features such as individual frequencies.

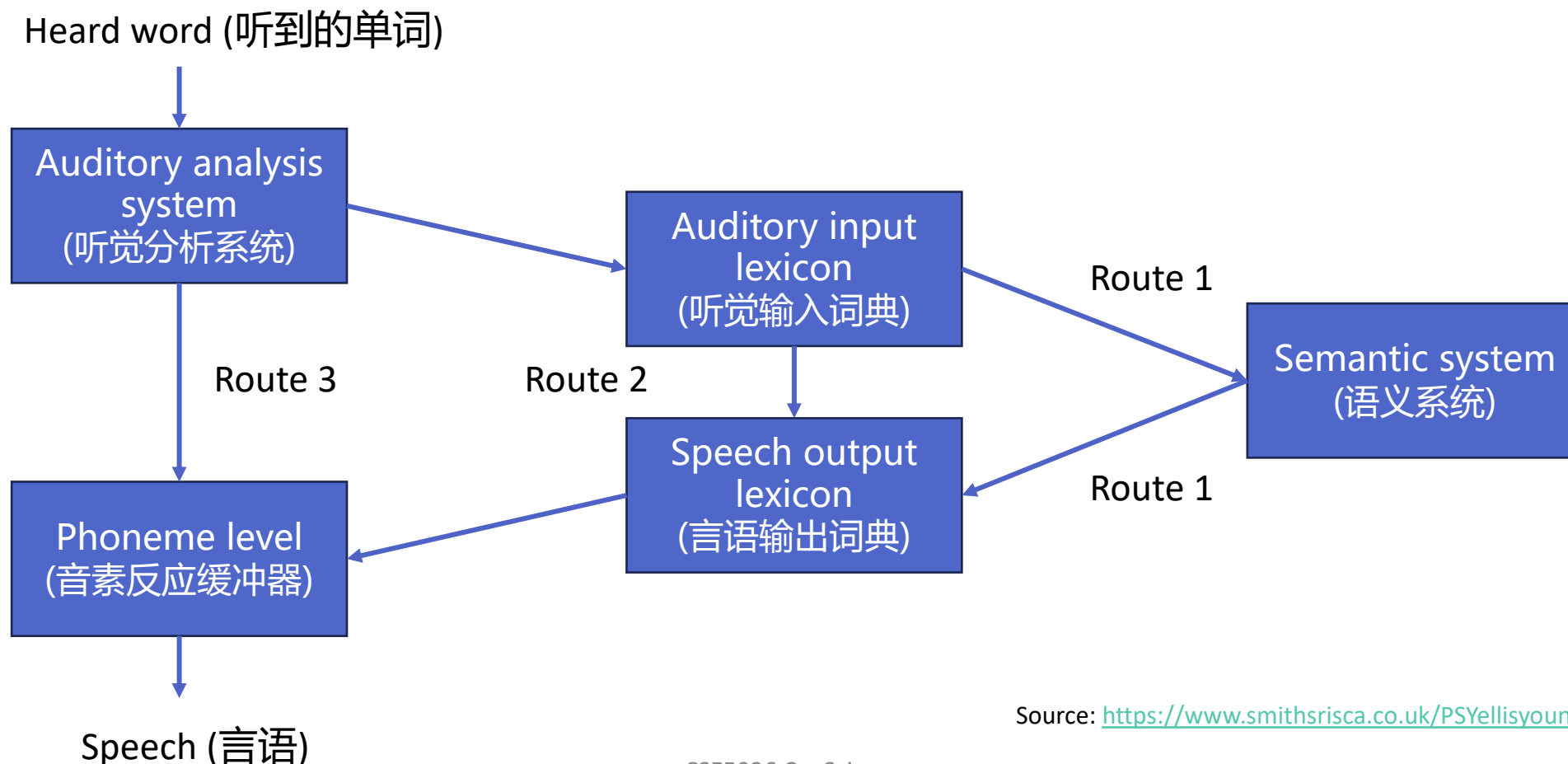


Overview

- Auditory Sensory System (听觉系统)
- Auditory Pathways (听觉神经通路)
- **Speech Perception (言语感知)**
- Speech Production (言语/语音生成)

Ellis & Young Model (口语通路模型)

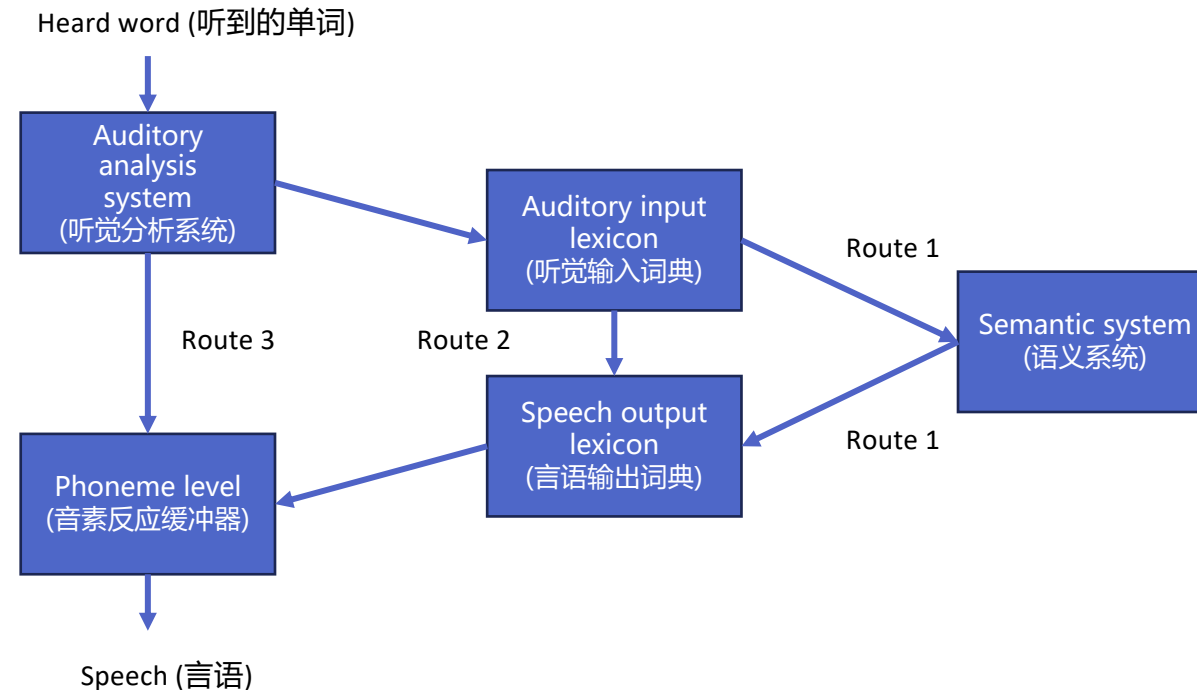
- Ellis & Young (1988): Speech perception and production



Sound Analysis and Word Recognition

- Auditory analysis system (听觉分析系统): extracts **phonemes** and acoustic features.
 - Phoneme (音位/音素): the smallest unit distinguishing word meanings.
 - Auditory input lexicon (听觉输入词典): recognizes familiar spoken words.
 - Meaning comes next, in the semantic system (语义系统).

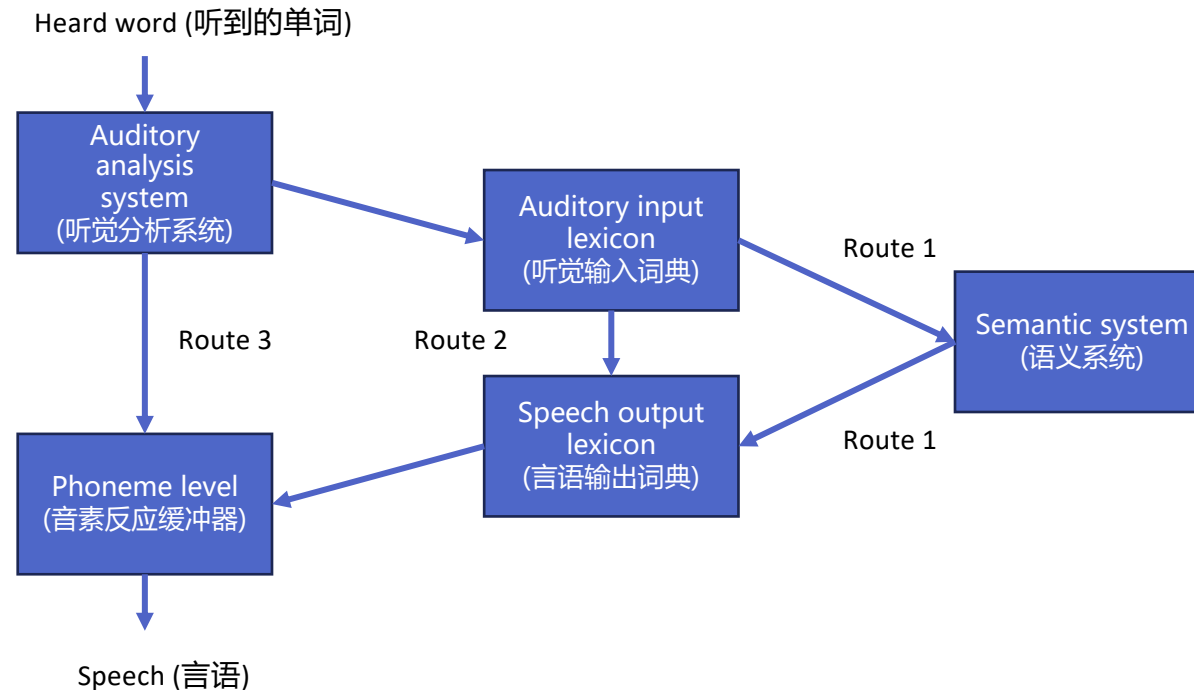
Unaspirated stop	/p/	(爸, bà), (波, bō)
Aspirated stop	/p ^h /	(怕, pà), (坡, pō)



Source: [Mandarin phoneme examples](#)

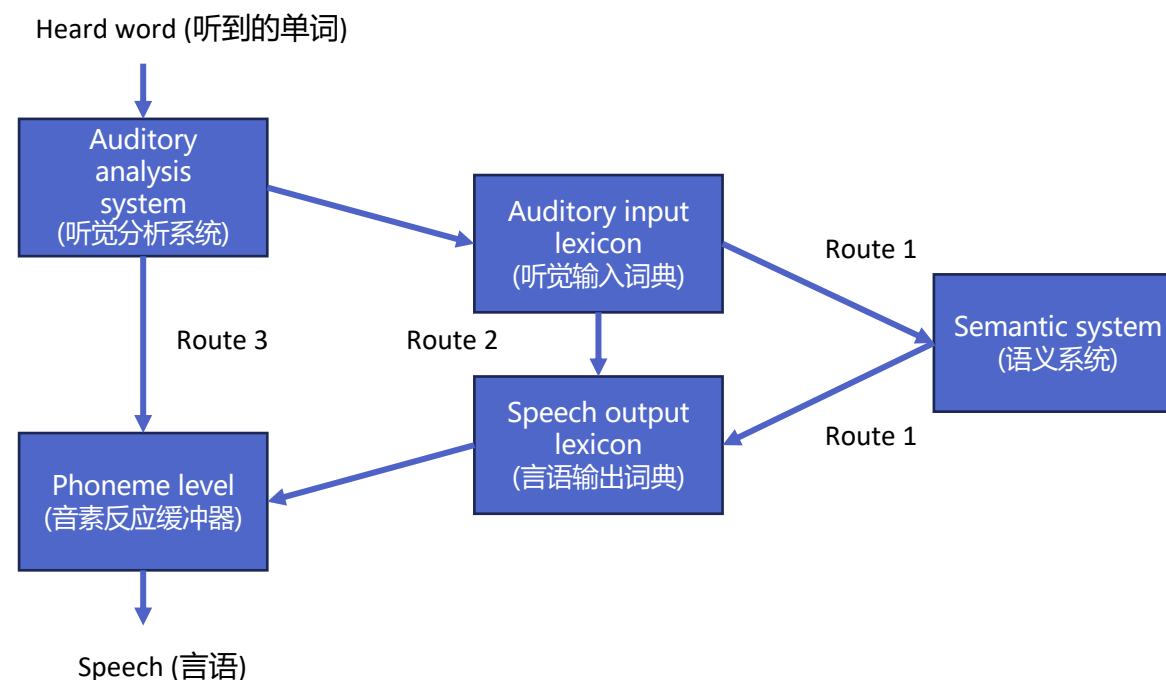
Meaning and Speech Output

- Semantic system (语义系统): stores word meanings.
 - An abstract functional module. Its implementation could be distributed, as in semantic embeddings.
- Speech output lexicon (言语输出词典): supplies stored spoken word forms.
- Phoneme level (音素反应缓冲器): holds the speech sounds needed for output.



Three Routes for Spoken Repetition

- Route 1 (semantic): recognizes a word form, accesses meaning, then retrieves speech output.
- Route 2 (lexical): repeats familiar words without accessing meaning.
- Route 3 (nonlexical): repeats unfamiliar words or nonwords via a direct link to the phoneme level.
- Word meaning deafness (词义性耳聋): impaired understanding of heard words with preserved repetition.



Overview

- Auditory Sensory System (听觉系统)
- Auditory Pathways (听觉神经通路)
- Speech Perception (言语感知)
- **Speech Production (言语/语音生成)**

Articulatory Phonetics (发音语音学)

- Articulatory phonetics studies how organs in the mouth, throat, and nose shape airflow from the lungs to produce speech sounds.
- Vocal organs (发声器官)
 - Vocal folds / cords (声带)
 - Closely spaced vocal folds vibrate as air passes through. Widely separated folds do not vibrate.
 - Voiced sounds (浊音) involve vocal-fold vibration, e.g., [b], [d], [g].
 - Unvoiced sounds (清音) lack vocal-fold vibration, e.g., [p], [t], [k].

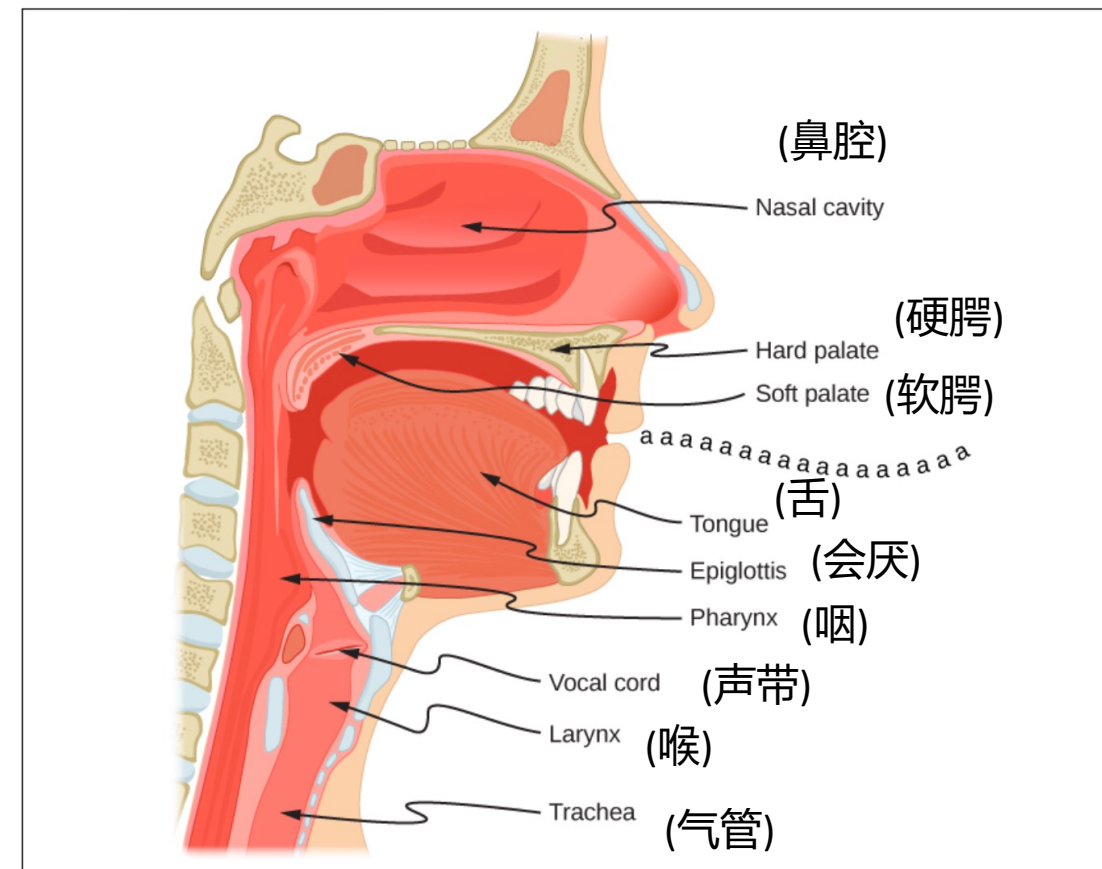
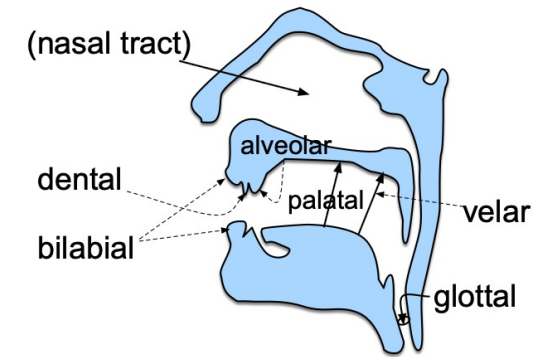


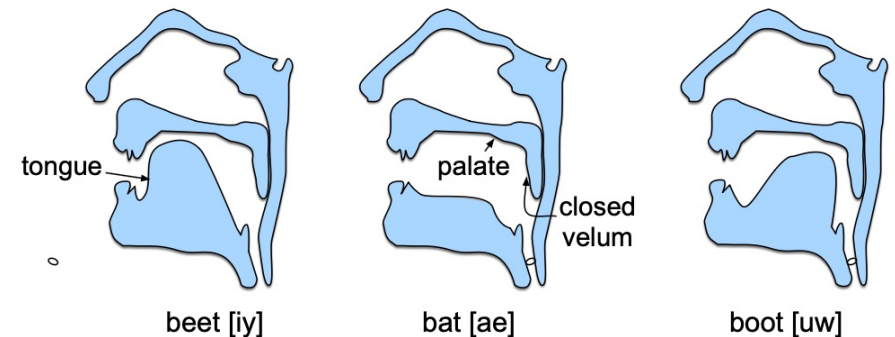
Figure 14.4 The vocal organs, shown in side view. (Figure from OpenStax University Physics, CC BY 4.0)

Consonants and Vowels (辅音与元音)

- Speech sounds fall into two broad classes: **consonants** and **vowels**.
- Consonants restrict or block airflow and may be voiced or unvoiced.
- Vowels involve less obstruction. They are usually voiced, *louder*, and *longer* than consonants.

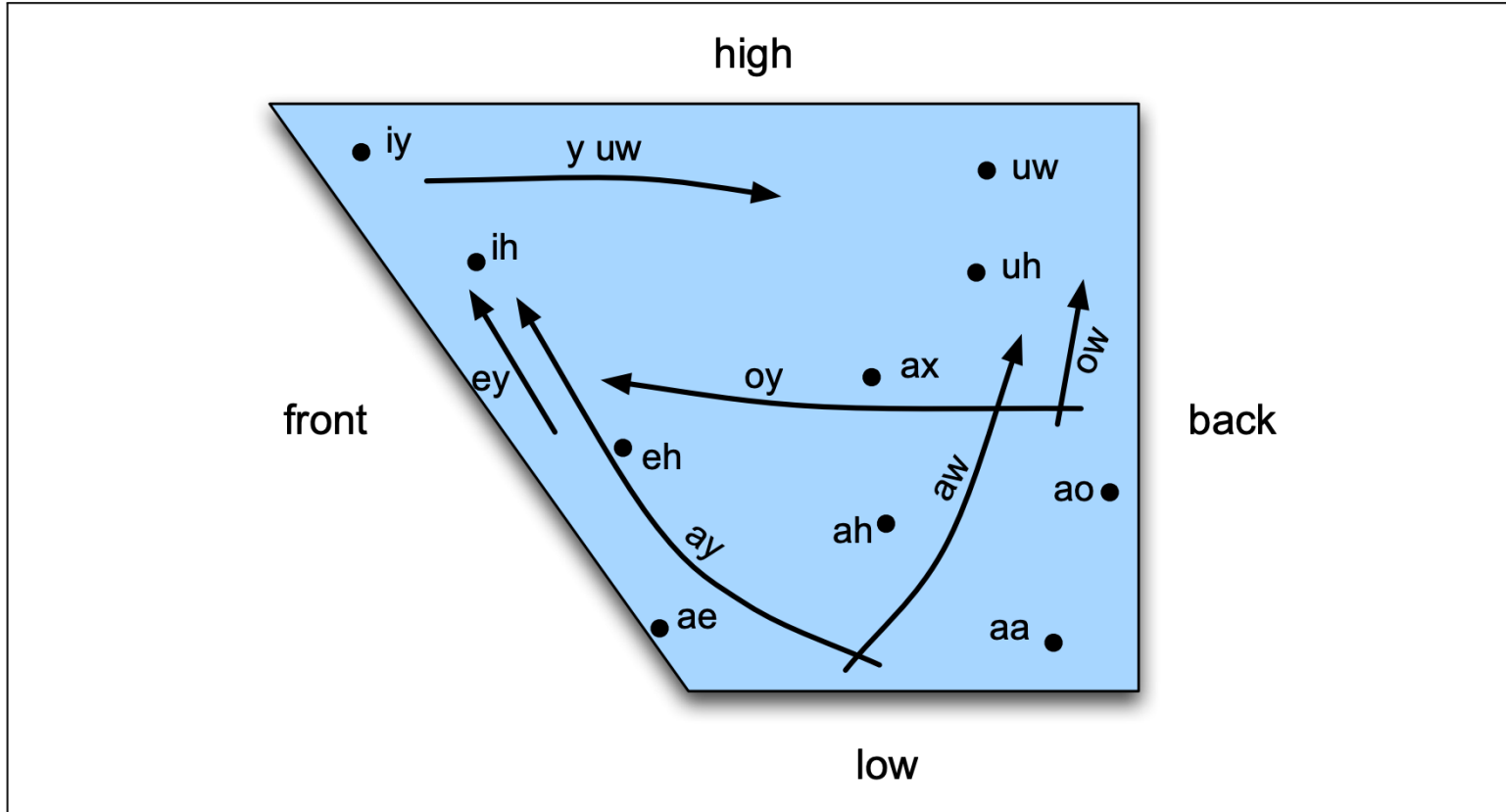


Main places of articulation for English **consonants**



Tongue positions for different **vowels**

“Vowel space” for English



- Dots roughly correspond to tongue positions.
- Paths: vowels in which the tongue positions change during production.

Figure 15.7 The schematic “vowel space” for English vowels.

Syllables (音节)

- Consonants and vowels combine to form **syllables** (音节).
- A syllable consists of a vowel-like sound (or sonorant) with its most closely associated consonants.

(猫薄荷)

 - Example: “*dog*” has one syllable, [d o g]. “*Catnip*” has two, [k ae t] and [n i p].
- **Nucleus** (核音): the central vowel.
- **Onset** (首音): initial consonant(s).
- **Coda** (尾音): final consonant(s).
- **Rime** (韵) = nucleus + coda

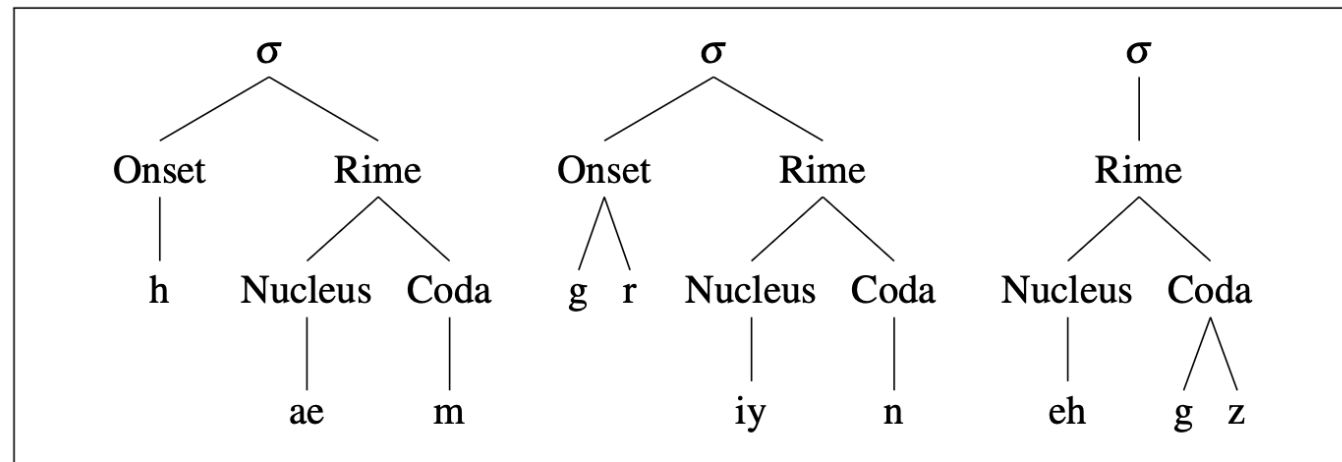


Figure 14.8 Syllable structure of *ham*, *green*, *eggs*. σ =syllable.

/'pra:sədi/

Prosody (韵律学)

(语调)

(节奏)

- Prosody (韵律学) concerns *intonation* and *rhythm*, especially how the **fundamental frequency (基频, F0)**, energy, and duration convey pragmatic, (语用) emotional, or conversational meaning.
- Energy relates to loudness (响度). **F0** contributes to perceived pitch (音高).
- Prosody can mark discourse structure, including the distinction between statements (陈述句) and questions (问句).
- **Prosodic prominence** involves pitch accent (音高重音), lexical stress (词汇重音), and vowel reduction to schwa (弱化元音/央元音).

/jwa:/

Prosodic Prominence (韵律突显)

- **Pitch accent (音高重音)**: a prominent word or syllable carries an accent in pitch.

- Example: accented words are underlined.

(14.1) I'm a little surprised to hear it characterized as happy.

- **Lexical stress (词汇重音)**: pitch accents occur on syllables that can bear stress.

- Underlined words carry accent. Bold syllables carry lexical stress.

(14.2) I'm a little surprised to hear it characterized as happy.

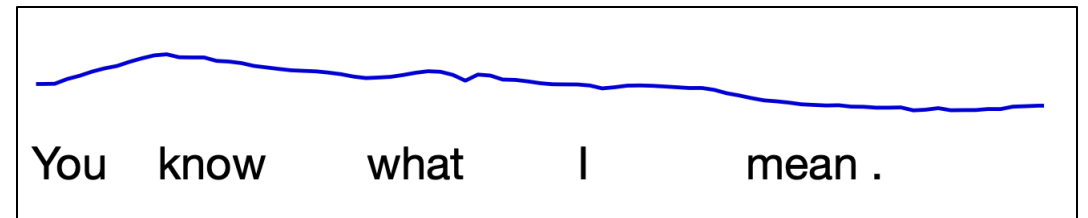
- Unstressed vowels may reduce to schwa (弱化元音/央元音).

Intonation (语调)

- Utterances with the same accents may still differ in prosody because they have different intonation patterns.
- Intonation (语调) is the *rise* and *fall* of **fundamental frequency (基频, F0)** over time.
- Example: statements and yes–no questions



A final rise in F0 indicates
a yes–no question (question rise)



A final fall in F0 indicates
a statement (final fall)

Basics of Speech Signal Processing

- Sound waves are characterized by frequency (频率) and amplitude (振幅).
- $y = A * \sin(2\pi \cdot f \cdot t)$: a sine wave with amplitude A and frequency f

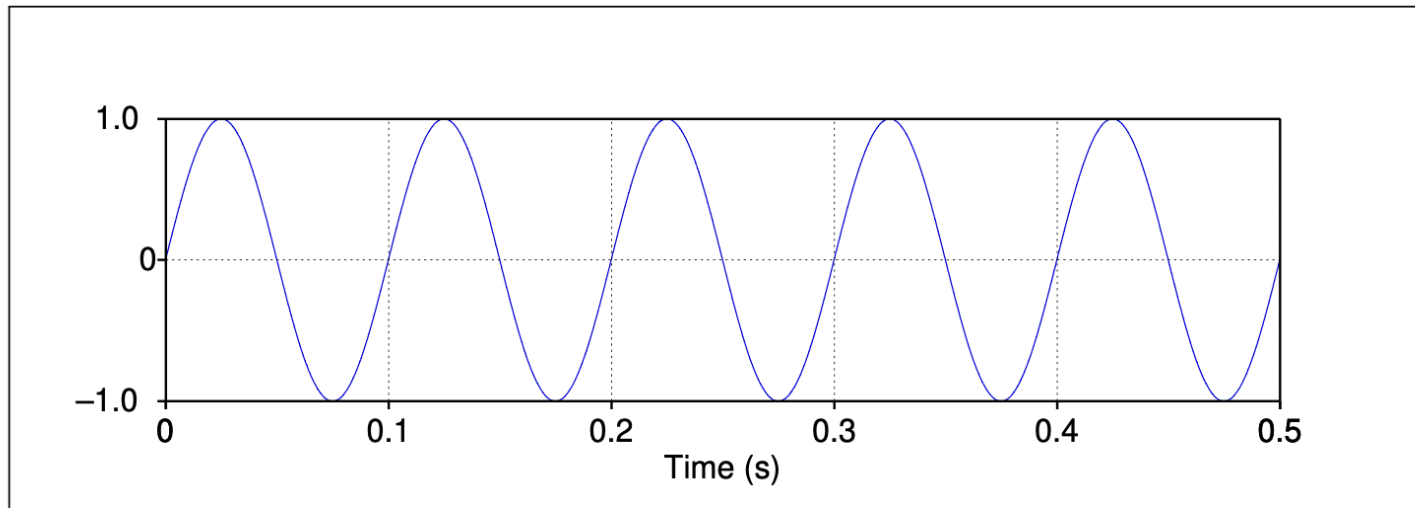


Figure 14.10 A sine wave with a frequency of 10 Hz and an amplitude of 1.

$$f = 10\text{Hz} \geq 10 \text{ cycles per sec}$$

Speech Waveforms

- A waveform plots air pressure as it changes over time.

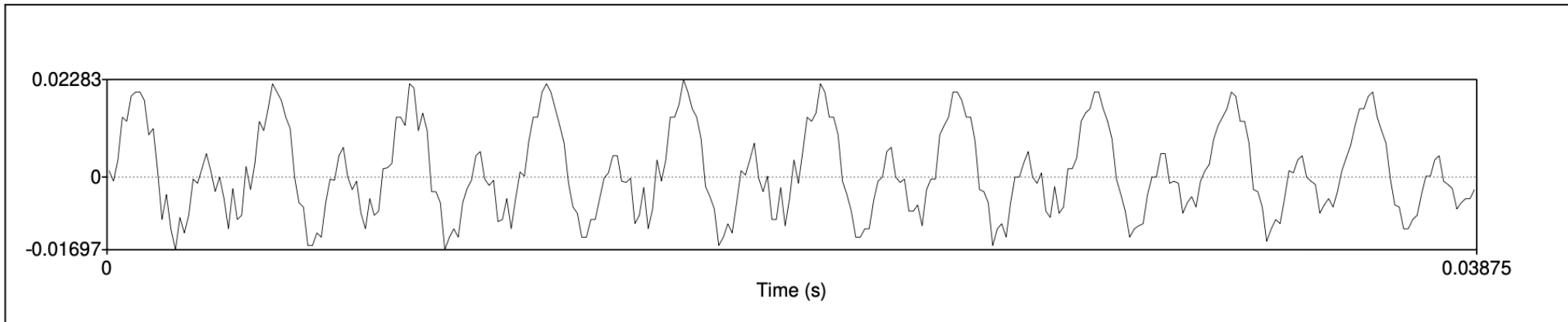


Figure 14.11 A waveform of the vowel [iy] from an utterance shown later in Fig. 14.15 on page 15. The y-axis shows the level of air pressure above and below normal atmospheric pressure. The x-axis shows time.

Positive values indicate **compression**. Negative values indicate rarefaction.

(压缩)

(稀疏)

From Sound Waves to Digital Signals

(模数转换)

(量化)

- Analog-to-digital conversion has two main steps: **sampling** and **quantization**.
- Sampling rate (采样率), f_s : the number of samples per second.
To avoid aliasing, signal frequencies must stay below the Nyquist frequency,
$$f_N = \frac{1}{2} f_s \cdot \text{(混叠)}$$
- Most speech information lies at frequencies ≤ 10000 Hz. Preserving that bandwidth requires a sampling rate of at least $f_s = 20000$ Hz.
- Traditional landline telephony: 8 kHz. A common audio sampling rate for online video: 48 kHz.

From Sound Waves to Digital Signals

- Even at an 8 kHz sampling rate, one second of speech requires 8,000 amplitude measurements. **Efficient storage matters.**
- **Quantization (量化)** stores amplitude as integers: 8-bit (−128 to 127) or 16-bit (−32,768 to 32,767).
- Traditional telephony: 8 kHz sampling and 8-bit quantization.

From Sound Waves to Digital Signals

- **Data compression**
- μ -law PCM algorithm (pulse code modulation, 脉冲编码调制):

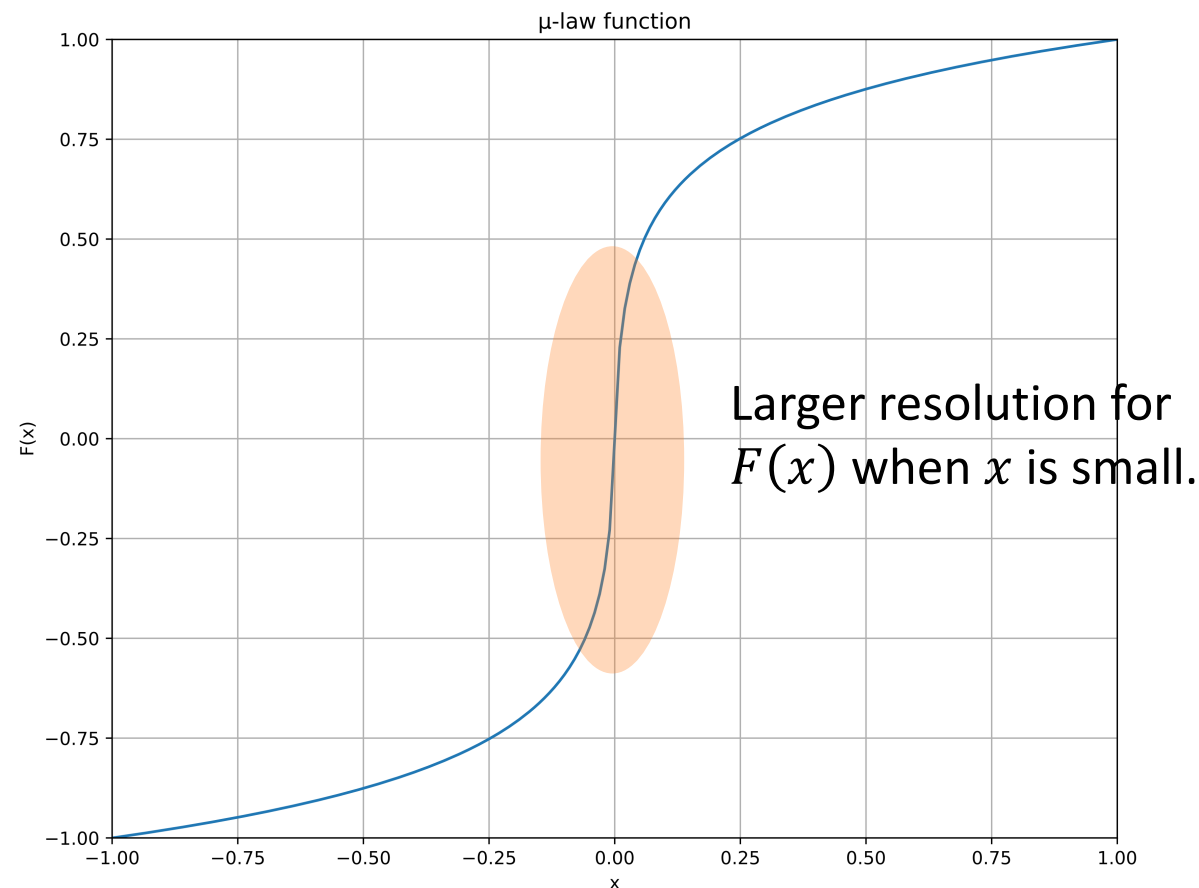
$$F(x) = \frac{\text{sgn}(x) \log(1 + \mu|x|)}{\log(1 + \mu)} \quad -1 \leq x \leq 1 \quad (\mu=255 \text{ for 8 bits})$$

- Intuition: human hearing is more sensitive at small intensities than large ones; the log represents small values with more faithfulness at the expense of more error on large values.

From Sound Waves to Digital Signals

- μ -law PCM algorithm

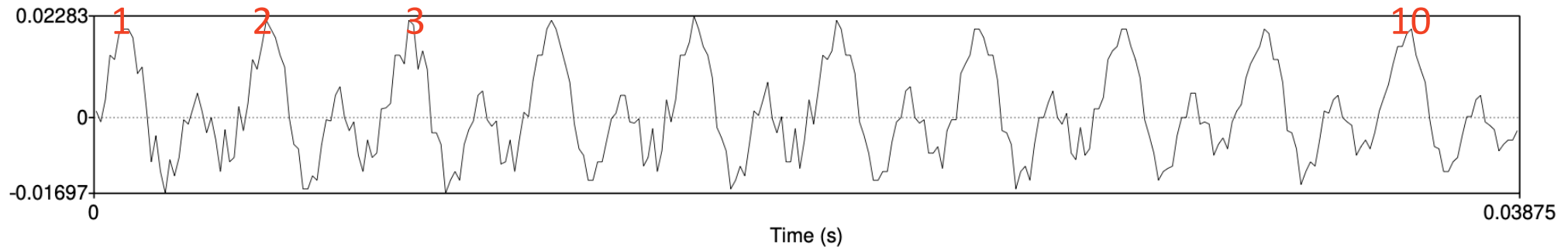
$$F(x) = \frac{\text{sgn}(x) \log(1 + \mu|x|)}{\log(1 + \mu)} \quad -1 \leq x \leq 1$$



Source: https://en.wikipedia.org/wiki/Mu-law_algorithm

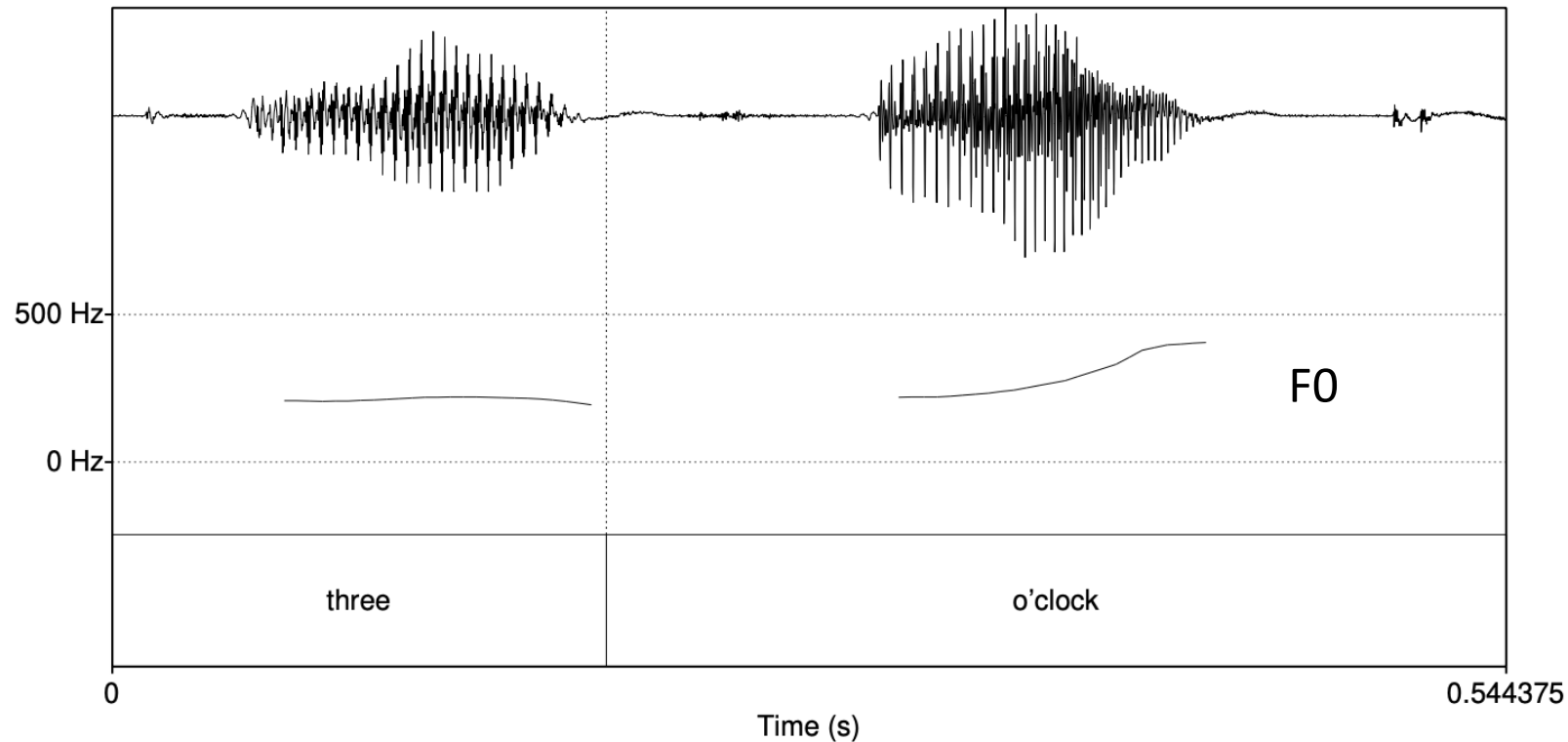
Frequency and Amplitude of Speech

- A speech waveform is not sinusoidal, but voiced speech shows **periodicity**.



- Ten similar peaks occur in 38.75 ms: $T = \frac{10}{0.03875} \approx 258 \text{ Hz}$
- Voicing comes from repeated opening and closing of the vocal folds (声带). Opening releases pressurized air from the lungs, creating a pressure pulse.
- Each major peak corresponds to one cycle of vocal-fold vibration. This repetition rate is the fundamental frequency (基频), **F0**.

Frequency and Amplitude of Speech



- A pitch track plots fundamental frequency (F0) over time.
- Here, F0 denotes a frequency value (0–500 Hz): obtained from automatic pitch counting.

For “Three o’clock?”, F0 rises at the end of the question.

Frequency and Amplitude of Speech

- Besides instantaneous amplitude, we often need a measure of amplitude over a time window.
- Signal power is the mean of the squared amplitudes:

$$\text{Power} = \frac{1}{N} \sum_{i=1}^N x_i^2$$

- Intensity in decibels (dB) expresses power relative to a reference corresponding to the human hearing threshold (听觉阈限).

$$\text{Intensity} = 10 \log_{10} \frac{1}{NP_0} \sum_{i=1}^N x_i^2$$

$P_0 = 2 \times 10^{-5} \text{ Pa}$
(reference pressure)

Pitch and Loudness (音高与响度)

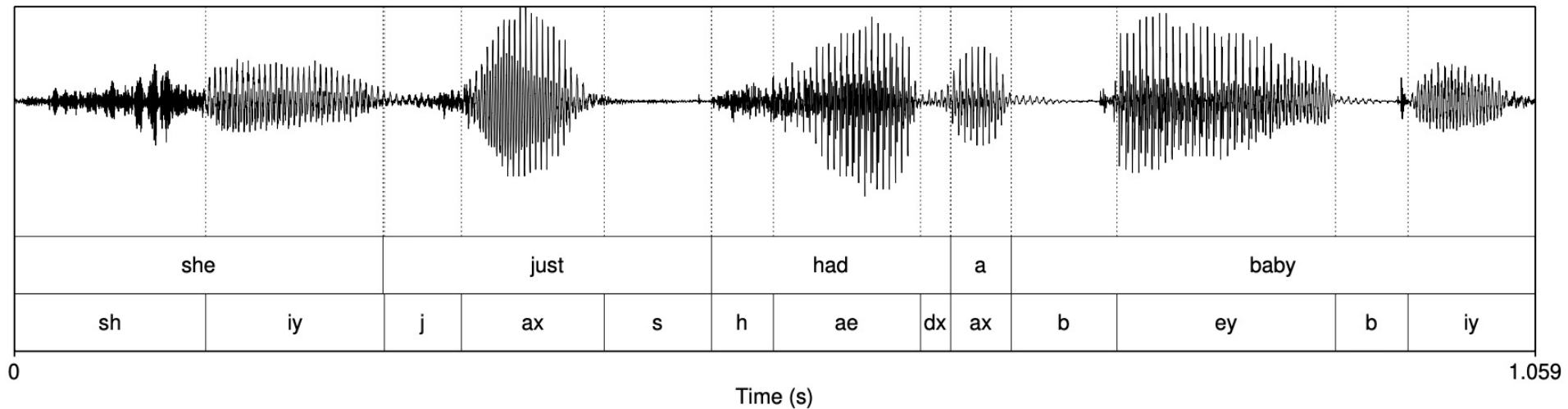
- Pitch (音高) relates to frequency. Loudness (响度) relates to intensity.
- Pitch is the perceptual counterpart of fundamental frequency (基频, F0).
 - Higher F0 generally produces a higher perceived pitch.
 - The relationship is *nonlinear* because auditory sensitivity varies across frequencies.
 - Between about 100 and 1,000 Hz, pitch perception is most accurate and approximately linear in frequency.
 - Above 1,000 Hz, the relationship is approximately logarithmic.
- **Mel scale** model (梅尔标度):

$$m = 1127 \ln\left(1 + \frac{f}{700}\right)$$

Equal perceived pitch intervals
correspond to equal differences in mel values.

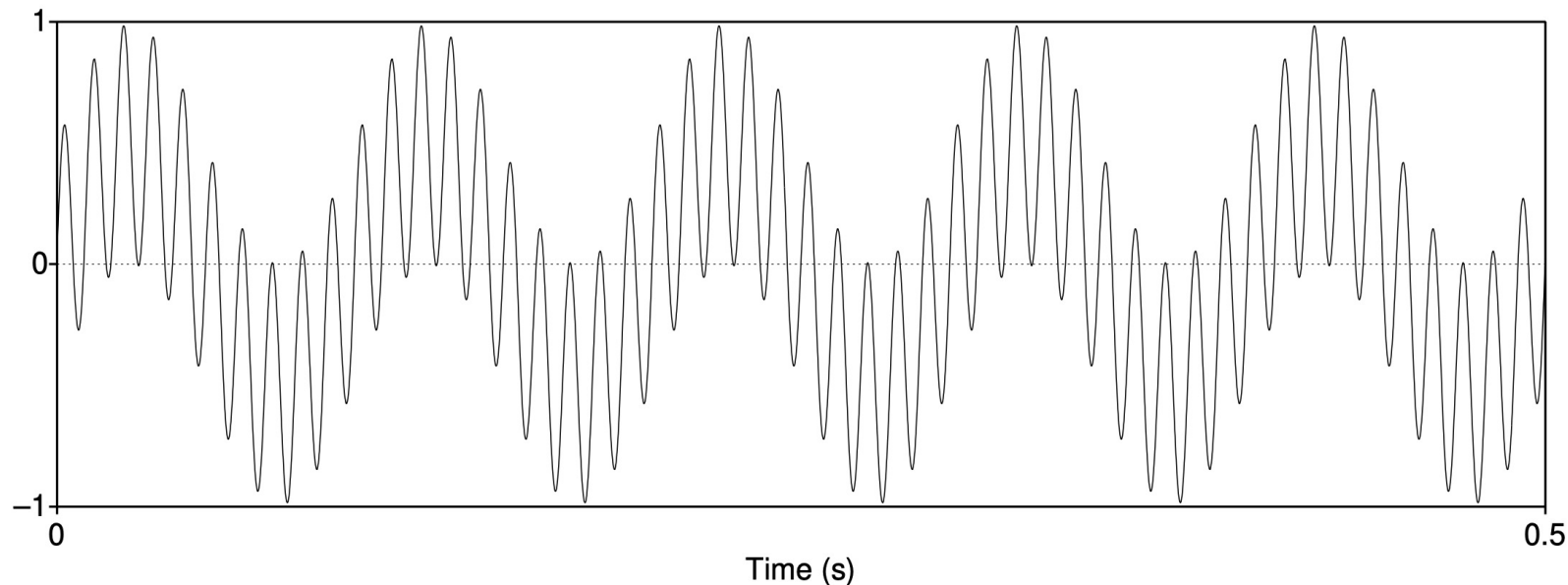
Interpreting Speech Waveforms

- Vowels (元音) are easier to identify because they are relatively long and loud.
 - The six vowels [iy], [ax], [ae], [ax], [ey], and [iy] show regular amplitude peaks, indicating voicing (浊音).
- Stop consonants (塞音) often show a near-silent closure followed by a small burst.
 - Two [b] sounds in “baby”



Spectra and Frequency Domain

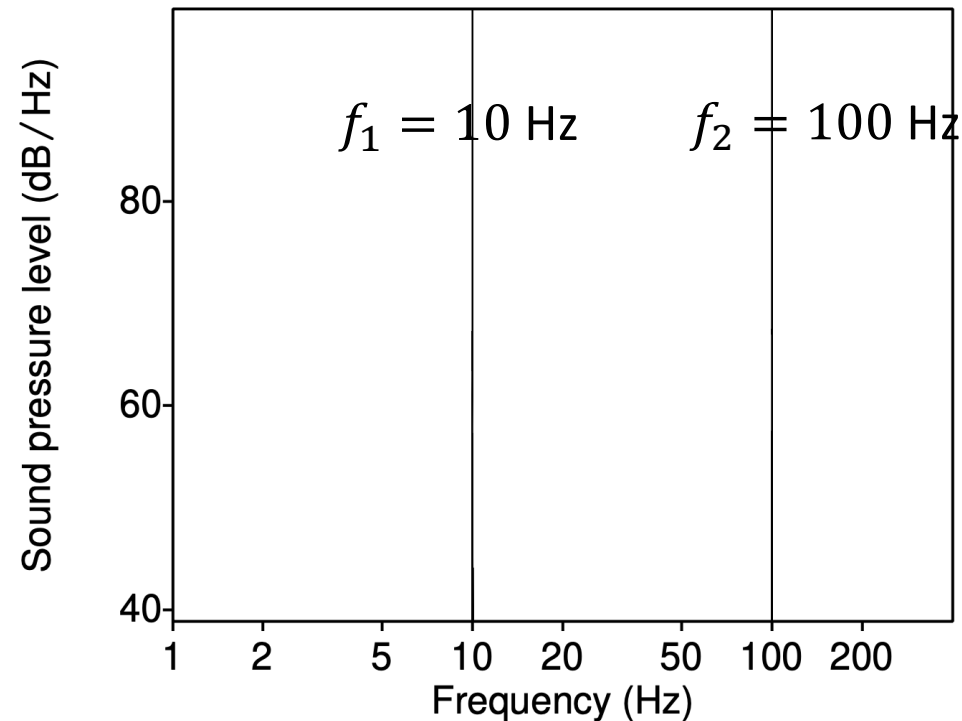
- **Fourier analysis:** Every complex wave can be represented as a sum of many sine waves of different frequencies.



$$y = A_1 \sin(2\pi \cdot f_1 \cdot t) + A_2 \sin(2\pi \cdot f_2 \cdot t) \quad f_1 = 10 \text{ Hz}, f_2 = 100 \text{ Hz},$$

Spectra and Frequency Domain

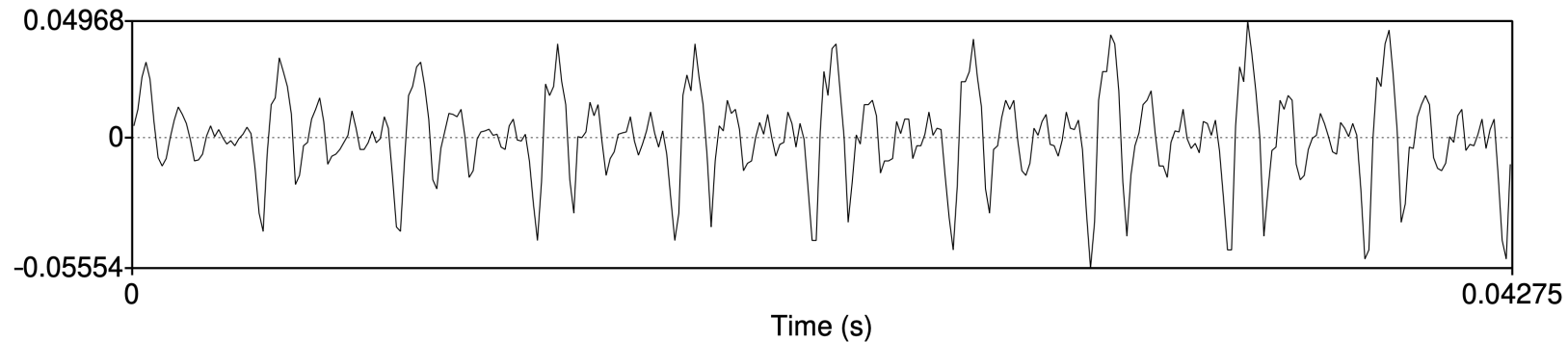
- **Spectrum:** A representation of each of the frequency components in a signal and their amplitudes.



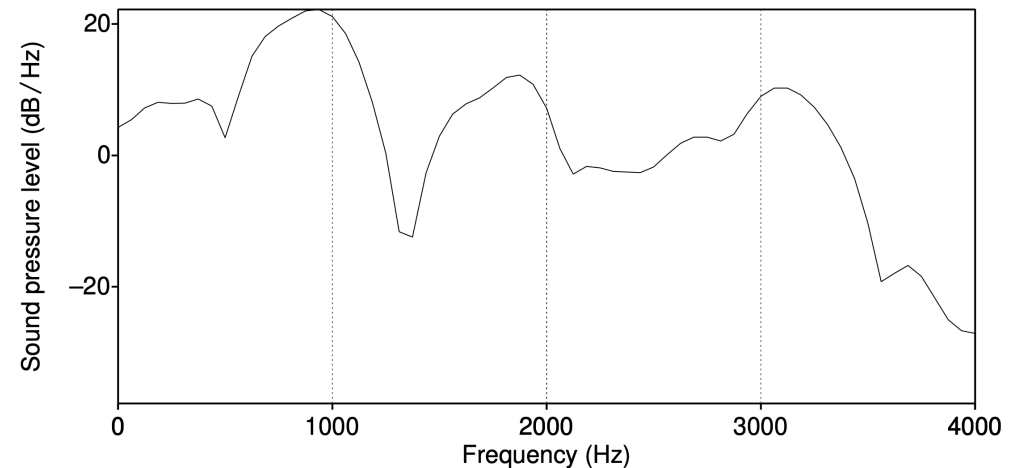
Spectrum of signal $y = A_1 \sin(2\pi \cdot f_1 \cdot t) + A_2 \sin(2\pi \cdot f_2 \cdot t)$

Spectra and Frequency Domain

- Example: This complex wave has a frequency f_1 of about 234 Hz: 10 cycles/0.0427 sec = 234 Hz.

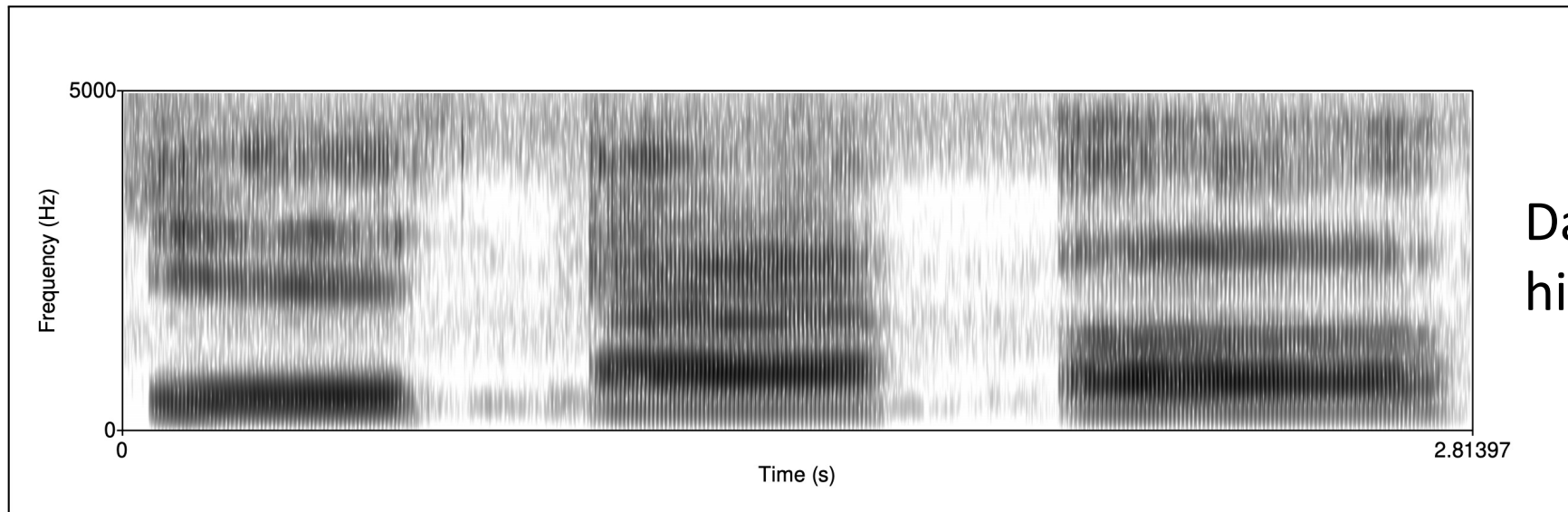


- A smoothed spectrum computed with a discrete Fourier transform (DFT)
 - There are roughly 4 tiny peaks within a cycle
 - $\Rightarrow$ A frequency f_2 of about 4 times f_1
 - $4 \times 234 = 936$ Hz



Spectrogram

- **Spectrogram:** A dynamic view of how spectrum changes over time.



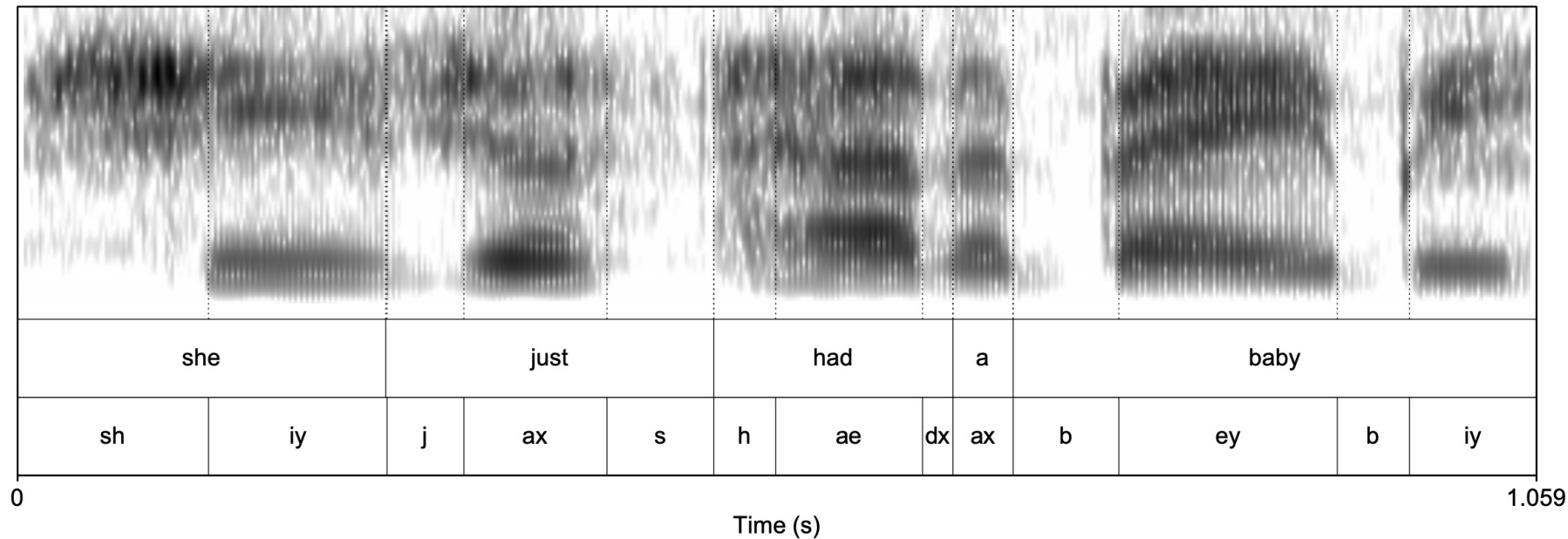
Darker points have higher amplitude.

Figure 15.21 Spectrograms for three American English vowels, [ih], [ae], and [uh]

- Spectrogram is a 3D view: time $\times$ frequency $\times$ amplitude

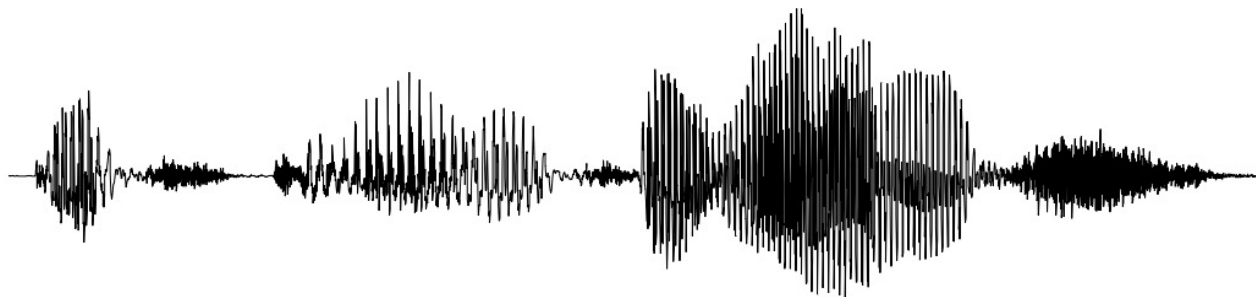
Spectrogram

- Each dark bar in a spectrogram is called a **formant** (共振峰) -- A frequency band that is particularly amplified.
- Since different vowels are pronounced with the vocal tract (声道) in different positions, they will produce different amplifications, thus different formants.



Automatic Speech Recognition (语音识别)

- ASR maps a speech waveform like this:



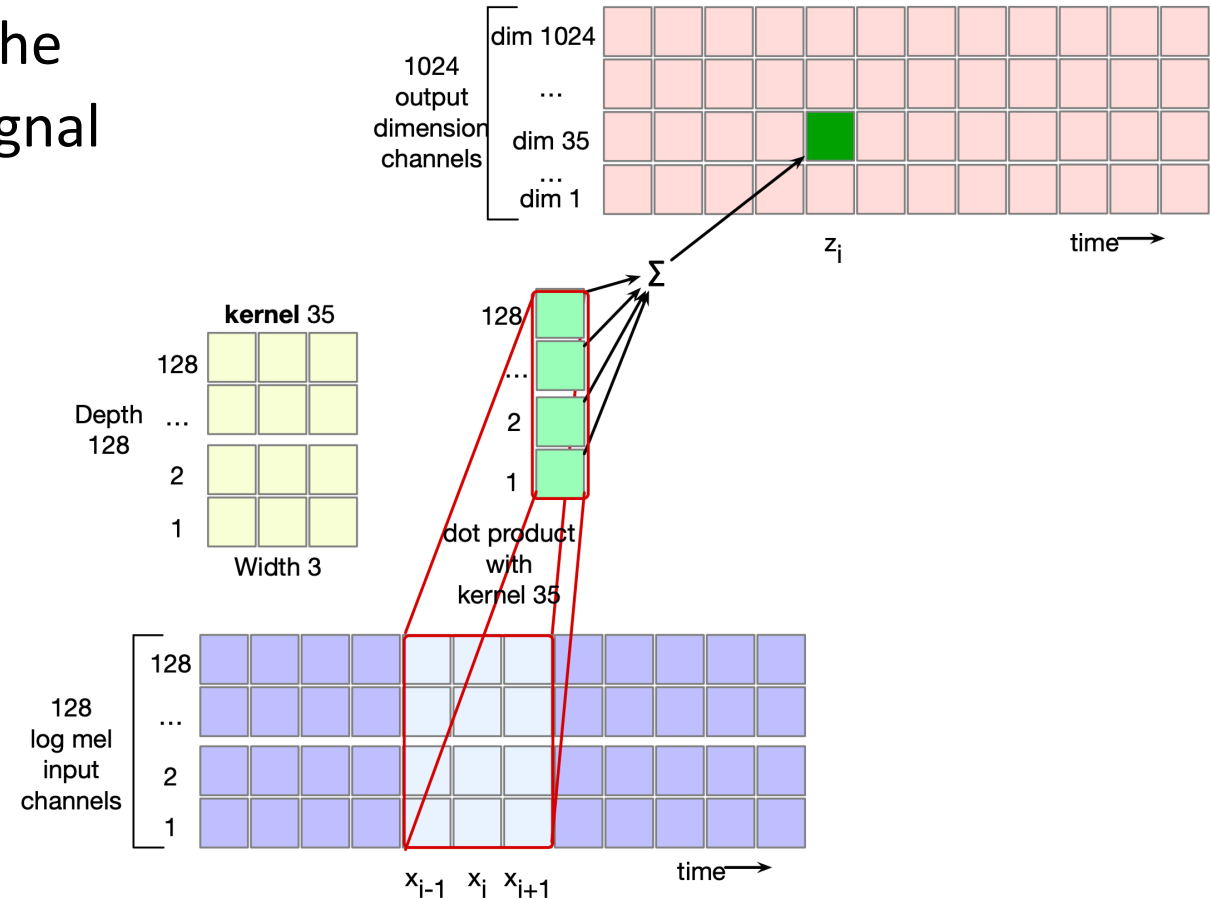
to the corresponding words: “It’s time for lunch!”

- A ConvNet can extract acoustic features for an encoder–decoder model.
- The reverse task is text-to-speech (TTS, 语音合成).

Automatic Speech Recognition

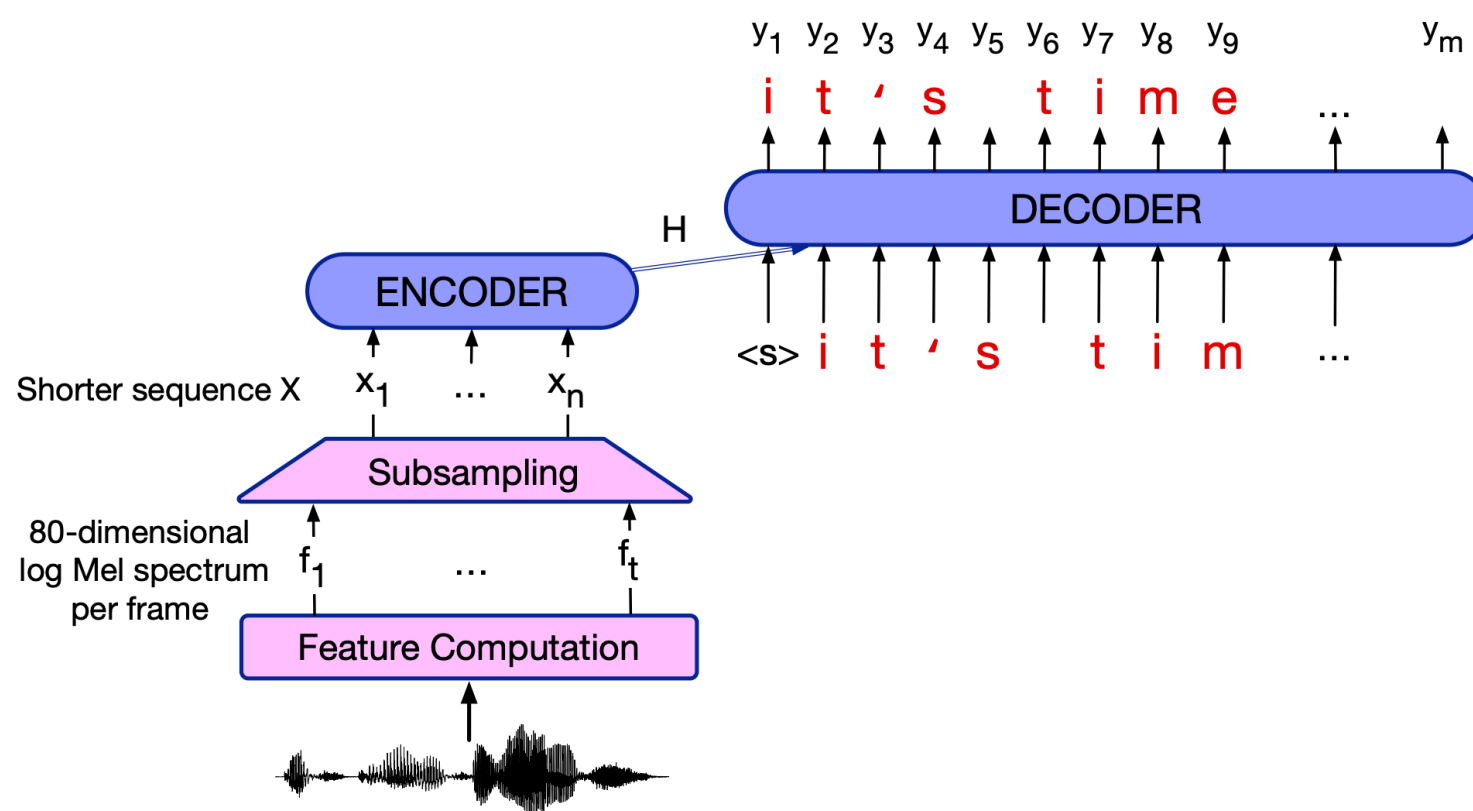
- Convolutional kernels applied to the **log mel** representation of input signal

$$mel(f) = 1127 \ln\left(1 + \frac{f}{700}\right)$$



Automatic Speech Recognition

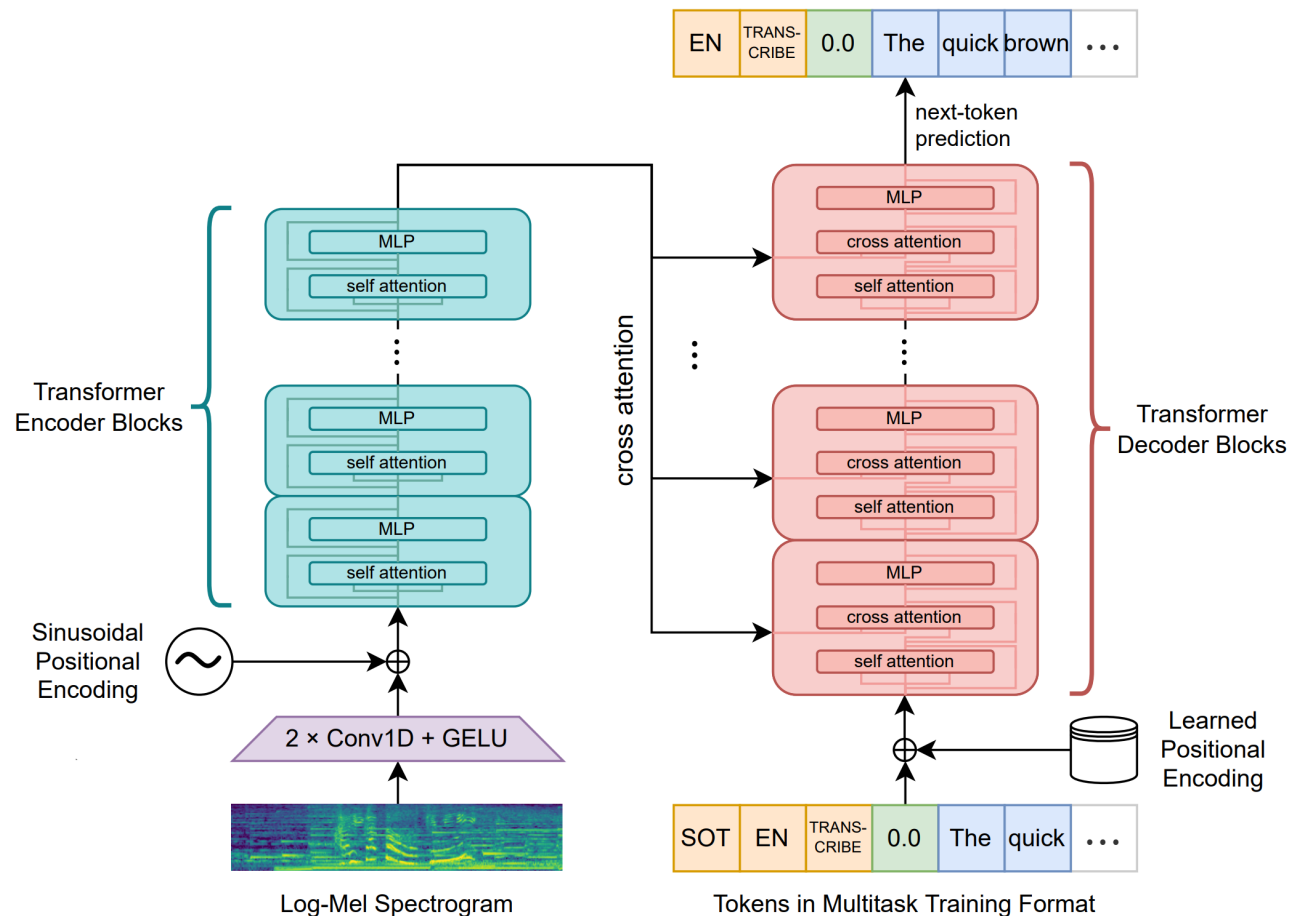
- Encoder-decoder architecture.



Output sequence Y can be either letters or tokens.

Automatic Speech Recognition

- OpenAI's Whisper architecture (2023)



Text-to-Speech (TTS)

- Modern TTS system is similar to language modeling.
- Use an **audio tokenizer** based on an **audio codec** (音频编码器) to induce discrete audio tokens to represent the speech.
- Then train a language model whose vocabulary includes both speech tokens and text tokens.

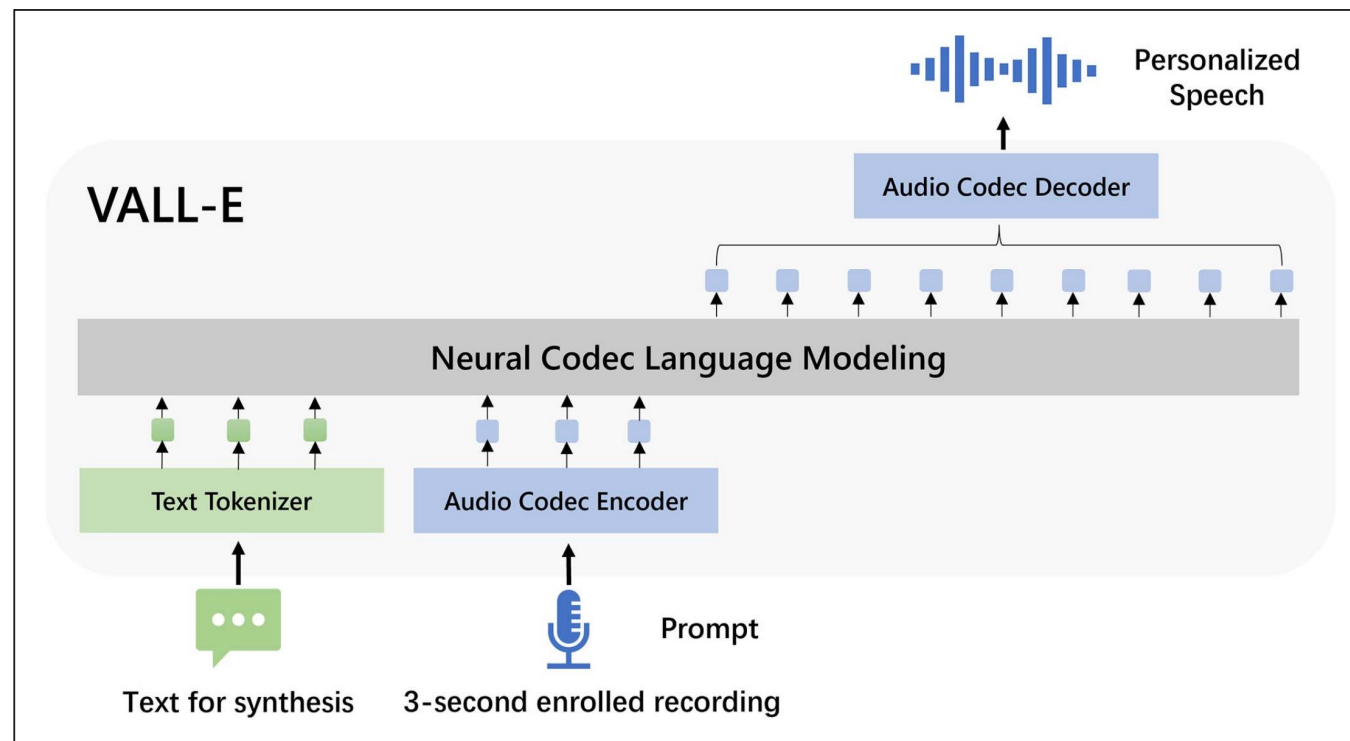


Figure 17.1 VALL-E architecture for personalized TTS (figure from [Chen et al. \(2025\)](#)).