语音学基础、发音与听觉感知

Phonetics, Speech Production and Perception

清华大学深圳研究生院

吴志勇

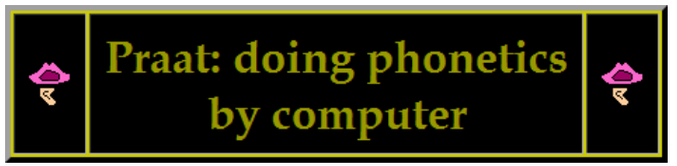
zywu@sz.tsinghua.edu.cn

MEDIA JOINT Research Center

学习要点

- 了解语音学的基本概念、术语
- 理解语音的产生过程
- 明确语音的感知机理

- 学习语音分析工具**Praat**的使用，利用**Praat**提取、理解、修改基频、语谱等声学特征参数，并生成修改后的语音



http://www.fon.hum.uva.nl/praat/

语音三要素

源滤波器模型

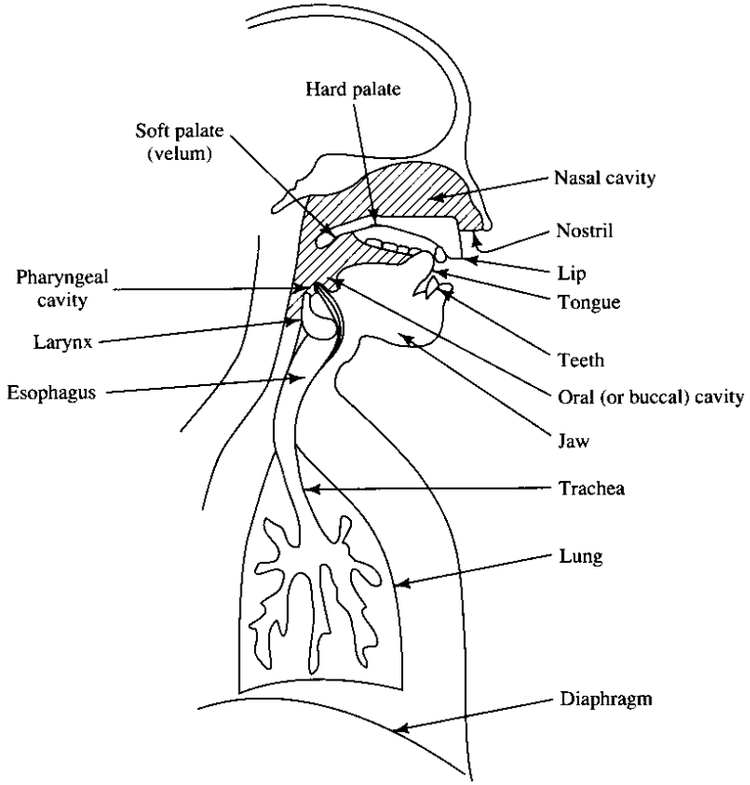
清音与浊音

语音的生成

SPEECH PRODUCTION

Speech Production: 语音产生

- **Speech Production**
  - Speech is produced by air-pressure waves emanating from the mouth and nostrils of a speaker
- **Vocal Tract: 声道**
  - The vocal tract is comprised of the organs and structures involved in the production of speech. These include the lungs, the trachea (also called the windpipe), the larynx, the oral cavity and the nasal cavity.
- **Articulators: 发音器官**
  - 肺 (lungs)
  - 气管 (trachea, windpipe)
  - 声带 (vocal cords, vocal folds, larynx)
  - 软腭 (soft palate, velum)
  - 硬腭 (hard palate)
  - 舌 (tongue)
  - 牙齿 (teeth)
  - 唇 (lips)
  - 鼻 (nostril)
  - 咽腔 (pharyngeal cavity, pharynx cavity)
  - 口腔 (oral cavity, buccal cavity)
  - 鼻腔 (nasal cavity)



(from: http://imp.lss.wisc.edu/~jrvalent/AIS/Grammar/Phonology/Phonol002a.html)

Speech Production: 语音产生

- **Functions of the Articulators:**
  各发音器官的功能
  - **肺** (lungs): 发音气流源头
  - **声带** (vocal cords, vocal folds, larynx): 受气流影响相互靠近收紧，发生振动，产生浊音 (voiced)；或者声带松弛使声门 (glottis) 开放，产生清音 (voiceless / unvoiced)
  - **软腭** (soft palate, velum): 具有阀门功能，打开时允许气流进入鼻腔 (nasal cavity)，关闭时禁止气流进入鼻腔
  - **硬腭** (hard palate): 口腔 (oral cavity) 顶部较长的硬表面。舌顶住硬腭时，发辅音 (consonant)
  - **舌** (tongue): 灵活的发音器官，远离硬腭发元音 (vowel) ; 靠近或接触硬腭或其他硬表面发辅音
  - **牙齿** (teeth): 发某些辅音时，用来顶住舌
  - **唇** (lips): 变圆或扁影响发元音的质量，或者完全紧闭，阻止气流从口腔发出

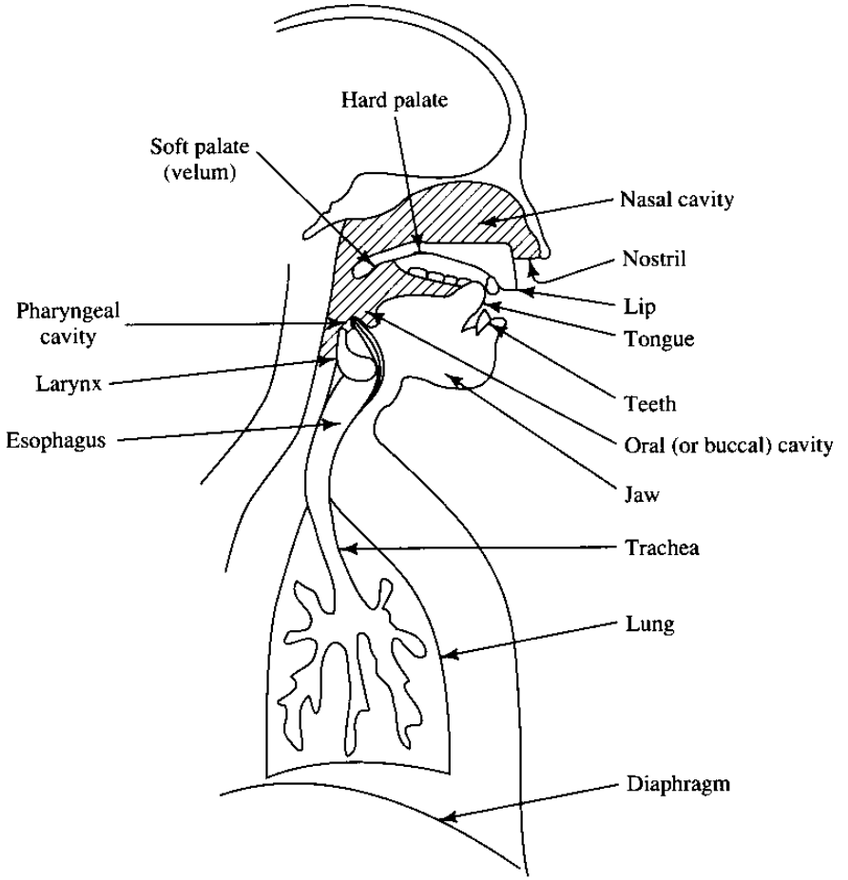
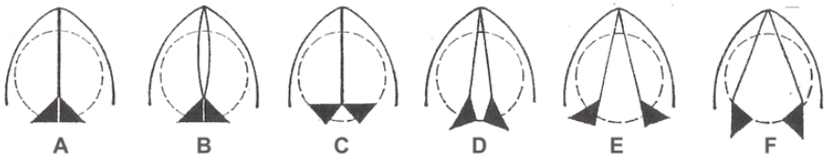


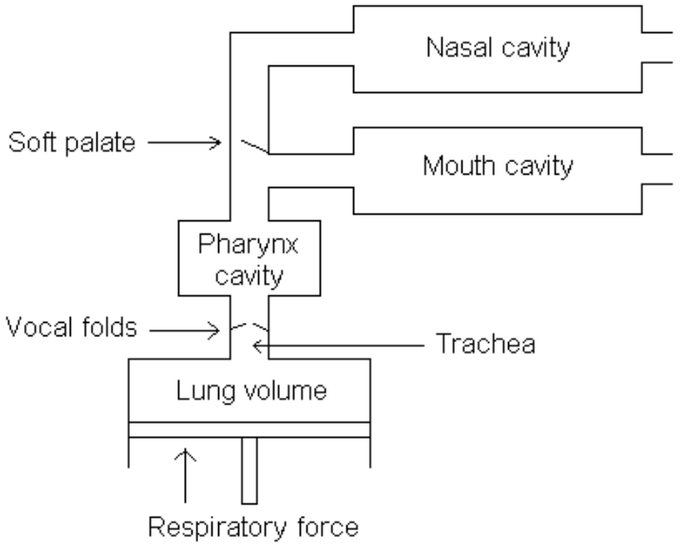
Diagram of the articulators (speech organs)

(from: http://imp.lss.wisc.edu/~jrvalent/AIS/Grammar/Phonology/Phonol002a.html)

Overview of Speech Generation

- **Overview of Speech Generation**
  - **Respiration** – Lungs provide the energy source
  - **Phonation** – Vocal folds convert the energy into audible sound
  - **Articulation** – Articulators transform the sound into intelligible speech





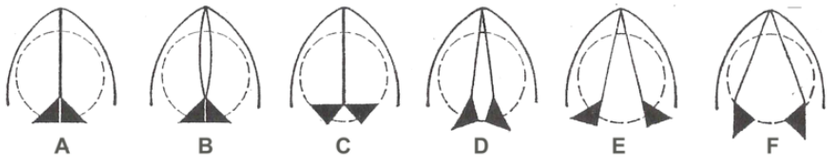
A simplified diagram of the vocal tract. Above the vocal folds are the various cavities that can be modified in size and shape to cause changes in the sound quality of the speech sounds.

- **由于声门 (glottis) 的肌肉张力，加上由肺部压迫出来的空气，就会造成声门的快速打开与关闭，这一疏一密的空气压力，即为语音源头，再经过声道、口腔、鼻腔的共振，就会产生不同声音。**
  - 声门振动的快慢，决定语音的基本频率（即音高）
  - 口腔、鼻腔、舌的位置、嘴型等，决定语音的内容（及音色）
  - 肺部压缩空气力量的大小，决定语音的音量（即音强）

(from: http://clas.mq.edu.au/phonetics/phonetics/introduction/index.html)

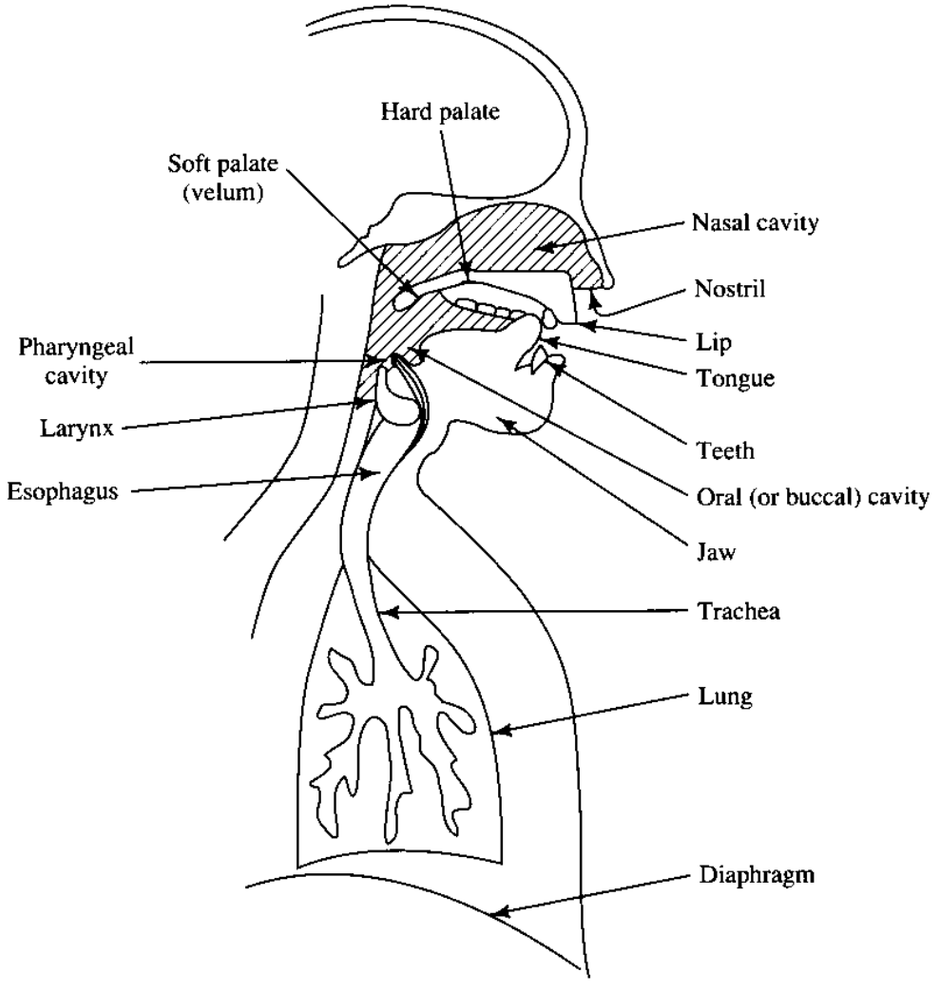
Overview of Speech Generation

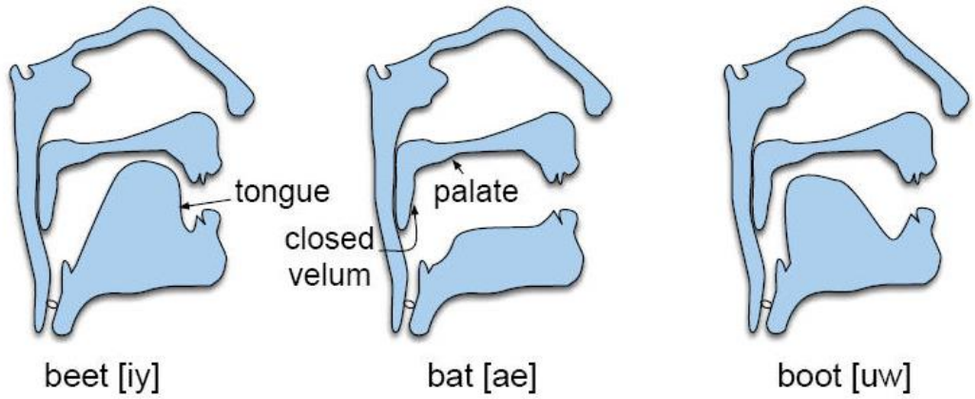
- 声带的振动

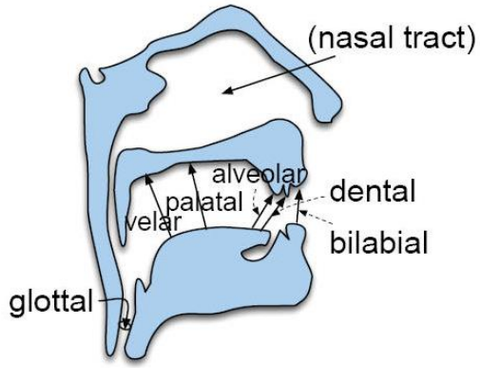


Overview of Speech Generation

- 声道的调制

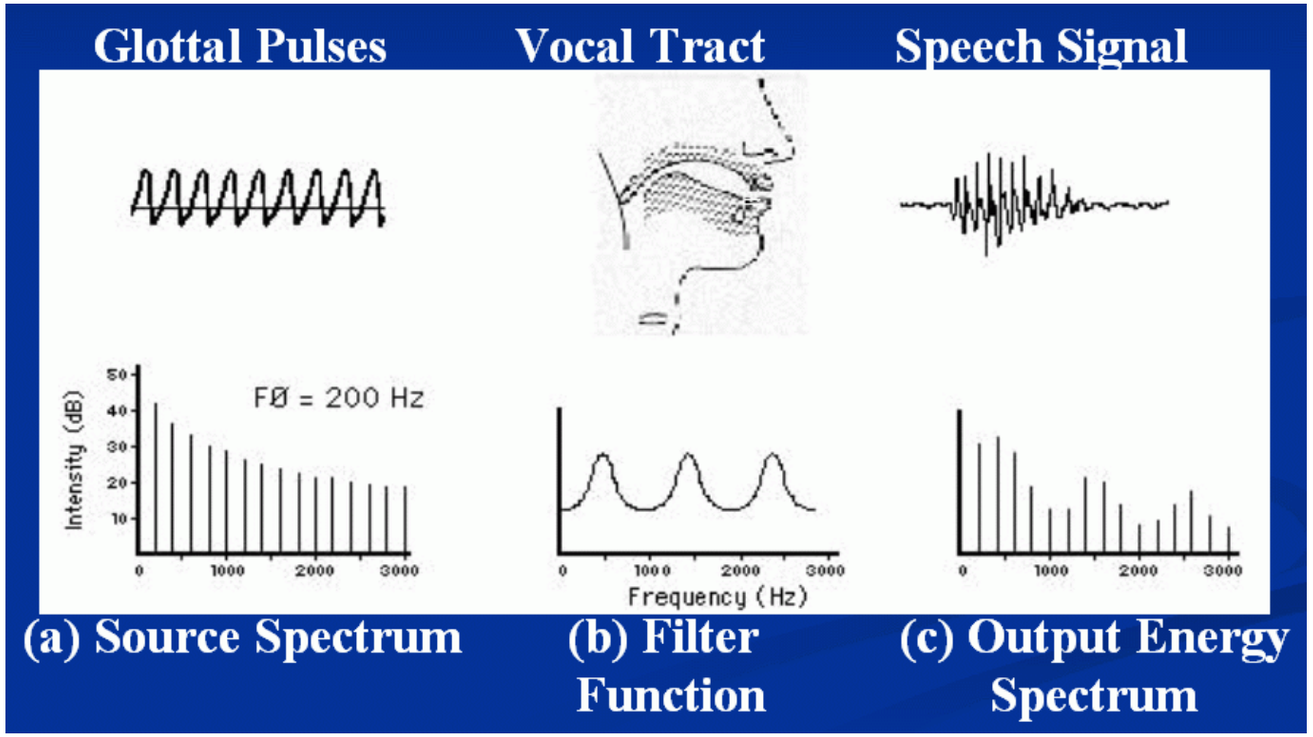






Speech Generation Model

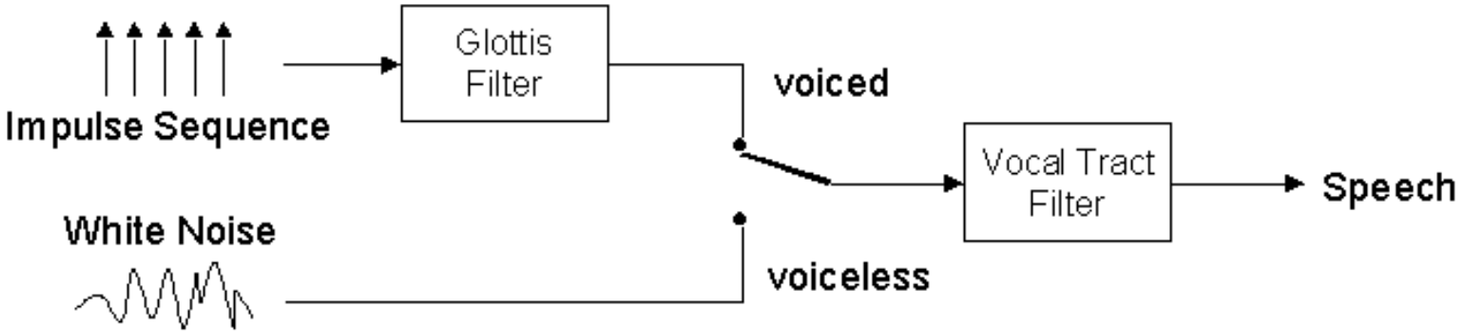
- **Source-Filter Model: 源-滤波器模型**
  - 语音的产生是由信号源（声门）的震动，经过滤波器（口腔、鼻腔、嘴型等）的调制而产生的



Source-filter model and the corresponding spectra

Speech Generation Model

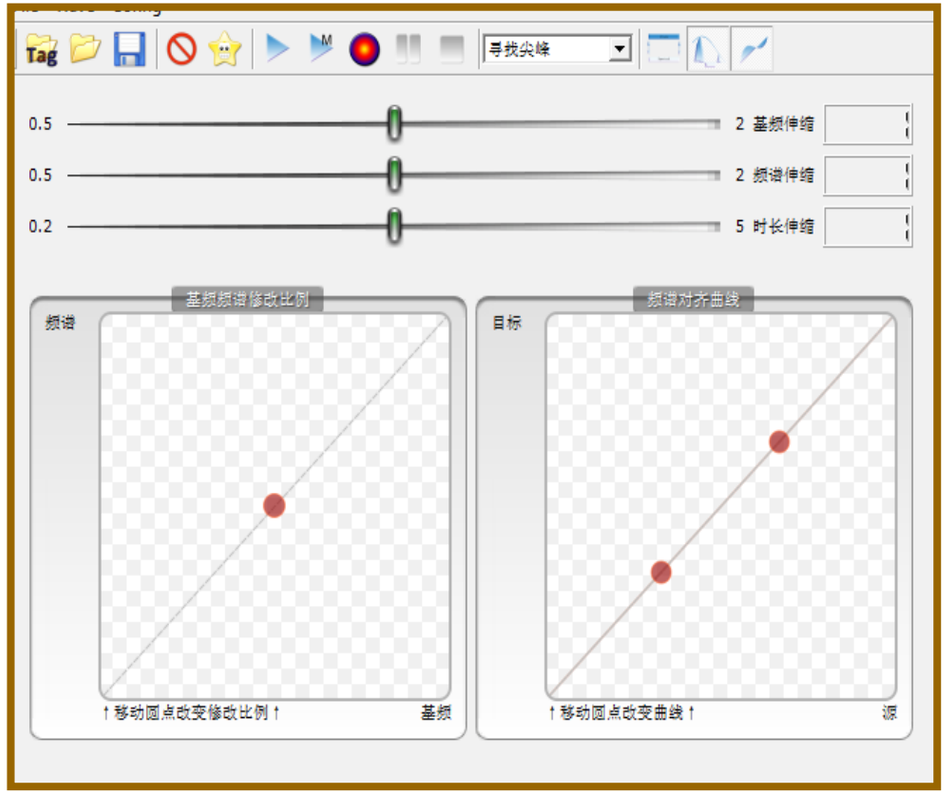
- **Source-Filter Model: 源-滤波器模型**
  - 语音的产生是由信号源（声门）的震动，经过滤波器（口腔、鼻腔、嘴型等）的调制而产生的



Block diagram representation of source-filter model. The vocal-tract filter is time-variant. A simplified model integrates the glottis filter into the vocal tract filter.

Acoustic Features of Speech

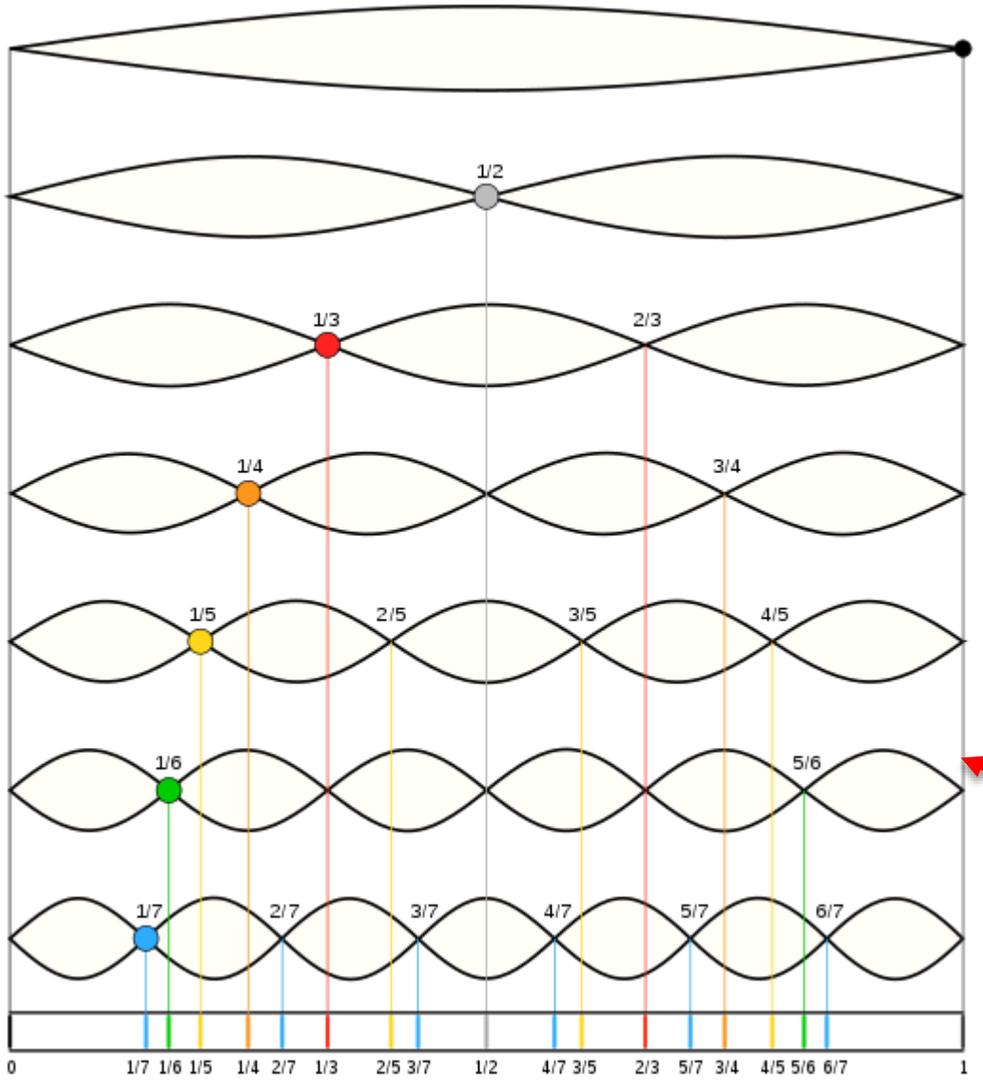
- **音高 Pitch (音调)**
  - 声音的高低。
    由声源振动频率(Frequency)
    决定。单位：赫兹Hz

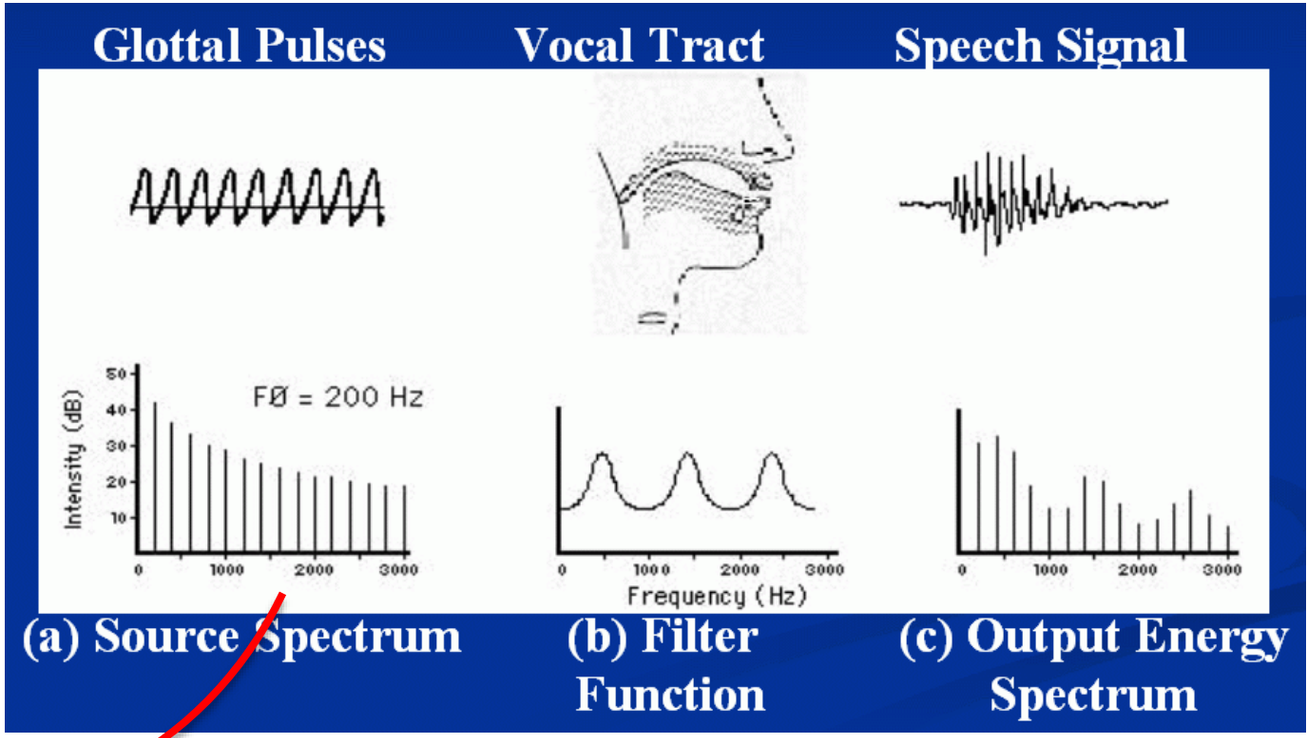


- **音色 Timbre**
  - 声音的特性。
    由声源、声道确定。

- **音强 Loudness (响度)**
  - 人主观的音量大小。
    由声音振幅(Amplitude)及人离声源的距离决定。单位：分贝dB

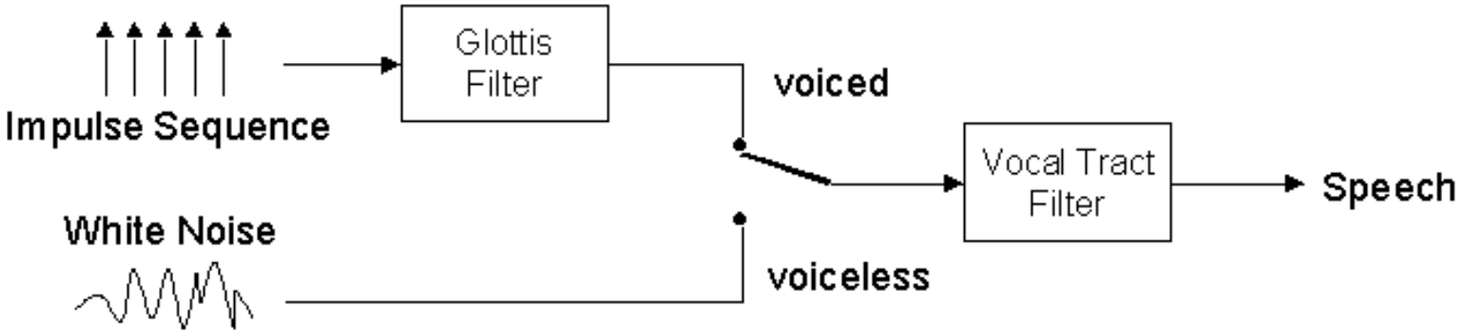
Harmonic（谐波）


1/2
1/3
2/3
1/4
3/4
1/5
2/5
3/5
4/5
1/6
5/6
1/7
2/7
3/7
4/7
5/7
6/7
0
1/7 1/6 1/5 1/4 2/7 1/3 2/5 3/7 1/2 4/7 3/5 2/3 5/7 3/4 4/5 5/6 6/7
1


Glottal Pulses
Vocal Tract
Speech Signal
FØ = 200 Hz
Intensity (dB)
50
40
30
20
10
0
1000
2000
3000
Frequency (Hz)
(a) Source Spectrum
(b) Filter Function
(c) Output Energy Spectrum

Voiced / Unvoiced Speech

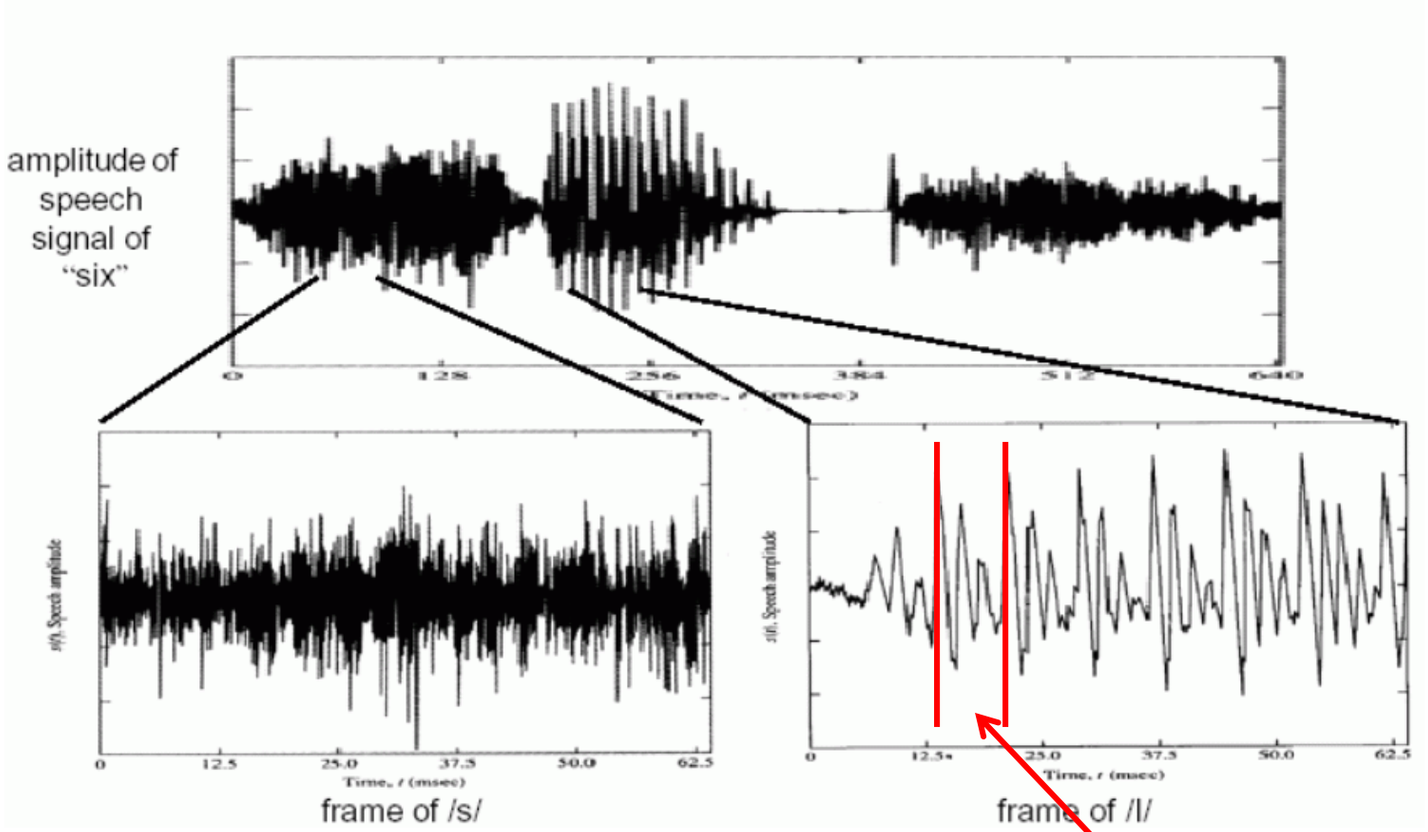
- **Source-Filter Model: 源-滤波器模型**
  - 语音的产生是由信号源（声门）的震动，经过滤波器（口腔、鼻腔、嘴型等）的调制而产生的



Block diagram representation of source-filter model. The vocal-tract filter is time-variant. A simplified model integrates the glottis filter into the vocal tract filter.

- **Voiced Speech: 浊音**
  - 声带振动引起，语音波形具有明显周期性，声带振动的频率称为基音频率或基频 (fundamental frequency, F_0)，人们可感受到稳定的音高存在
- **Voiceless / Unvoiced Speech: 清音**
  - 声带不振动，波形类似白噪声，人们无法感受到稳定的音高存在

Voiced / Unvoiced Speech



- **清音Unvoiced Speech: /s/, /k/**
- **浊音Voiced Speech: /i/**

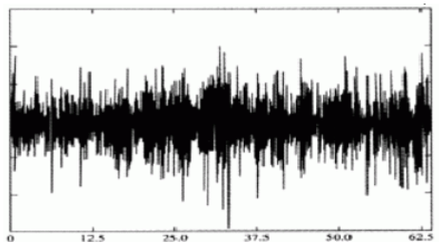
浊音的波形呈现周期性
声带振动的频率：基频 F_0

Hints:
• Put your hand on your throat, you can feel the vibration of the glottis.

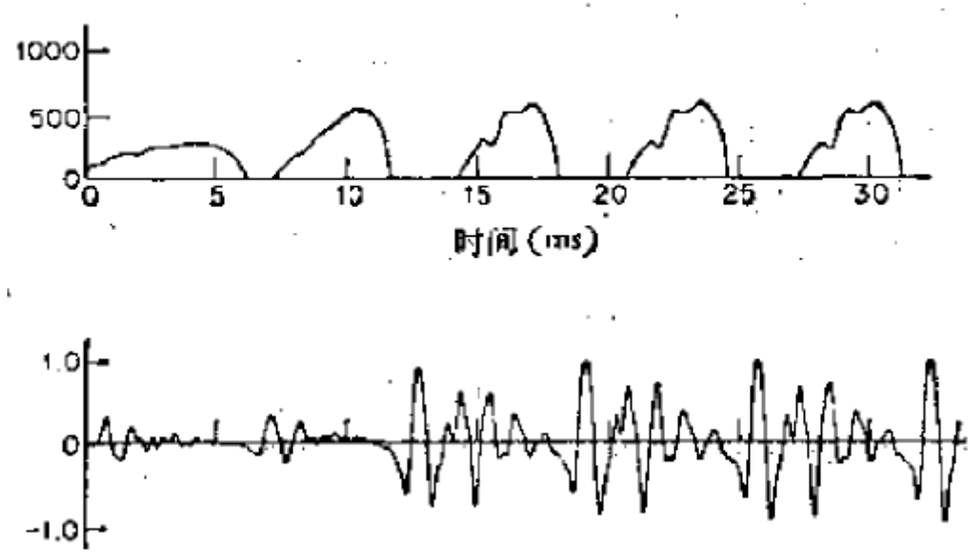
Visual Speech

Voiced / Unvoiced Speech

- 清音 Unvoiced Speech
  - 激励信号是一个随机噪声信号



- 浊音 Voiced Speech
  - 激励信号是一周期为 T_0（基音周期）的斜三角形的脉冲串



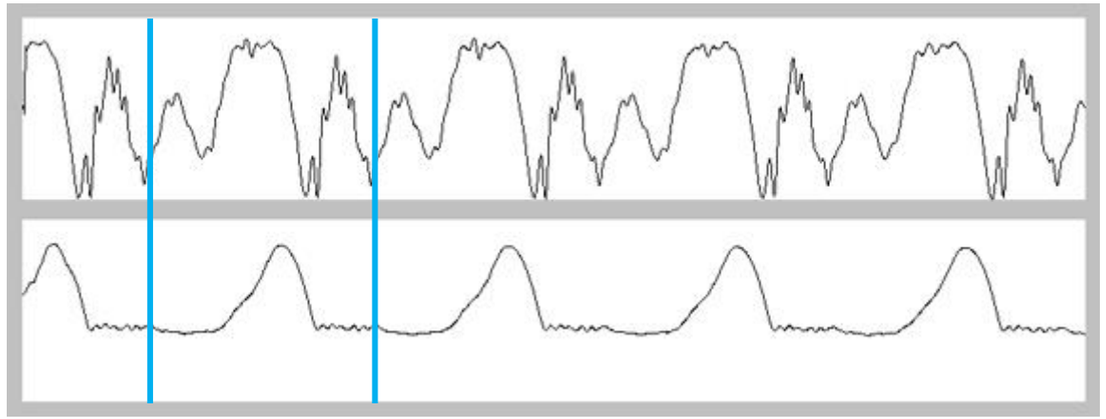
元音 /a/ 的声门波（上）与语音波形（下）

(from: L. R. Rabiner, Digital Processing of Speech Signals, 科学出版社)

Voiced / Unvoiced Speech

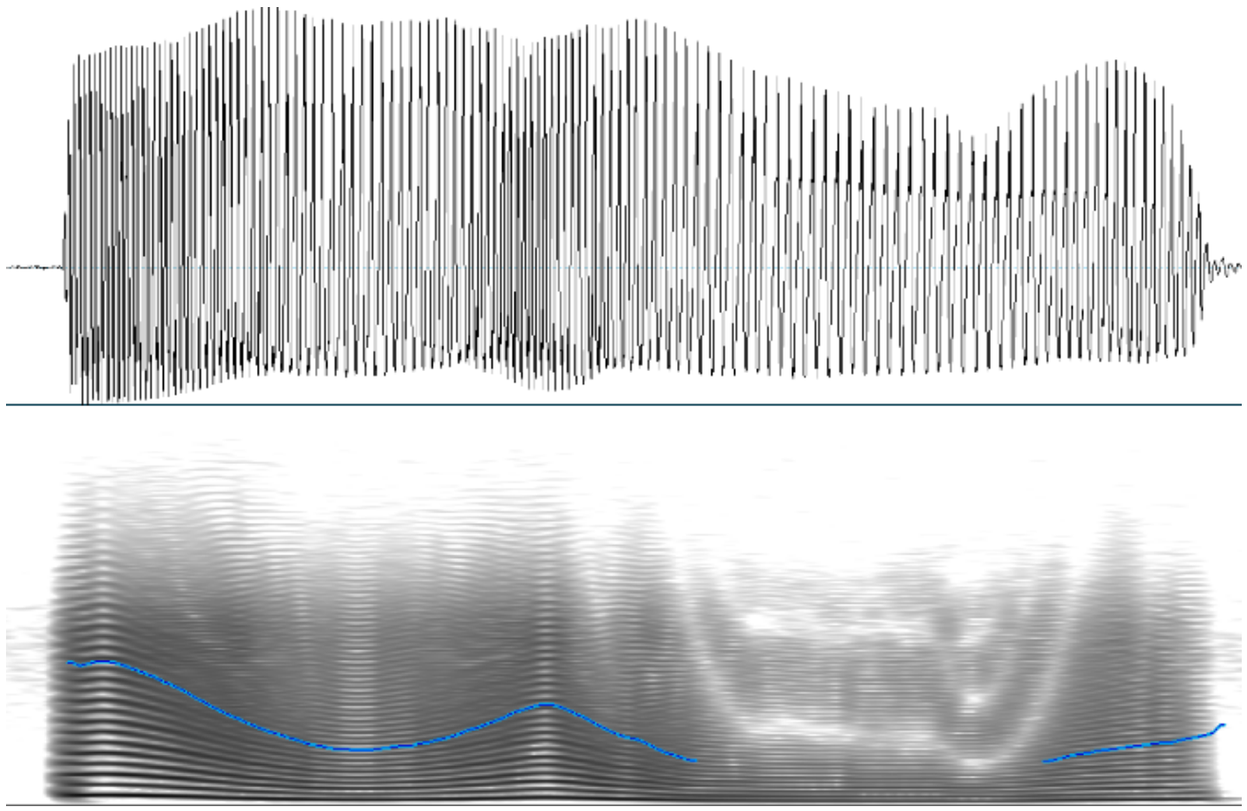
- **Glottal Waveform / EGG Signal: 声门波**

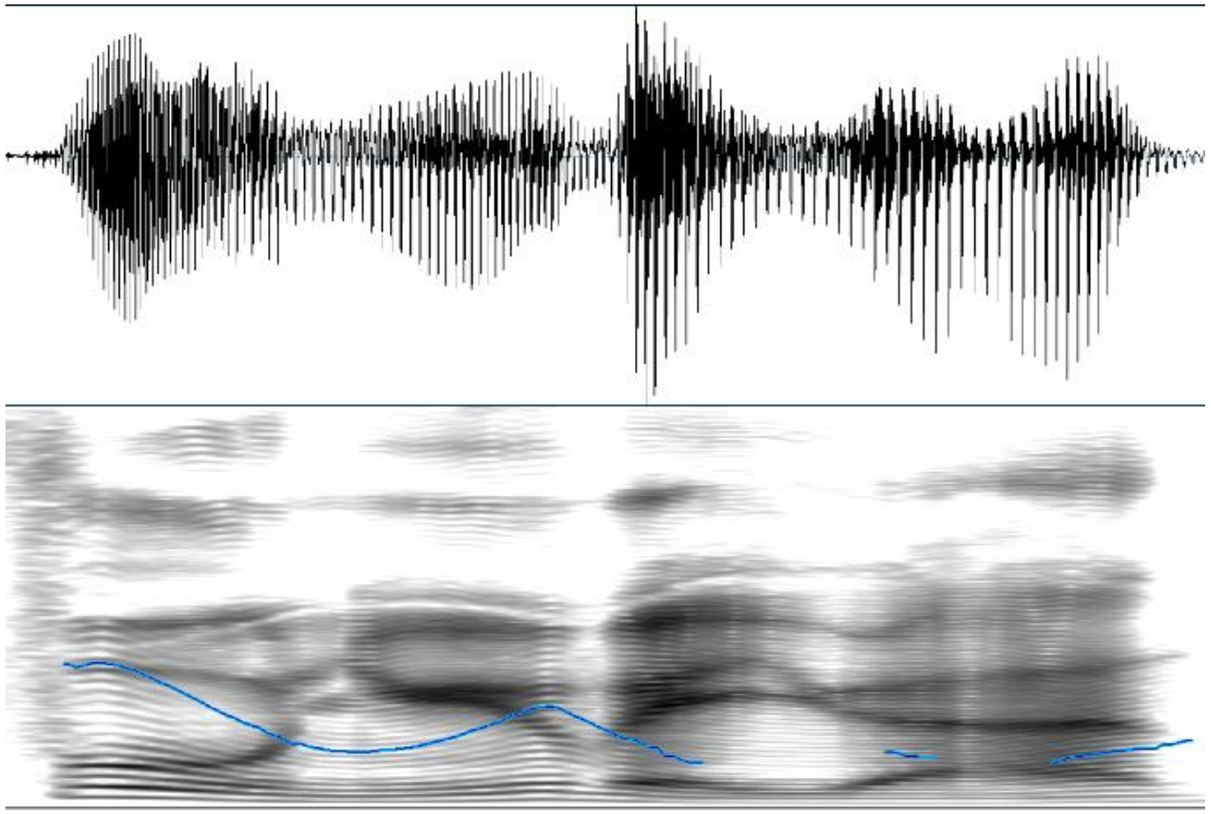
Electroglottograph (EGG)



Praat

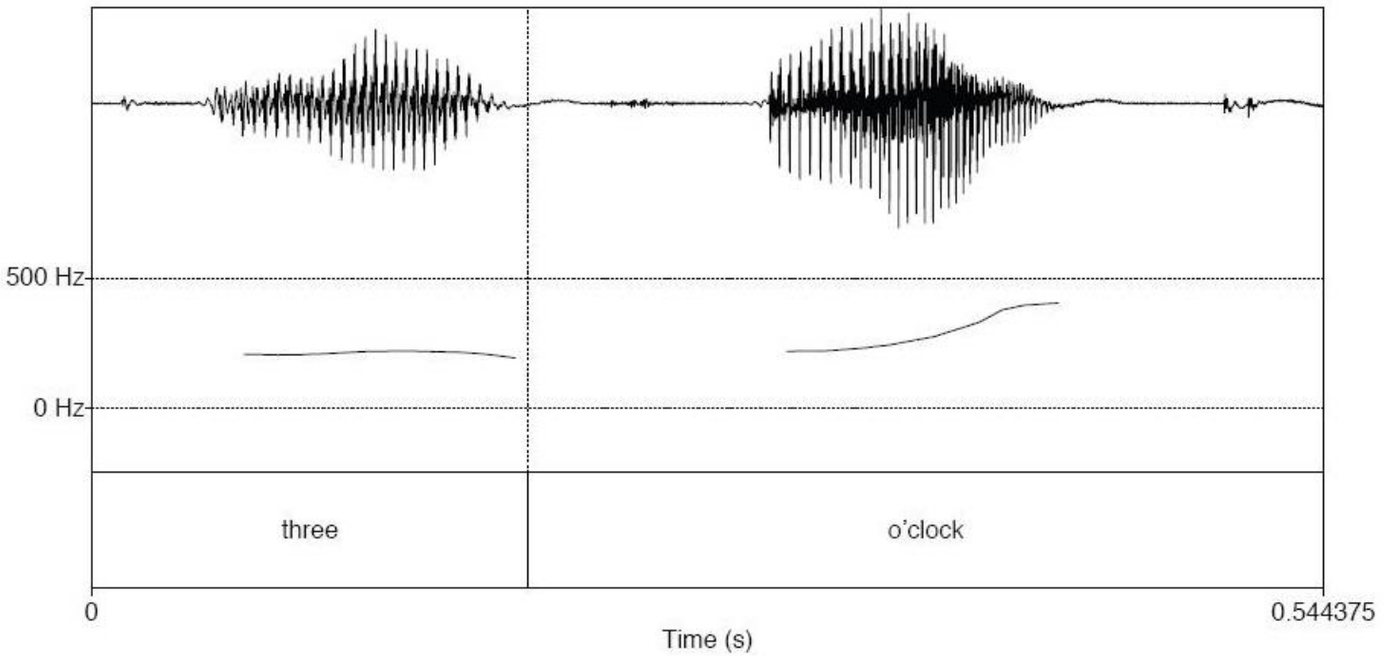
EGG Signal vs. Waveform





F0 Related Features

- **F_0: 基频 / 基音频率**
  - F_0 is strictly on the acoustic, measurable, production side.

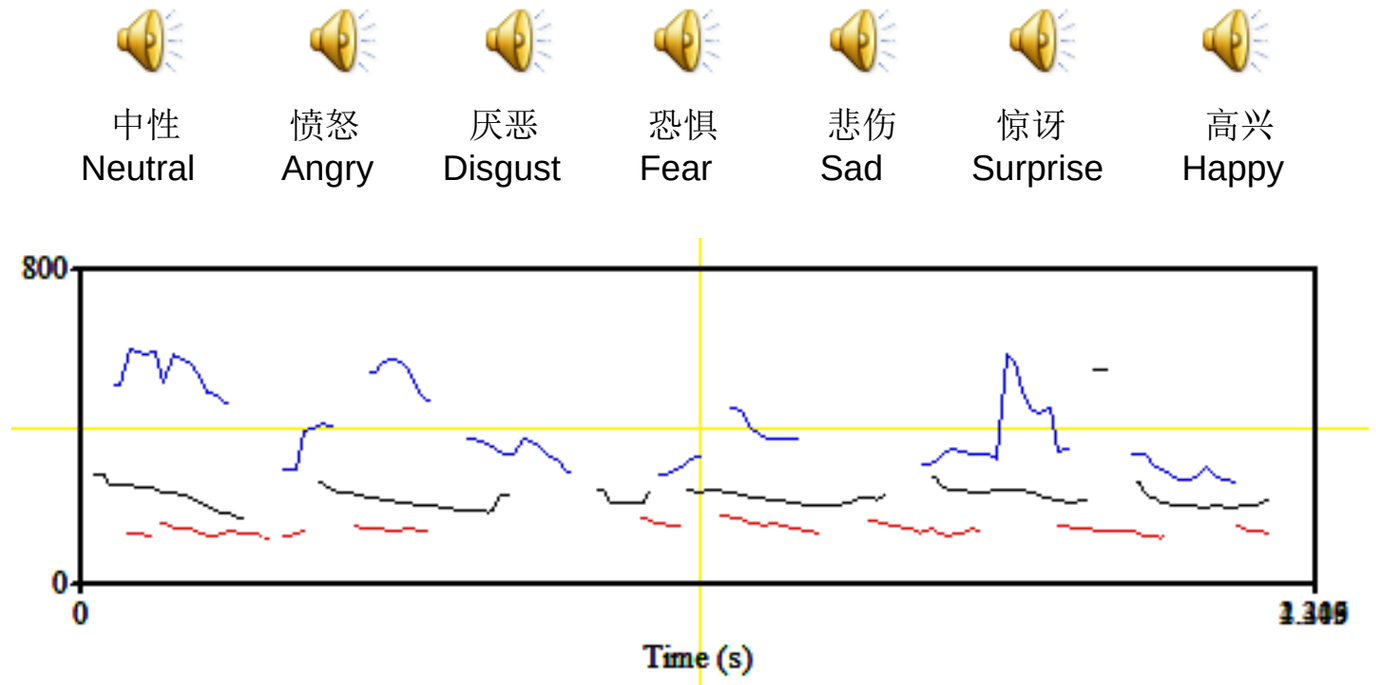


- **Pitch: 音高**
  - Pitch refers to the *impression* created in the hearer on the auditory side.
  - Pitch represents the perceived fundamental frequency of a sound.
  - All languages use pitch to express *emotional* and other *paralinguistic* information, and to convey *emphasis*, contrast, and other such features in what is called *intonation*.
  - Pitch is one of the three important features related to ***speech prosody***.

F0 Related Features

- **Pitch: 音高**
  - Pitch is one of the three important features related to *speech prosody* (语音韵律).
  - All languages use pitch to express *emotional* and other *paralinguistic information* (副语言学信息), and to convey emphasis, contrast, and other such features in what is called *intonation* (语调).

Acoustic features related to speech prosody:
- pitch
- duration
- energy



Praat

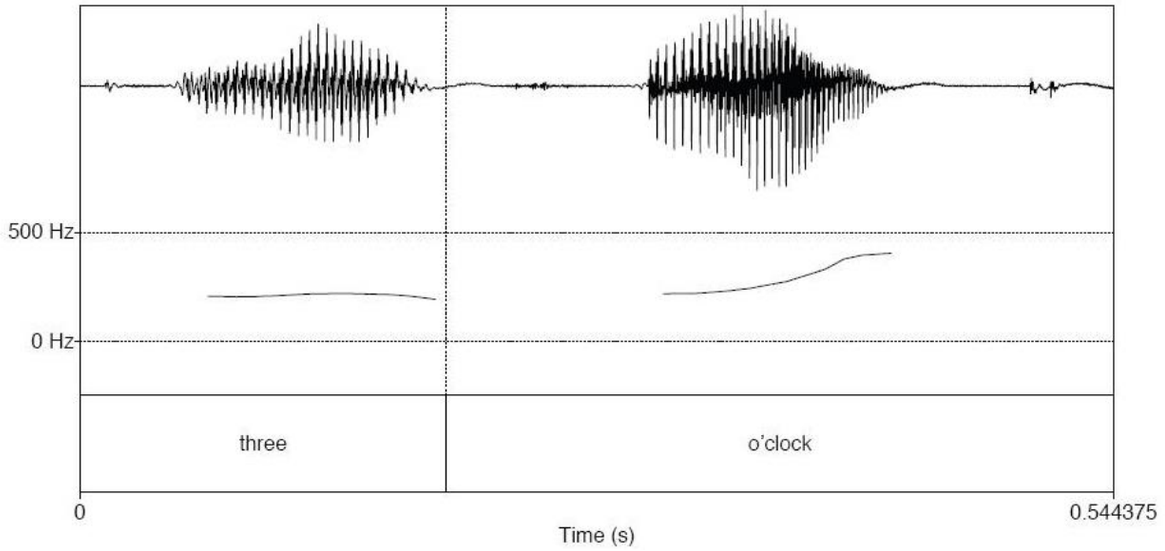
F0 Related Features

- **Pitch Range: 音高范围**
  - Different speakers have different ranges of fundamental frequency.

	F_0 avg	F_0 min	F_0 max
Men	125	80	200
Women	225	150	250
Children	300	200	500

The approximate pitch ranges and average values (Hz)

- **Pitch Contour: 音高曲线**
  - The pitch contour of a sound is a function or curve that tracks the perceived pitch of the sound over time.

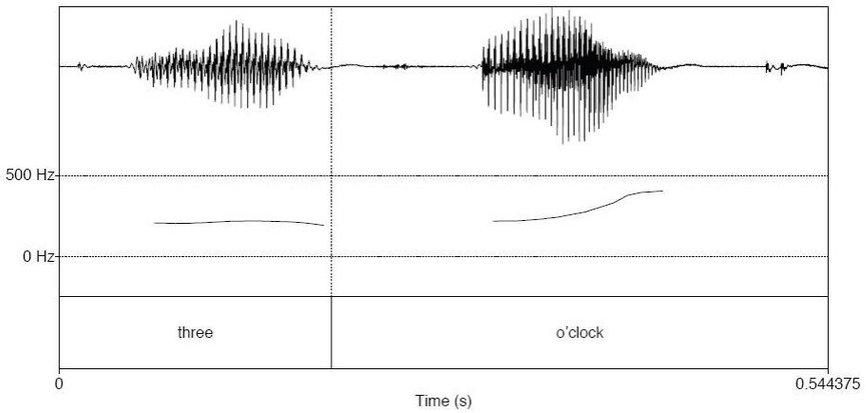


Pitch contour of the question "Three o'clock?", shown below the waveform. Note the rise in F_0 at the end of the question. Note the lack of pitch contour during the very quiet part (silence and unvoiced part).

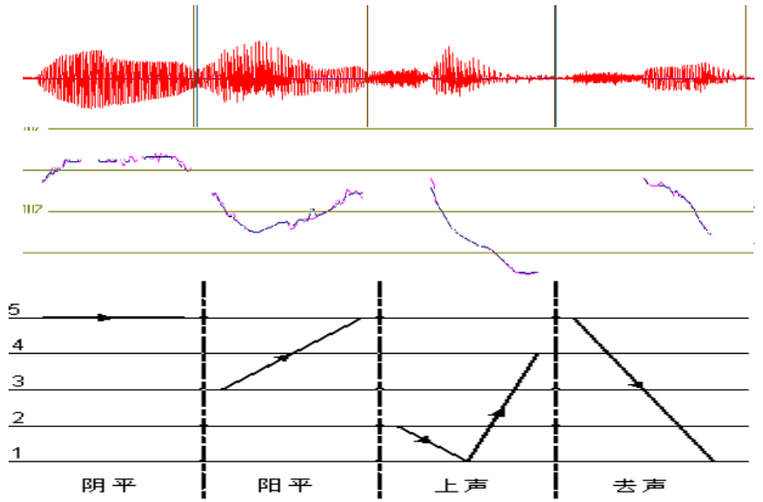
F0 Related Features

- **Tone: 声调**
  - Tone is the use of pitch in language to distinguish lexical or grammatical meaning.

- **Intonation: 语调**
  - Intonation is variation of pitch, which carries the information of prosody.



Pitch contour of the question “Three o’clock?”, shown below the waveform. Note the rise in F_0 at the end of the question. Note the lack of pitch contour during the very quiet part (silence and unvoiced part).

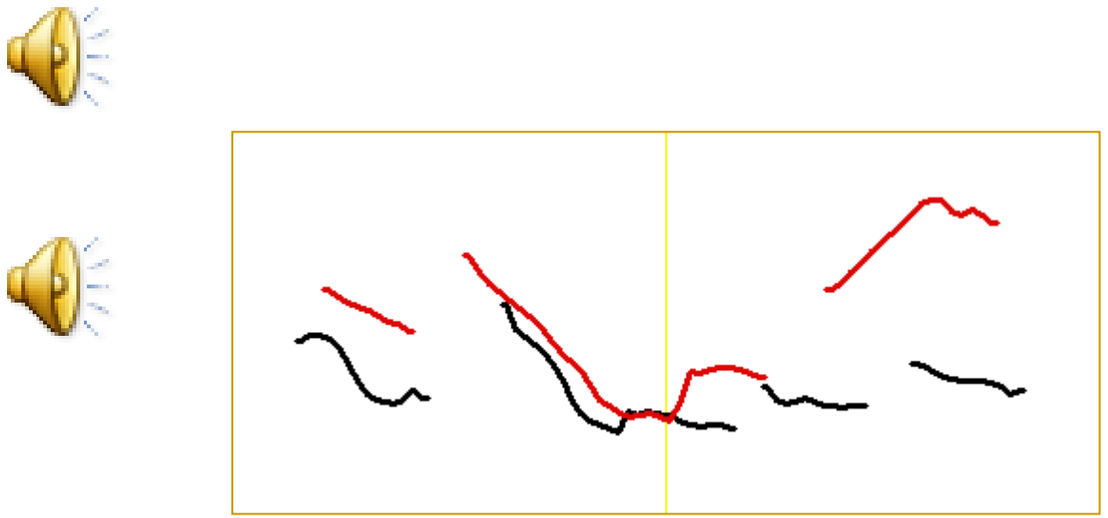


妈(ma1) 麻(ma2) 马(ma3) 骂(ma4)

汉语是有调语言，声调具有表意作用

F0 Related Features

- **Intonation: 语调**
  - Intonation is variation of pitch, which carries the information of prosody.



Praat

音位/音素
视位/视素
元音/辅音
音节

语音学基础
PHONETICS

Phonetics and Phonology

- **Phonetics: 语音学**
  - The study of speech sounds and their production, classification and transcription.
  - 语音学是研究言语声音(即语音)的学科。狭义的语音学对应英语中 phonetics 一词，关切的重点在具体语音本质以及产生语音的方法。
- **Phonology: 音韵学/音系学**
  - The study of the distribution and patterning of speech sounds in a language and of the tacit rules governing pronunciation.
  - 音韵学(音系学)研究音位或语音区别特征在某种语言中运作的抽象规则和语音的系统。
- **广义的语音学是指这两大方面研究的总合**
  - Articulatory phonetics: 发音语音学
    - 从生理角度，研究发音器官（如唇、齿、舌、声门等）如何彼此协调动作，以发出语音
  - Acoustic phonetics: 声学语音学
    - 从声学角度，研究语音物理现象，如语音频率、时长、振幅等
  - Auditory phonetics: 听觉语音学
    Psycholinguistics: 心理语言学
    - 研究语音的感知历程，大脑和人耳的处理等

语音学教程, 林焘, 王理嘉著,
北京大学出版社, 1992

Phoneme: 音位/音素

- **音位/音素：**不可分割的、最小的*音位学*单位
- A **phoneme** is the smallest posited linguistically distinctive unit of sound.
- Phonemes carry ***no semantic content*** themselves. In theoretical terms, phonemes are not the physical segments themselves, but cognitive abstractions or categorizations of them.

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[iy]	[i]	lily	[l ih l iy]
[ih]	[ɪ]	lily	[l ih l iy]
[ey]	[eɪ]	daisy	[d ey z iy]
[eh]	[ɛ]	pen	[p eh n]
[ae]	[æ]	aster	[ae s t axr]
[aa]	[ɑ]	poppy	[p aa p iy]
[ao]	[ɔ]	orchid	[ao r k ix d]
[uh]	[ʊ]	wood	[w uh d]
[ow]	[oʊ]	lotus	[l ow dx ax s]
[uw]	[u]	tulip	[t uw l ix p]
[ah]	[ʌ]	buttercup	[b ah dx axr k ah p]
[er]	[ɝ]	bird	[b er d]
[ay]	[aɪ]	iris	[ay r ix s]
[aw]	[aʊ]	sunflower	[s ah n f l aw axr]
[oy]	[oɪ]	soil	[s oy l]
Reduced and uncommon phones			
[ax]	[ə]	lotus	[l ow dx ax s]
[axr]	[ɚ]	heather	[h eh dh axr]
[ix]	[ɨ]	tulip	[t uw l ix p]
[ux]	[ʉ]	dude[1]	[d ux d]

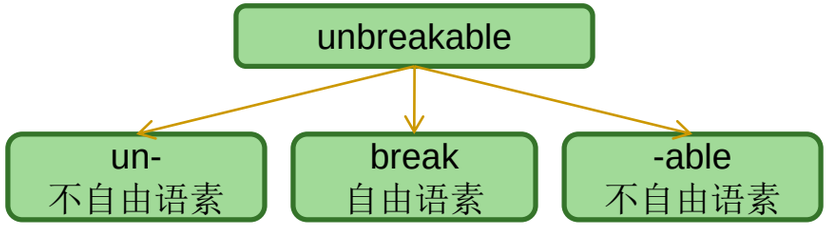
ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]
[t]	[t]	tea	[t iy]
[k]	[k]	cook	[k uh k]
[b]	[b]	bay	[b ey]
[d]	[d]	dill	[d ih l]
[g]	[g]	garlic	[g aa r l ix k]
[m]	[m]	mint	[m ih n t]
[n]	[n]	nutmeg	[n ah t m eh g]
[ng]	[ŋ]	baking	[b ey k ix ng]
[f]	[f]	flour	[f l aw axr]
[v]	[v]	clove	[k l ow v]
[th]	[θ]	thick	[th ih k]
[dh]	[ð]	those	[dh ow z]
[s]	[s]	soup	[s uw p]
[z]	[z]	eggs	[eh g z]
[sh]	[ʃ]	squash	[s k w aa sh]
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]
[ch]	[tʃ]	cherry	[ch eh r iy]
[jh]	[dʒ]	jar	[jh aa r]
[l]	[l]	licorice	[l ih k axr ix sh]
[w]	[w]	kiwi	[k iy w iy]
[r]	[r]	rice	[r ay s]
[y]	[j]	yellow	[y eh l ow]
[h]	[h]	honey	[h ah n iy]
Less commonly used phones and allophones			
[q]	[ʔ]	uh-oh	[q ah q ow]
[dx]	[ɾ]	butter	[b ah dx axr]
[nx]	[ɾ̃]	winner	[w ih nx axr]
[el]	[l̩]	table	[t ey b el]

Hints:
• ***ARPAbet***:
美式英语音素表，46个音素。

Morpheme: 语素

- **语素：**最小的、***具有语义***的结构单元，是最小的***语法单位***，是最小的***语音语义结合体***
- A **morpheme** is the smallest structural and linguistic unit with ***semantic meaning***.
- In spoken language, morphemes are composed of phonemes.
- In written language, morphemes are composed of graphemes (the smallest units of written language).

- **英语语素**



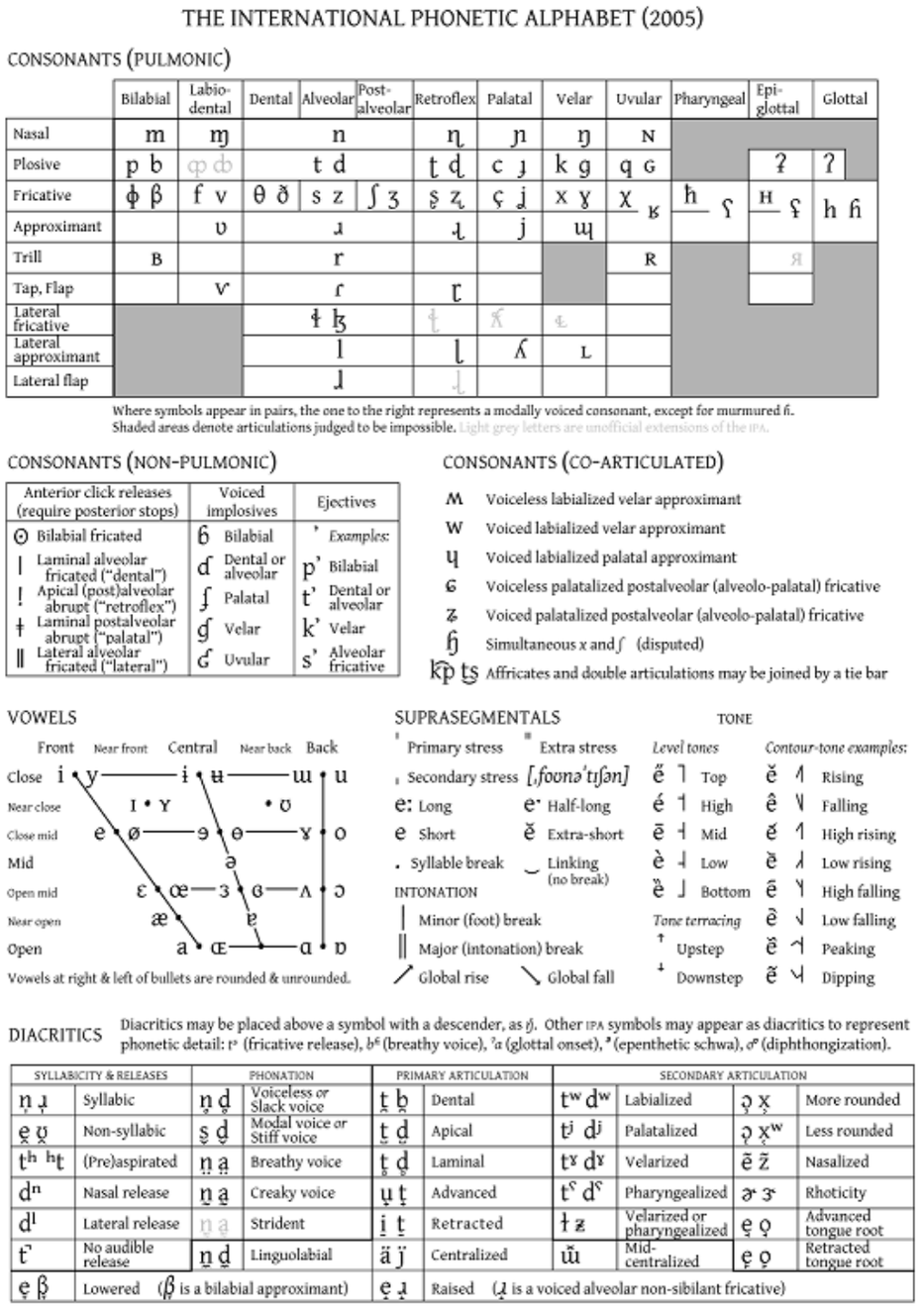
- **汉语语素**
  - 单音节语素
    - 天、地、人、中、去、大、了、吗……
  - 双音节语素
    - 萝卜、蜻蜓、蜘蛛、吩咐、徘徊、芙蓉……
  - 多音节语素
    - 巧克力、奥林匹克、罗曼蒂克、凡士林……
  - 非音节语素
    - 儿化音中，如“花儿、鸟儿”中的“儿”不是一个音位，因此是非音节语素

(from: wikipedia)

IPA: 国际音标

- **IPA**
  - The International Phonetic Alphabet (IPA) is a system of *phonetic notation* based on the Latin alphabet, devised by the International Phonetic Association as a standardized representation of the sounds of spoken language.
  - The IPA is used by foreign language students and teachers, linguists, speech pathologists and therapists, singers, actors, lexicographers, and translators.

(from: wikipedia)



THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ			
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʡ	ʔ
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	ʜ ʢ	h ɦ
Approximant		ʋ		ɹ		ɻ	j	ɰ				
Trill	ʙ			r					ʀ		ʡ	
Tap, Flap		ⱱ		ɾ		ɽ						
Lateral fricative				ɬ ɮ								
Lateral approximant				l		ɭ	ʎ	ʟ				
Lateral flap				ɺ								

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ɦ.
Shaded areas denote articulations judged to be impossible. Light grey letters are unofficial extensions of the IPA.

CONSONANTS (NON-PULMONIC)

Anterior click releases (require posterior stops)	Voiced implosives	Ejectives
ʘ Bilabial fricated	ɓ Bilabial	ʼ Examples:
ǀ Laminal alveolar fricated ("dental")	ɗ Dental or alveolar	pʼ Bilabial
ǃ Apical (post)alveolar abrupt ("retroflex")	ʄ Palatal	tʼ Dental or alveolar
ǂ Laminal postalveolar abrupt ("palatal")	ɠ Velar	kʼ Velar
ǁ Lateral alveolar fricated ("lateral")	ʛ Uvular	sʼ Alveolar fricative

CONSONANTS (CO-ARTICULATED)

ʍ Voiceless labialized velar approximant
w Voiced labialized velar approximant
ɥ Voiced labialized palatal approximant
ɕ Voiceless palatalized postalveolar (alveolo-palatal) fricative
ʑ Voiced palatalized postalveolar (alveolo-palatal) fricative
ʃ͡ʒ Simultaneous x and ʃ (disputed)
k͡p t͡s Affricates and double articulations may be joined by a tie bar

VOWELS

Front · Near front · Central · Near back · Back

Close: i y · ɨ ʉ · ɯ u
Near close: ɪ ʏ · ʊ
Close mid: e ø · ɘ ɵ · ɤ o
Mid: ə
Open mid: ɛ œ · ɜ ɞ · ʌ ɔ
Near open: æ · ɐ
Open: a ɶ · ɑ ɒ

Vowels at right & left of bullets are rounded & unrounded.

SUPRASEGMENTALS

ˈ Primary stress · ˌ Secondary stress [ˌfoʊnəˈtɪʃən]
eː Long · eˑ Half-long
e Short · ĕ Extra-short
. Syllable break · ‿ Linking (no break)

INTONATION

| Minor (foot) break
‖ Major (intonation) break
↗ Global rise · ↘ Global fall

TONE

Level tones		Contour-tone examples:	
e̋	˥ Top	ě	˩˥ Rising
é	˦ High	ê	˥˩ Falling
ē	˧ Mid	ě	˦˥ High rising
è	˨ Low	ȅ	˩˧ Low rising
ȅ	˩ Bottom	ȇ	˥˦ High falling
		ȇ	˧˩ Low falling
	Tone terracing	ȅ	˩˥˦ Peaking
	↑ Upstep	ȅ	˥˦˩ Dipping
	↓ Downstep		

DIACRITICS Diacritics may be placed above a symbol with a descender, as ŋ̊. Other IPA symbols may appear as diacritics to represent phonetic detail: tˢ (fricative release), bᵉ (breathy voice), ʔᵃ (glottal onset), ᵊ (epenthetic schwa), oᵘ (diphthongization).

SYLLABICITY & RELEASES		PHONATION		PRIMARY ARTICULATION		SECONDARY ARTICULATION			
n̩ ɹ̩	Syllable	n̥ d̥	Voiceless or Slack voice	t̪ b̪	Dental	tʷ dʷ	Labialized	ɔ̹ x̹	More rounded
e̯ u̯	Non-syllabic	s̬ d̬	Modal voice or Stiff voice	t̺ d̺	Apical	tʲ dʲ	Palatalized	ɔ̜ xʷ	Less rounded
tʰ ʰt	(Pre)aspirated	n̤ a̤	Breathy voice	t̻ d̻	Laminal	tˠ dˠ	Velarized	ẽ z̃	Nasalized
dⁿ	Nasal release	n̰ a̰	Creaky voice	u̟ t̟	Advanced	tˤ dˤ	Pharyngealized	ɚ ɝ	Rhoticity
dˡ	Lateral release	n̼ a̼	Strident	i̠ t̠	Retracted	ɬ̪ ʒ	Velarized or pharyngealized	e̘ o̘	Advanced tongue root
t̚	No audible release	n̪ d̪	Linguolabial	ä ï	Centralized	ɯ̽	Mid-centralized	e̙ o̙	Retracted tongue root
e̞ β̞	Lowered (β̞ is a bilabial approximant)			e̝ ɹ̝	Raised (ɹ̝ is a voiced alveolar non-sibilant fricative)				

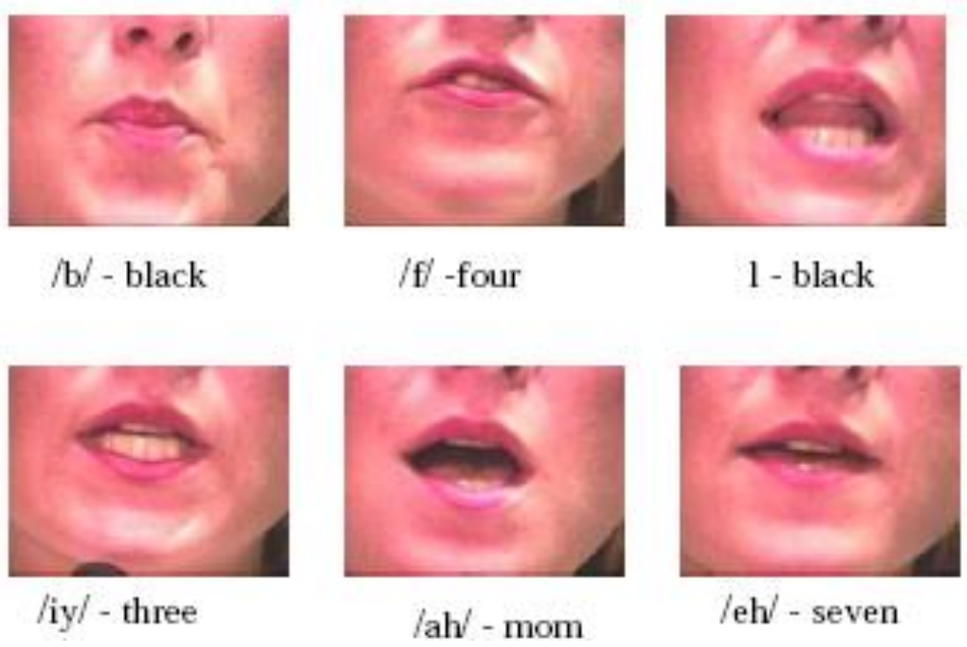
Viseme: 视位/视素

- **Viseme: visual phoneme**
  - A **viseme** is a representational unit used to classify speech sounds in the visual domain, corresponding to the phoneme in the aural domain.
  - 语音在视觉域的最小单元，同听觉域的音素相关
  - A **viseme** describes the particular facial and oral positions and movements that occur alongside the voicing of phonemes.
  - 描述了在发某个特定音素时对应的可视发音器官的形状、位置、动作。

Chen, T. (2001). *Audiovisual speech processing.* IEEE Signal Processing Magazine, 9–31.

Viseme: 视位/视素

- **Viseme: visual phoneme**
  - Phonemes and visemes do not always share a one-to-one correspondence.
    - Several phonemes share the same viseme: look the same on the face when produced
      - Such as /k/, /g/, /ŋ/, (viseme: /k/), or /tʃ/, /ʃ/, /dʒ/, /ʒ/ (viseme: /ch/)
    - There could *be differences in timing and duration* during actual speech
  - Some sounds which are hard to distinguish acoustically are clearly distinguished by the face (Chen 2001).
    - For example, acoustically speaking English /l/ and /r/ could be quite similar (especially in clusters, such as 'grass' vs. 'glass'). Yet visual information can show a clear contrast.
    - This is demonstrated by the more frequent mishearing of words on the telephone than in person (e.g., /d/ for 'dog', /b/ for 'boy').



Chen, T. (2001). *Audiovisual speech processing.* IEEE Signal Processing Magazine, 9–31.

Viseme: 视位/视素

- **Bimodal Processing:** 双模态处理
  - Speech is best understood as bimodal (aural and visual)
  - Comprehension can be compromised if one of these two domains is absent (McGurk and MacDonald 1976).
- **Lip Reading:** 唇读
  - The comprehension of speech by visemes alone is known as speechreading or "lip reading".
- **MkGurk Effect**
  - /ba/, /ga/ or /da/?

/ba/+/ga/->/da/

A visual /ga/ combined with an audio /ba/ is often heard as /da/

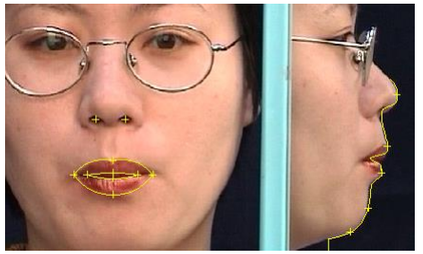
链接

McGurk, H. and J. MacDonald (1976, December). *Hearing lips and seeing voices*. Nature, 746–748.

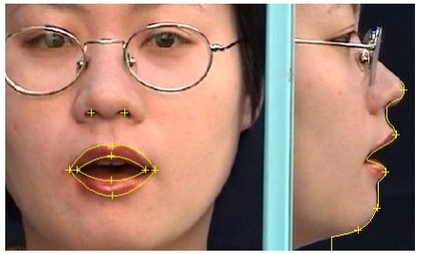
Viseme: 视位/视素

- 汉语视位

视位号	对应声韵母	视位号	对应声韵母	视位号	对应声韵母
0	NA(自然状态)	7	z, c, s	14	i, -i(资韵)
1	b, p, m	8	a, ang	15	o
2	F	9	ai, an	16	ou
3	d, t, n, l	10	ao	17	u
4	g, k, h	11	e, eng	18	ü
5	j, q, x	12	ei, en	19	-i(知韵)
6	zh, ch, sh, r	13	er		

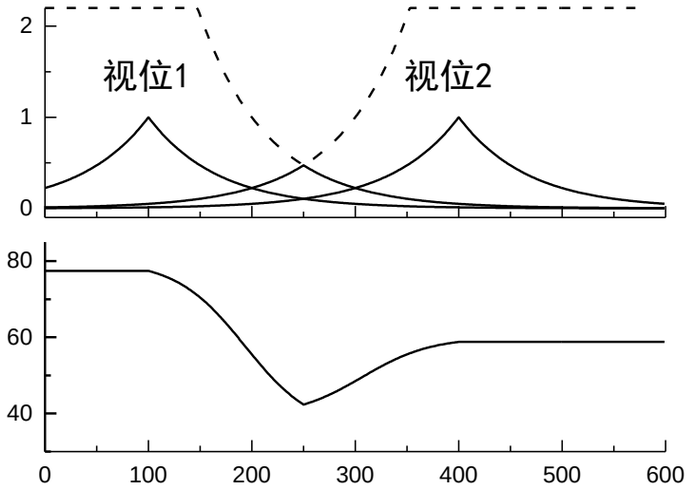


/b/



/a/

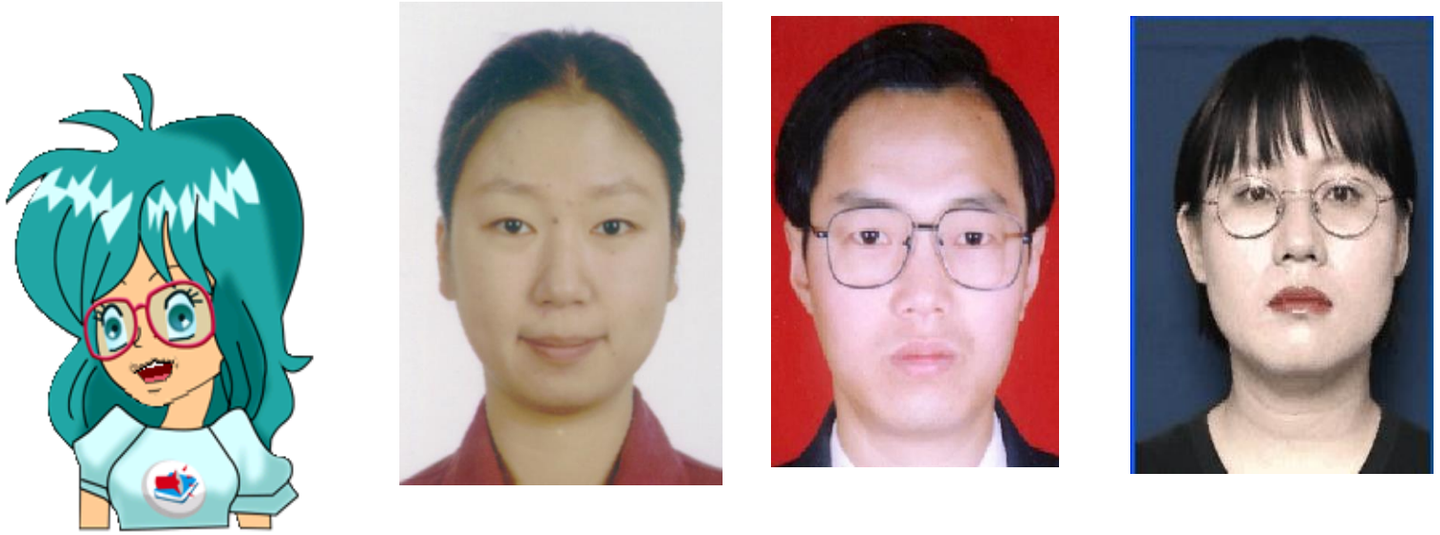
Static Viseme: 静态视位

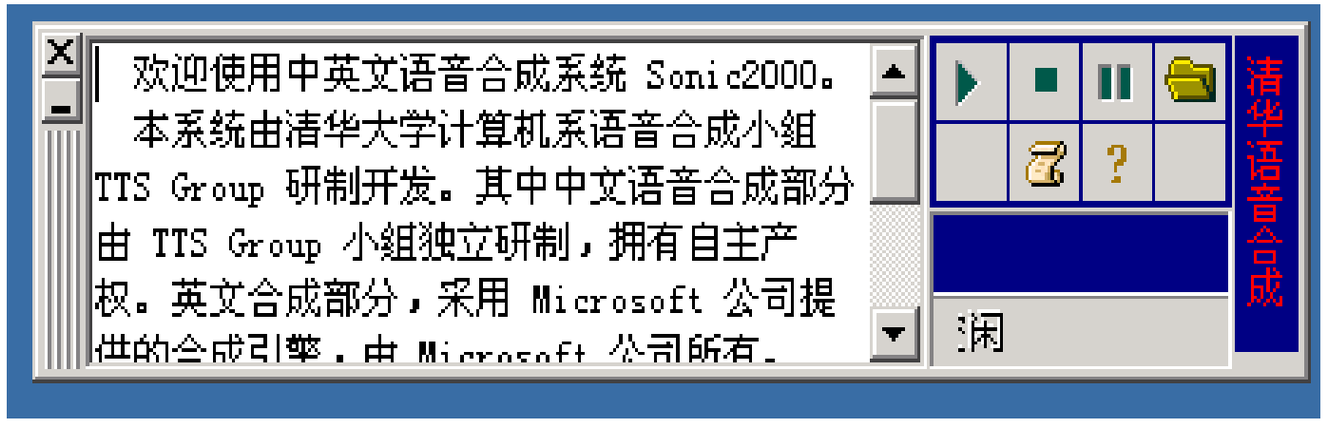


Dynamic Viseme: 动态视位

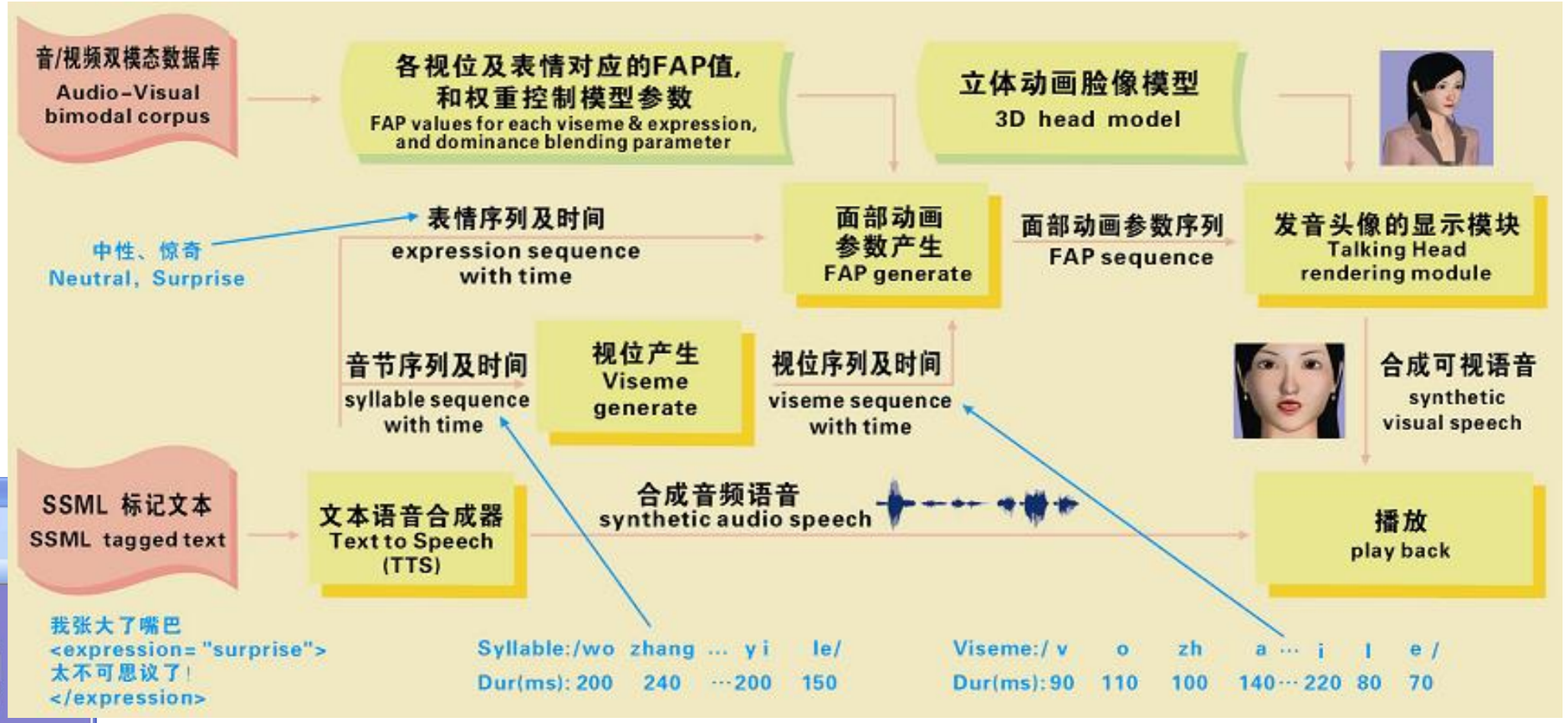
TTVS：可视语音合成

Text-to-Visual Speech Synthesis





TTVS：可视语音合成



Crystal:
Chinese Text-to-Visual-Speech Synthesis

Articulatory Phonetics: 发音语音学

- **音素的分类**
  - 音素可以分为元音 (vowel)和辅音 (consonant)

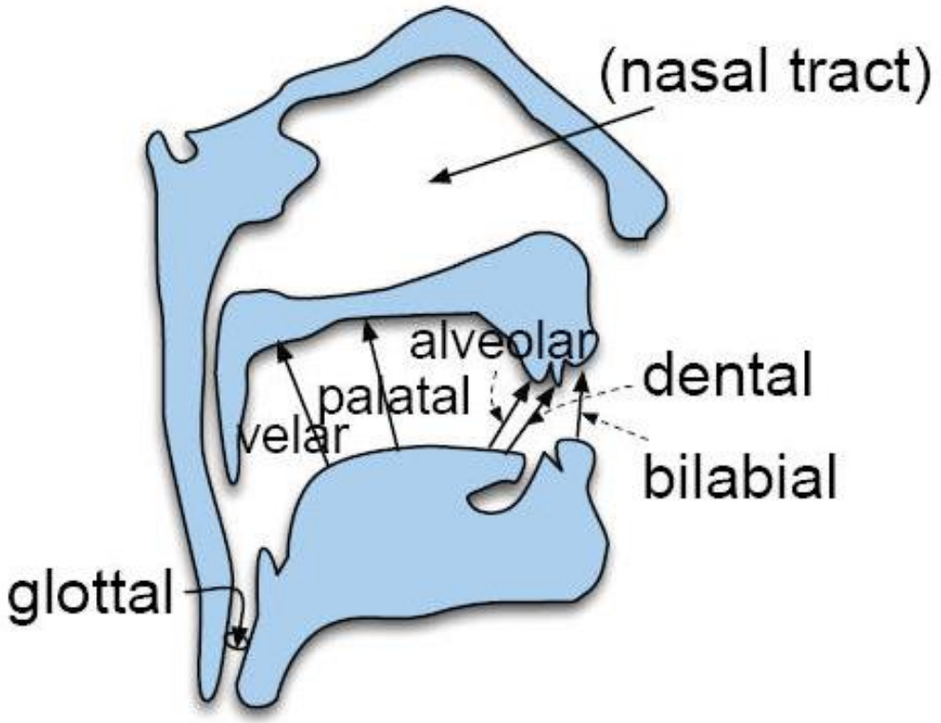
- **Vowel: 元音**
  - 声腔开放，气流较为顺畅的通过，通常为浊音，比辅音声音洪亮且持续时间长
  - 如：[aa], [ae], [ao], [ih] …

- **Consonant: 辅音**
  - 由限制或阻挡气流产生，可以是浊音或清音
  - 如：[p], [b], [t], [d], [k] …
  - 辅音分类：
    - 发音部位：气流受到阻碍的部位
    - 发音方法：阻碍气流以及解除阻碍的方法

Categories of Consonants: 辅音分类

- **Place of Articulation: 按发音部位分**

  由气流受到最大阻挡的位置确定的辅音



  - 双唇音 (labial):
    [p] in *plot*, [b] in *bear*, [m] in *mother*
  - 齿间音 (dental):
    [th] in *thing*, [dh] in *though*
  - 唇齿音 (labiodental):
    [v] in *clove*, [f] in *flour*
  - 齿龈音 (alveolar):
    [s] in *soup*, [z] in *eggs*, [t] in *tea*, [d] in *dill*
  - 硬腭音 (palatal): [sh] in *shrimp*, [ch] in *china*, [zh] in *asian*
  - 软腭音 (velar): [k] in *cuckoo*, [g] in *goose*, [ŋ] in *kingfisher*
  - 声门音/喉音 (glottal): [q] in *uh-oh*

Categories of Consonants: 辅音分类

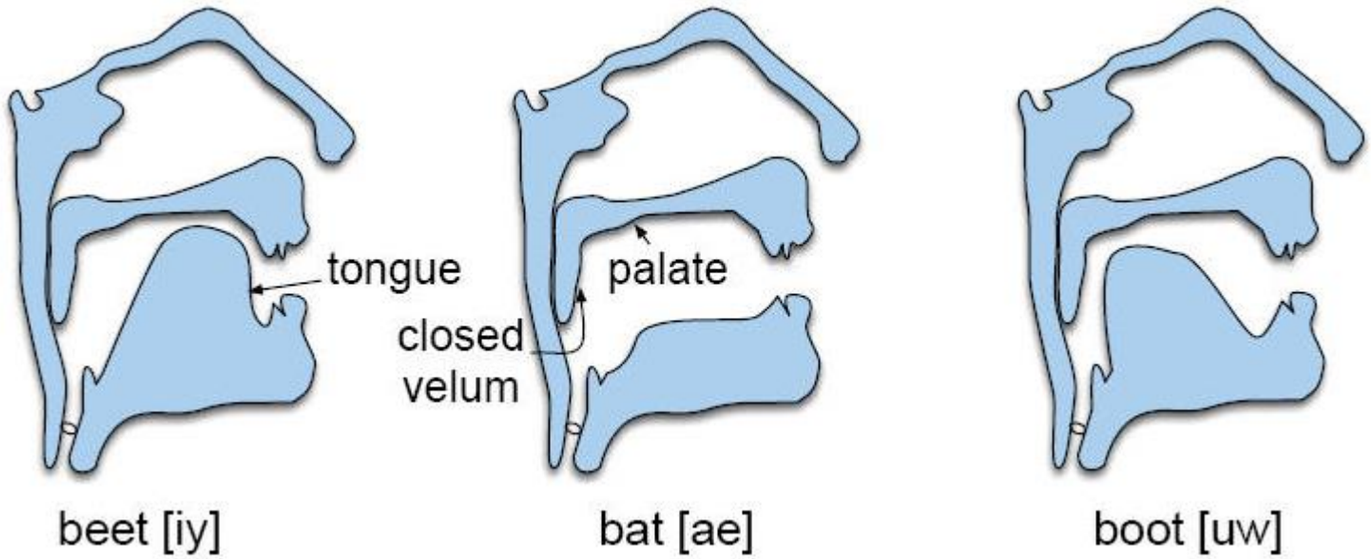
Manner of Articulation: 按发音方法分

描述气流如何受到阻碍而产生的辅音，例如完全阻隔或部分阻隔

- **爆破音** (plosive, stop):
  完全阻隔 (complete blocked) → 释放 (release)
  voiced stopes: [b], [d], [g]; unvoiced stops: [p], [t], [k]
- **鼻音** (nasal):
  降低软腭，让空气进入鼻腔， [n], [m], [ng]
- **摩擦音** (fricatives):
  部分阻隔，嘶嘶声 (hissing)。
  唇齿摩擦音[f], [v]; 齿摩擦音 [th], [dh];
  齿龈摩擦音 [s], [z]; 腭龈摩擦音 [sh], [zh]
- **无擦通音** (approximants):
  两个发音器官很靠近，但不接触，气流通过。
  [y] in *yellow*, [w] in *wood*
- **闪音/拍音** (tap/flap):
  舌快速接触齿龈，[dx] in *lotus* [l ow dx ax s]

Categories of Vowels: 元音分类

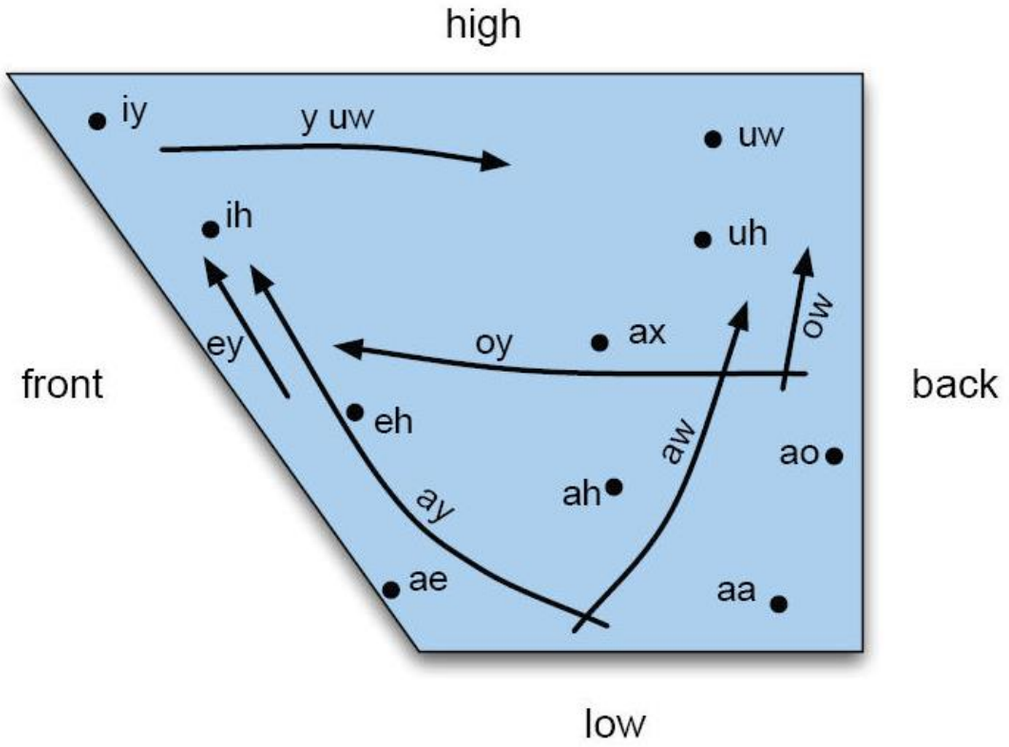
- **Place of Articulation: 按发音部位分**
  - **高度** (height)：舌头最高位置的高度
  - **前后** (frontness or backness)：舌头高位位于口腔的前后位置
  - **圆度** (rounded)：嘴唇形状



Positions of the tongue for three English vowels, high front [iy], low front [ae], and high back [uw]; tongue positions modeled after Ladefoged (1996)

Categories of Vowels: 元音分类

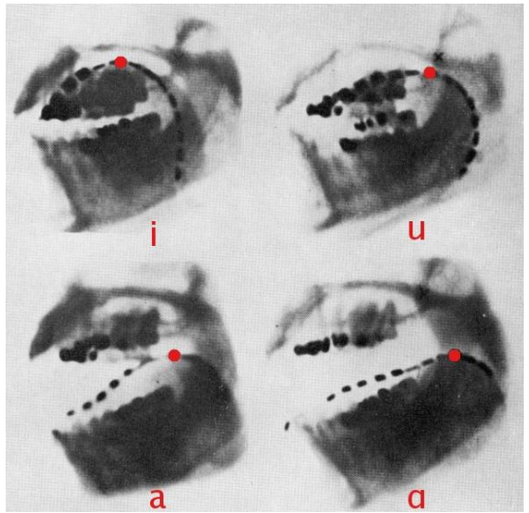
- **Place of Articulation: 按发音部位分**
  - **高度** (height)：舌头最高位置的高度
  - **前后** (frontness or backness)：舌头高位位于口腔的前后位置
  - **圆度** (rounded)：嘴唇形状



Schematic characterization of the vowel height and front or back position of different vowels.

Categories of Vowels: 元音分类

- **Vowel Chart: 元音舌位图**



VOWELS

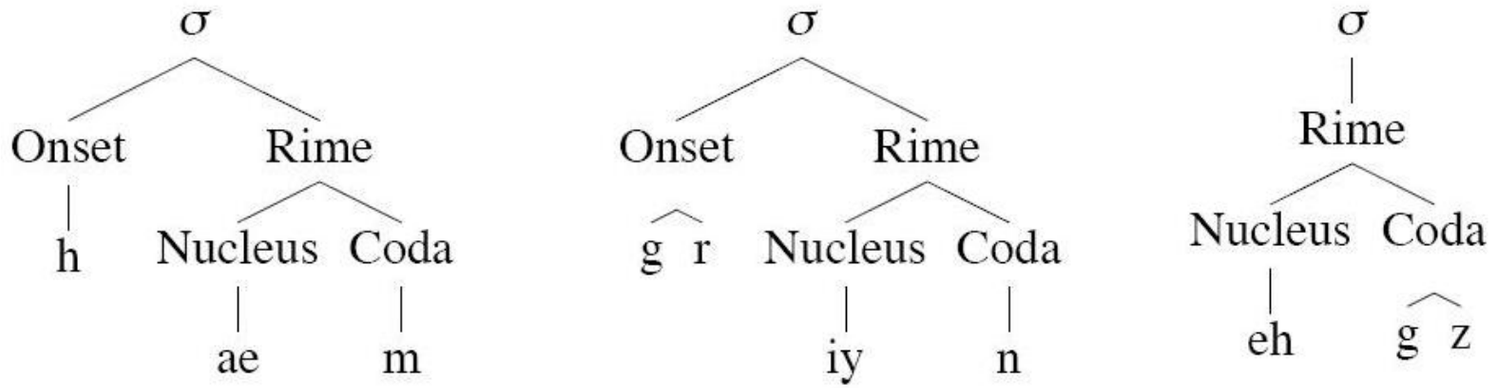
		前元音 Front	次前元音 Near front	央元音 Central	次后元音 Near back	后元音 Back
闭	Close	i	y	ɨ	ʉ	ɯ	u
次闭	Near close		ɪ • ʏ		• ʊ	
半闭	Close mid	e	ø	ɘ	ɵ	ɤ	o
中	Mid			ə			
半开	Open mid	ɛ	œ	ɜ	ɞ	ʌ	ɔ
次开	Near open		æ		ɐ		
开	Open		a	Œ-a		ɑ	ɒ

Vowels at right & left of bullets are rounded & unrounded.

当符号成对出现时，右边的代表圆唇元音

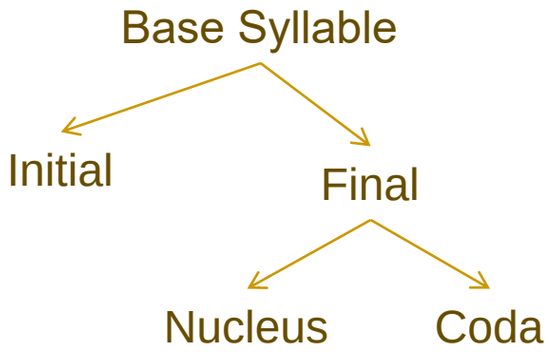
Syllable: 音节

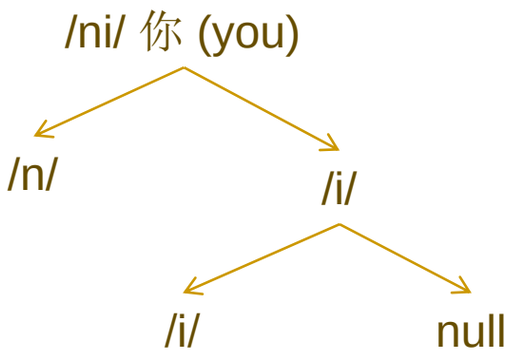
- **Consonants and vowels combine to make a syllable**
  - A vowel-like (or sonorant) sound together with some of the surrounding consonants that are most closely associated with it.
  - 在元音之前的辅音，叫作音节头(onset)或称声母(initial)
  - 在音节头之后的元音及随后的子音被叫作韵母(rime/rhyme/final)
  - 而韵母里的元音叫作音节核(nucleus)
  - 随后的可有可无的子音叫做音节尾(coda)

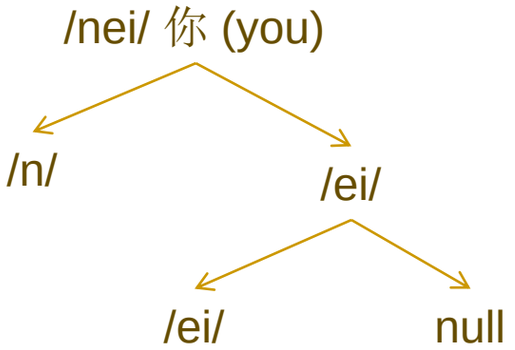


Syllable: 音节

- **Chinese Syllable: 汉语音节**
  - 汉语是有调语言
  - 在中文中，一个汉字读音为一个带调音节(tone syllable)
    - 普通话的带调音节约为1300多个
  - 如去掉声调，称为基础音节/无调音节(base syllable)
    - 普通话的无掉音节约为400多个







语谱

语谱图

宽带语谱图/窄带语谱图

共振峰

语音信号频域分析

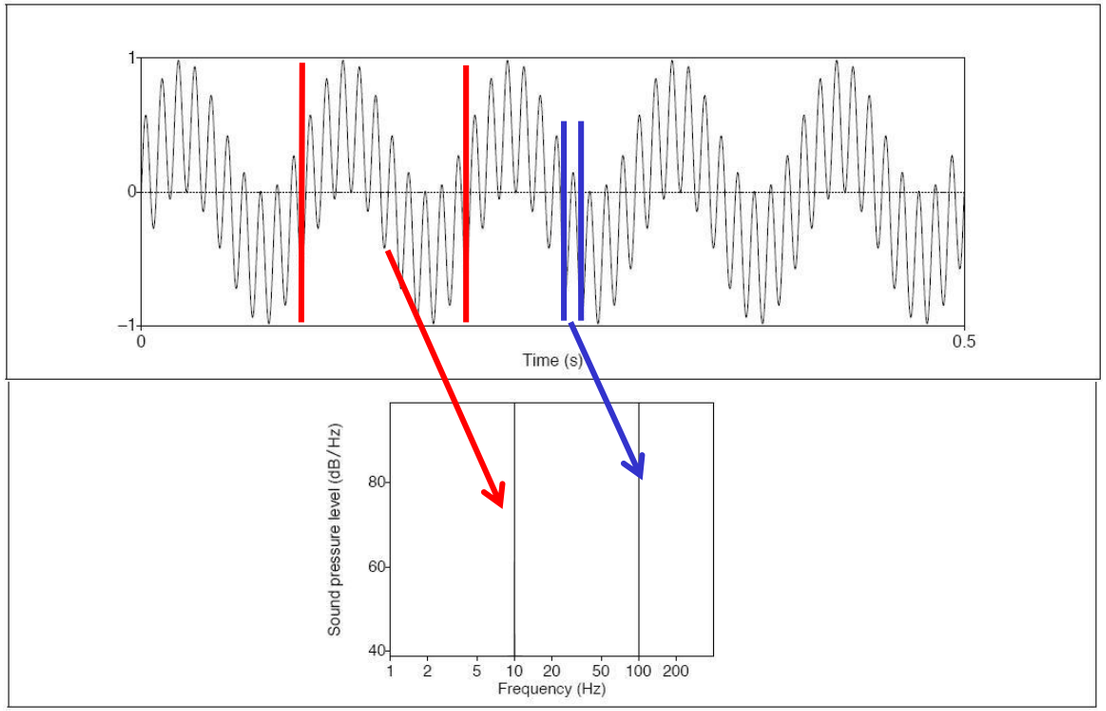
FREQUENCY DOMAIN ANALYSIS

Frequency Domain Analysis: 频域分析

- **Importance of Frequency Information**
  - The frequency information of the waveform are very important for some applications such as speech recognition, speaker identification, etc.
- **Spectrum: 语谱**
  - The spectrum of a signal is a representation of each of its frequency components and their amplitudes.
- **Spectrogram: 语谱图**
  - A spectrogram is a way of envisioning how the different frequencies that make up a waveform change over time.
- **Formant: 共振峰**
  - A formant is a frequency band that is particularly amplified by the vocal tract.

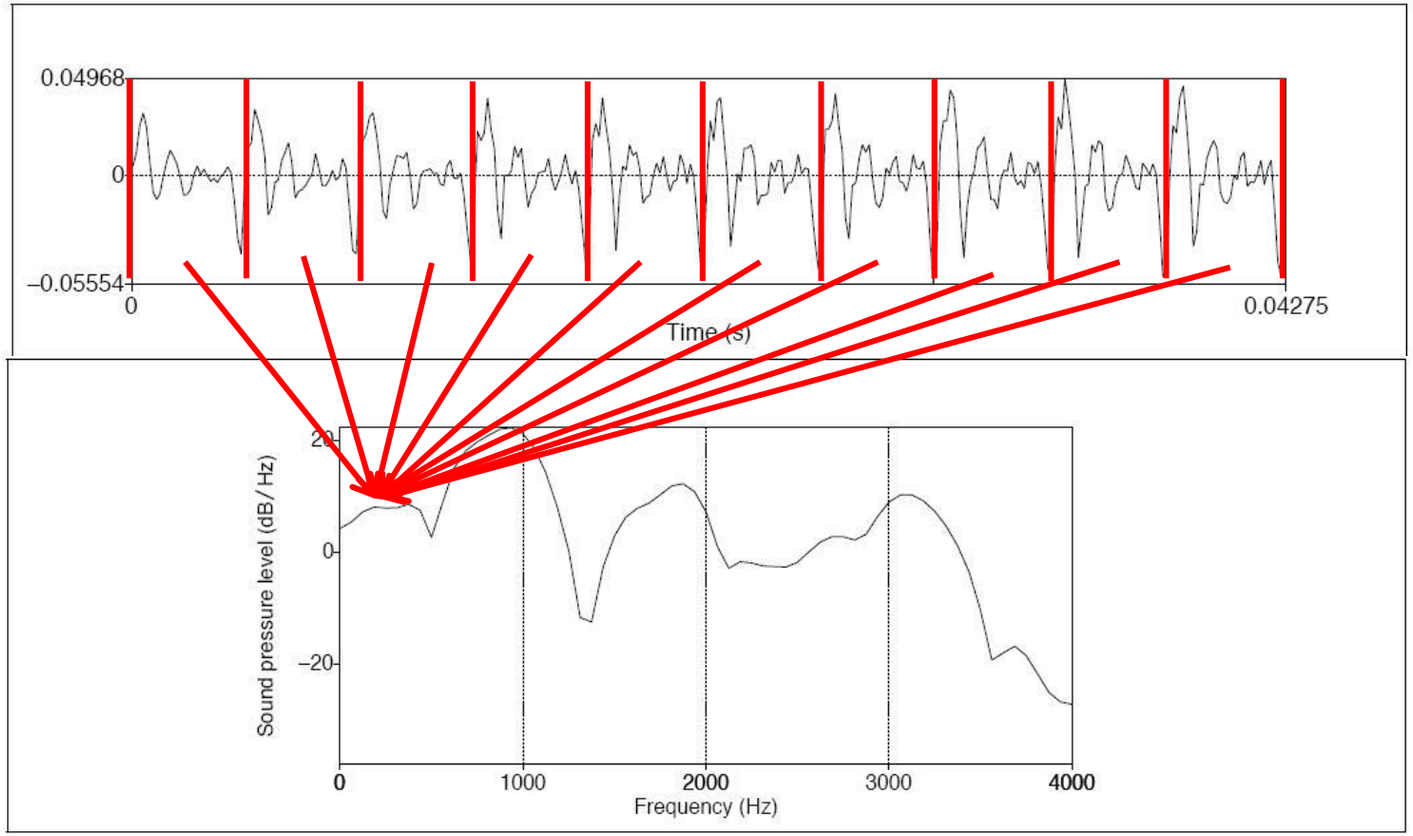
Spectrum: 语谱

- **Spectrum: 语谱**
  - The spectrum of a signal is a representation of each of its frequency components and their amplitudes.
  - 语音信号的频域波形，描述信号包含的频率成分和它们的幅度
- **Fourier Analysis: 傅立叶分析，语谱分析工具**
  - 每个复杂的波形都是由不同频率的正弦波组合而成



Spectrum: 语谱

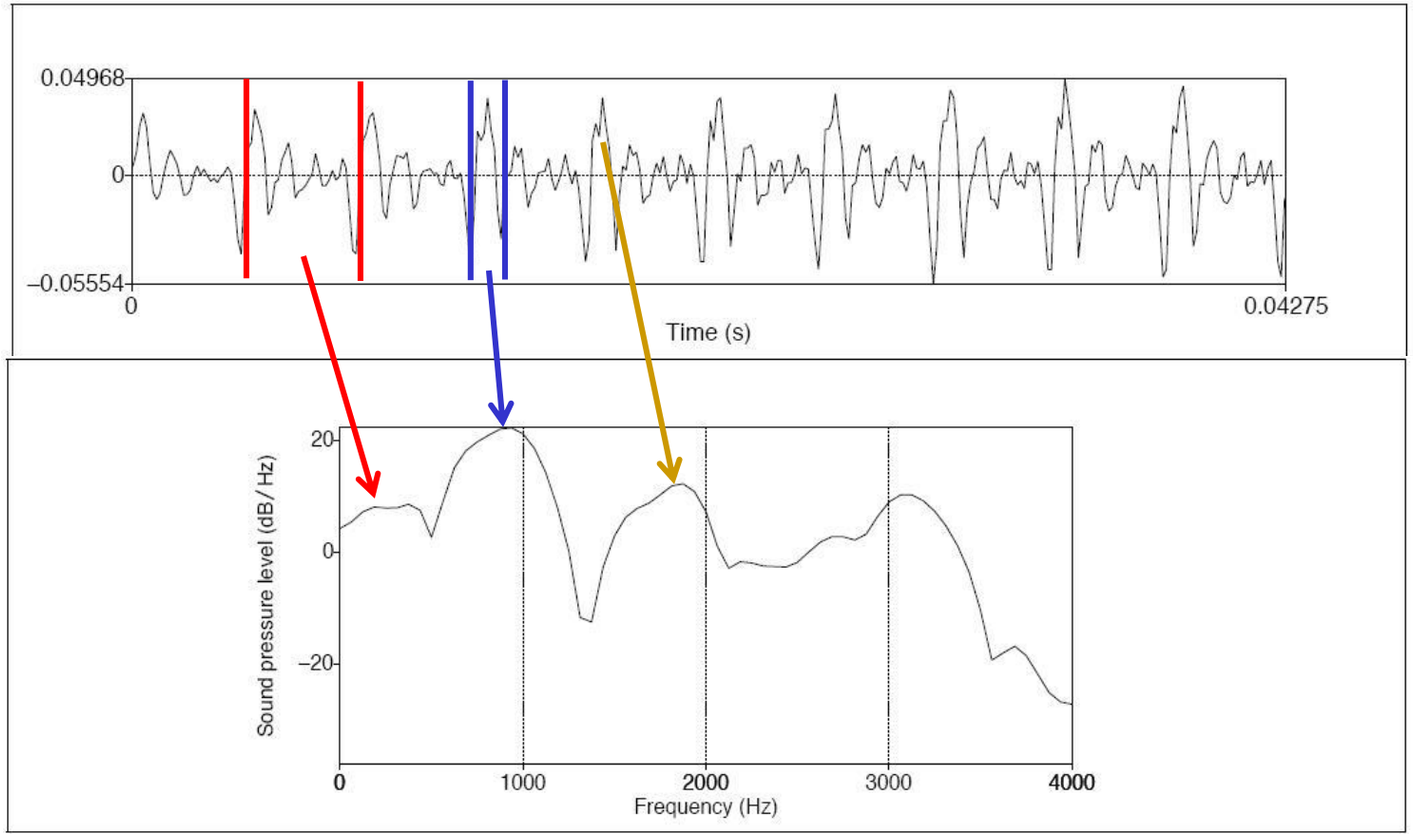
- **Spectrum: 语谱**
  - 语谱可以用来区分不同的音素 (phoneme)
  - 不同音素具有不同的频谱特征 (spectral signature)



The waveform for the vowel /ae/ from *hɑ̲d*, and its spectrum computed from DFT

Spectrum: 语谱

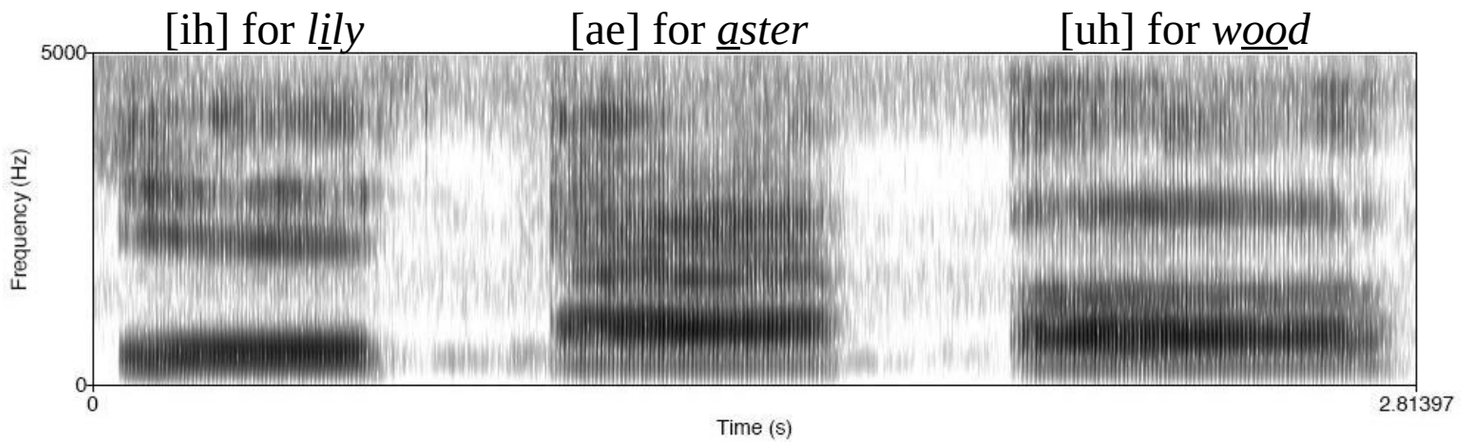
- **Spectrum: 语谱**
  - 语谱可以用来区分不同的音素 (phoneme)
  - 不同音素具有不同的频谱特征 (spectral signature)



The waveform for the vowel /ae/ from *hɑ̲d*, and its spectrum computed from DFT

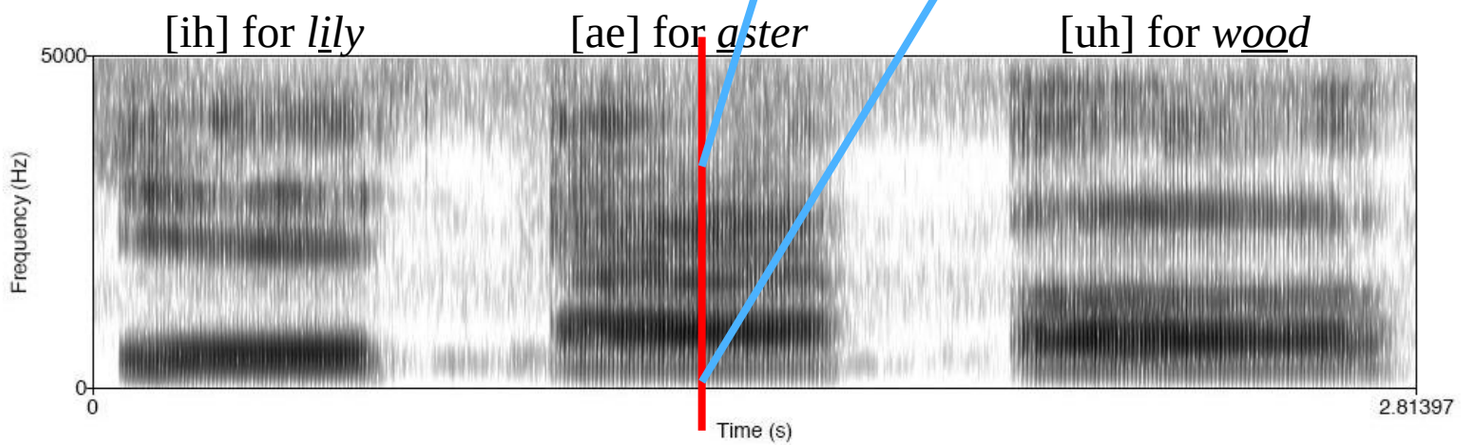
Spectrogram: 语谱图

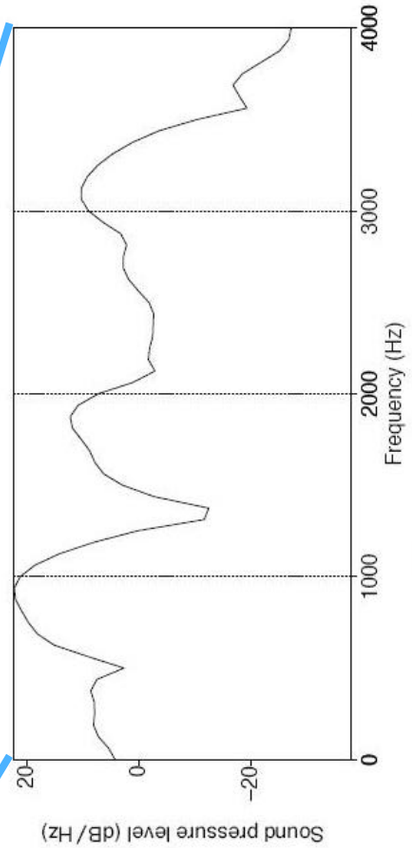
- **Spectrogram: 语谱图**
  - A spectrogram is a way of envisioning how the different frequencies that make up a waveform change over time.
  - The *x*-axis shows time, as it did for the waveform.
  - The *y*-axis now shows frequencies in Hertz.
  - The darkness of a point on a spectrogram corresponding to the amplitude of the frequency component.
    - Very dark points have high amplitude, light points have low amplitude.
  - The spectrogram is a useful way of visualizing the three dimensions (time x frequency x amplitude).



Spectrogram: 语谱图

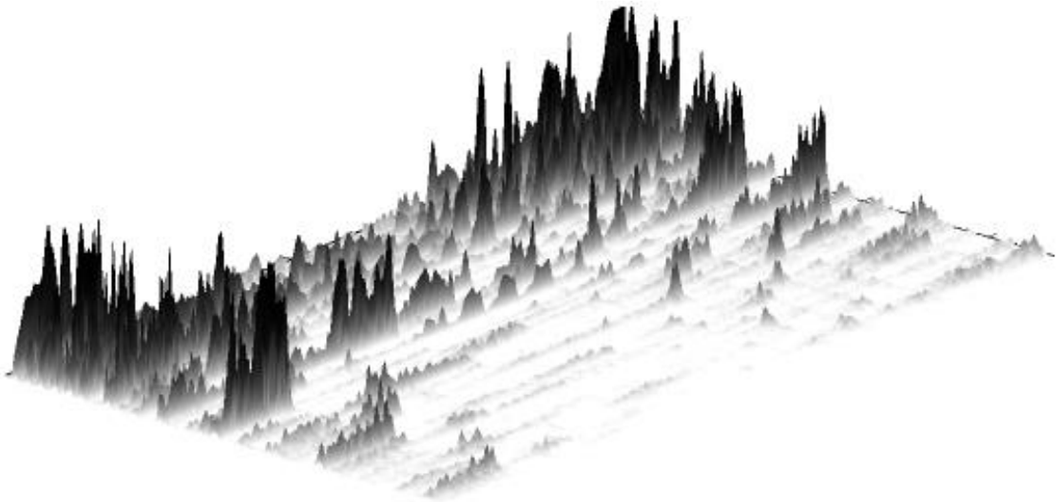
- **Spectrogram: 语谱图**
  - A spectrogram is a way of envisioning how the different frequ waveform change over time.
  - The *x*-axis shows time, as it did for the waveform.
  - The *y*-axis now shows frequencies in Hertz.
  - The darkness of a point on a spectrogram corresponding to the uency component.
    - Very dark points have high amplitude, light points have lo
  - The spectrogram is a useful way of visualizing the three dime cy x amplitude).





Spectrogram: 语谱图

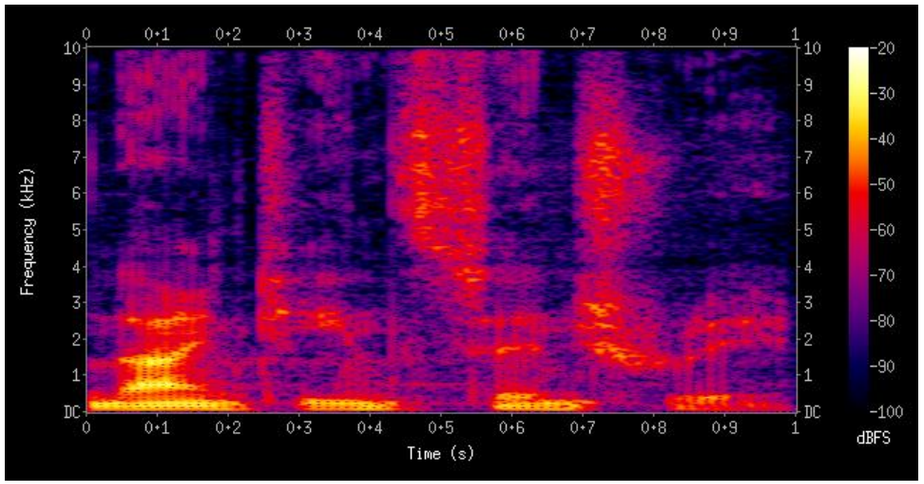
- **Spectrogram**
  - 描述信号包含的频率成分和它们的幅度



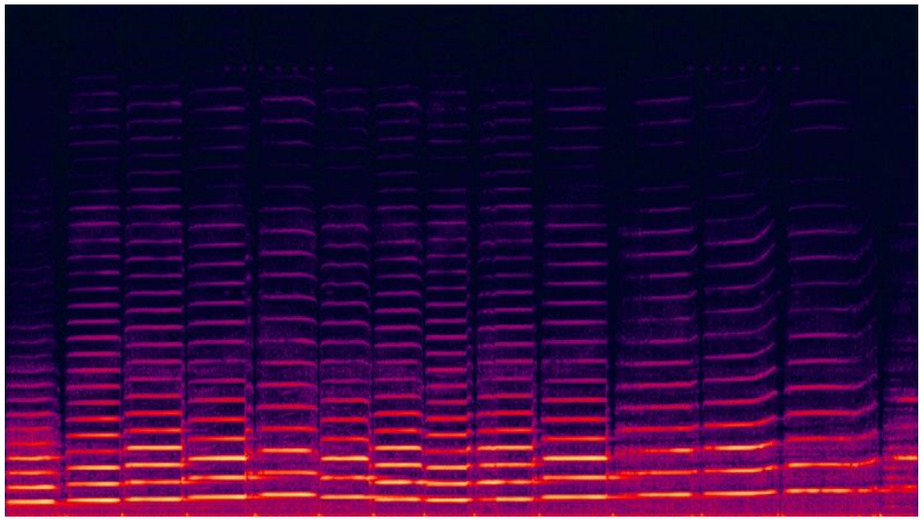
3D surface spectrogram of a part from a music piece.

(from: wikipedia)

CoolEdit



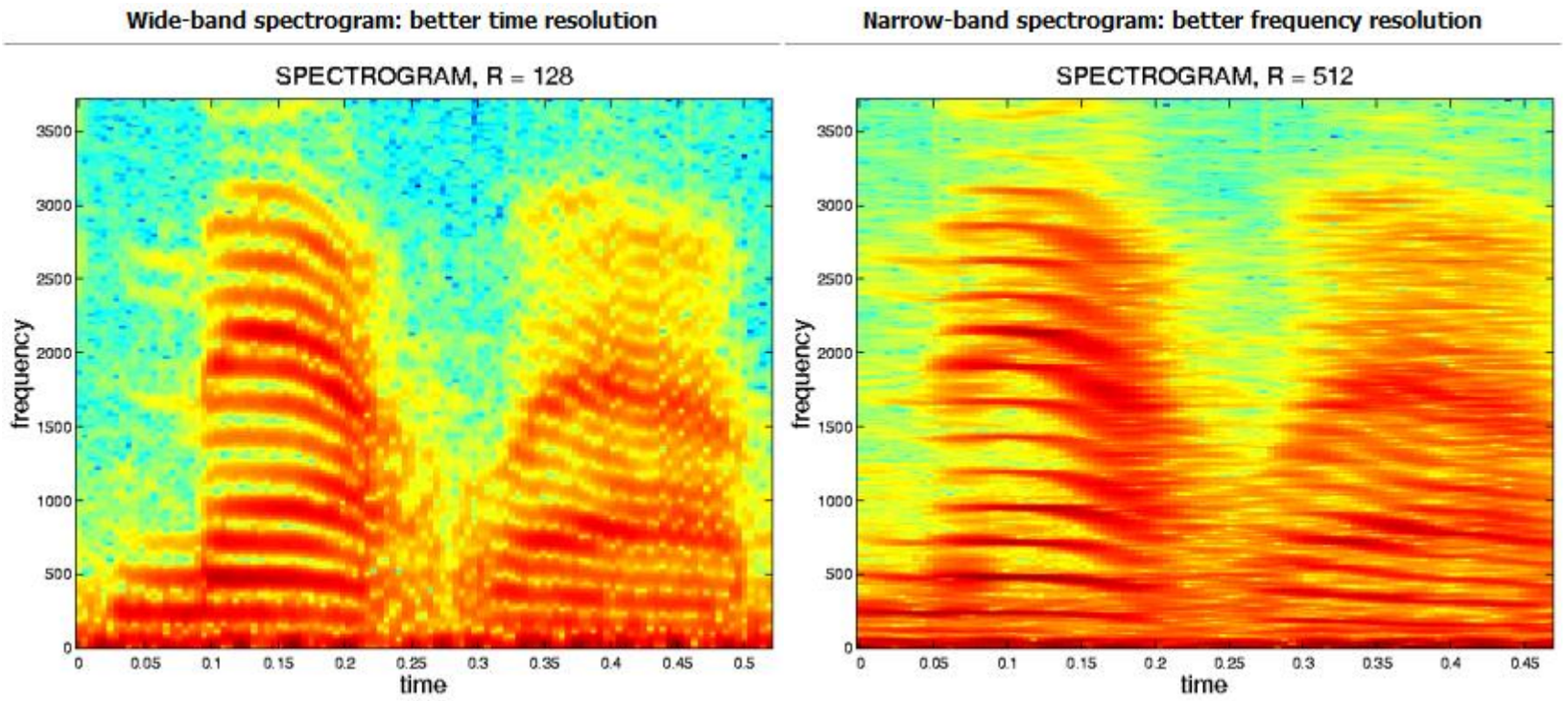
Spectrogram of a male voice saying ‘nineteenth century’.



Spectrogram of violin playing.

Spectrogram: 语谱图

- **Wide-band Spectrogram: 宽带语谱图**
  - 频率分辨率取300-400Hz，时间分辨率2-5ms，良好的时间分辨率，频率分辨率较差
- **Narrow-band Spectrogram: 窄带语谱图**
  - 频率分辨率取50-100Hz，时间分辨率5-10ms，良好的频率分辨率，时间分辨率较差



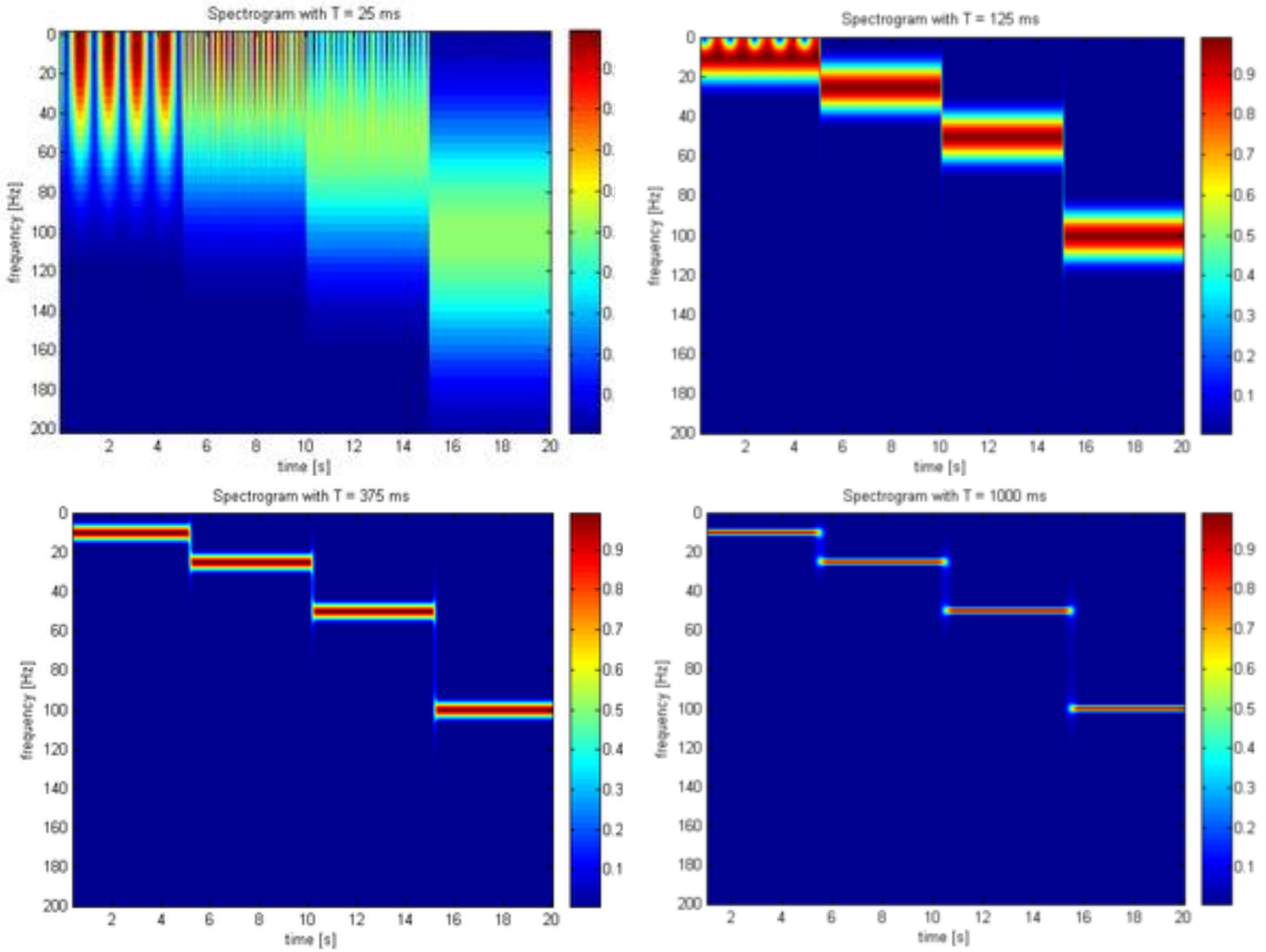
(from: http://cnx.org/content/m10570/latest/)

praat

Spectrogram: 语谱图

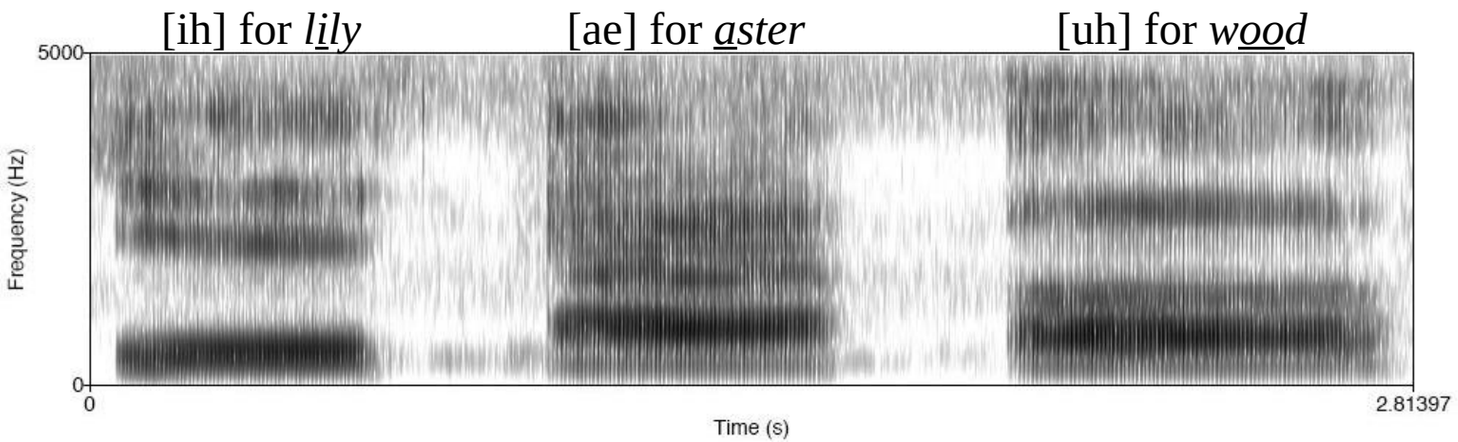
- **Resolution Issues**
  - 两种分辨率均受窗函数的影响
    - Wide window gives better frequency but poor time resolution.
    - Narrow window gives good time but poor frequency resolution.
  - Explanation
    - Frequency Resolution:
      Frequency space between
      2 consecutive coefficients: fs/N

$$x(t) = \begin{cases} \cos(2\pi 10t); & 0 \le t < 5s \\ \cos(2\pi 25t); & 5 \le t < 10s \\ \cos(2\pi 50t); & 10 \le t < 15s \\ \cos(2\pi 100t); & 15 \le t < 20s \end{cases}$$



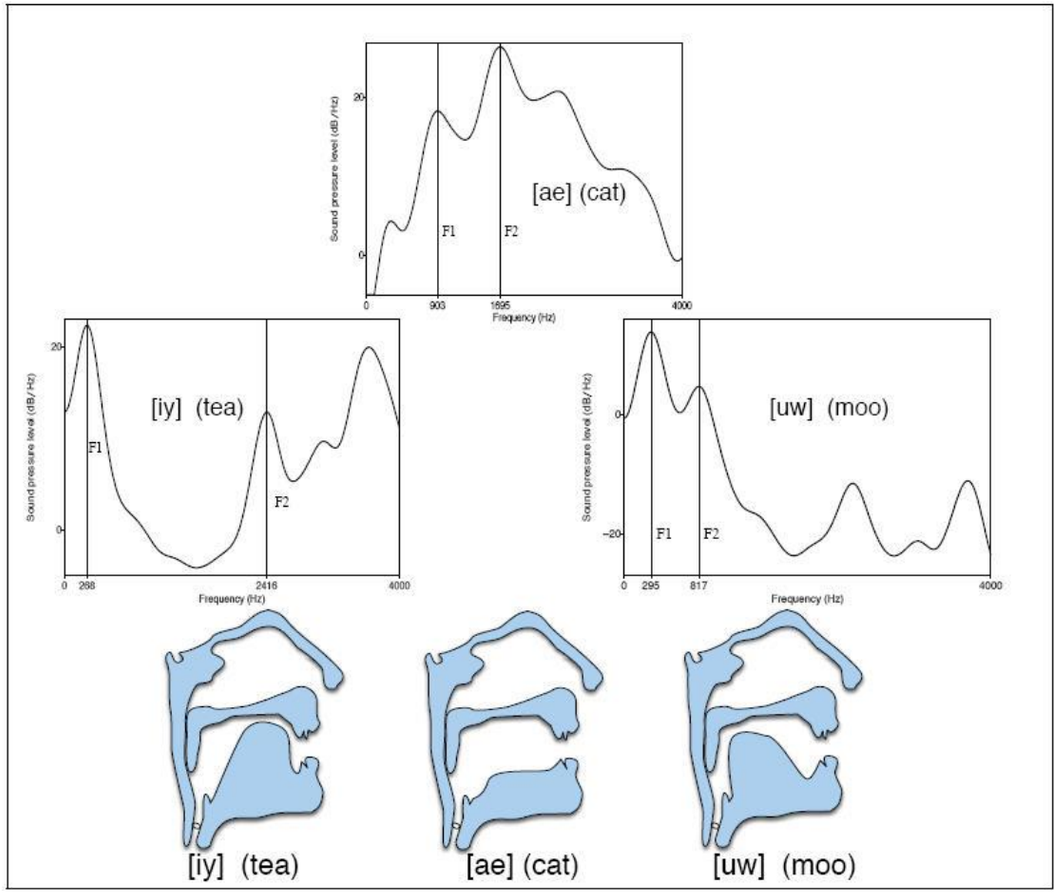
Formant: 共振峰

- **Formant: 共振峰**
  - 是指在声音的频谱中能量相对集中的一些区域（语谱峰值）。
  - 共振峰不但是音质的决定因素，而且反映了声道（共振腔）的物理特征。
  - 声音在经过共振腔时，受到腔体的滤波作用，使得频域中不同频率的能量重新分配，一部分因为共振腔的共振作用得到强化，另一部分则受到衰减，得到强化的那些频率
    在时频分析的语谱图上表现为浓重的黑色条纹。
  - 由于能量分布不均匀，强的部分
    犹如山峰一般，故而称之为共振峰。



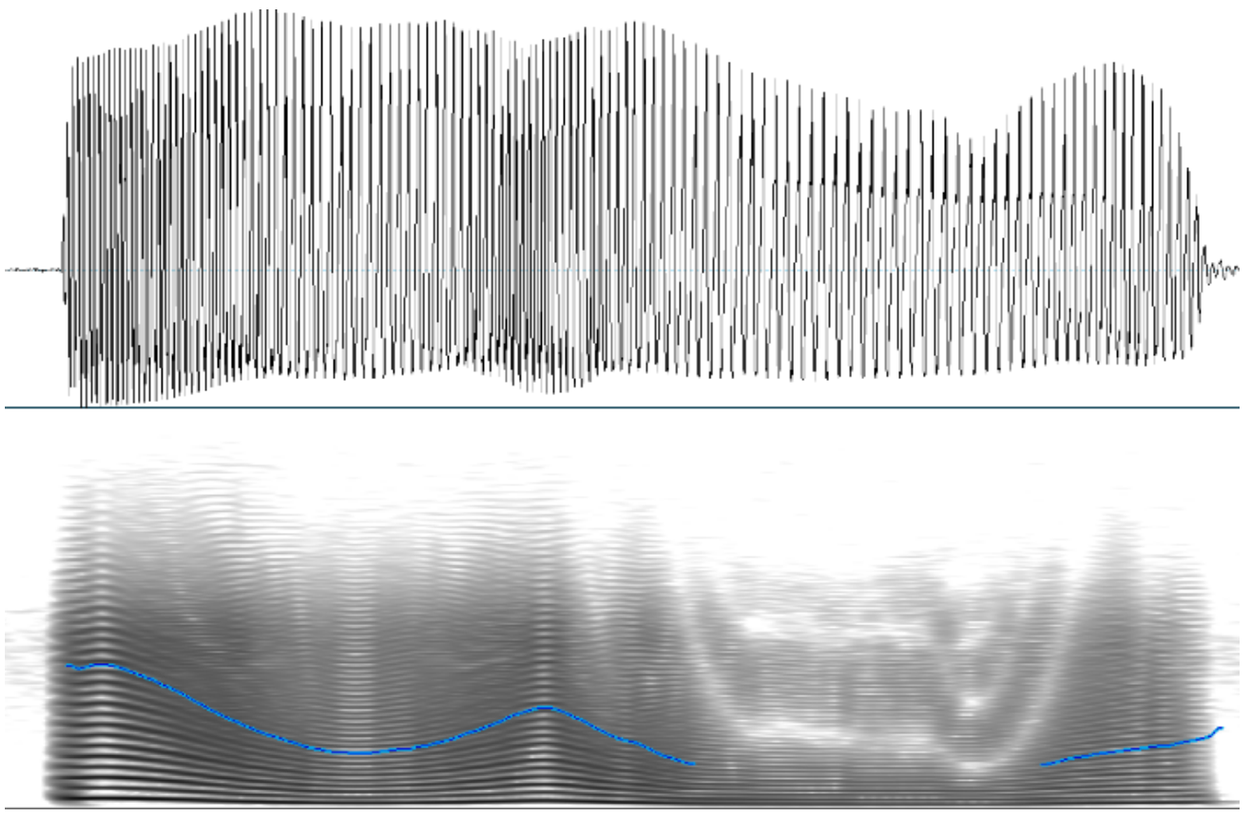
Formant: 共振峰

- **共振峰是被声道特别放大的频带**
  - 由于不同元音在声道内不同位置产生，不同元音会产生不同种类的放大或共振。
- **第一和第二共振峰 (F_1和F_2) 对于区分不同元音尤为重要**

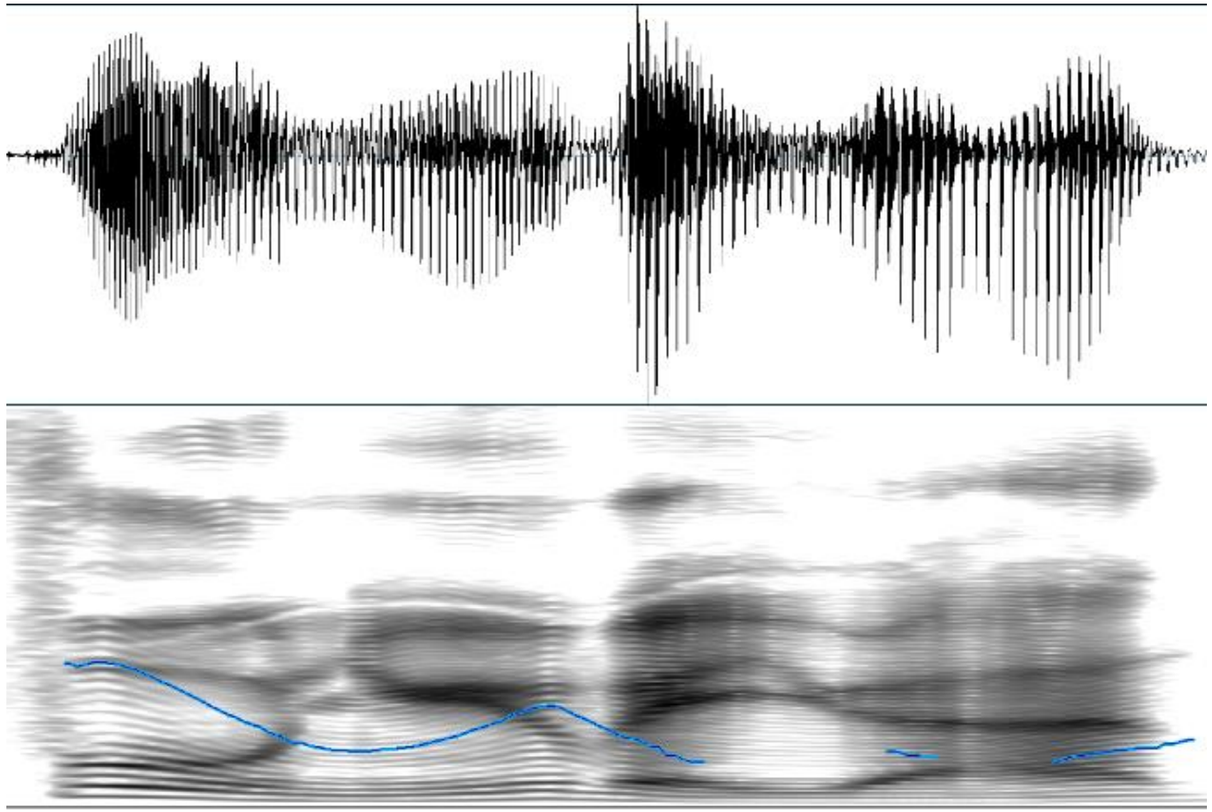


Formant: 共振峰

- 共振峰与谐波的关系



左下：（只有声带振动的）谐波语谱图



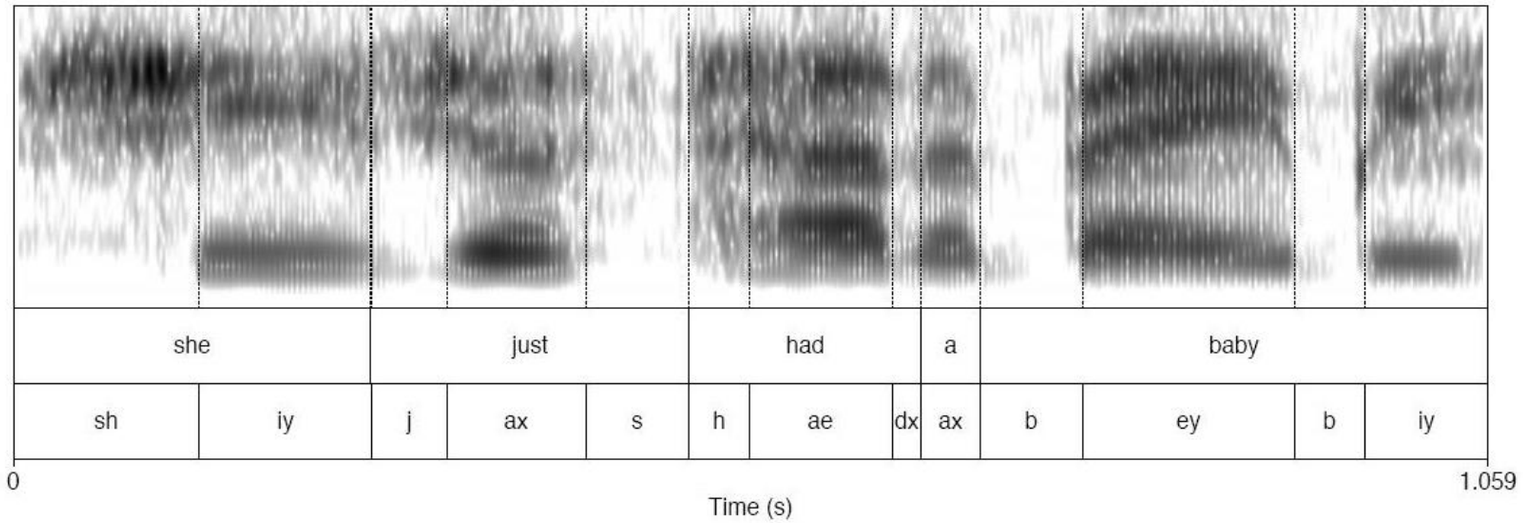
右下：（谐波经过声道调制后的）具有共振峰的语谱图

Phonetic Transcription: 语音学标注

- **Phonetic Resources**
  - Phonetically annotated corpus, e.g., TIMIT, Switchboard, etc.
  - Time-aligned transcription

Hints:
- Forced-alignment of speech corpus
- 语料库自动标注的研究

From Switchboard:



From TIMIT:

she	had	your	dark	suit	in	greasy	wash	water	all	year
sh iy	hv ae dcl	jh axr	dcl d aa r kcl	s ux q	en	gcl g r iy s ix	w aa sh	q w aa dx axr q	aa l	y ix axr

人耳频率响应

等响曲线

听域/听阈

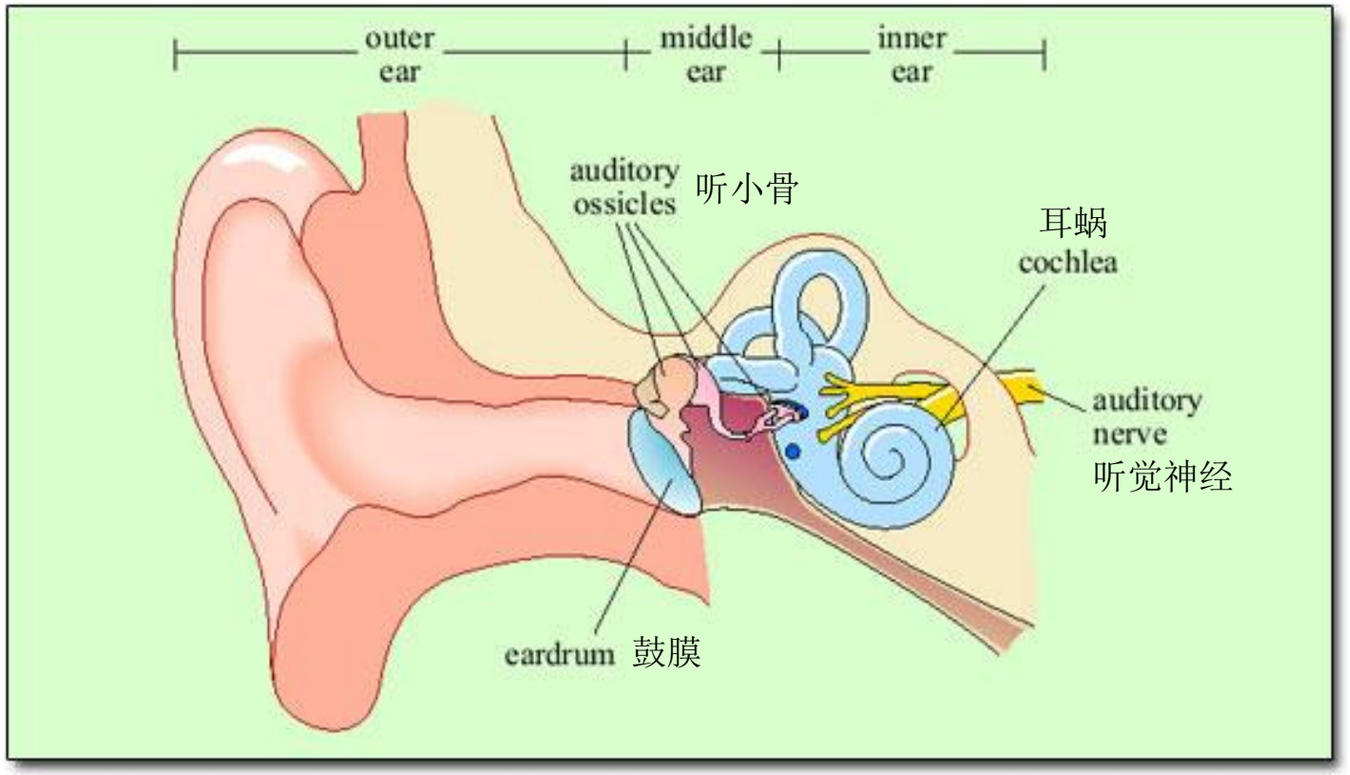
掩蔽效应

语音信号听觉感知

SPEECH PERCEPTION

Speech Perception: 听觉感知

- **人耳：频谱分析仪**
  - 外耳：声源定位；对声音进行放大
  - 中耳：通过听小骨进行声阻抗变换，放大声压；包含内耳
  - 内耳：机械振动向神经发放信号转换，把声音刺激变成神经冲动



Speech Perception: 听觉感知

- **Physical vs. Perceptual: 物理特性 vs. 听觉特性**

Physical Quantity 物理量	Perceptual Quantity 感知量
Intensity 声强	Loudness 响度
Fundamental Frequency 基频	Pitch 音高或音调
Spectral Shape 频谱形状	Timbre 音色或音品
Onset/offset Time 破发时间	Timing 时序
Phase Difference in Binaural Hearing 双耳听觉上的相位差	Location 定位

声音三要素（主观心理量）

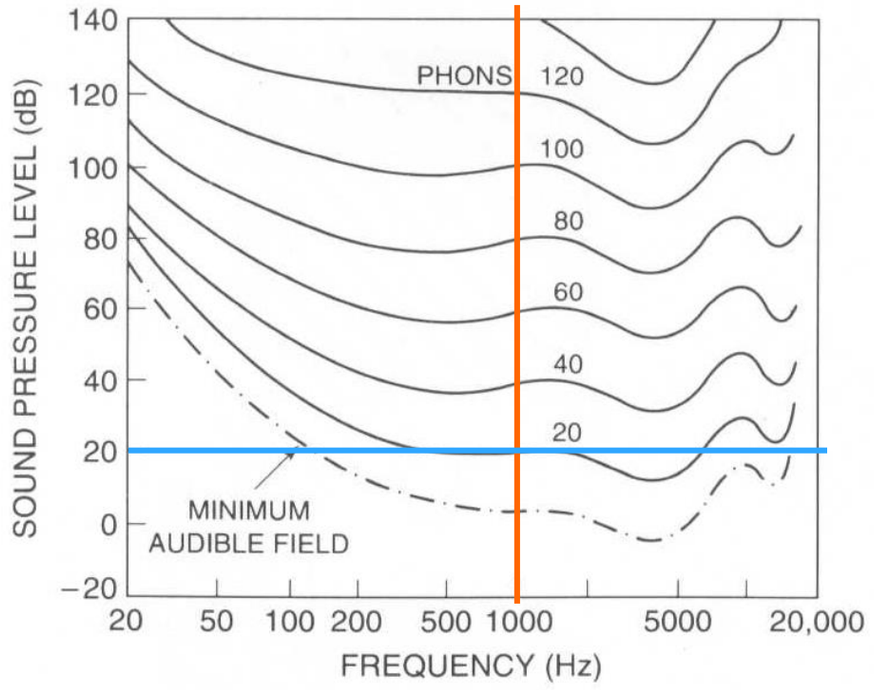
- **Timbre: 音色/音品**
  - 由声音波形的谐波频谱和包络决定。
  - 声音波形的基频所产生的听得最清楚的音称为**基音**，各次谐波的微小振动所产生的声音称为**泛音**。单一频率的音称为**纯音**，具有谐波的音称为**复音**。每个基音都有固有的频率和不同响度的泛音，借此可以区别其他具有相同响度和音调的声音。
  - 声音波形各次谐波的比例和随时间的衰减大小决定了各种声源的**音色特征**，其包络是每个周期波峰之间的连线，包络的陡缓影响声音强度的瞬态特性。
- **Binaural Hearing: 双耳听觉**
  - 定位作用。时间和声强对高低频具有不同作用。低频声音主要靠双耳之间的时差进行定位，高频声音主要靠双耳声强差来定位。

Equal Loudness Curve: 等响曲线

- **响度级: Loudness Level**
  - 人主观感觉不同频率成分的纯音（Pure Tones）强弱的物理量
  - 单位：方(Phon)，数值上等于 1kHz 纯音的声压级
  - 人耳对不同频率的纯音的强弱的辨别能力是不一样的

- **等响曲线**
  - 表示人听到同样响度的声音时，其声压级（SPL: Sound Pressure Level）与频率的关系
  - 反映了**响度级**与**频率**、声压之间的关系

频率范围在 3kHz~4kHz 附近，等响曲线的变化对应的声压级最小，也即，此时人耳的分辨率最灵敏。

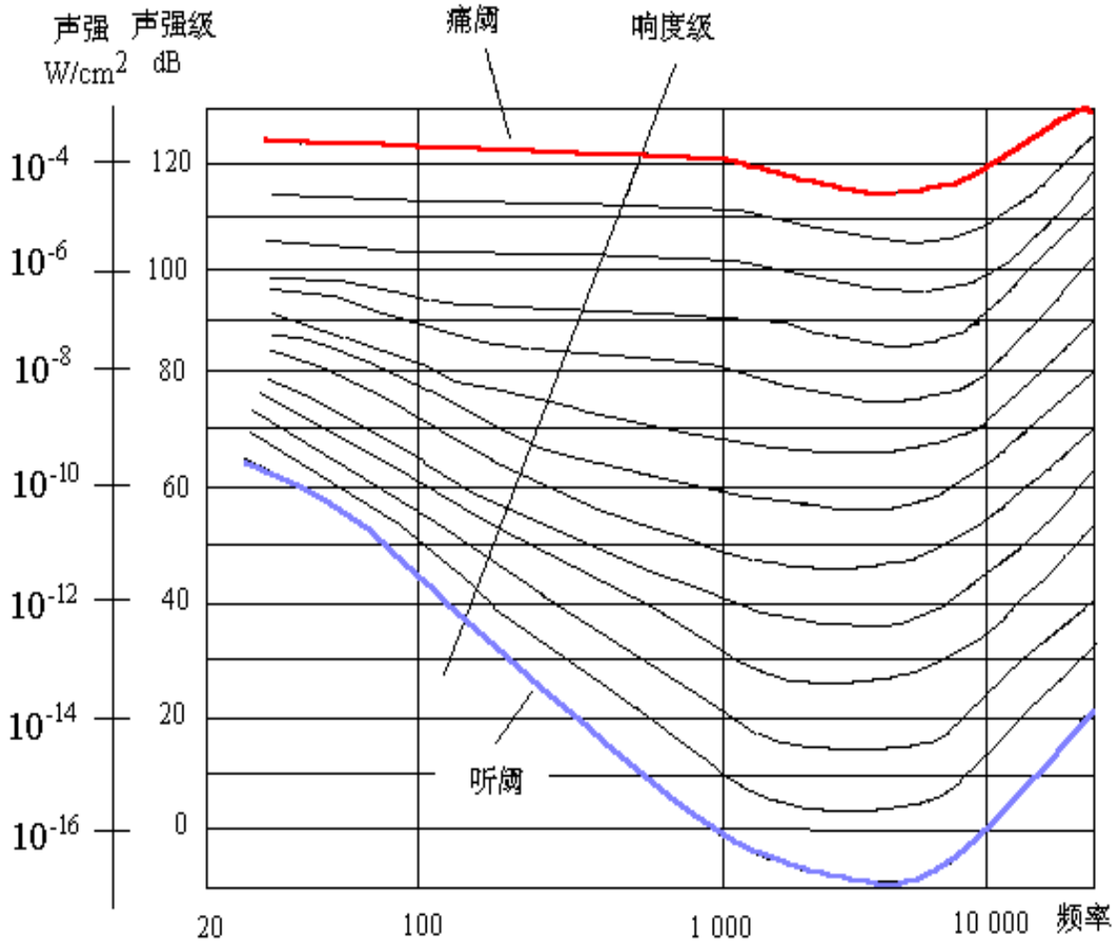


Frequency response characteristics of the human auditory system as a function of loudness

(from: http://en.wikipedia.org/wiki/Psychoacoustic_model)

Auditory Threshold: 听阈

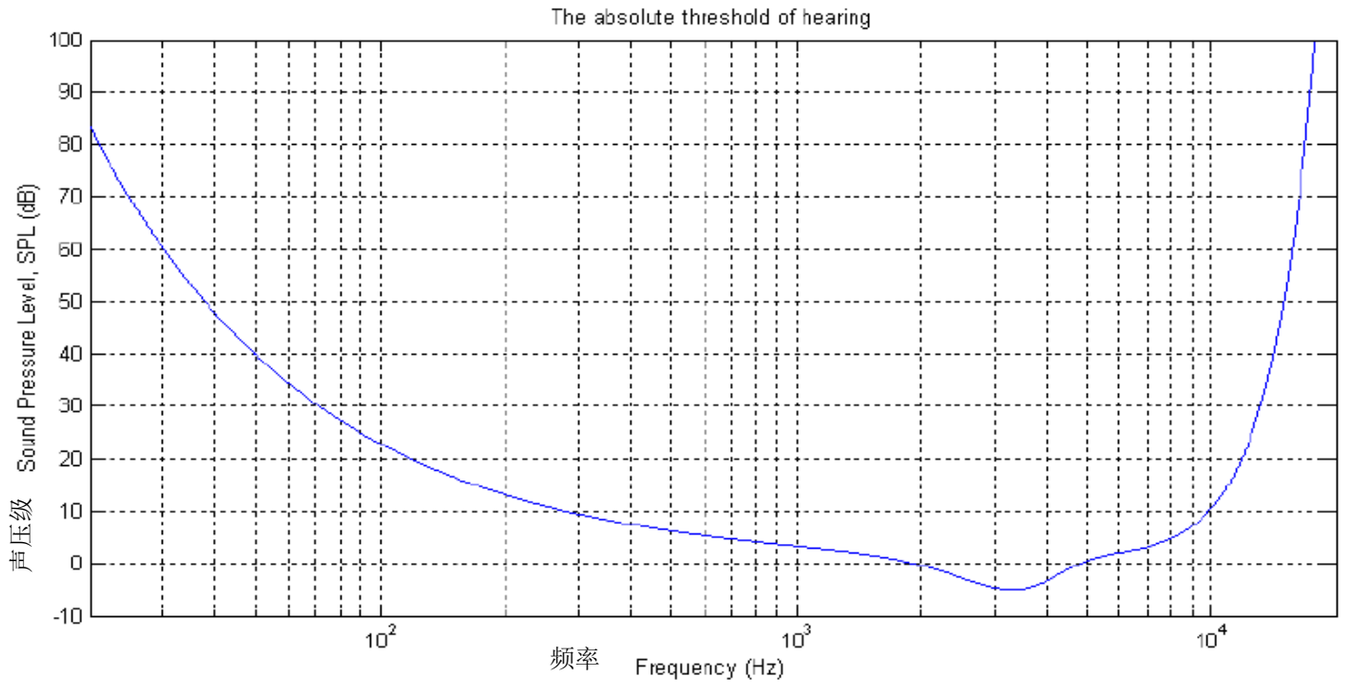
- **听阈**
  - 在没有噪声的环境下，声音的某一个频率点（纯音），信号能够产生听觉感知的最低能量幅度
  - 人耳刚好能听见某一频率点声音时的响度（下限）
- **痛阈**
  - 声音强到使人耳感到疼痛时的响度（上限）
- **听觉范围**
  - Intensity Range of Audible Sounds
  - 人耳可听声音的强度范围是0-120dB（声压级）
  - 人耳听觉范围位于听阈和痛阈之间



(from: [1])

The Absolute Threshold of Hearing

- **绝对听觉门限**
  - 也即：**听阈**
  - 在没有噪声的环境下，针对声音的某一频率点（纯音），信号能够产生听觉感知的最低能量幅度
  - 又称：**静音门限** (The Threshold in Quiet)
  - 与频率 f 的关系



纯音的听阈与频率有关： 1KHz纯音的听觉门限（听阈）约为4dB， 10KHz 时约为 15dB。

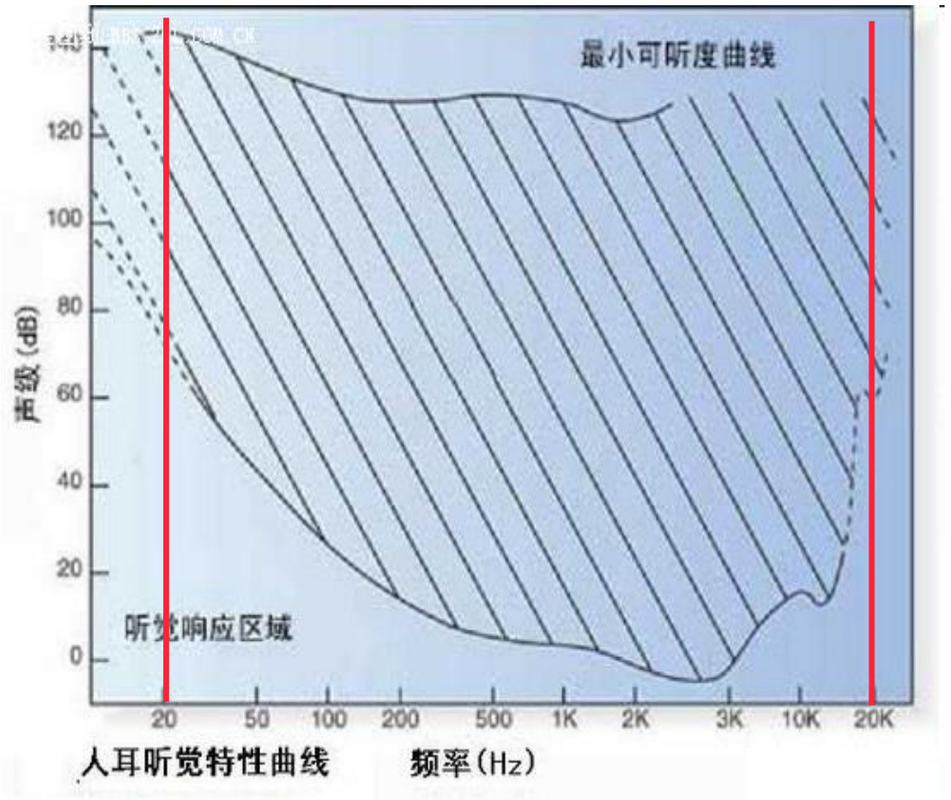
$$T_q(f) = 3.64(f/1000)^{-0.8} - 6.5e^{-0.6(f/1000-3.3)^2} + 10^{-3}(f/1000)^4$$

(from: [4])

Auditory Zone: 听域

- **Auditory Zone: 听域**
  - 人可听声音的频率范围为16.4Hz-10KHz -16KHz-20KHz

人耳的听觉特性

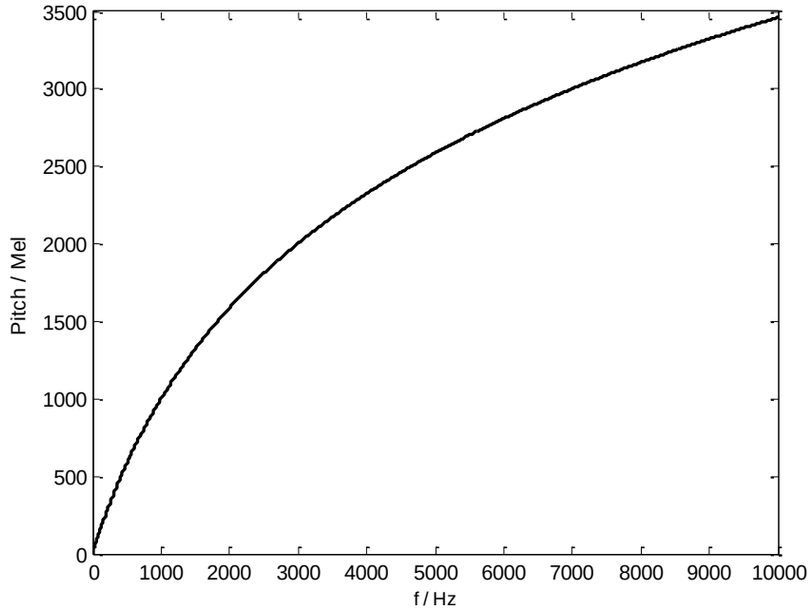


Frequency Response: 人耳频率响应

- **Auditory Zone: 听域**
  - 人可听声音的频率范围为16.4Hz-10KHz -16KHz-20KHz

人耳听到的频率范围.mp4

- **Pitch: 音高人**
  - 人耳对不同频率的主观感受
  - 人耳对频率的感知是非线性的，近似为对数函数 (logarithm)
- **Mel Domain: 美尔域**
  - Mel: 音高的度量单位



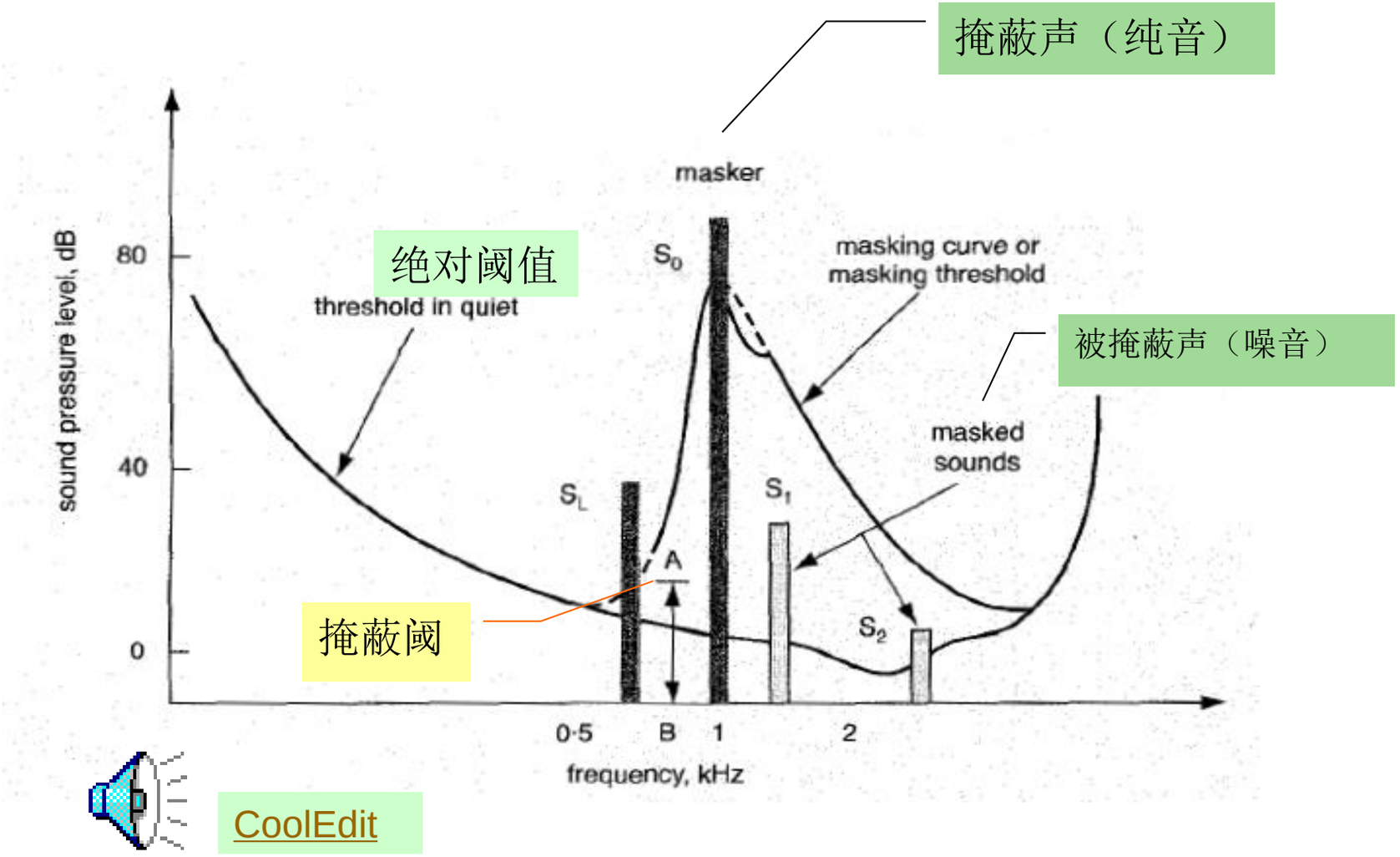
$$P_{\text{Mel}} \cong (1000/\lg 2) \times \lg(1 + 0.001 f_{\text{Hz}})$$

Auditory Masking: 掩蔽效应

- **Auditory masking** occurs when the perception of one sound is affected by the presence of another sound (Gelfand 2004)
  - 是一种心理学现象，是由人耳对声音频率分辨机制决定的。是指一个较强声音的附件，相对较弱的声音不易被人耳察觉，即被强音所掩蔽。
- **同时掩蔽（频率掩蔽）**
  - 一个强纯音会掩蔽其附近频率同时发声的弱纯音
- **异时掩蔽（时域掩蔽）**
  - 在时间上相邻的声音之间也有掩蔽现象
  - 掩蔽阈值是时间、频率和声压级的函数

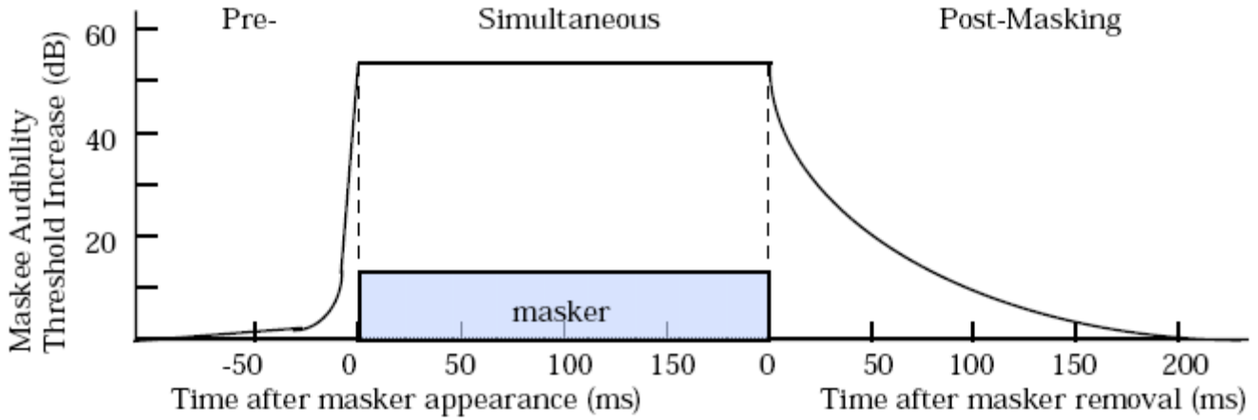
Auditory Masking: 掩蔽效应

- 同时掩蔽（频率掩蔽）



Auditory Masking: 掩蔽效应

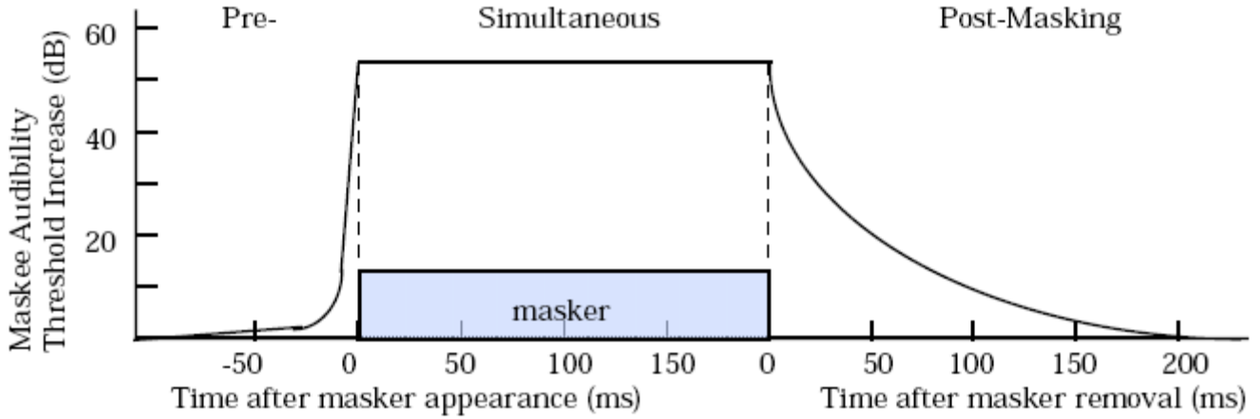
- **异时掩蔽（时域掩蔽） / Non-simultaneous Masking**
  - 在时间上相邻声音的掩蔽现象
  - 若在很短时间内出现两个声音，无论声音出现的先后次序，声压级大的声音 (masker) 会掩蔽声压级小的声音 (maskee)



(from: [4])

Auditory Masking: 掩蔽效应

- **时域掩蔽 / 异时掩蔽 / Non-simultaneous Masking**
  - 后向掩蔽: post-masking
    - 所影响的时间较长，根据掩蔽音声压级及时长的不同，后向掩蔽的持续时间约为50-300ms
  - 前向掩蔽: pre-masking
    - 所影响的时间较短，仅约5-20ms左右



(from: [4])

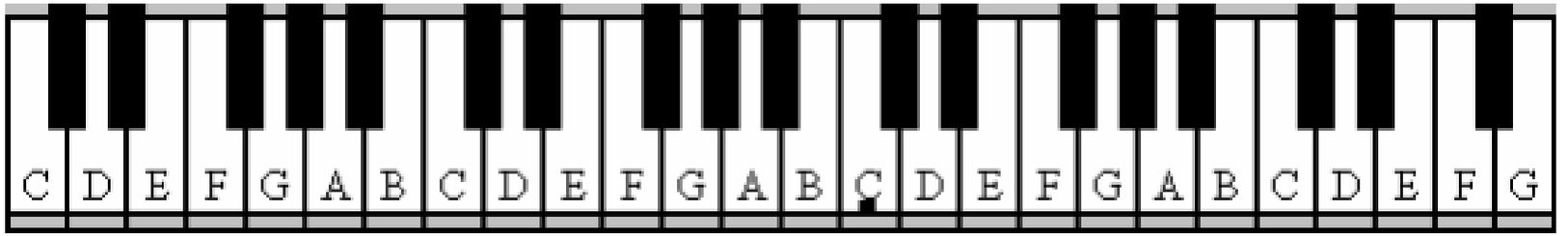
Musical Pitch: 音乐音高

- **Semitones: 半音**
- **12音阶**
  - 每个全音阶 (Octaves，8度) 包含12个半音 (semitones)，对应键盘上的七个白键和五个黑键
- **每向上相隔一个全音阶，频率会变成两倍**
  - 例如：中央la音是440Hz (69 semitones)，向上平移一个全音阶之后，频率就变成880Hz (81 semitones)。$f_1 = 2f_2$
  - 由基本频率转换成半音的公式如下：
    - semitone = 69 + 12 * log2(frequency / 440)

Musical Pitch: 音乐音高

Pitch Name(音名)

每一個 tone(樂音) 都有一個名字，稱爲 “pitch name(音名)”。以下是鋼琴上白色鍵的 pitch name(音名)：



在上圖中，所有白鍵加起來只有 7 個 pitch names(音名)。從 A 音 到 G 音，就不斷重覆。而任何兩個同名的鍵之間的距離，都叫做一個 “octave(8 度)”。例如，從一個 A 音 到下一個 A 音，相距了一個 “octave(8 度)”。

Interval(音程)

“Interval(音程)” 是兩個不同的音之間的距離。在白鍵上， B 和 C 、 E 和 F 是挨緊著的；而在 C、D 和 E ， F、G、A 和 B 之間卻隔著一個黑鍵。通常，兩個緊挨著的 tones(樂音) 之間的距離稱爲 “semitone(半音)”，把中間隔著一個鍵(不論是黑鍵或是白鍵)的兩個 tones(樂音) 之間的距離稱爲 “tone” or “whole tone (全音)”。

Q&A