Spoken Language Structure

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

- X. Huang et. al., Spoken Language Processing, Chapter 2
- 王小川教授，語音訊號處理，Chapters 2~3

Introduction

- Take a button-up approach to introduce the basic concepts from sound to phonetics (語音學) and phonology (音韻學)
  - Syllables (音節) and words (詞) are followed by syntax (語法) and semantics (語意), which form the structure of spoken language processing

- Topics covered here
  - Speech Production
  - Speech Perception
  - Phonetics and Phonology
  - Structural Features of the Chinese Language

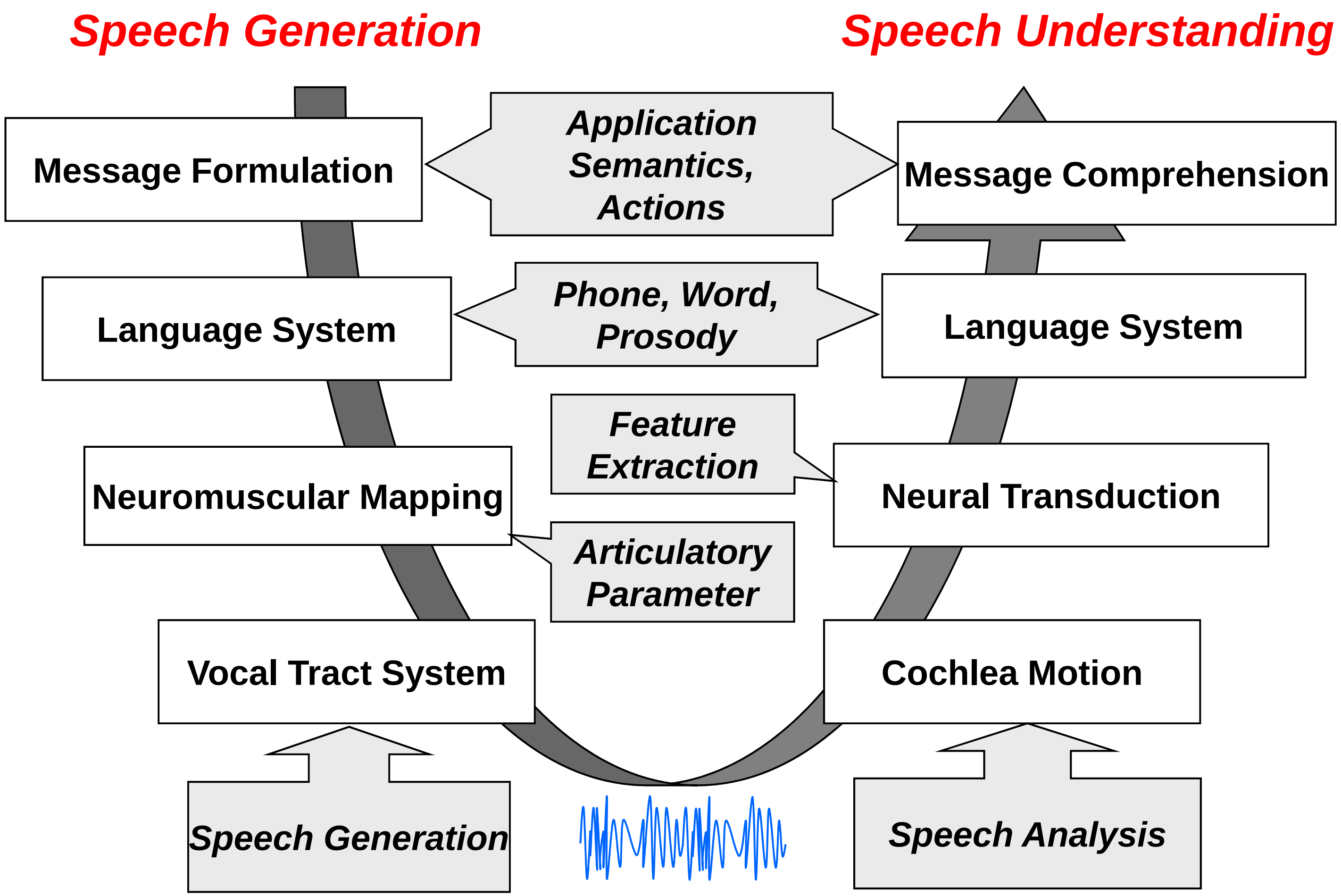
Determinants of Speech Communication

- Spoken language is used to communicate information from a speaker to a listener. Speech production and perception are both important components of the speech chain

- Speech signals are composed of analog sound patterns that serve as the basis for a discrete, symbolic representation of the spoken language – phonemes, syllables and words

- The production and interpretation of these sounds are governed by the ***syntax*** and ***semantics*** of the language spoken

Determinants of Speech Communication (cont.)



Computer Counterpart

- The Speech Production Process
  - **Message formulation:** create the concept (message) to be expressed
  - **Language system:** convert the message into a sequence of words and find the pronunciation of the words (or the phoneme sequence).
    - Apply the prosodic pattern: duration of phoneme, intonation (語調) of the sentence, and the loudness of the sounds
  - **Neuromuscular (神經肌肉) Mapping:** perform articulatory (發聲的) mapping to control the vocal cords, lips, jaw, tongue, etc., to produce the sound sequence

Computer Counterpart (cont.)

- The Speech Understanding Process
  - **Cochlea (耳蝸) motion:** the signal is passed to the cochlea in the inner ear, which performs the frequency analysis as a filter bank
  - **Neural transduction:** converts the spectral signal into activity signals on the auditory nerve, corresponding to a feature extraction component

  It's unclear how neural activity is mapped into the language system and how message comprehension (理解) is achieved in the brain

Sound

- Sound is a longitudinal (縱向的) pressure wave formed of compressions (壓縮) and rarefactions (稀疏) of air molecules (微粒), in a direction parallel to that of the application of energy

- *Compressions* are zones where air molecules have been forced by the application of energy into a tighter-than-usual configuration

- *Rarefactions* are zones where air molecules are less tightly packed

Sound (cont.)

- The alternating configurations of compression and rarefaction of air molecules along the path of an energy source are sometimes described by the graph of a ***sine wave***
- The use of the sine graph is only a notational convenience for charting local pressure variations over time

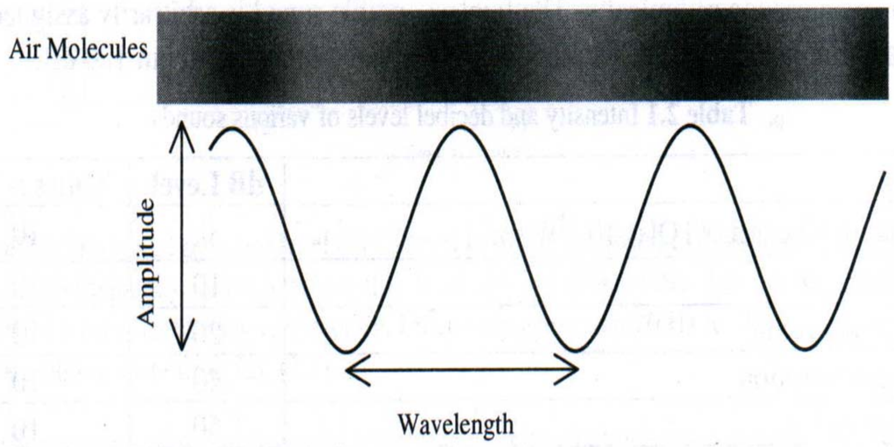


Figure 2.2 Application of sound energy causes alternating compression/rarefaction of air molecules, described by a sine wave. There are two important parameters, amplitude and wavelength, to describe a sine wave. Frequency [cycles/second measured in Hertz (Hz)] is also used to measure of the waveform.

Measures of Sounds

- *Amplitude* is related to *the degree of displacement of the molecules* from their resting position
  - Measured on a logarithm scale in *decibels* (dB, 分貝)
  - A decibel is a means for comparing the intensity (強度) of two sounds:
    $$10 \log_{10}(I / I_0). \; I, I_0 \text{ are two intensity levels}$$
  - The intensity is proportional to the **square** of the sound pressure *P.* The **Sound Pressure Level** (SPL) is a measure of the absolute sound pressure *P* in dB
    $$SPL(dB) = 20 \log_{10}(P / P_0)$$
  - The reference 0 dB corresponds to the threshold of hearing, which is P_0=0.00002 µ*bar* for a tone of 1KHz
    - E.g., the speech conversation level at 3 feet is about 60dB SPL; a jackhammer's level is about 120 db SPL

Measures of Sound (cont.)

- Absolute threshold of hearing: is the maximum amount of energy of a pure tone that cannot be detected by a listener in a noise free environment

Table 2.1 Intensity and decibel levels of various sounds.

Sound	dB Level	Times > TOH
Threshold of hearing (TOH: $10^{-12} W/m^2$)	0	10^0
Light whisper	10	10^1
Quiet living room	20	10^2
Quiet conversation	40	10^4
Average office	50	10^5
Normal conversation	60	10^6
Busy city street	70	10^7
Acoustic guitar – 1 ft. away	80	10^8
Heavy truck traffic	90	10^9
Subway from platform	100	10^{10}
Power tools	110	10^{11}
Pain threshold of ear	120	10^{12}
Airport runway	130	10^{13}
Sonic boom	140	10^{14}
Permanent damage to hearing	150	10^{15}
Jet engine, close up	160	10^{16}
Rocket engine	180	10^{18}
Twelve ft. from artillery cannon muzzle ($10^{10} W/m^2$)	220	10^{22}

in sound pressure level

The absolute threshold of hearing is the maximum amount of energy of a pure tone that cannot be detected by a listener in a noise free environment. The absolute threshold of hearing is a function of frequency that can be approximated by

$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4 \quad (dB\ SPL) \tag{2.3}$$

and is plotted in Figure 2.3.

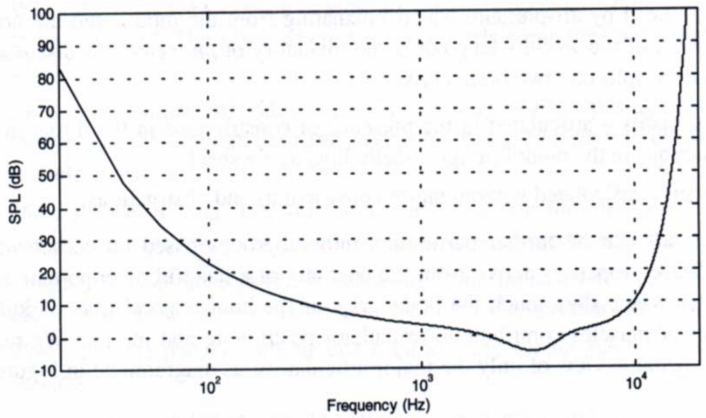


Figure 2.3 The sound pressure level (SPL) level in dB of the absolute threshold of hearing as a function of frequency. Sounds below this level are inaudible. Note that below 100 Hz and above 10 kHz this level rises very rapidly. Frequency goes from 20 Hz to 20 kHz and is plotted in a logarithmic scale from Eq. (2.3).

Speech Production
– *Articulation*

- Speech
  - – Produced by air-pressure waves emanating (發出) from the mouth and the nostrils(鼻孔)
  - – The inventory of **phonemes** (音素) are the basic units of speech and split into two classes
    - **Consonant** (子音/輔音)
      - – Articulated (發音) when constrictions (壓縮) in the throat or obstructions (阻塞) in the mouth
    - **Vowel** (母音/元音)
      - – without major constrictions and obstructions

Speech Production
– *Articulation* (cont.)

- Human speech production apparatus
  - **Lungs** (肺): source of air during speech
  - **Vocal cords** (larynx,喉頭): when the vocal folds (聲帶) are held close together and oscillate one another during a speech sound, the speech sound is said to be **voiced** (<=>**unvoiced**)
  - **Soft Palate** (Velum,軟顎): allow passage of air through the nasal cavity
  - **Hard palate** (硬顎): tongue placed on it to produce certain consonants
  - **Tongue**(舌): flexible articulator, shaped away from palate for vowel, closed to or on the palate or other hard surfaces for consonant
  - **Teeth**: braces (支撐) the tongue for certain consonants
  - **Lips**(嘴唇): round or spread to affect vowel quality, closed completely to stop the oral air flow for certain consonants (*p,b,m*)

Speech Production
– *Articulation* (cont.)

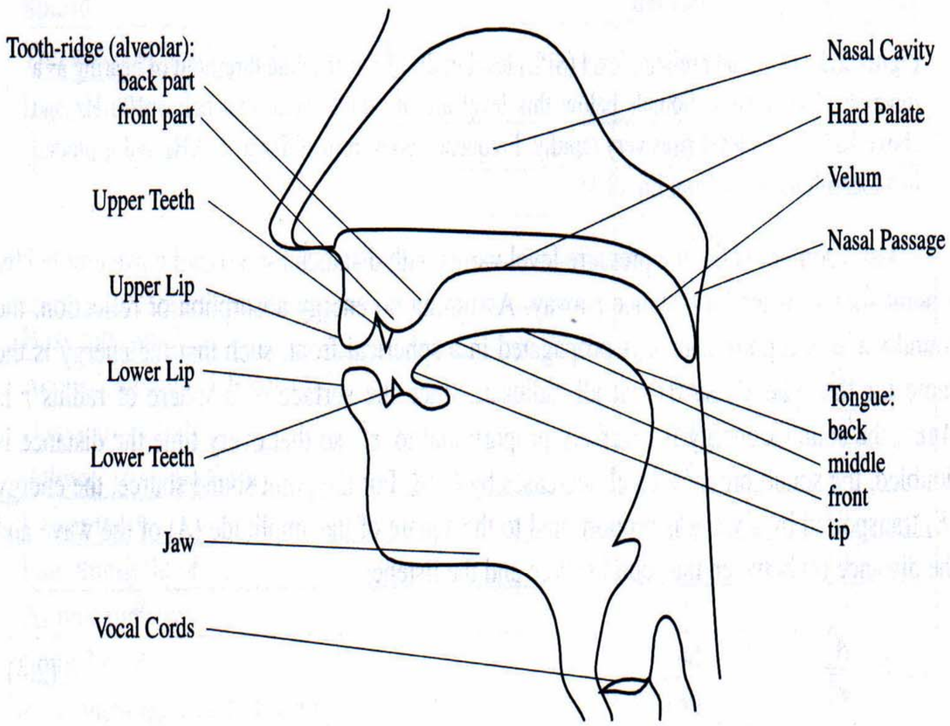


Figure 2.4 A schematic diagram of the human speech production apparatus.

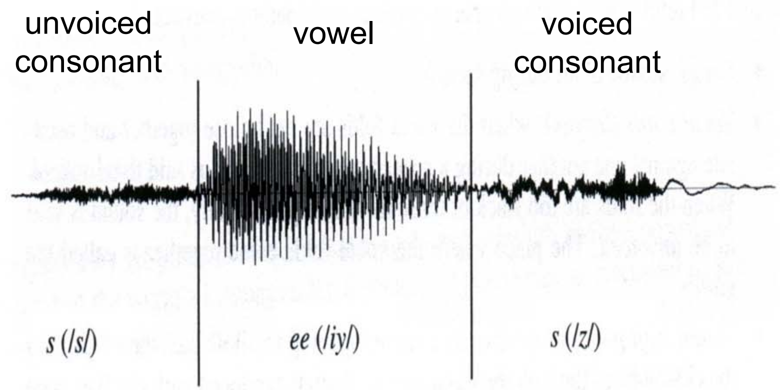


Figure 2.5 Waveform of *sees*, showing a voiceless phoneme /s/, followed by a voiced sound, the vowel /iy/. The final sound, /z/, is a type of voiced consonant.

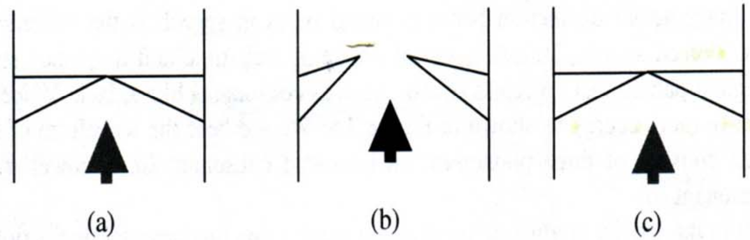


Figure 2.6 Vocal fold cycling at the larynx. (a) Closed with sub-glottal pressure buildup; (b) trans-glottal pressure differential causing folds to blow apart; (c) pressure equalization and tissue elasticity forcing temporary reclosure of vocal folds, ready to begin next cycle.

Speech Production
- The Voicing Mechanisms

- Voiced sounds
  - Including vowels, have a roughly **regular pattern** in both time and frequency structures than voiceless sounds
  - Have more energy
  - Vocal folds vibrate during phoneme articulation (otherwise is **unvoiced**)
    - Vocal folds' vibration (60H ~ 300 Hz, cycles in sec.)
    - 男生分佈較低，女生分佈較高
    - The greater mass and length of adult male vocal folds as opposed to female
  - In psychoacoustics, the distinct vowel timbres (of a sound of a instrument, 音質/色) is determined by how the tongue and lips shaping the oral resonance (共鳴/振) cavity

Speech Production
- *The Voicing Mechanisms* (cont.)

- Voiced sounds (cont.)
  - The rate of cycling (open and closing) of vocal folds in the larynx during phonation of voiced sounds is called the fundamental frequency (基頻)
    - The fundamental frequency contributes more than any other single factor to the perception of ***pitch*** in speech
    - A prosodic feature for use in recognition of tonal languages (e.g., Chinese) or as a measure of speaker identity or authenticity

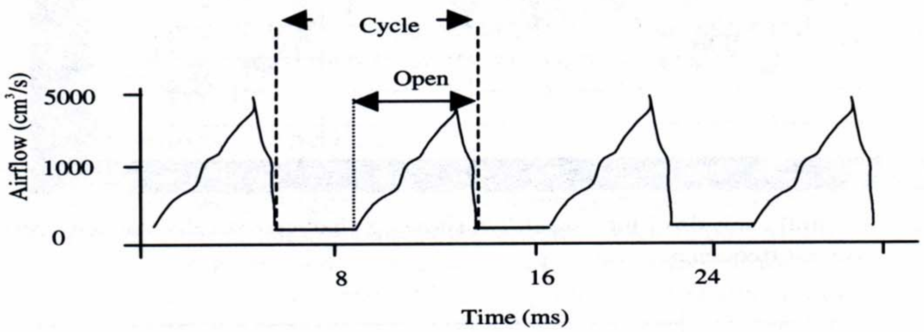


Figure 2.7 Waveform showing air flow during laryngeal cycle.

Speech Production

- Pitch

The term *pitch* is often used interchangeably with fundamental frequency. However, there is a subtle difference. Psychoacousticians (scientists who study the perception of sound) use the term *pitch* to refer to the *perceived* fundamental frequency of a sound, *whether or not that sound is actually present in the waveform.* Speech transmitted over the commercial phone lines, for example, are usually bandlimited to about 300–3000 Hz. Nevertheless, a person who is phonating at 110 Hz will be *perceived* as phonating at 110 Hz by the listener, even though the fundamental frequency of the received waveform cannot be less than 300 Hz. In this case, the psychoacoustician would say that the pitch of the received speech waveform is 110 Hz, while the lowest frequency in the signal is 330 Hz. This quirk of the human auditory system requires that we be careful with these terms. Nevertheless, with this caution, we will routinely use the word "pitch" to mean "fundamental frequency" in this book, since it is conventional to do so. Since we will not be concerned with perceptual phenomena, this will not cause ambiguities to arise.

Speech Production
- *Formants*

- The resonances (共振/共鳴) of the cavities that are typical of particular articulator configurations (e.g. the different vowel timbres) are called formants (共振峰)

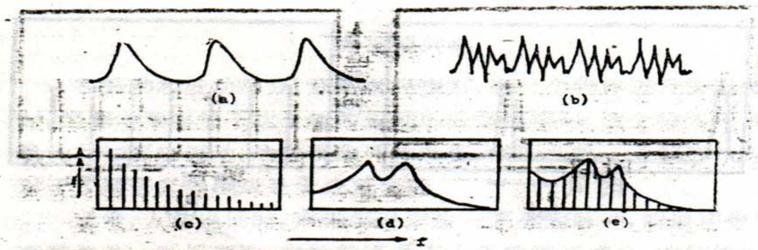


图 3.8 语音的声腔调节示意图:(a) 声带脉冲, (b) 幅射波,(c) 声带谱,(d) 声腔共振包络,(e) 语音频谱

(此图采自 G. Fant, Acoustic Theory of Speech Production)

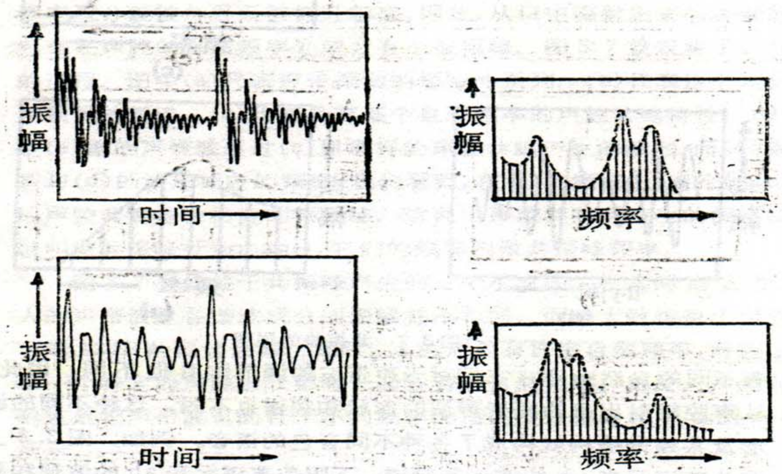


*图 3.9 英语元音 uh(上图)和 ah(下图) 的波形及频谱

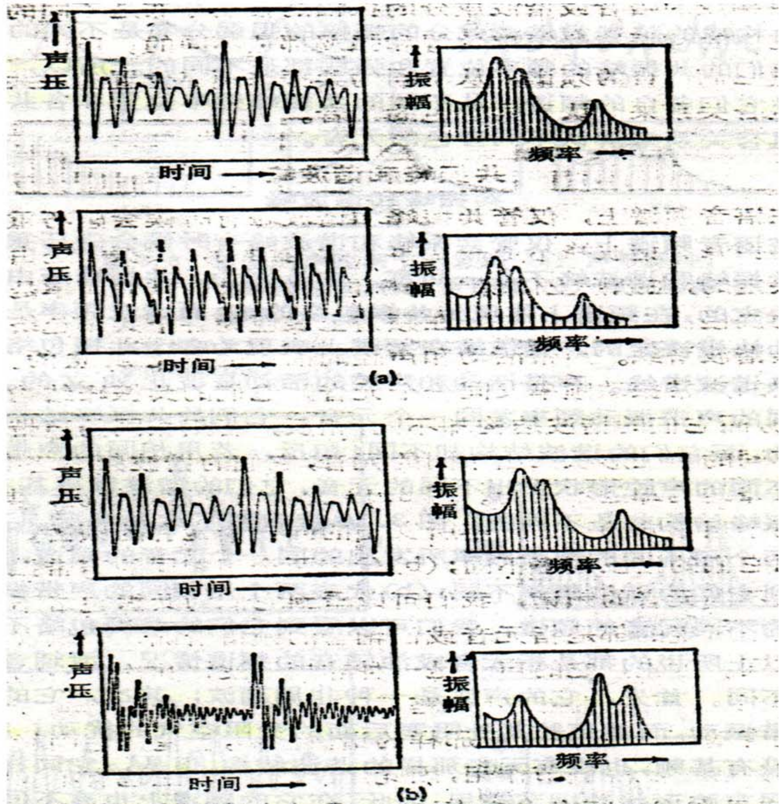


*图 3.10 (a) 用不同的声带频率所发的同一元音的波形和频谱
(b) 用相同的声带频率所发的不同元音的波形和频谱

Speech Production
- *Formants* (cont.)

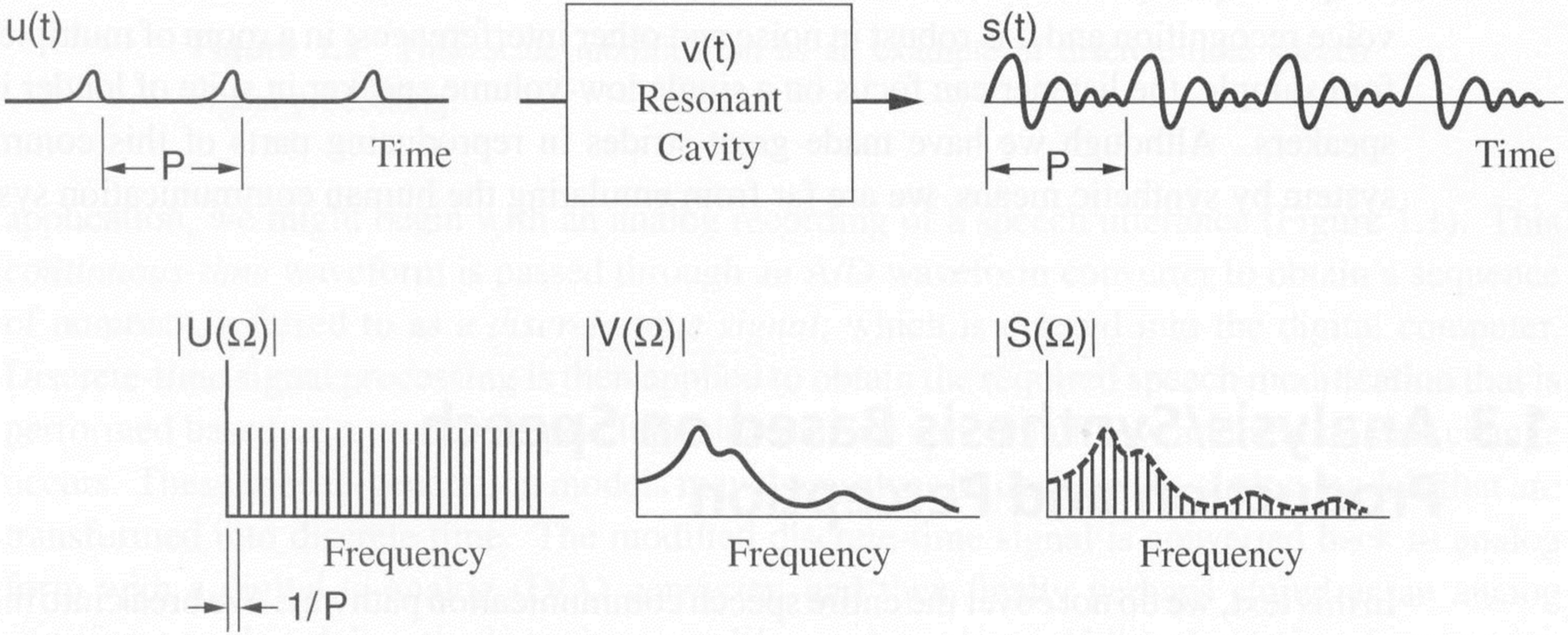


Figure 1.2 Speech production mechanism and model of a steady-state vowel. The acoustic waveform is modeled as the output of a linear time-invariant system with a periodic impulse-like input. In the frequency domain, the vocal tract system function spectrally shapes the harmonic input.

Speech Production
- *Formants* (cont.)

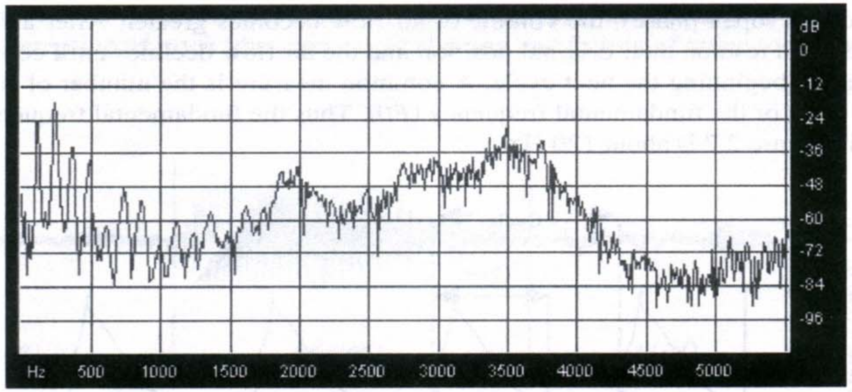


Figure 2.8 A spectral analysis of the vowel /*iy*/, showing characteristically uneven distribution of energy at different frequencies.

Spectrum

頻譜

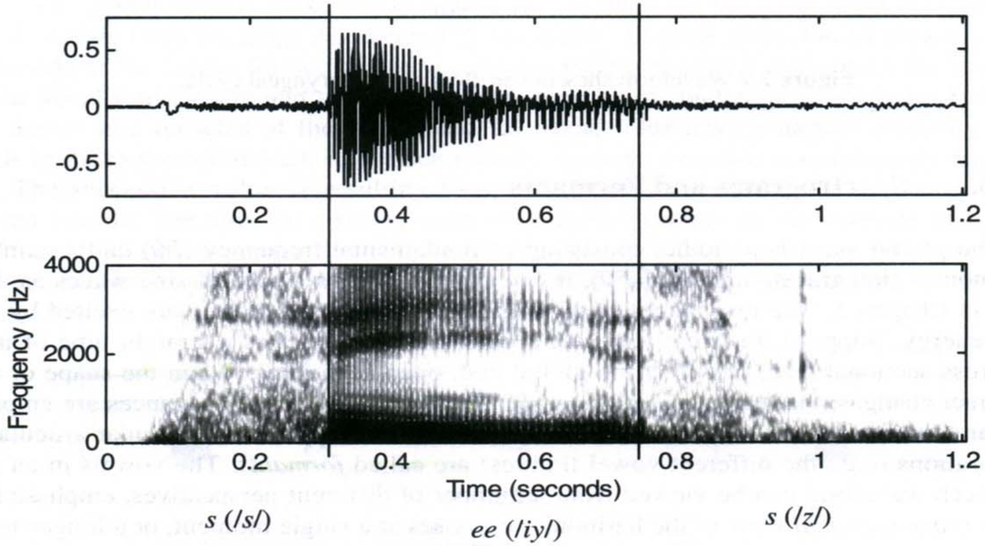


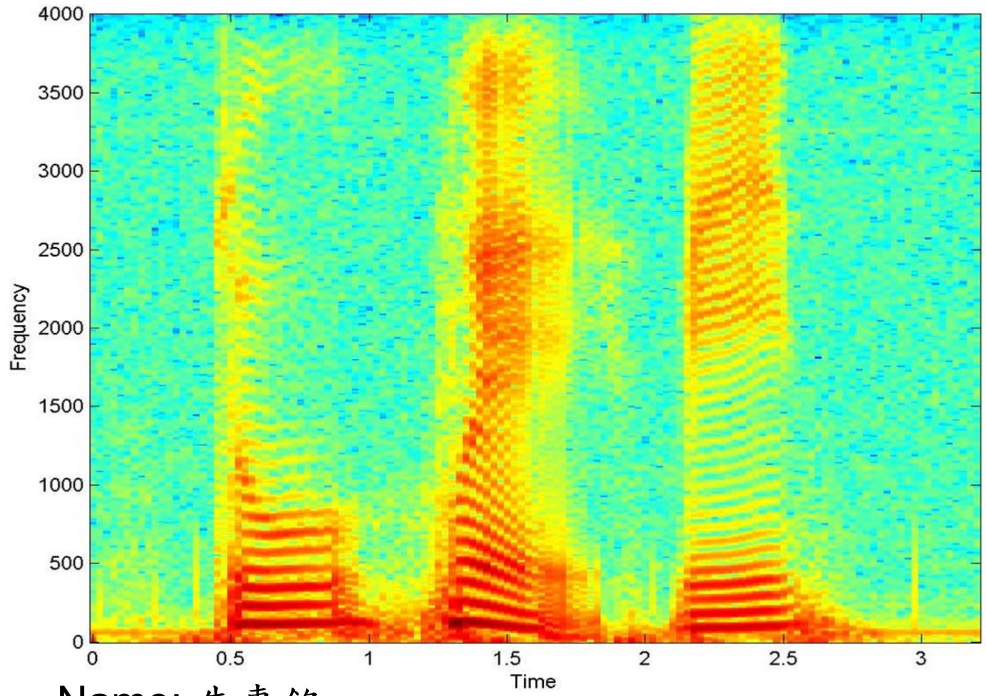
Figure 2.9 The spectrogram representation of the speech waveform *sees* (approximate phone boundaries are indicated with heavy vertical lines).

Spectrogram

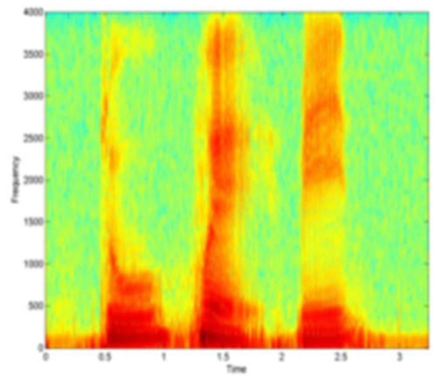
聲譜圖

Speech Production
- *Formants* (cont.)

- Narrowband Spectrogram
  - Both pitch harmonic and format information can be observed



Name: 朱惠銘
1024-point FFT, 400 ms/frame, 200 ms/frame move



100 ms/frame,
50 ms/frame move

Wide-band spectrograms：shorter windows (<10ms)
- Have good time resolution

Narrow-band spectrograms：Longer windows (>20ms)
- The harmonics can be clearly seen

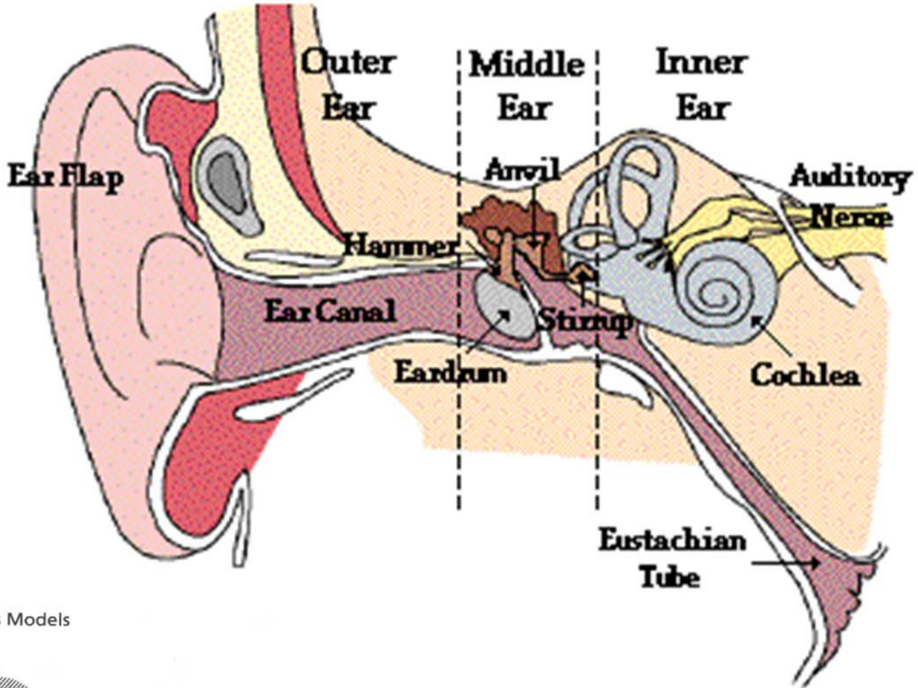
Speech Perception

Physiology of the Ear

- The ear processes an acoustic pressure signal by
  - First transforming it into a mechanical vibration pattern on the basilar membrane (基底膜)
  - Then representing the pattern by a series of pulses to be transmitted by the auditory nerve
- Physiology of the Ear
  - When air pressure variations reach the eardrum from the outside, it vibrates, and transmits the vibrations to bones adjacent to its opposite side
  - Then the energy is transferred by mechanical action of the stapes into an impression on the membrane stretching over the oval window
  - The cochlea can be roughly regarded as a set of filter banks, whose outputs are ordered by location
    - Frequency-to-place transformation

Speech Perception

Physiology of the Ear (cont.)



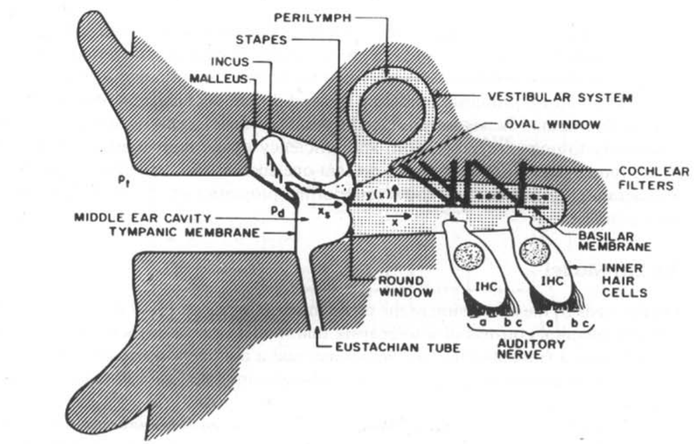
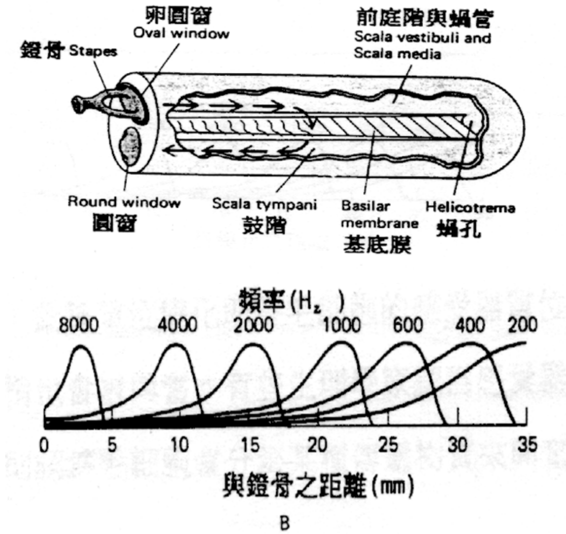


Figure 3.48 Expanded view of the middle and inner ear mechanics.



Speech Perception

Physiology of the Ear (cont.)

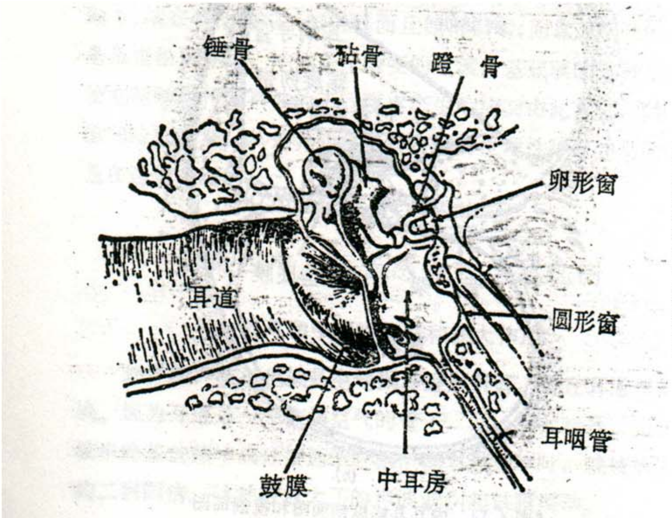
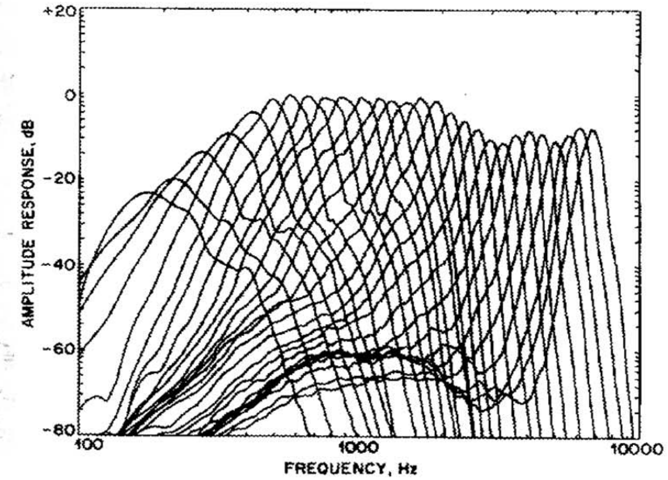


*图 4.13 中耳听小骨



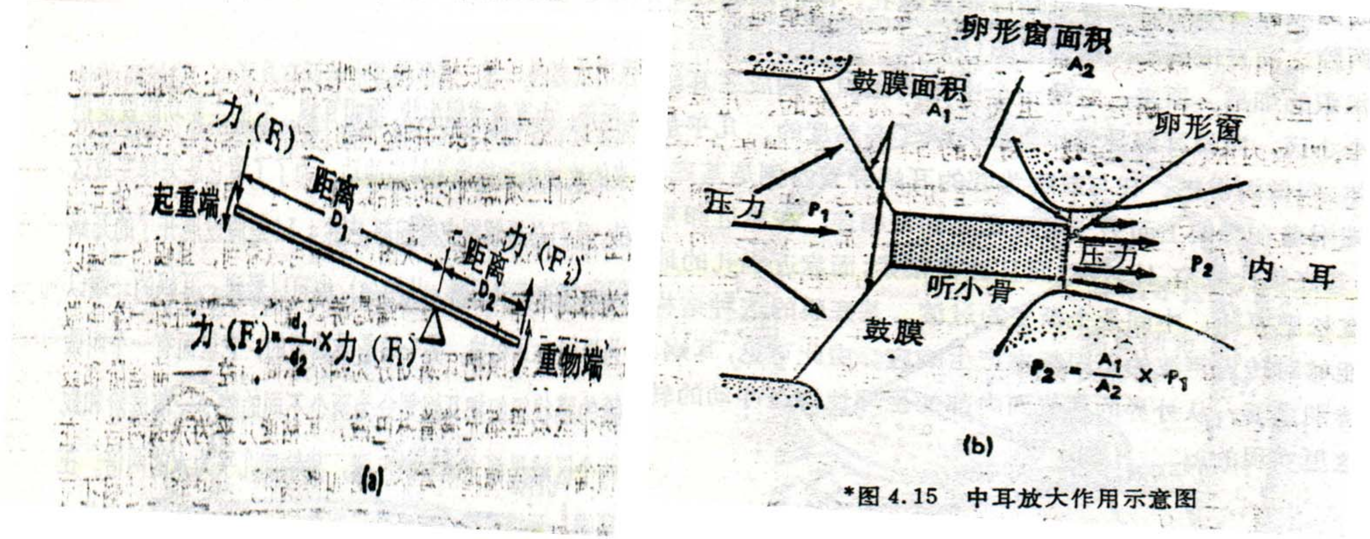


*图 4.15 中耳放大作用示意图

Speech Perception

Physical vs. Perceptual Attributes

- Non-uniform equal loudness perception of tones of varying frequencies
  - Tones of different pitch have different perceived loudness
  - Sensitivity of the ear varies with the frequency and the quality of sound
  - Hear sensitivity reaches a maximum around 4000 Hz

Table 2.2 Relation between perceptual and physical attributes of sound.

Physical Quantity	Perceptual Quality
Intensity	Loudness
Fundamental frequency	Pitch
Spectral shape	Timbre
Onset/offset time	Timing
Phase difference in binaural hearing	Location

Speech Perception

Physical vs. Perceptual Attributes

- Non-uniform equal loudness perception

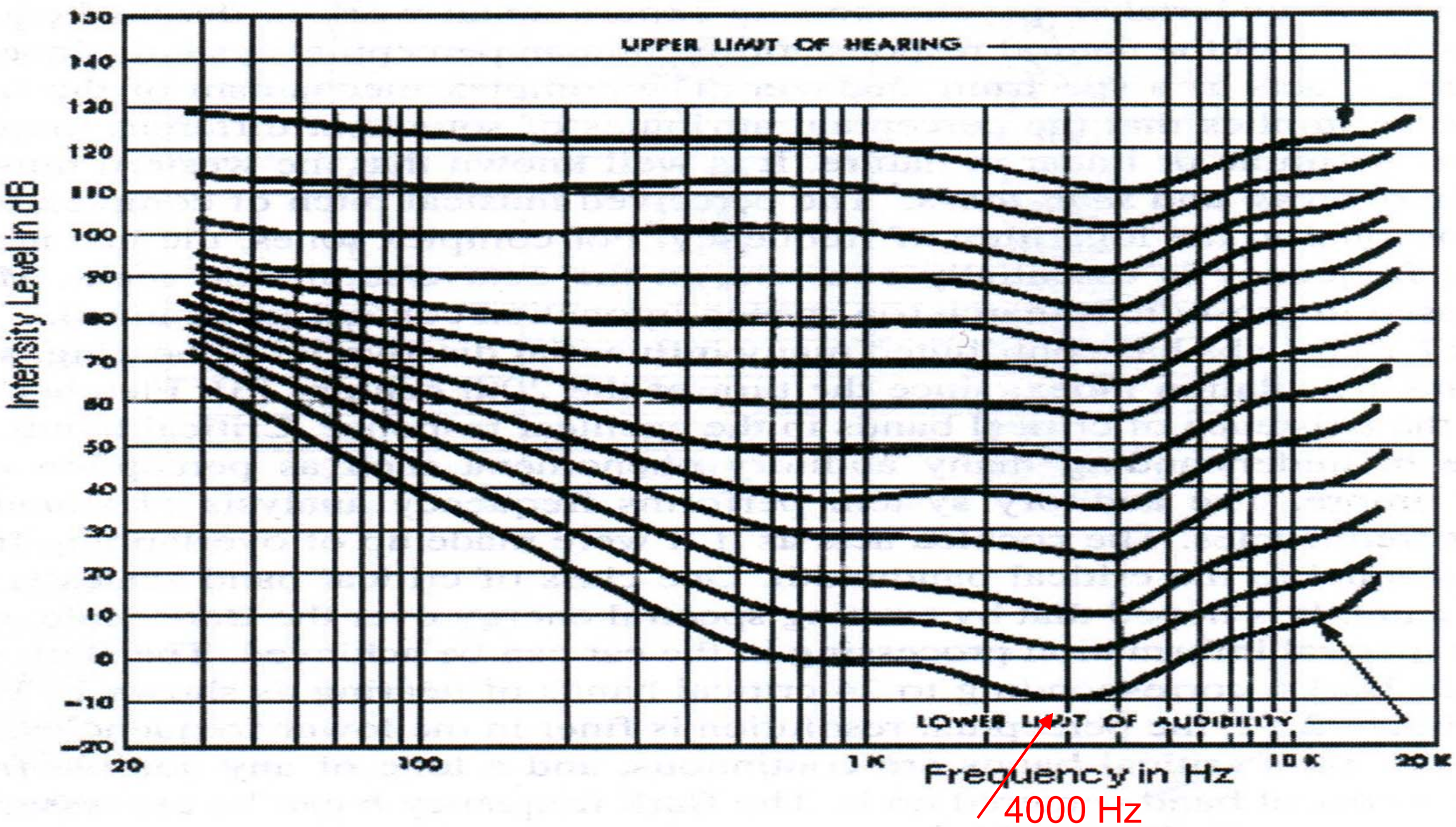


Figure 2.11 Equal-loudness curves indicate that the response of the human hearing mechanism is a function of frequency and loudness levels. This relationship again illustrates the difference between physical dimensions and psychological experience (after ISO 226).

Speech Perception

Physical vs. Perceptual Attributes (cont.)

- Masking: when the ear is exposed to two or more different tones, it's a common experience that one tone may *mask* others
  - An **upward shift** in the hearing threshold of the weaker tone by the louder tone
  - A pure tone masks of higher frequency more effectively than those of lower frequency
  - The greater the intensity of the masking tone, the broader the range of frequencies it can mask

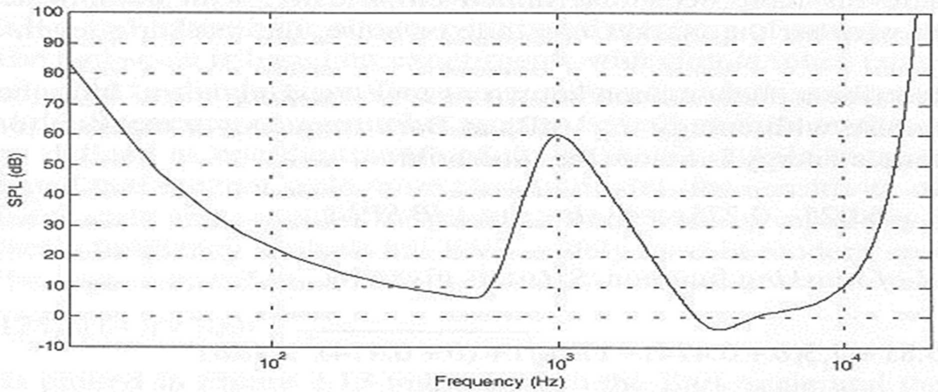


Figure 2.15 Absolute threshold of hearing and spread of masking threshold for a 1 kHz sine-wave masker with a 69 dB SPL. The overall masked threshold is approximately the largest of the two thresholds.

Speech Perception

Physical vs. Perceptual Attributes (cont.)

- The sense of localization attention (Lateralization)
  - Binaural listening greatly enhances our ability to sense the direction of the sound source
  - Time and intensity cues have different impacts for low frequency and high frequency, respectively
    - Low-frequency sounds are lateralized mainly on the basis of interaural **time** differences
    - High-frequency sounds are lateralized mainly on the basis of interaural **intensity** differences
- The question of distinct voice quality
  - Speech from different people sounds different, e.g., different fundamental frequencies, different vocal-tract length
  - **The concept of timbre** (音質) is defined as that the attribute of auditory sensation by which a subject can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar

Speech Perception

Frequency Analysis

- Researchers undertook psychoacoustic (心理聲學) experimental work to derive frequency scales that attempt to model the natural response of the human perceptual system (*the cochlea acts as a spectrum analyzer*)
  - The perceptual attributes of sounds at different frequencies may not be entirely simple or linear in natural

- **Bark Scale**: Fletcher's work (1940) pointed to the existence of critical bands in the cochlear response
  - The **cochlea** acts as if it were made up of **overlapping filters** having bandwidth equal to the critical bandwidth
  - One class of critical band scales is called ***Bark frequency scale*** (24 critical bands)

Speech Perception

Frequency Analysis (cont.)

- **Bark Scale**: (cont.)
  - Treat spectral energy over the Bark scale, a more natural fit with spectral information processing in the ear can be achieved
  - The perceptual resolution (解析度) is finer in the lower frequencies
  - The critical bands are continuous such that a tone of any audible frequency always finds a critical band centered on it

$$b(f) = 13\arctan(0.00076f) + 3.5\arctan\left[\left(\frac{f}{7500}\right)^2\right]$$

Speech Perception

Frequency Analysis (cont.)

- **Bark Scale**: (cont.)

Table 2.3 The Bark frequency scale.

Bark Band #	Edge (Hz)	Center (Hz)
1	100	50
2	200	150
3	300	250
4	400	350
5	510	450
6	630	570
7	770	700
8	920	840
9	1080	1000
10	1270	1170
11	1480	1370
12	1720	1600
13	2000	1850
14	2320	2150
15	2700	2500
16	3150	2900
17	3700	3400
18	4400	4000
19	5300	4800
20	6400	5800
21	7700	7000
22	9500	8500
23	12000	10500
24	15500	13500

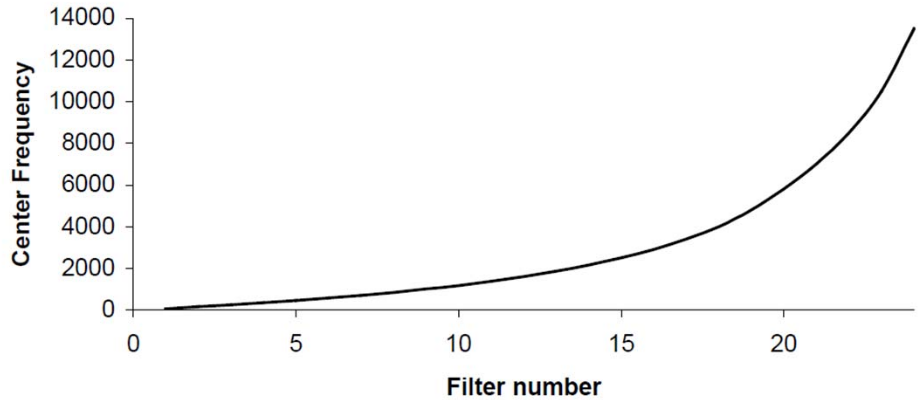
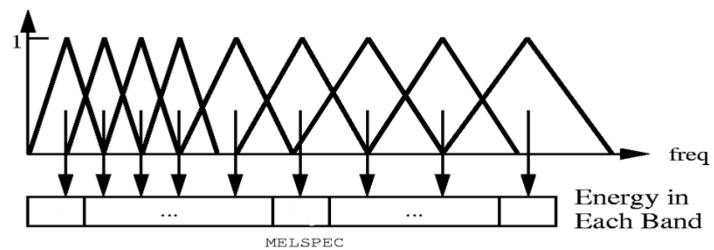


Figure 2.12 The center frequency of 24 Bark frequency filters as illustrated in Table 2.3.



Mel-Scale Filter Bank

Speech Perception

Frequency Analysis (cont.)

- **Mel Frequency Scale (Mel): linear below 1 KHz and logarithmic above**
  - Model the sensitivity of the human ear
  - **Mel**: a unit of measure of perceived pitch or frequency of a tone
- Steven and Volkman (1940)
  - Arbitrarily chose the frequency 1,000 Hz as “1,000 mels”.
  - Listeners were then asked to change the physical frequency until the pitch they perceived was twice the reference, then 10 times, and so on; and then half the reference, 1/10, and so on
    - These pitches were labeled 2,000, 10,000 mels and so on; and 500 and 100 mels, and so on
  - Determine a mapping between the real frequency scale (Hz) and the perceptual frequency (Mel)
  - Have been widely used in modern speech recognition system

Speech Perception

Frequency Analysis (cont.)

- **Mel Frequency Scale** (cont.)

$$Mel\left(f\right) = 1125 \ \ln\left(1 + \frac{f}{700}\right)$$

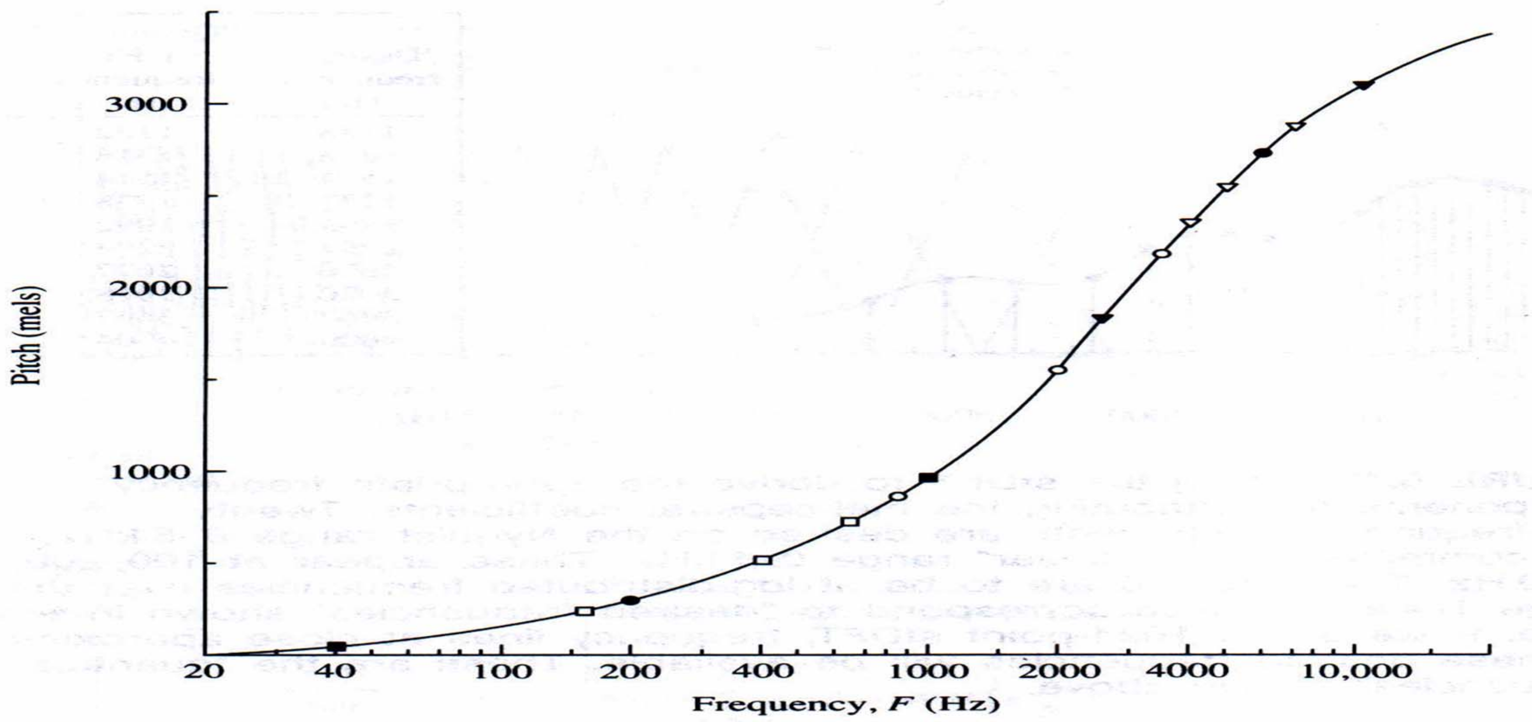


FIGURE 6.12. The mel scale. After Stevens and Volkman (1940).

Speech Perception

Frequency Analysis (cont.)

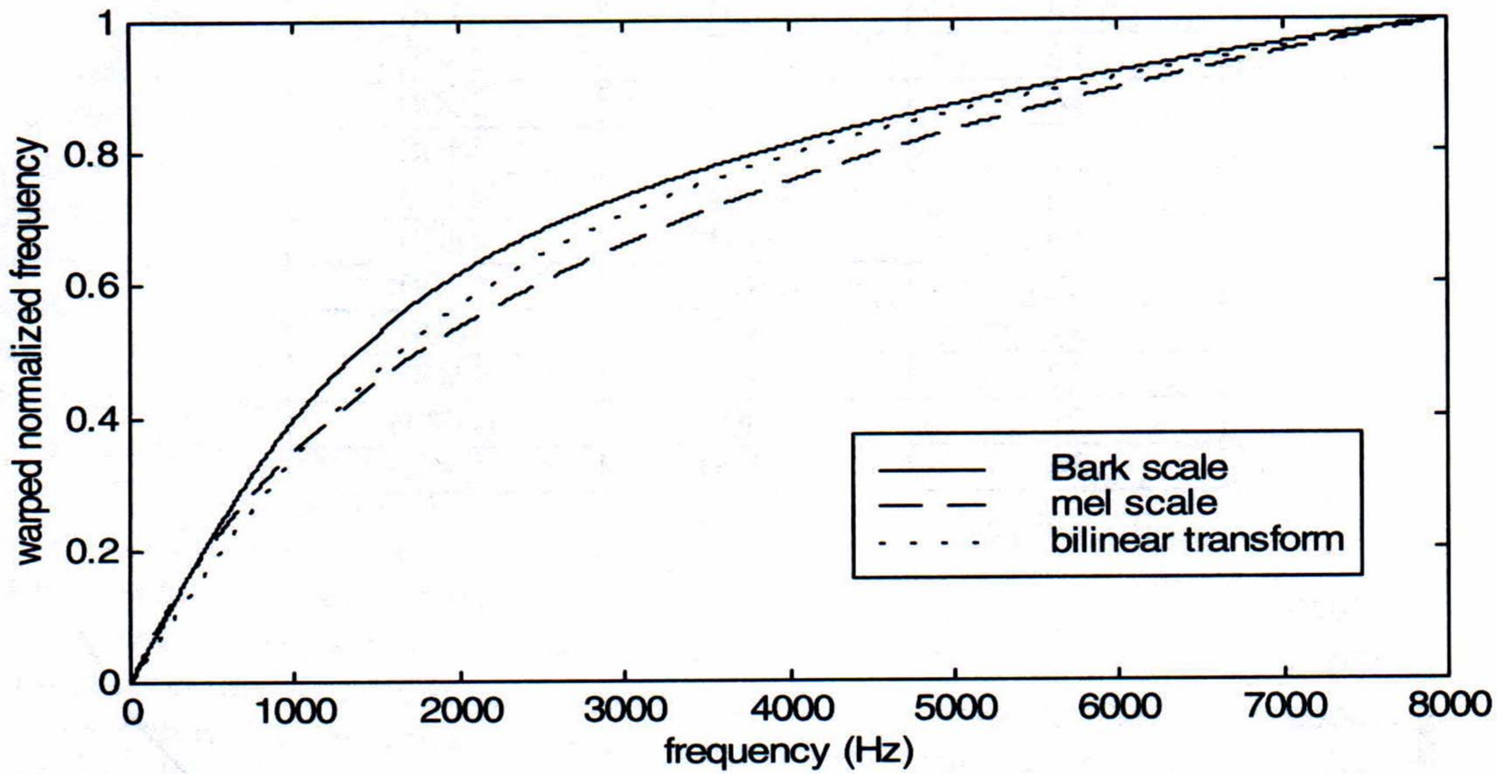


Figure 2.13 Frequency warping according to the Bark scale, ERB scale, mel-scale, and bilinear transform for $\alpha = 0.6$: linear frequency in the x-axis and normalized frequency in the y-axis.

Phonetics and Phonology

- Phonetics (語音學): The study of speech sounds and their production, classification, and transcription

- Phonology (音韻學): The study of the distribution and patterning of speech sounds in a language and of the tacit rules governing the speech pronunciation

Phoneme and Phone

- **Phoneme and Phone**
  - In speech science, the term *phoneme* (音素/音位) is used to denote any of the *minimal units of speech sound* in a language that can serve to distinguish one word from another
    - E.g., m*ea*n /iy/ and m*a*n /ae/
  - The term *phone* is used to denote a phoneme's acoustic realization
    - E.g., phoneme ***/t/*** has two very different acoustic realizations in the word *sa**t*** and *me**t**er.* We had better treat them as ***two different phones*** when building a spoken language system
    - E.g., phoneme */l/* : like and sail

Phoneme and Phone

- Phoneme and phone interchangeably used to refer to the speaker-independent and context-independent units of meaningful sound contrast
  - The set of phonemes will differ in realization across individual speakers

Table 2.4 English phonemes used for typical spoken language systems.

Phonemes	Word Examples	Description
iy	*feel, eve, me*	front close unrounded
ih	*fill, hit, lid*	front close unrounded (lax)
ae	*at, carry, gas*	front open unrounded (tense)
aa	*father, ah, car*	back open unrounded
ah	*cut, bud, up*	open-mid back unrounded
ao	*dog, lawn, caught*	open-mid back round
ay	*tie, ice, bite*	diphthong with quality: aa + ih
ax	*ago, comply*	central close mid (schwa)
ey	*ate, day, tape*	front close-mid unrounded (tense)
eh	*pet, berry, ten*	front open-mid unrounded
er	*turn, fur, meter*	central open-mid unrounded rhoti-
ow	*go, own, tone*	back close-mid rounded
aw	*foul, how, our*	diphthong with quality: aa + uh
oy	*toy, coin, oil*	diphthong with quality: ao + ih
uh	*book, pull, good*	back close-mid unrounded (lax)
uw	*tool, crew, moo*	back close round
b	*big, able, tab*	voiced bilabial plosive
p	*put, open, tap*	voiceless bilabial plosive
d	*dig, idea, wad*	voiced alveolar plosive
t	*talk, sat*	voiceless alveolar plosive &
t	*meter*	alveolar flap
g	*gut, angle, tag*	voiced velar plosive
k	*cut, ken, take*	voiceless velar plosive
f	*fork, after, if*	voiceless labiodental fricative
v	*vat, over, have*	voiced labiodental fricative
s	*sit, cast, toss*	voiceless alveolar fricative
z	*zap, lazy, haze*	voiced alveolar fricative
th	*thin, nothing, truth*	voiceless dental fricative
dh	*then, father, scythe*	voiced dental fricative
sh	*she, cushion, wash*	voiceless postalveolar fricative
zh	*genre, azure*	voiced postalveolar fricative
l	*lid*	alveolar lateral approximant
l	*elbow, sail*	velar lateral approximant
r	*red, part, far*	retroflex approximant
y	*yacht, yard*	palatal sonorant glide
w	*with, away*	labiovelar sonorant glide
hh	*help, ahead, hotel*	voiceless glottal fricative
m	*mat, amid, aim*	bilabial nasal
n	*no, end, pan*	alveolar nasal
ng	*sing, anger*	velar nasal
ch	*chin, archer, march*	voiceless alveolar affricate: t + sh
jh	*joy, agile, edge*	voiced alveolar affricate: d + zh

Vowels

- The tongue shape and positioning on the oral cavity do not form a major constriction (壓縮) of air flow during vowel articulation
  - Variations of tongue placement give each vowel its distinct character by changing the resonances (the positions of formants)
    - Just as different sizes and shapes of bottles give rise to different acoustic effects when struck
  - The linguistically important dimensions of the tongue movements are generally the ranges [**front <-> back**] and [**high <-> low**]

- F1 and F2
  - The primary energy entering the pharyngeal (咽) and oral (口腔) cavities in vowel production vibrates at the fundamental frequency. The major resonances of the oral and pharyngeal cavities for vowels are called F1 and F2

Vowels (cont.)

- F1 and F2 (cont.)
  - The major resonances of these two cavities for vowels are called **F1** and **F2**, the first and second formants
    - **Determined by** the tongue placement and oral tract shape in vowels
    - **Determine** the characteristic timbre or quality of the vowel
  - English vowels can be described by the relationship of F1 and F2 to one another
  - F2 is determined by **the size of the and shape of the oral portion**, forward of the major tongue extrusion(擠壓)
  - F1 corresponds to **the back or pharyngeal portion** of the cavity (the cavity from the glottis (聲門) to the tongue extrusion), which is longer than the forward part. Its resonance would be lower
  - **Rounding the lips** has the effect of extending the front-of-tongue cavity, thus lowering F2

Vowels (cont.)

- The characteristic F1 and F2 values are ideal locations for perception

Table 2.5 Phoneme labels and typical formant values for vowels of English.

Vowel Labels	Mean F1 (Hz)	Mean F2 (Hz)
iy (*feel*)	300	2300
ih (*fill*)	360	2100
ae (*gas*)	750	1750
aa (*father*)	680	1100
ah (*cut*)	720	1240
ao (*dog*)	600	900
ax (*comply*)	720	1240
eh (*pet*)	570	1970
er (*turn*)	580	1380
ow (*tone*)	600	900
uh (*good*)	380	950
uw (*tool*)	300	940

嘴唇愈成圓形或愈開

Vowels (cont.)

- The tongue hump (彎曲、隆起) is the major actor in vowel articulation. The most important secondary vowel mechanism for English and many other language is lip rounding
- E.g. */iy/* (*see*) and */uw/* (*blue*)
  - When you say */iy/*, your tongue will be in the high/front position and your lips will be flat, slightly open, and somewhat spread
    - Lower F1 and Higher F2
  - When you say */uw/*, your tongue will be in the high/back position and your lips begin to round out, ending in a more puckered (縮攏的) position
    - Higher F1 and Lower F2

Vowels (cont.)

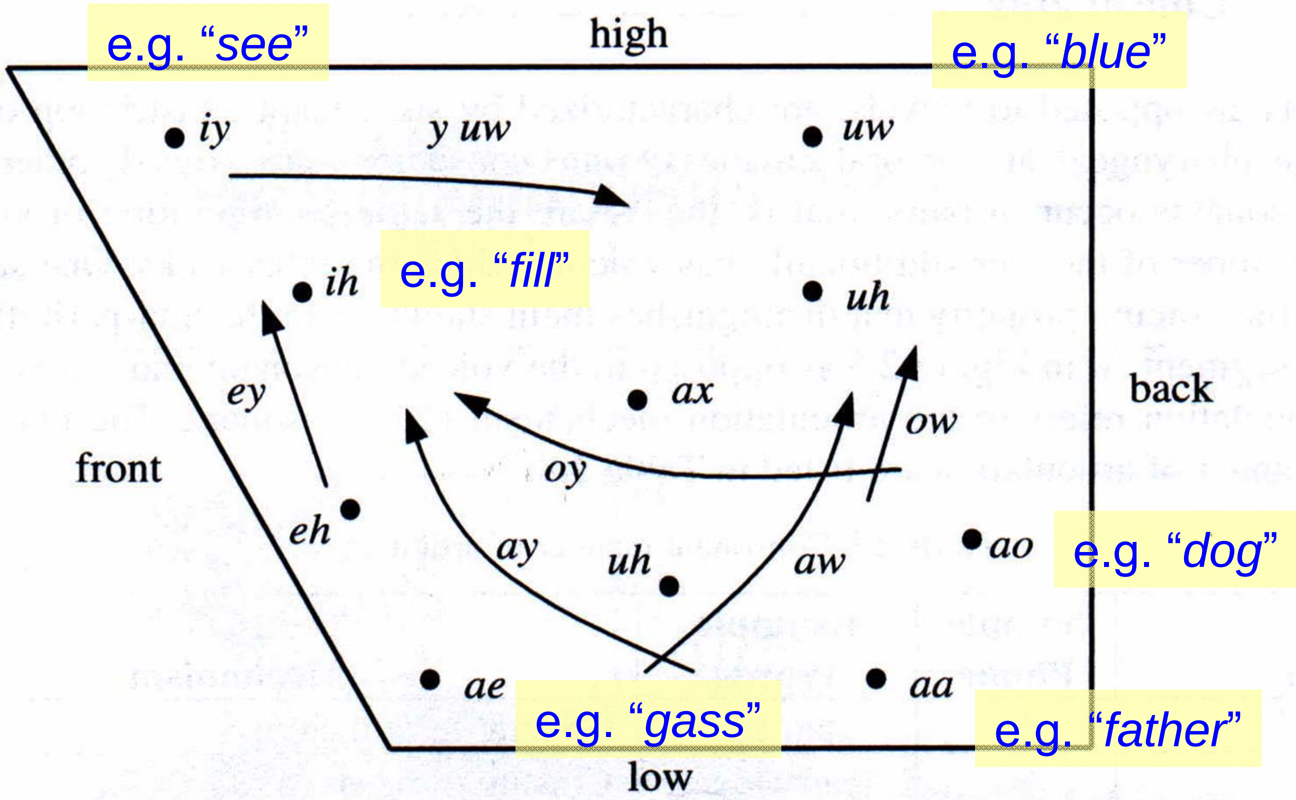


Figure 2.18 Relative tongue positions of English vowels [24].

Vowels (cont.)

- Diphthongs(雙母音)
  - A special class of vowels that combine two distinct sets of F1/F2 values

Table 2.6 The diphthongs of English.

Diphthong Labels	**Components**
ay (**tie**)	*/aa/* ➔ */iy/*
ey (**ate**)	*/eh/* ➔ */iy/*
oy (**coin**)	*/ao/* ➔ */iy/*
aw (**foul**)	*/aa/* ➔ */uw/*

Vowels (cont.)

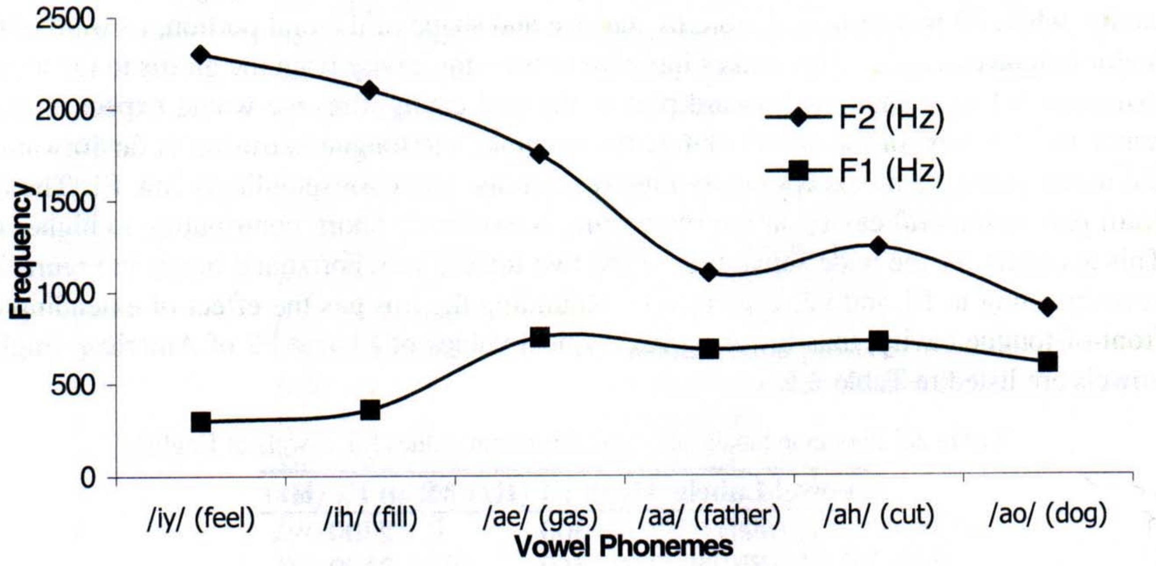


Figure 2.17 F1 and F2 values for articulations of some English vowels.

The major articulator for English vowels is the middle to rear portion of the tongue.

- Note: not only tongue hump (彎曲、隆起) but also lip rounding is the two major actor in vowel articulation for most languages

Consonants

- Characterized by significant constriction (壓縮) or obstruction (阻塞) in the pharyngeal and/or oral cavities
  - Some consonants are voiced; others are not
  - Many consonants occur in pairs, i.e., sharing the same configuration of articulators and one member of the pair additionally has voicing while the other lacks (e.g. /*z, s*/)

Table 2.8 Consonant manner of articulation.

	Manner	Sample Phone	Example Words	Mechanism
破裂音	Plosive	*/p/*	tat, tap	Closure in oral cavity
鼻音	Nasal	*/m/*	team, meet	Closure of nasal cavity
摩擦音	Fricative	*/s/*	sick, kiss	Turbulent airstream noise
捲舌音	Retroflex liquid	*/r/*	rat, tar	Vowel-like, tongue high and curled back
舌邊音	Lateral liquid	*/l/*	lean, kneel	Vowel-like, tongue central, side airstream
滑音	Glide	*/y/,/w/*	yes, well	Vowel-like

Consonants (cont.)

- Plosives (破裂音)
  - E.g., */b, p/*, */d, t/*, /g, k/
  - Consonant that involve complete blockage of oral cavity
- Fricatives (摩擦音)
  - E.g., */z, s/*
  - Consonants that involve nearly complete blockage of oral cavity
- Nasals (鼻音)
  - E.g., */m, n, ng/*
  - Consonants that let the oral cavity significantly constricted, velar (軟顎) open, voicing and air pass through the nasal cavity
- Retroflex liquids (捲舌音)
  - E.g., /r/
  - The tip of the tongue is curled back slightly

Consonants (cont.)

- Lateral liquids (舌邊音)
  - E.g., /l/
  - Air stream flows around the sides of the tongue
- Glides (滑音)
  - E.g. /y, w/
  - Be a little shorted and lack the ability to be stressed, usually at the initial position within a syllable (e.g., yes, well)

Consonants (cont.)

- Semi-vowels
  - Have voicing without complete constriction or obstruction of the vocal tract
  - Include the liquid group /r, l/ and glide group /y, w/
  - {vowels, semi-vowels}: sonorant (響音)
- Non-sonorant consonants
  - Maintain some voicing before or during the obstruction until the pressure differential across the glottis (聲門) to disappear, due to the closure
  帶聲的子音
  - E.g., */b, d, g, z, zh, v/* (voicing) and their counterparts
    /p, t, k, s, sh, f/ 不帶聲的子音

Consonants (cont.)

**图 4.10 普通话 b、d、g 和 s、sh、x 的X光舌位图

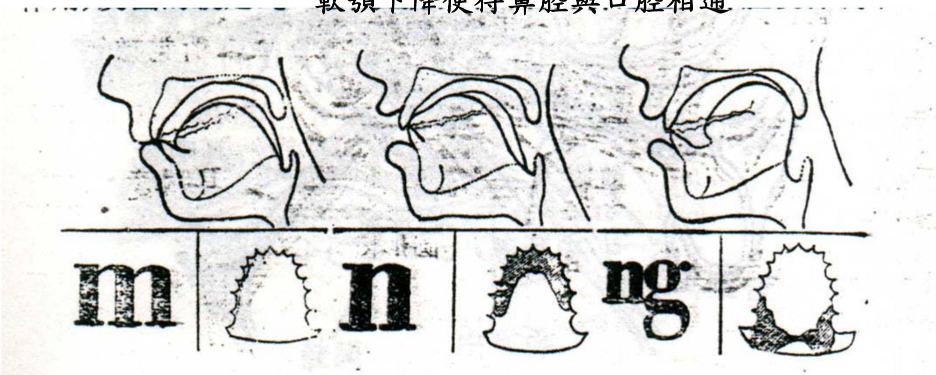


**图 4.11 普通话里 m、n、ng 的X光舌位图

Phonetic Typology (語音的類型)

- **Length**: Japanese vowels have a characteristic distinction of the length that can be hard for non-natives to perceive and use when learning the language
  - The word ***kado*** (corner) and ***kaado*** (card) are spectrally identical, differing in their durations
  - Length is phonemically distinctive for Japanese

- **Pitch:**
  - The primary dimension lacks in English
  - Many Asia and Africa language are tonal
    - E.g. Chinese
  - For tonal language, they have lexical meaning contrasts cued by pitch
    - E.g. Mandarin Chinese has four primary tones

Phonetic Typology (cont.)

- **Pitch**: (cont.)
  - Though English don’t make systematic use of pitch in its inventory of word contrasts, we always see with any possible phonetic effect:
    - Pitch is systematically viewed in English to signal a speaker’s emotions, intentions and attitudes
    - Pitch has some linguistic function in signaling grammatical structure as well

Table 2.11 The contrastive tones of Mandarin Chinese.

Tone	Shape	Example	Chinese	Meaning
1	High level	*ma*	妈	mother
2	High rising	*ma*	麻	numb
3	Low rising	*ma*	马	horse
4	High falling	*ma*	骂	to scold

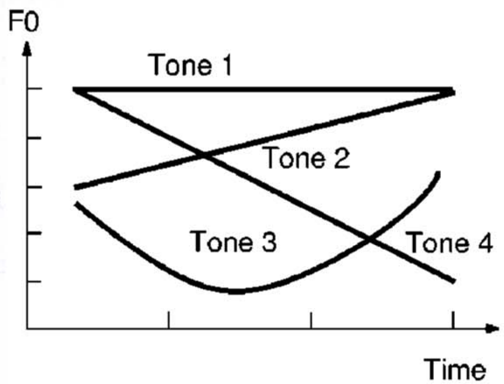
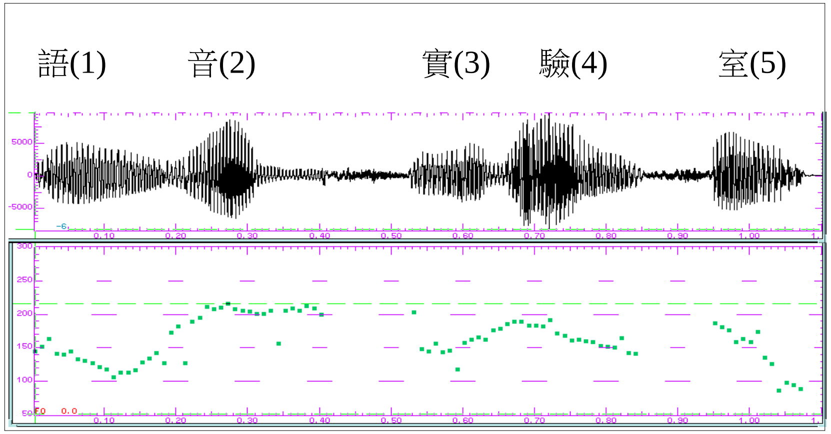


Figure 1: Pitch patterns of four lexical tones.

Phonetic Typology (cont.)



	Tone 1	Tone 2	Tone 3	Tone 4	neutral tone
number of models	4	6	6	4	3
typical tone concatenation combinations	1 1-(2) (3)-1 (3)-1-(2)	2 2-(2) (1)-2 (1)-2-(2) (3)-2 (3)-2-(2)	3 3-(1) (1)-3 (1)-3-(1) (3)-3 (3)-3-(1)	4 4-(1) (3)-4 (3)-4-(1)	5 (1)-5 (3)-5

The Allophone: Sound and Context

- Phonetic units should be correlated with potential meaning distinctions
  - *mean /m iy n/* and *men /m eh n/*
- However, the fundamental meaning-distinguishing sound is often modified in some systematic way by its phonetic neighbors
  - **Coarticulation**: the process by which the neighbor sounds influence one another
  - **Allophone:** when the variations resulting from coarticulatory processes can be consciously perceived, the modified phonemes are called allophones
  - E.g. :
    - *p* in (*pin*, */p ih n/*) produces a notice puff (噴出) of air, called aspiration (送氣), but loses its aspiration in (*spin*, */s p ih n/*)
    - A vowel before a voicing consonant, .e.g., *bad /d/*, seems typically longer than the same vowel before the unvoiced counterpart, in this case *bat /t/*

The Allophone: Sound and Context (cont.)

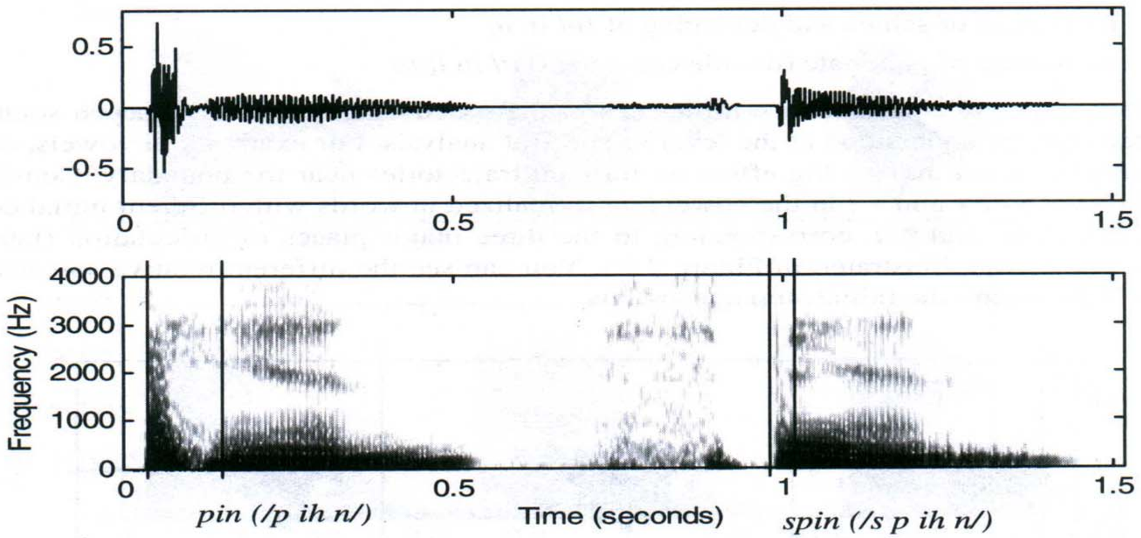


Figure 2.22 Spectrogram: bursts of *pin* and *spin*. The relative duration of a *p*-burst in different phonetic contexts is shown by the differing width of the area between the vertical lines.

Structural Features of Chinese Language

- Not Alphabetic (字母的)
- At Least 10,000 Commonly Used Characters (字)
  - Almost all morphemes (詞素) with their own meaning
  - All monosyllabic
- Unlimited Number of Words (詞) , at Least 100,000 Commonly Used , Each Composed of One to Several Characters (字)
  - The meaning of the word can be directly or partly related, or even completely irrelevant to the meaning of the component characters

    書店，大學，和尚，光棍
- Chinese is a Tonal Language
  - 4 lexical tones, 1 neutral tone (the number is for Mandarin)

Adapted from Prof. Lin-shan Lee

Structural Features of Chinese Language (cont.)

- About 1,335 Syllables Only (the number is for Mandarin)
  - About 408 base-syllables if differences in tone disregarded (the number is for Mandarin)
- Large Number of Homonym Characters (同音字) Sharing the Same Syllable
- Monosyllabic Structure of Chinese Language
  - Each syllable stands for many characters with different meaning
  - Combination of syllables (characters) gives unlimited number of words
  - Small number of syllables carries plurality (多重性) of linguistic information
- Almost Each Character with Its Own Meaning, thus Playing Some Linguistic Role Independently

Adapted from Prof. Lin-shan Lee

Structural Features of Chinese Language (cont.)

- No Natural Word Boundaries in a Chinese Sentence

  電腦科技的進步改變了人類的生活和工作方式

  - Word segmentation not unique
  - Words not well defined
  - Commonly accepted lexicon not existing
- Open Vocabulary Nature with Flexible Wording Structure
  - New words easily created everyday

    電 (electricity) + 腦 (brain)→電腦 (computer)
  - Long word arbitrarily abbreviated

    臺灣大學 (Taiwan University) →臺大
  - Name/title

    李登輝總統 (President T.H. Lee) →李總統登輝
  - Unlimited number of compound words

    高 (high) + 速 (speed) + 公路 (highway) →高速公路(freeway)

Adapted from Prof. Lin-shan Lee

Structural Features of Chinese Language (cont.)

- Difficult for Word-based Approaches Popularly Used in Alphabetic Languages
  - Serious out of vocabulary (OOV) problem

- Considering Phonetic Structure of Mandarin Syllables
  - INITIAL / FINAL's
  - Phone-like-units / phonemes

- Different Degrees of Context Dependency
  - intra-syllable only
  - intra-syllable plus inter-syllable
  - right context dependent only
  - both right and left context dependent

Structural Features of Chinese Language (cont.)

- Examples
  - 22 INITIAL’s extended to 113 right-context-dependent INITIAL’s
  - 33 phone-like-units extended to 145 intra-syllable right-context-dependent phone-like-units, or 481 with both intra/inter-syllable context dependency
  - 4,606 triphones with intra/inter-syllable context dependency

<table>
<tr><td colspan="5">Syllables (1,345)</td></tr>
<tr><td colspan="4">Base-syllables (408)</td><td rowspan="5">Tones
(4+1)</td></tr>
<tr><td rowspan="2">INITIAL’s
(21)</td><td colspan="3">FINAL’s (37)</td></tr>
<tr><td>Medials
(3)</td><td>Nucleus
(9)</td><td>Ending
(2)</td></tr>
<tr><td>Consonants
(21)</td><td colspan="3">Vowels plus Nasals
(12)</td></tr>
<tr><td colspan="4">Phones (31)</td></tr>
</table>