第一章 语音产生与感知机理

天津大学 智能与计算学部
王 龙标
longbiao_wang@tju.edu.cn

课程信息

- 课程名称：语音信息处理
- 课堂讲授：24学时
- 上机：16学时
- 成绩考评
  - 平时成绩（作业、出席）：10%
  - 上机实验：30%
  - 考试：60%（选择题、判断题？、填空题、简述题、算法题、综合题）

扫码进本课程群

- 后续实践课程：《语音与语言理解综合实践》

建议参考书

- 《语音信号处理（第3版）》 清华大学出版（2019年出版）韩纪庆、张磊、郑铁然 编著

- 《听觉信息处理前沿研究》 上海交通大学出版社（2021年出版）党建武、俞凯编

- 《语音信号处理理论与实践》高等教育出版社（2023年出版）王龙标主编；党建武、于强编

建议其它参考书

- 《解析深度学习：语音识别实践》 电子工业出版（2016年出版）主编：俞栋；译者：俞凯/钱彦旻

- Rabiner, L., & Juang, B-H. (1993) Fundamentals of Speech Recognition, Prentice-Hall.

- Rabiner, L., & Shafer, R. (2010) Theory and Applications of Digital Speech Processing, Prentice Hall.

其它资料

- 《语音信息处理》实验教学方案
- 《语音信息处理》实验环境准备和操作指南
- 《语音信息处理》课件

本课程的学习目标

- 从语音产生、感知和信号处理三个方面研究和理解人类语音的本质及其属性
- 语音的本质属性
  - 物理属性：语音是由语言信息调制的声波
  - 社会属性：语音是传达人类意图的重要方式之一（称为言语）
  - 科学属性：语音是言语产生和感知机理相互作用的媒介
- 两种研究方法:
  - 科学方法: 回答为什么（产生和感知的机理）
  - 工程方法: 回答怎么样（识别和合成的方法）
- 本课程始终融合和贯通两种研究方法

本课程的学习内容

- 语音产生与感知机理
- 语音信号处理的基础
- 语音识别
- 语音合成
- 语音增强
- 声纹识别
- 语音信息处理实践

本课程的教学日历

- 讲课：语音产生与感知机理（9.11）
- 讲课：语音信号处理的基础（9.12、9.14）
- 上机：语音特征提取算法实现（9.19）
- 讲课：语音识别（9.21、9.25、9.26）
- 上机：语音识别系统实现及验证（9.28、10.7、10.9）
- 讲课：语音合成（10.10、10.12）
- 讲课：语音增强（10.17、10.19）
- 上机：语音增强算法实现（10.23、10.24）
- 讲课：声纹识别（10.26、10.31（兼整体复习））
- 上机：声纹识别系统实现及验证（11.2）
- 期末考试：11.19 10:00-12:00（考场：46-A209）
- 实验课教室：（上午）47教第二机房
  （下午）47教第八机房

第一章　语音产生与感知机理

- 1.1 语音产生的机理
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

第一章　语音产生与感知机理

- 1.1 语音产生的机理
  - 1.1.1 有声语言（语音）的形成（了解）
  - 1.1.2 语音的发音器官及其发音机理（基本概念）
  - 1.1.3 发音运动及其范畴化（基本概念）
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

声音与语音

- 声音(Sound)的定义（维基百科）

  声音是振动产生的声波，是通过介质（空气、固体、液体）传播并能被人或动物听觉器官所感知的波动现象。

- 语音（Speech）的定义

  语音是声音的一种，是人在言语交流过程中由发音器官产生和由听觉系统感知的语言载体。因此，语音信息处理的研究，长期以来一直是围绕着人的发音（产生）和听觉（感知）机理、及相关的语言学概念展开的。

注：当我们考虑有声语言（Spoken language）的“语言信息”时，用“言语”的称谓，当把它作为物理信号进行处理时，称之为“语音”。在不引起混淆的情况下，一般不做严格地区分。

声音的产生、传播与感知

- 声音的产生

  声音是由振动产生的，它引起介质（如空气等）粒子的简谐（正弦规律）运动，形成波动。

- 声音的传播

  空气分子的简谐振动，使周围的空气产生疏密变化，形成疏密相间的纵波，进行传播。

- 声音的感知

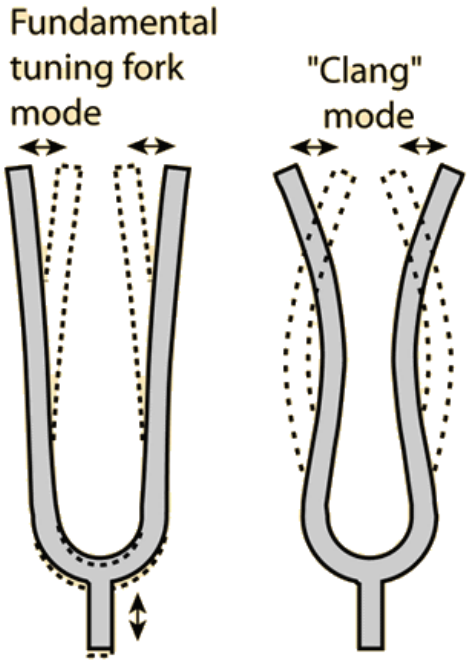
  耳蜗将中耳传来的机械运动转为淋巴液的波动，所含频率在基底膜对应位置引起谐振将其分解并传入大脑进行感知。语音感知过程会融入其他知识对声音进行综合判断。

声音的产生（基本概念）

- 两端关闭或开放的声管、两端固定的弦的谐振频率 $f_n = \frac{nc}{2l}$

f_n为第n次谐振频率，l为管长，c为音速。

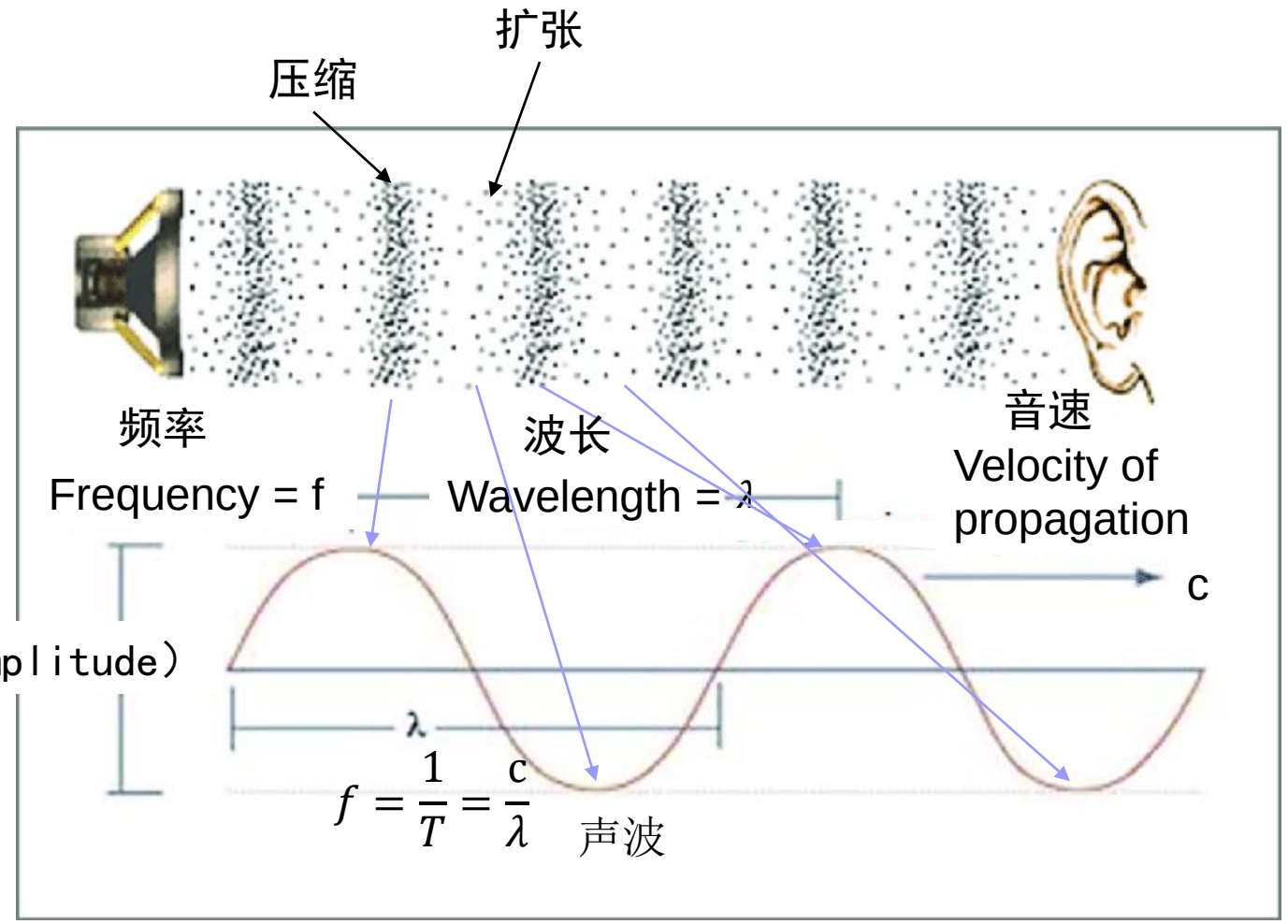
谐振频率现象是指一物理系统在特定频率和波长下，比其他频率和波长以更大的振幅做振动的情形。



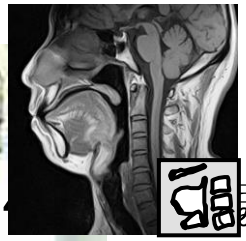
- 一端关闭另一端开放的声管或一段固定另一端自由的谐振频率

$$f_n = \frac{(2n-1)c}{4l}$$

声音的传播


扩张
压缩
频率
Frequency = f
波长
Wavelength = λ
音速
Velocity of propagation
c
振幅（Amplitude）
λ
$f = \frac{1}{T} = \frac{c}{\lambda}$
声波

语音及语言的形成



书面文字
（Written language）

书写语言大约在[illegible]完成

象形文字

象形文字大约出现在一万年前

有声语言
（Spoken language）

语音大约出现在4万到5万年前

手势语言

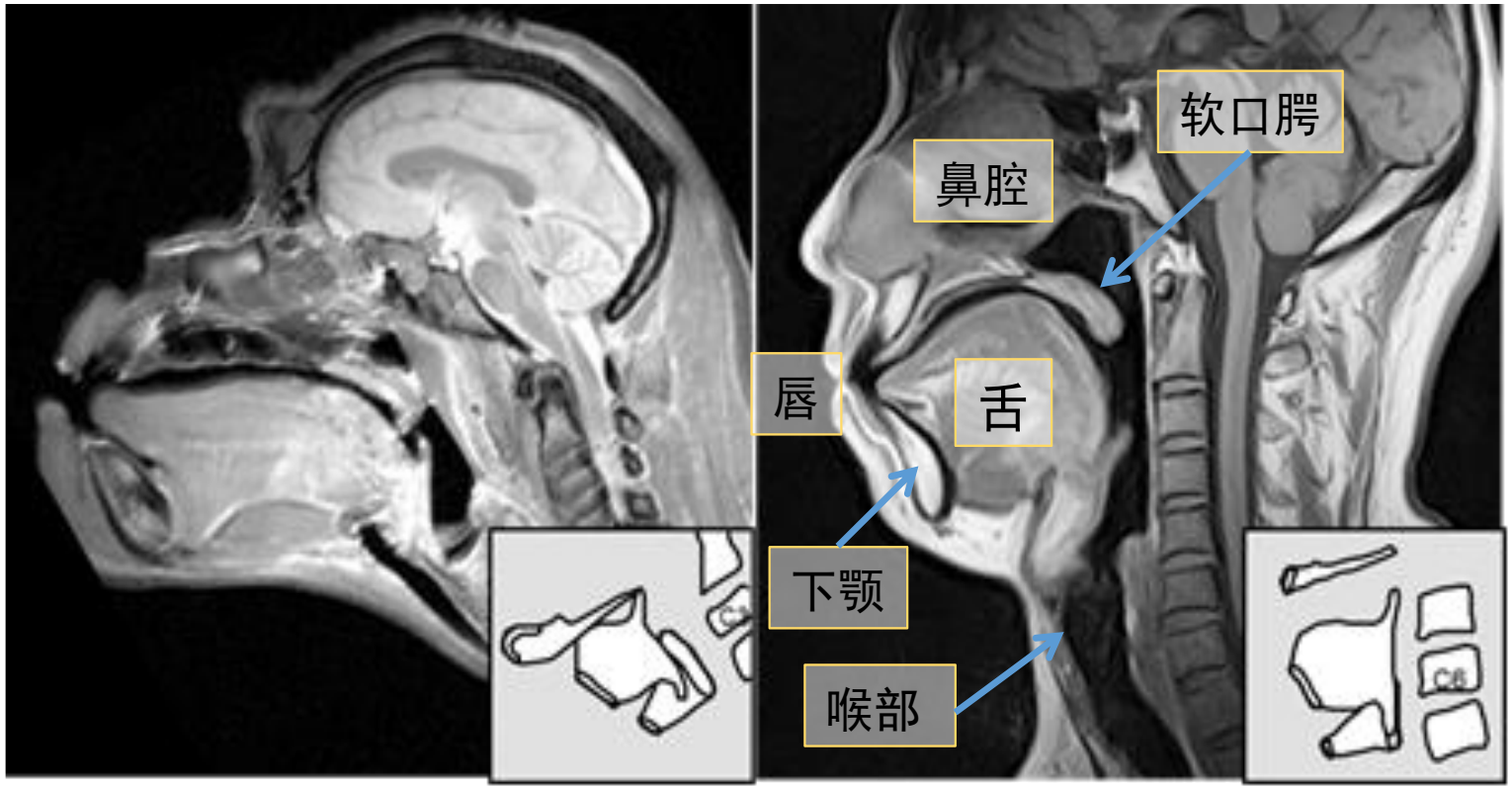
出现在人类开始站立，双手解放

使用声音和动作进行交流对人类和动物来说是很常见的

发音器官在形态学上的进化

类人猿：面部突出，舌头扁平，喉部气囊，舌骨和喉粘结

人类：口腔短，咽腔长，弯曲声道，舌骨和喉分离



(Hayama, et al., 1996)

第一章　语音产生与感知机理

- 1.1 语音产生的机理
  - 1.1.1 有声语言的形成
  - 1.1.2 语音的发音器官及其发音机理
  - 1.1.3 发音运动及其范畴化
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

语音的发音器官

- **发音器官**

肺、气管、喉（包括声带）、咽、鼻、口。它们共同形成一条形状复杂的管道

- **声带（Vocal Cords）** 10~14mm

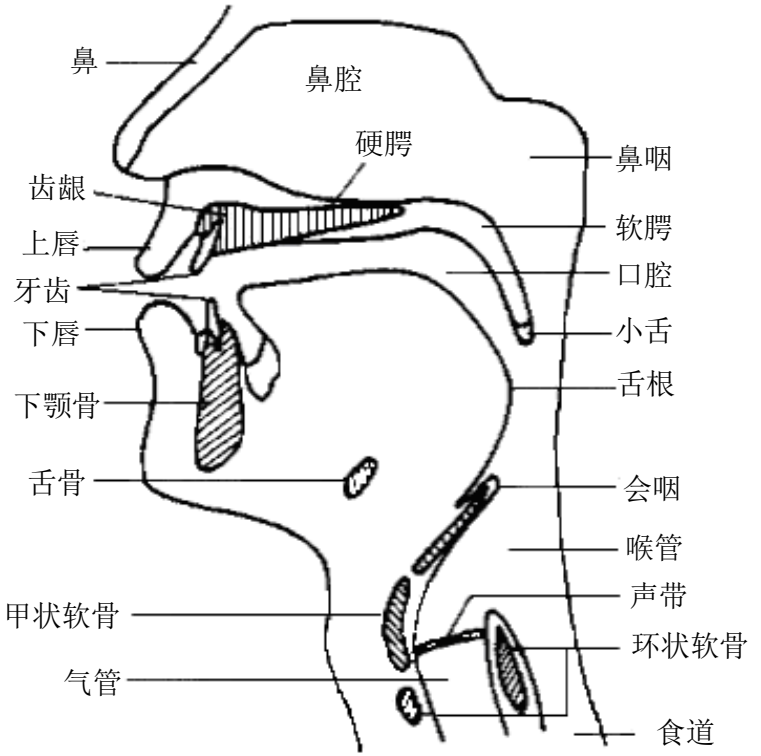
在喉部的从喉结到杓状软骨之间的韧带褶

- **声门（Glottis）**

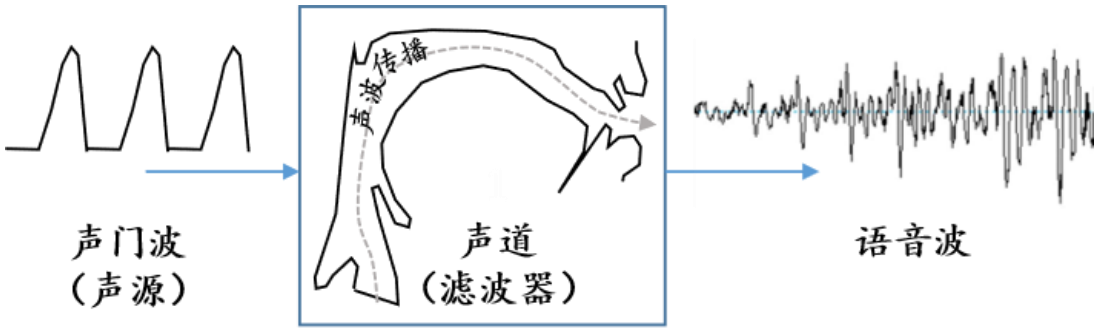
两个声带之间形成一个开闭自如的声门

- **声道（vocal tract）**

由咽腔、口腔和鼻腔三个空气腔体组成



语音产生的原理图



- 人的发声过程包括两个步骤
  - 声门/声带产生不同频率的声音
    准周期气流脉冲或白噪声
  - 声道对声源的调制作用

声源（激励源）的产生

- 声源
  - 肺部产生气流，通过猴头时冲击声带，使声带产生振动
  - 能量：绝大多数来源于正常呼吸时肺部呼出的稳定气流

- 声带
  - 既是一个阀门又是一个振动部件
  - 呼吸时左右两声带打开（声门开）
  - 在说话的时候合拢，肺部气流经气管形成冲击“打开-闭合-打开-闭合-…”声门，从而冲击声带产生振动，然后通过声道响应变成语音

声道调制

- 声道
  - 咽、口腔和鼻腔
  - 从声门延伸至口唇的非均匀截面的声管，约17cm（成年男性）

- 功能
  - 谐振腔：放大某一频率而衰减其他频率分量
  - 谐振频率/共振频率（formant frequency）：当激励的频率达到它声道的固有频率，则声道会以最大的振幅来振荡（共鸣）。由每一瞬间的声道外形决定，又称为共振峰，是声道的重要声学特征

调音

- 调音（Articulation）

  为了发出各种各样的声音，需要调整声道的形状，称之为调音

- 调音运动（Articulation Movement）

  声道各部分的动作

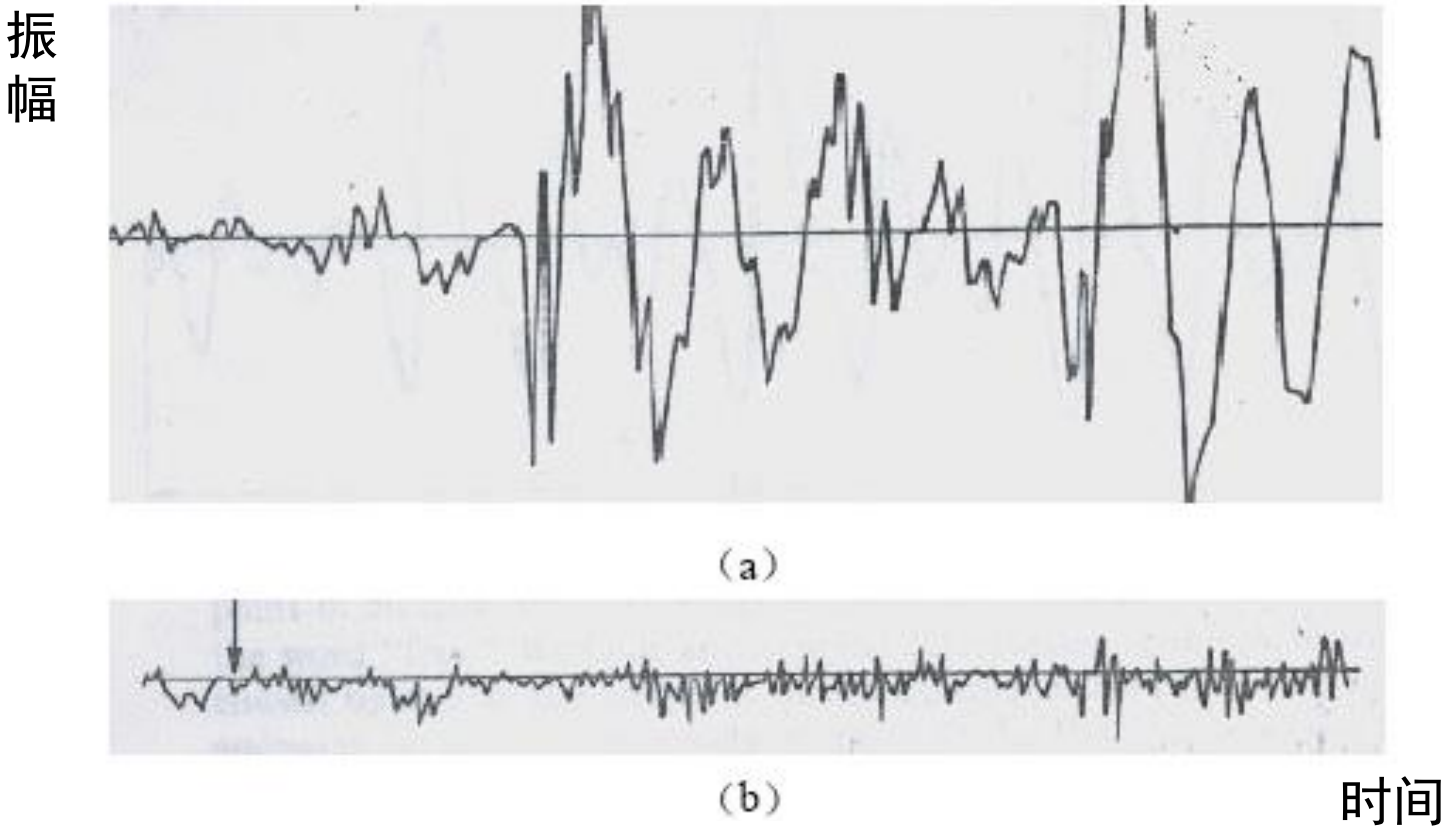
- 调音器官（Articulation Organ）

  调音用的声道的各部分器官，包括舌、颚、唇和嘴等声道中可以自由活动的部分

浊音、清音与爆破音

- 浊音（voiced sounds）：声道打开，声带在先打开后关闭，气流经过使声带要发生张驰振动，变为准周期振动气流。浊音的激励源被等效为准周期的脉冲信号。
  - 如发/U/、/d/、/i/等音
- 清音（unvoiced sounds）：声带不振动，而在某处保持收缩，气流在声道里收缩后高速通过产生湍流，再经过主声道（咽、口腔）的调整最终形成清音。清音的激励源被等效为一种白噪声信号。
  - 如发/ʃ/音
- 爆破音（plosive sounds）：声道关闭之后产生压缩空气然后突然打开声道所发出的声音。
  - 如发/ tʃ/音时

浊音与清音的波形



（a）浊音波形；　（b）清音波形

语音的基本特性

- 振幅（Amplitude）
  - 波振动的大小，一般用dB表示（声压级的标准测量单位）
- 基音周期（Pitch Period）
  - 声带开启-闭合一次的时间
- 基频（Pitch Frequency: F0）
  - 基音周期的倒数，声带振动的基本频率
- 音调
  - 声带振动的频率（即基音）决定了声音频率的高低，频率快则音调高，否则音调低
  - 人的基音范围
    - 80~500 HZ，儿童和青年女性偏高，成年男性偏低

第一章　语音产生与感知机理

- 1.1 语音产生的机理
  - 1.1.1 有声语言的形成
  - 1.1.2 语音的发音器官及其发音机理
  - 1.1.3 发音运动及其范畴化
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

发音运动指令的范畴化

- 舌头大约有8000个运动神经元单位

- 语音的发音姿势需要满足一定的条件
  （1）发音姿势要简单易学；
  （2）发音姿势要能够产生稳定的声学特性

- 人类的语音发音运动被范畴化和符号化，而动物（鸟类除外）的叫声无法符号化；每一个范畴声学特征对应一组范畴化的运动指令

- 发音运动指令的范畴化：一个相对独立的语音单元，即使其发音姿势在一定范围内变动，其声学特性应该稳定不变。

音素及音位的基本概念

- 人类有大约7000种语言，包括约200个元音和600个辅
  - 音素（Phone）：被范畴化成为生成上可区分的最小语音单元
  - 区分标准：物理和生理属性
- 但是，对于一种特定的语言，大约有40个音素
- 若干相近的音素归类为一个音位（Phoneme）
  - 音位：听觉上可以区分的最小语音单元
  - 区分标准：社会和语言属性

比如，汉语拼音中的z、c、s和zh、ch、sh在普通话中是两组不同的音位，但在南方许多方言中它们是同一组音位

语音生成上的最小语音单元：音素

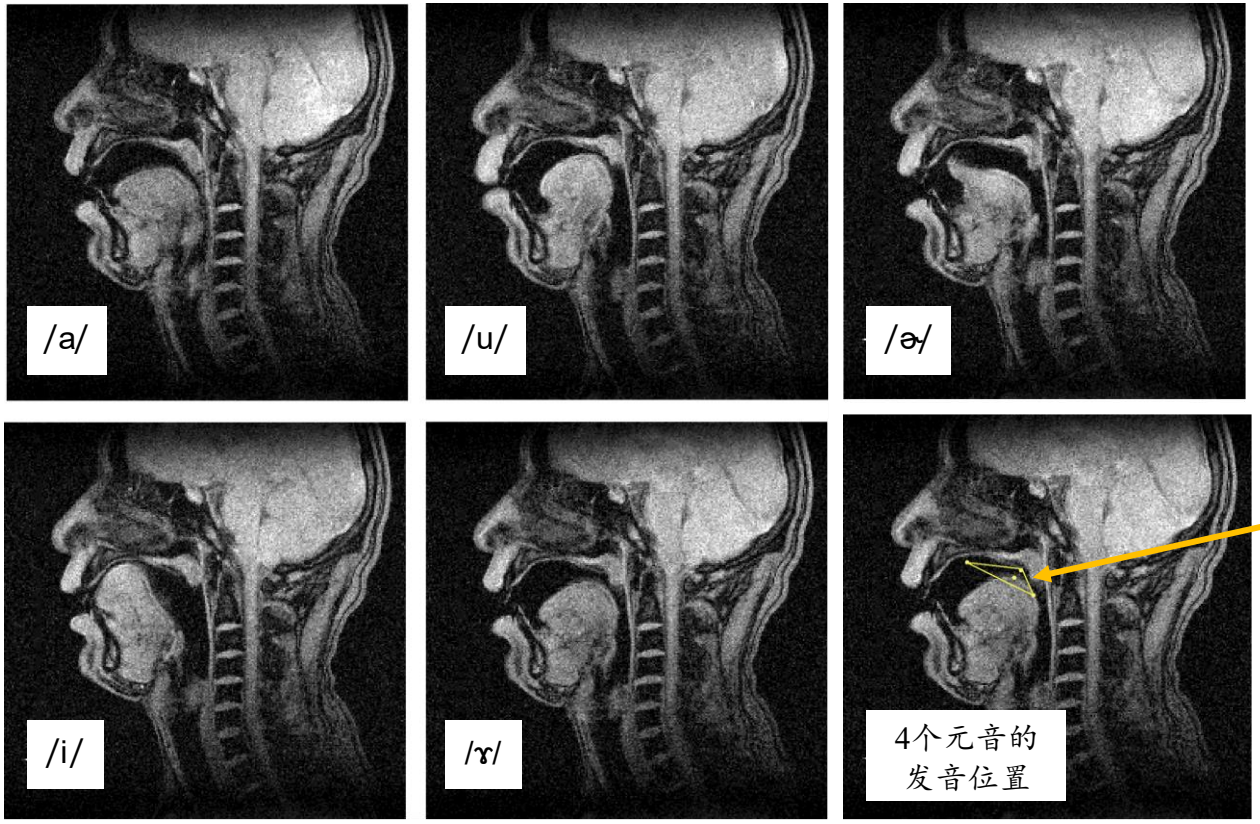
元音(vowel)：声音由相对开放的声道结构产生的语音，有声带的振动，但没有可听见的摩擦音。

辅音(consnant)：由于发音器官的形变而使气流（声波）在声道部分或完全受阻而产生的语音。

元音与辅音的可能组合

一个语言单元由单个的不间断声音组成。该声音可由单元音，双元音，或元音化辅音单独形成，或者以这些音为中心先后由一个或多个辅音包围而形成。

汉语基本元音的磁共振(MRI)成像（了解）



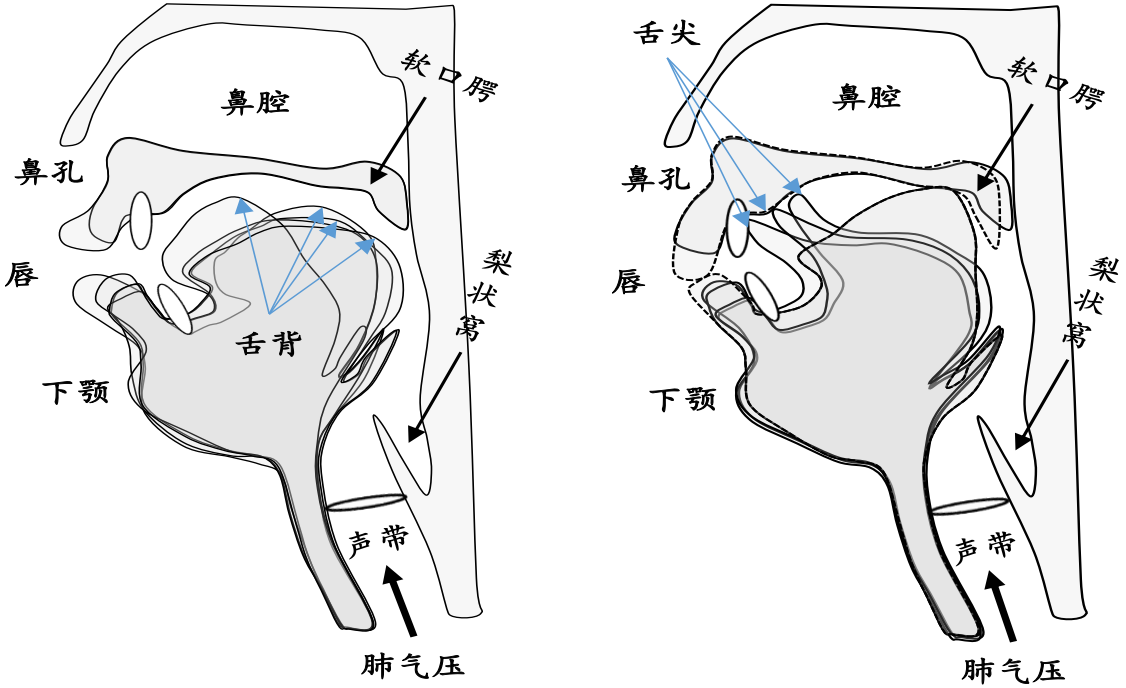
/ɚ/为汉语拼音/er/

汉语的元音三角形/a, i, u, e/

舌体运动为主、唇部变形为辅

IPA符号/ɤ/为汉语拼音/e/

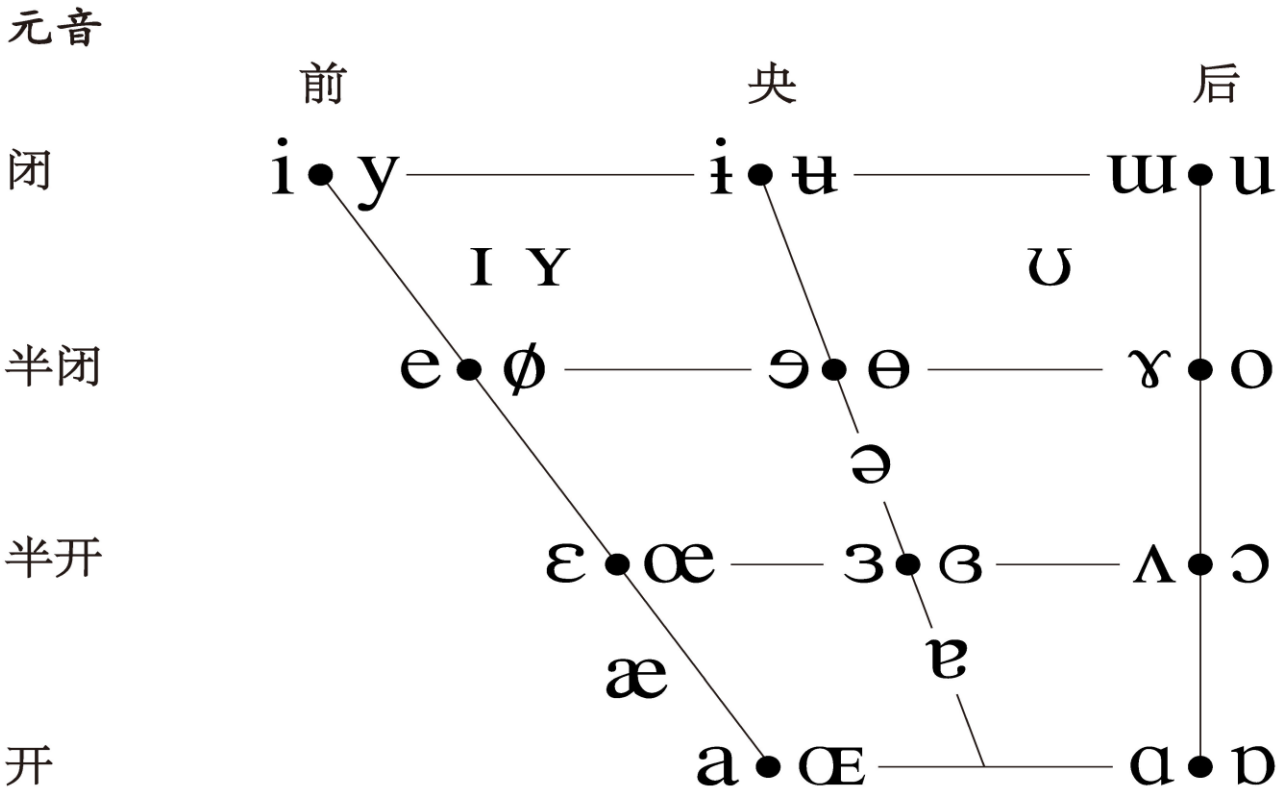
从MRI中提取的舌头轮廓（元音）（了解）



（a）元音　　（b）辅音

- 元音的发音姿势，舌体前后上下大幅度运动，嘴唇有前突运动
- 可以用元音发音时舌背最高点（舌位）位置和唇部形状来粗略地描述元音的发音特征

国际音标学协会（IPA）定义的元音舌位（了解）



元音发音时声波从喉部到唇部辐射，中途不受阻碍。

汉语元音的拼音和国际音标（了解）

发音方式（唇部形状）	前			后	
	非圆唇		圆唇	非圆唇	圆唇
闭	-i [ɿ]/[ʅ]	i [i]	ü [y]		u [u]
半闭				e [ɤ]	o [ɔ]
半开		ê [ɛ]		er [ɚ]	
开		a [ä]			
	舌面元音	舌尖元音			

发音部位（舌位前后）

汉语拼音/zi, ci, si/和/zhi, chi, shi/的韵母

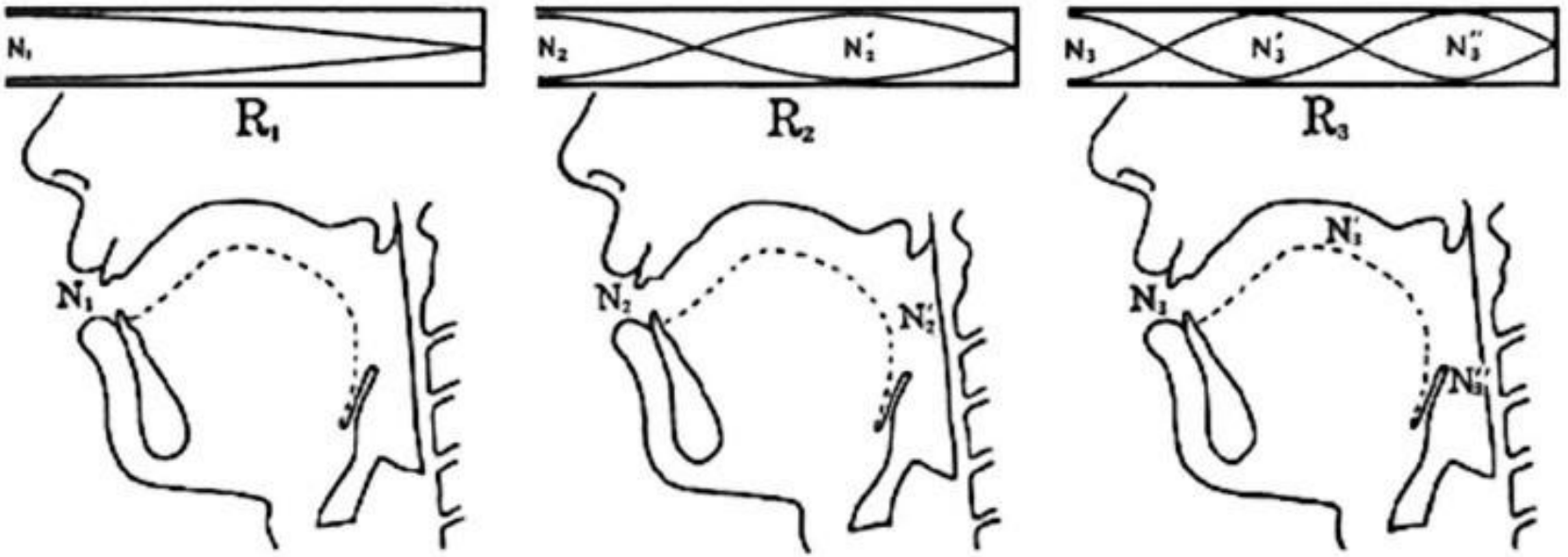
（汉字“二”的发音）是汉语特有的

横向表示发音部位，竖向表示发音方式。

元音的共振峰频率

共振峰=声道共振产生的波峰 （同时受声源特性的影响）

用单端封闭管中的驻波解释声道共振 $f_n = \frac{(2n-1)c}{4l}$

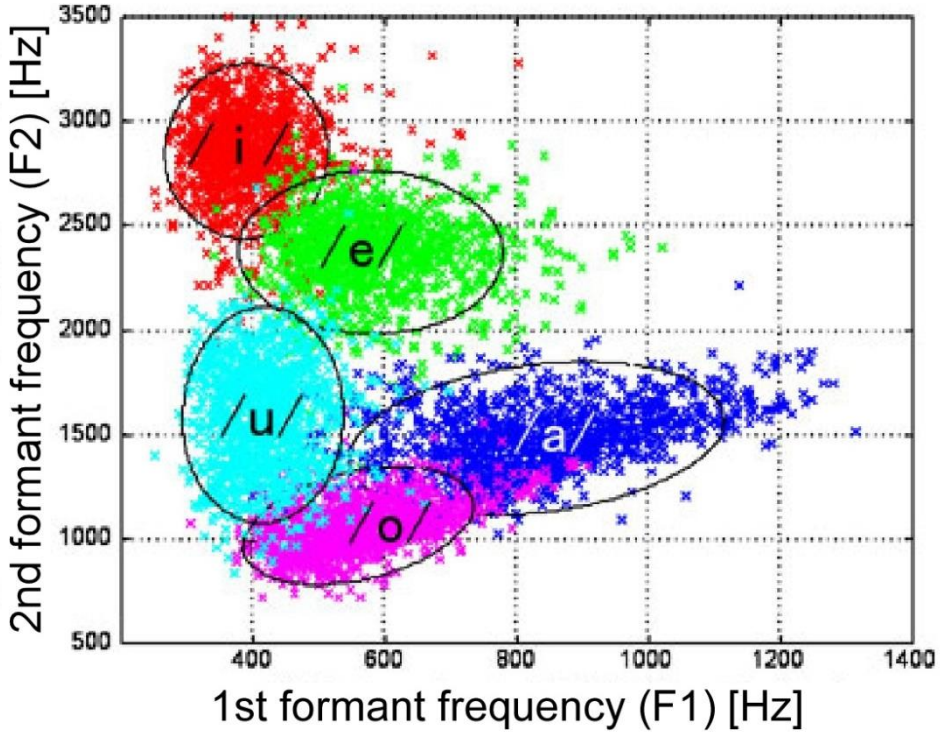


在均匀管(上图)和均匀声道(下图)中最大音量流点(N_*)的分布 （Chiba & Kajiyama， 1942)

辅音一般没有共振峰，除非半元音可以被认为有共振峰

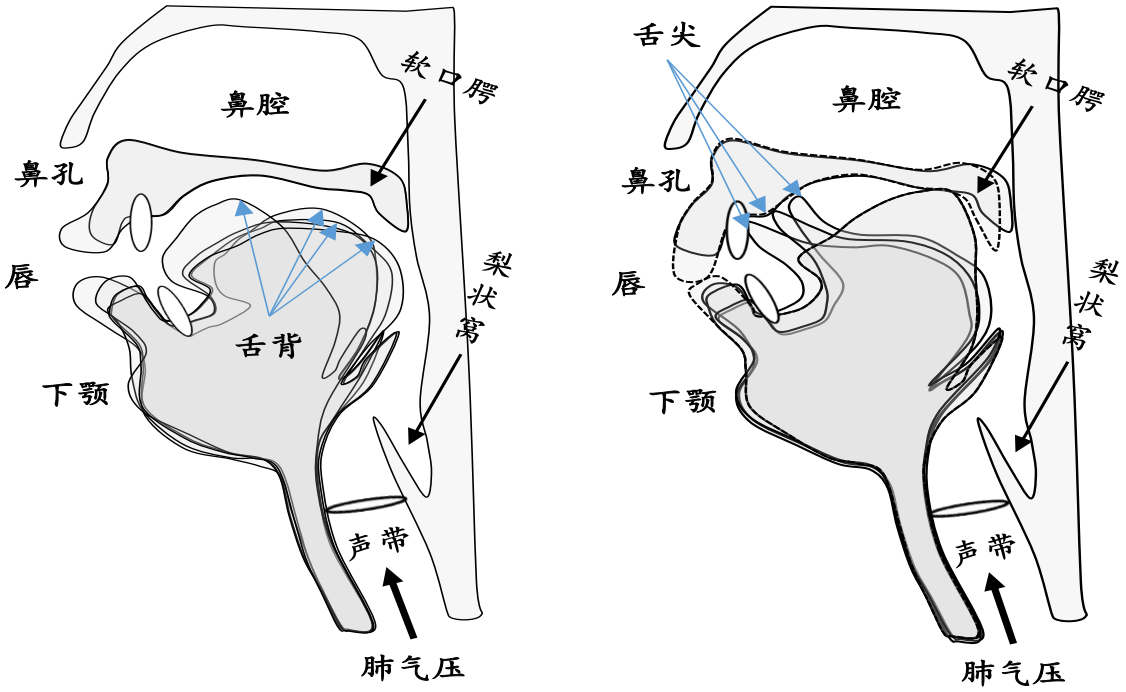
元音：F1-F2图

在声学上，元音的特征主要取决于共振峰频率(F1和F2)。



F1-F2平面图

从MRI中提取的舌头轮廓（辅音）（了解）



（a）元音　　（b）辅音

- 相对于元音，辅音的发音运动复杂得多
- 大部分辅音的发音和舌尖与舌前部相关
- 描述调音特性的元素：发音部位、发音方式和有无浊音源

辅音国际音标表（了解）

辅音（肺部气流）

	双唇	唇齿	齿	龈	龈后	卷舌	硬腭	软腭	小舌	咽	声门
塞音	p b			t d		ʈ ɖ	c ɟ	k g	q ɢ		ʔ
鼻音	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
颤音	ʙ			r					ʀ		
拍音或闪音		ⱱ		ɾ		ɽ					
擦音	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
边擦音				ɬ ɮ							
边音		ʋ		ɹ		ɻ	j	ɰ			
边近音				l		ɭ	ʎ	ʟ			

横向表示发音部位，竖向表示发音方式。每个辅音有发音位置和方式表示，为阴影显示此组合无法发音。每个格子中右侧的符号表示浊辅音，左边是清辅音。

汉语辅音的拼音和国际音标（了解）

发音部位 / 发音方法		双唇	唇齿	龈	齿	龈后	硬腭	软腭	
塞音	不送气	**b** [p]			**d** [t]			**g** [k]	清音
	送气	**p** [p^h]			**t** [t^h]			**k** [k^h]	
塞擦音	不送气			**z** [ʦ]		**zh** [ʧ]	**j** [ʨ]		
	送气			**c** [$ʦ^h$]		**ch** [$ʧ^h$]	**q** [$ʨ^h$]		
擦音			**f** [f]	**s** [s]		**sh** [ʃ]	**x** [ɕ]	**h** [x]	
						r [ʐ]			浊音
鼻音		**m** [m]			**n** [n]			**ng** [ŋ]	
边音					**l** [l]				

竖向表示发音部位，横向表示发音方式。每个辅音有发音位置和方式表示。括号外是汉语拼音符号，括号中是对应的国际音标符号

音节（Syllable）

音节是语言的最基本单位。它至少包括一个元音，有或没有辅音，构成一个词的整体或部分。

一个音节可分为两个部分：开头和韵脚。韵脚由一个音核及其后面的辅音组成。

英语音节 = 音节首+音节核（+音节尾）
汉语音节 = 声母+韵母(元音+韵母) （字的发音）
日语音节 = 音节首+音节核 （字或者子字的发音）

语音识别的建模单元：音素、音节、字等

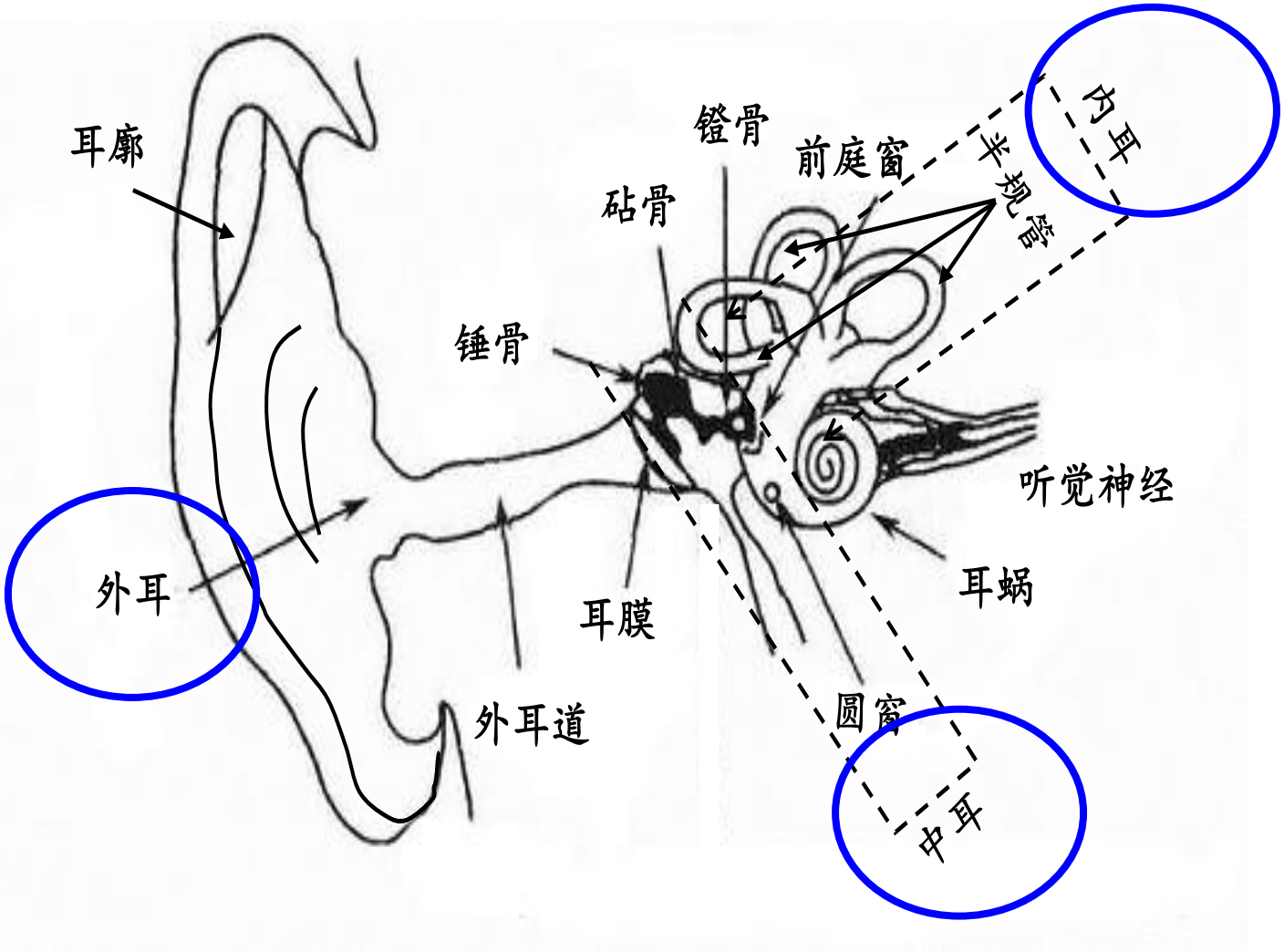
第一章　语音产生与感知机理

- 1.1 语音产生的机理
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

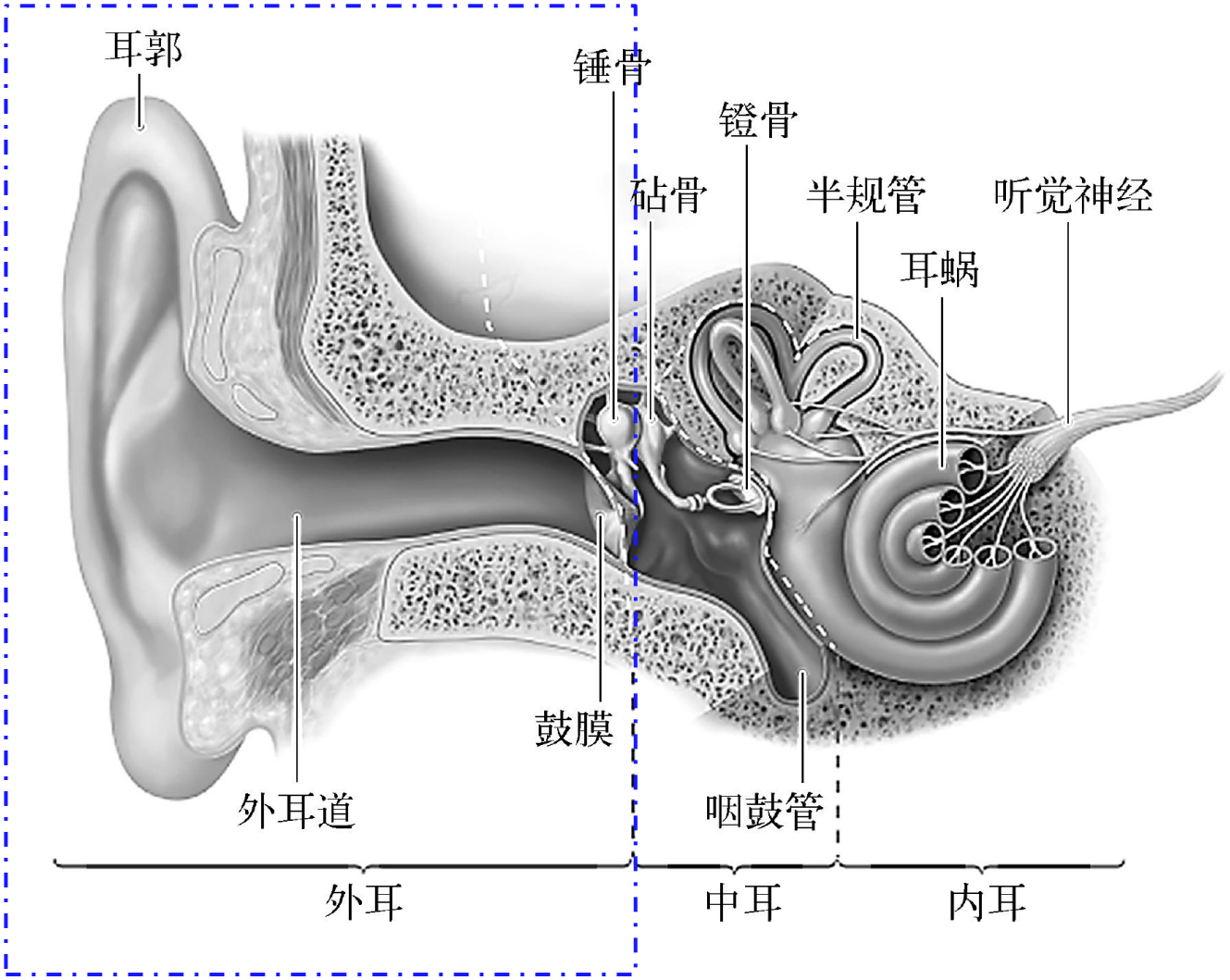
第一章　语音产生与感知机理

- 1.1 语音产生的机理
- 1.2 语音感知的机理
  - 1.2.1 听觉器官的构造及其功能（**基本概念**）
  - 1.2.2 听觉感知机理与听觉特性（**了解**）
- 1.3 语音产生与感知的相互作用

外耳、中耳和内耳的结构视图



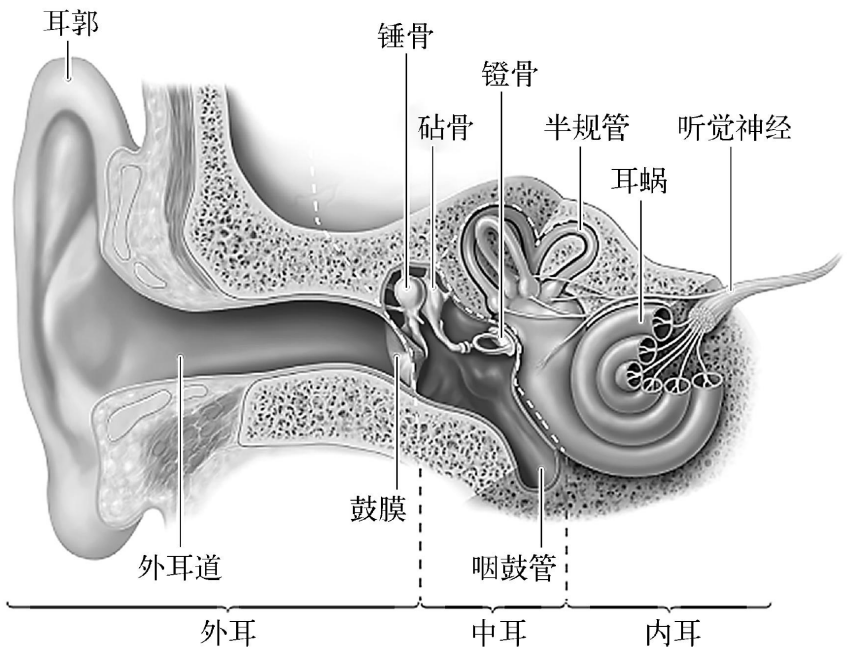
外耳、中耳和内耳的结构视图（续）



外耳(Outer ear)

耳廓的形状可使频谱峰压点在5.5 kHz的纯音提高10分贝(dB)的增益。

外耳道长约2.5 cm，
常温下音速约为350 m/s



外耳道的初次和二次谐振频率大约为多少Hz？

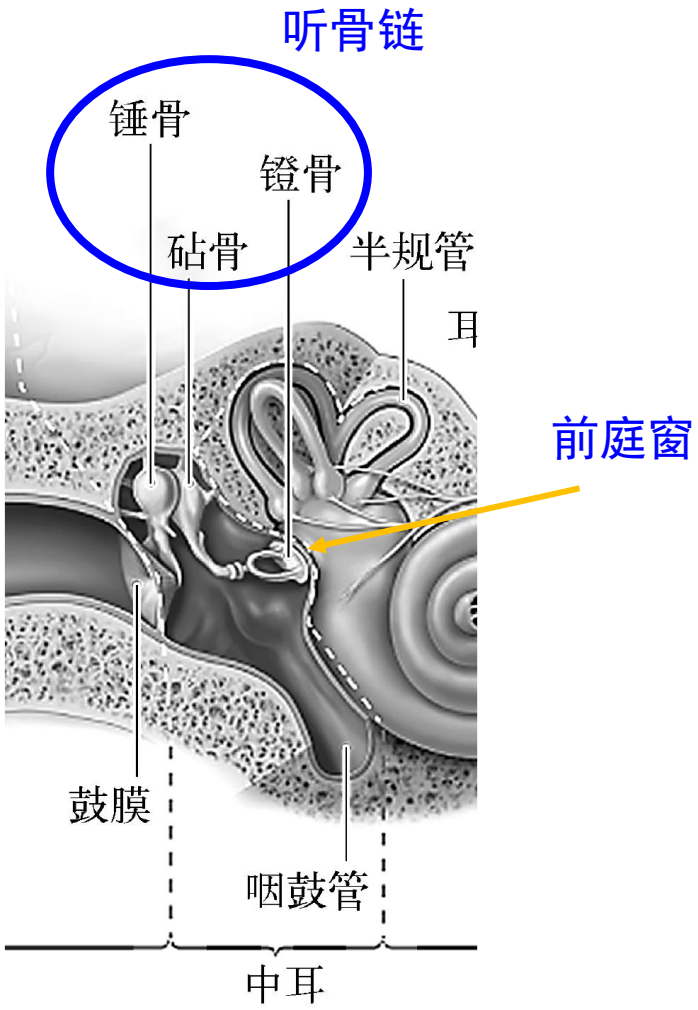
$$f_n = \frac{(2n-1)c}{4l}$$

初次谐振频率大约位于3.5 kHz附近，其2次谐振频率将达10 kHz前后。

- 外耳包括耳廓和外耳道，主要起集声的作用。
- 外耳道是声波传导的通路，一端开放，另一端由鼓膜封闭。

中耳(Middle ear)

- 中耳位于外耳和内耳之间，主要由鼓膜、听骨链、中耳肌和咽鼓管等结构组成。
- 声波从外耳道进入，作用于鼓膜，后者随之产生相应的振动。
- 鼓膜和听骨链形成一个力学系统，其功能是放大来自外耳的声波振动，并传输给内耳。



耳鼓膜振动传给锤骨，锤骨将振动经过砧骨和镫骨传递到耳蜗的前庭窗，中耳将空气振动转换为机械振动传递给内耳。

听骨链对500-2000 Hz的声波产生较大的谐振，起到带通滤波器的作用。

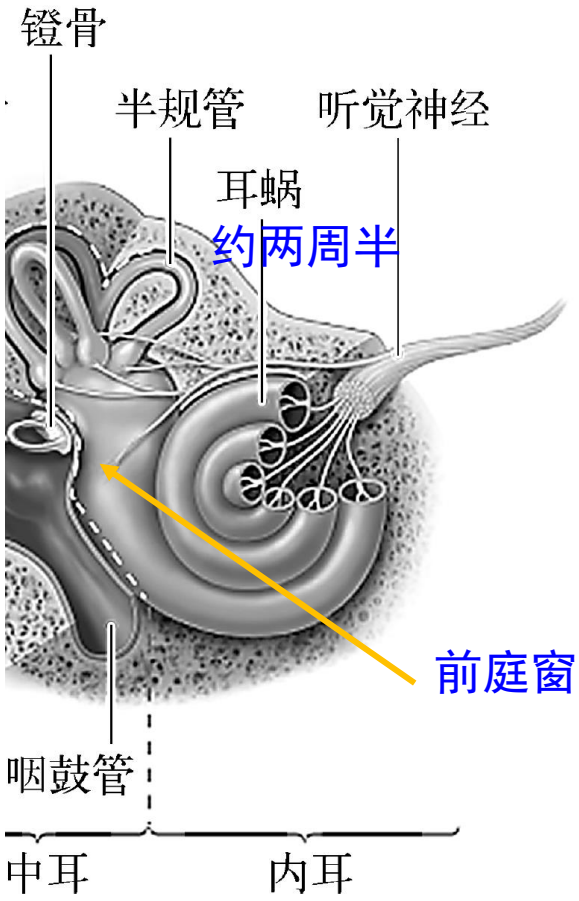
内耳(Inner ear)

- 内耳位于耳朵最深处，其最主要的结构是骨迷路，由耳蜗(Cochlea)和前庭系统构成。

前庭系统：平衡觉的末梢器官

耳蜗：听觉功能（将来自外耳和中耳的机械振动转换为神经电信号传递给大脑）。

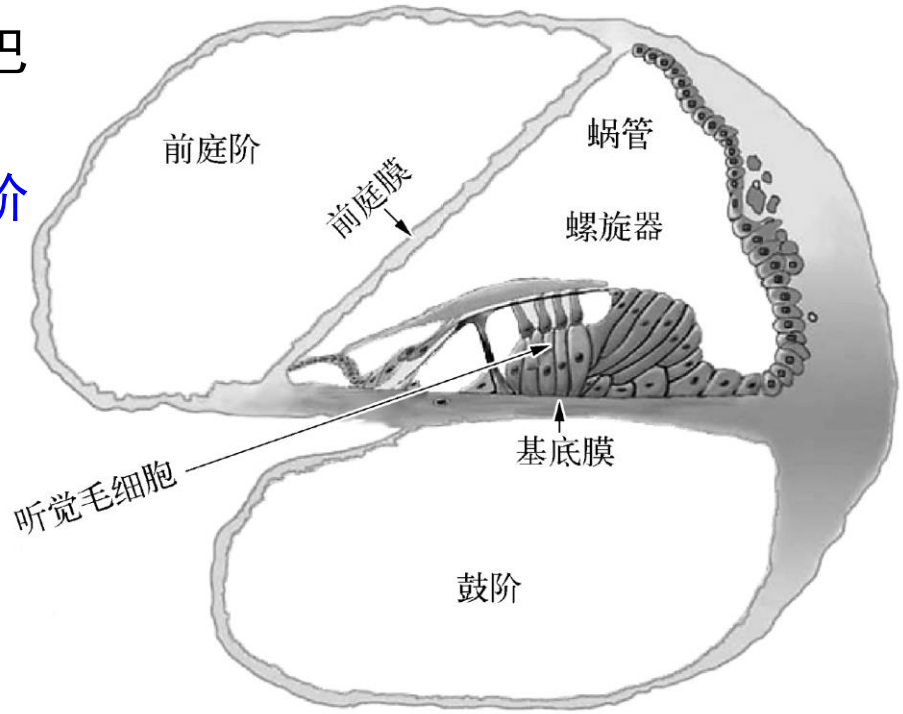
内耳兼有听觉和感受位置变化的双重功能



耳蜗横截面

耳蜗由三个充满淋巴液的空腔组成：

前庭阶、蜗管、鼓阶



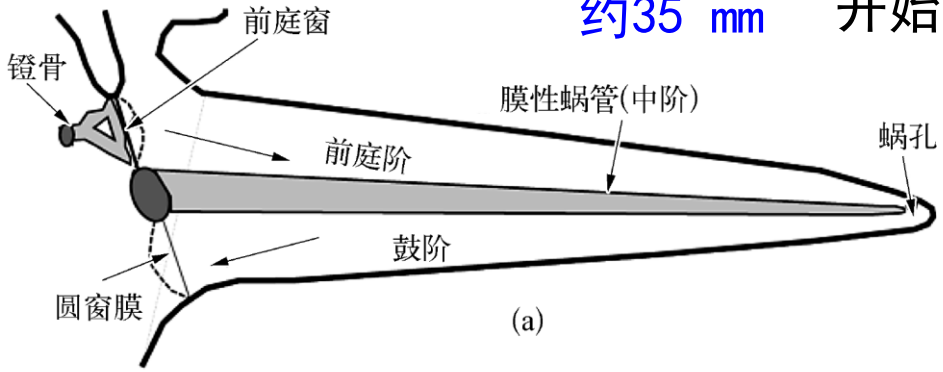
毛细胞损伤会导致听觉灵敏度下降，甚至导致感觉神经性耳聋等问题。

内毛细胞：将耳蜗内液体的声音振动转化为电学信号，并通过听觉神经传递到听觉脑干，再到听觉皮层。

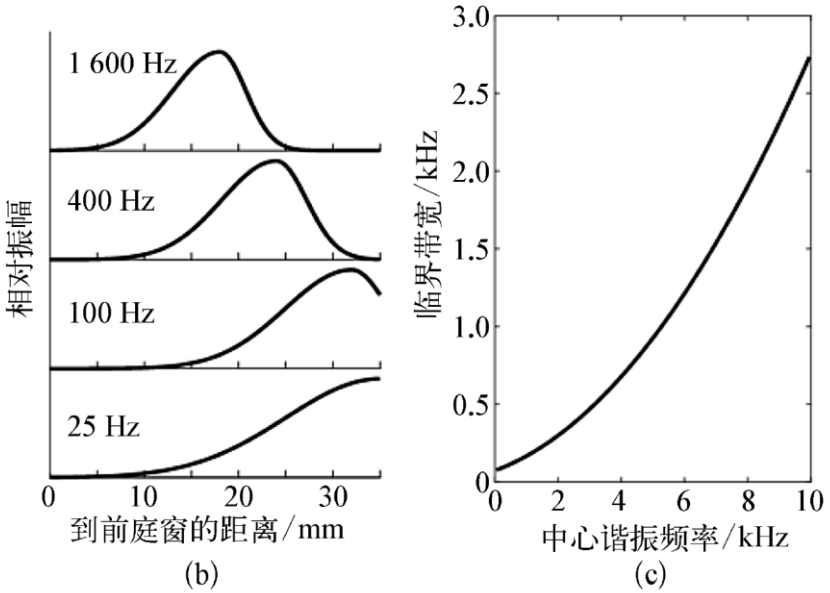
外毛细胞：机械地把传入耳蜗的低水平声音放大，也会对强度过大的声音可以进行能动的抑制。

耳蜗的声音分析功能

基底膜的波动从耳蜗基部开始，依次向蜗顶移动



(a) 耳蜗和基底膜的示意图



(b) 基底膜谐振位置与频率成分抽取的关系（滤波器组）

高频的频率分辨率急速下降。

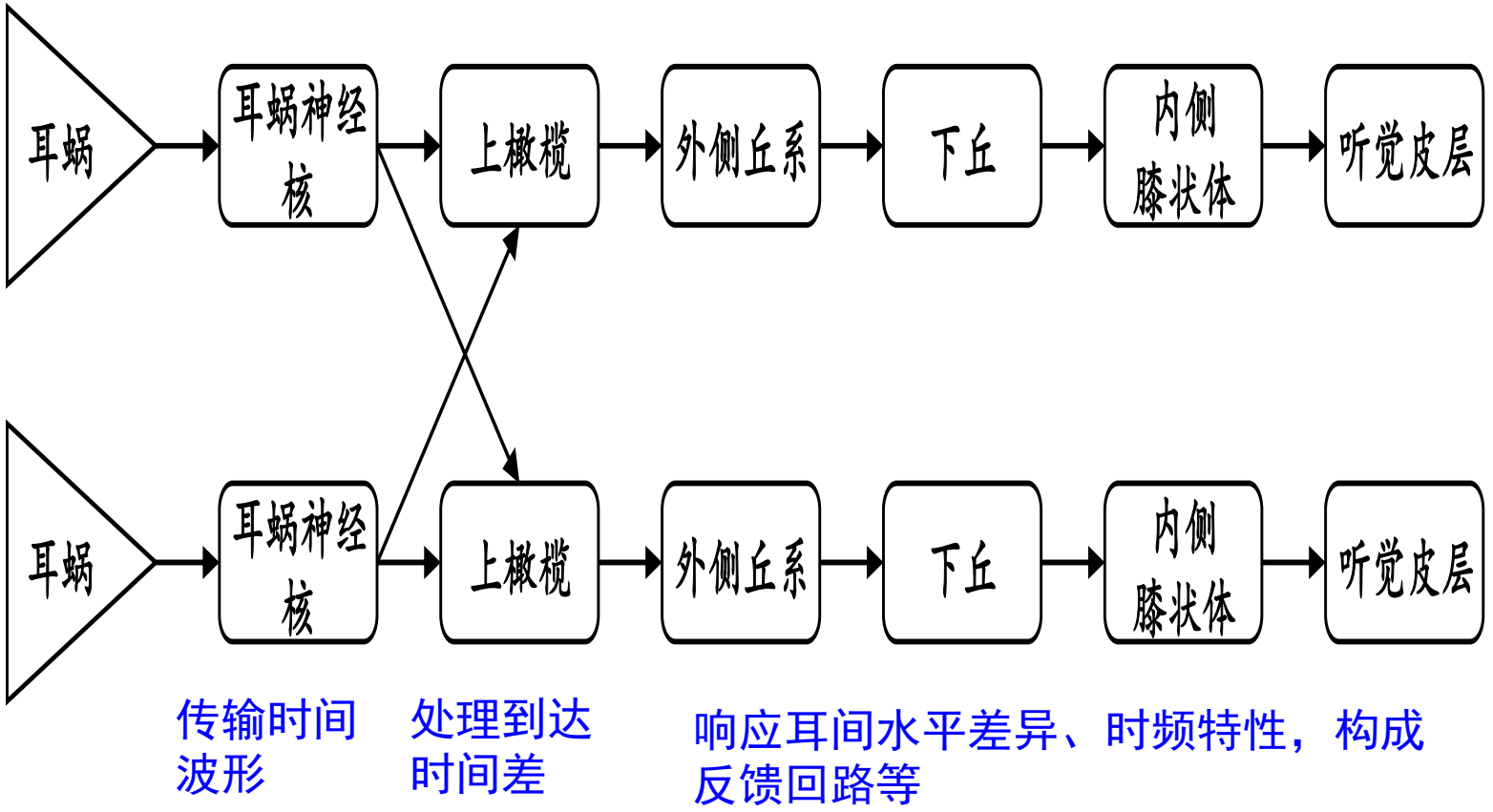
(c) 中心谐振频率和临界带宽的关系（巴克尺度）

人的听觉范围：20~20000 Hz，受基底膜谐振频率范围的制约。

第一章　语音产生与感知机理

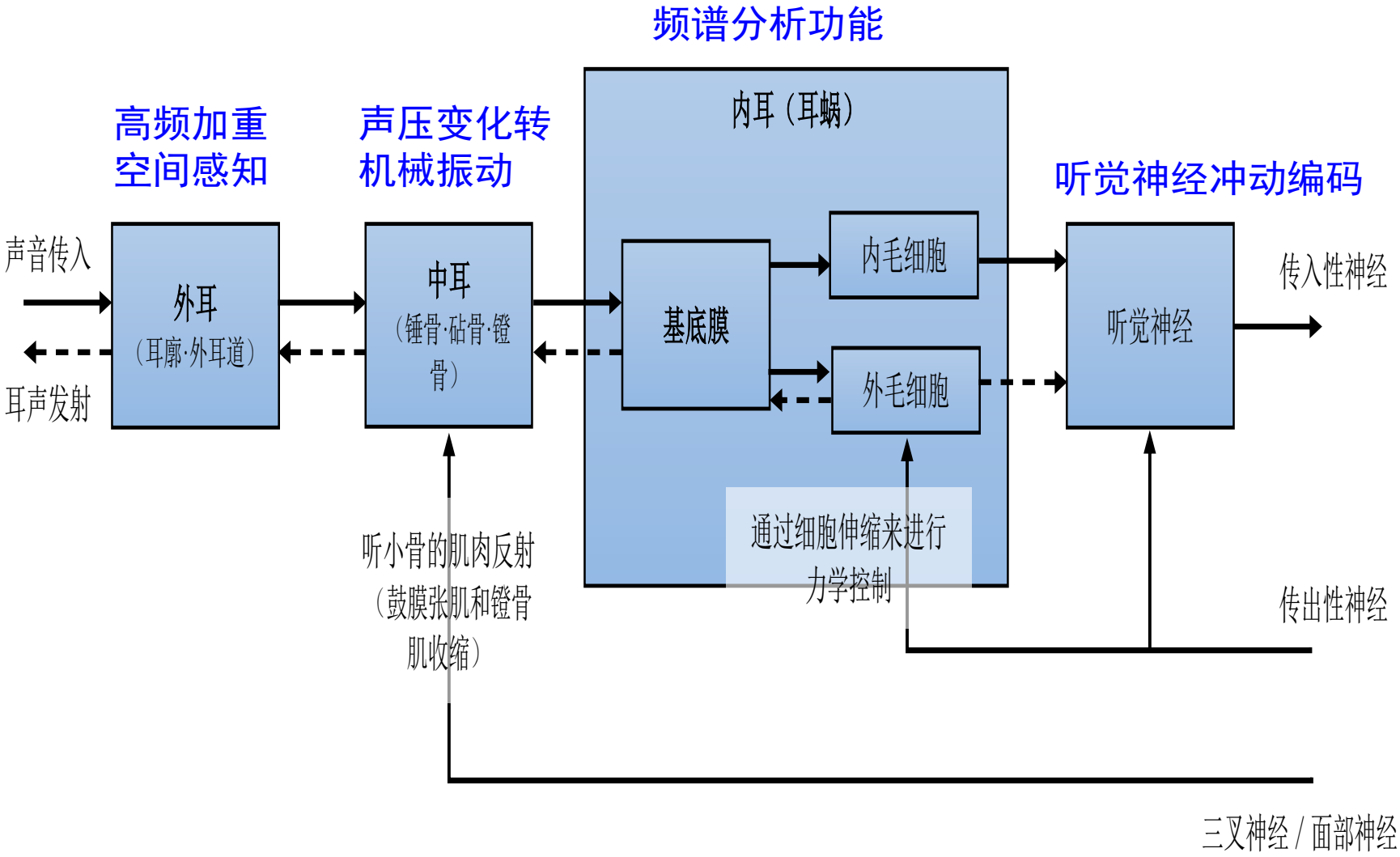
- 1.1 语音产生的机理
- 1.2 语音感知的机理
  - 1.2.1 听觉器官的构造及其功能
  - 1.2.2 听觉感知机理与听觉特性
- 1.3 语音产生与感知的相互作用

听觉传导路径



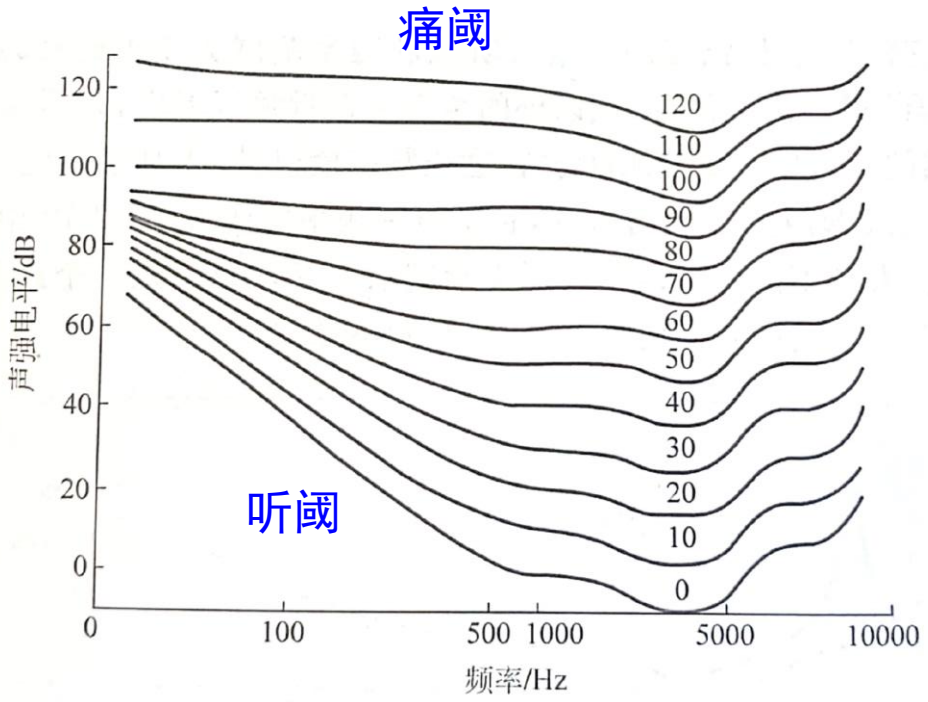
- 听觉中枢神经系统在以基膜上的听觉滤波器组的频率分析机制为基础上，在脑干中进行多级并行信息处理。

听觉末梢系统


频谱分析功能
高频加重
空间感知
声压变化转
机械振动
内耳（耳蜗）
听觉神经冲动编码
声音传入
外耳
（耳廓·外耳道）
中耳
（锤骨·砧骨·镫骨）
基底膜
内毛细胞
外毛细胞
听觉神经
传入性神经
耳声发射
听小骨的肌肉反射
（鼓膜张肌和镫骨肌收缩）
通过细胞伸缩来进行力学控制
传出性神经
三叉神经 / 面部神经

响度（基本概念）

- 响度（Loudness）
  - 响度：反映一个人主观感觉不同频率成分的声音强弱的物理量，单位为方（phon）。
  - 是一种主观心理量，主观感觉到的声音强弱的一种衡量标准，它与频率有关。一样的音强，不一样的频率，则响度也会有所不同。



等响度曲线

听阈：声音小到人耳刚刚能听见时的大小

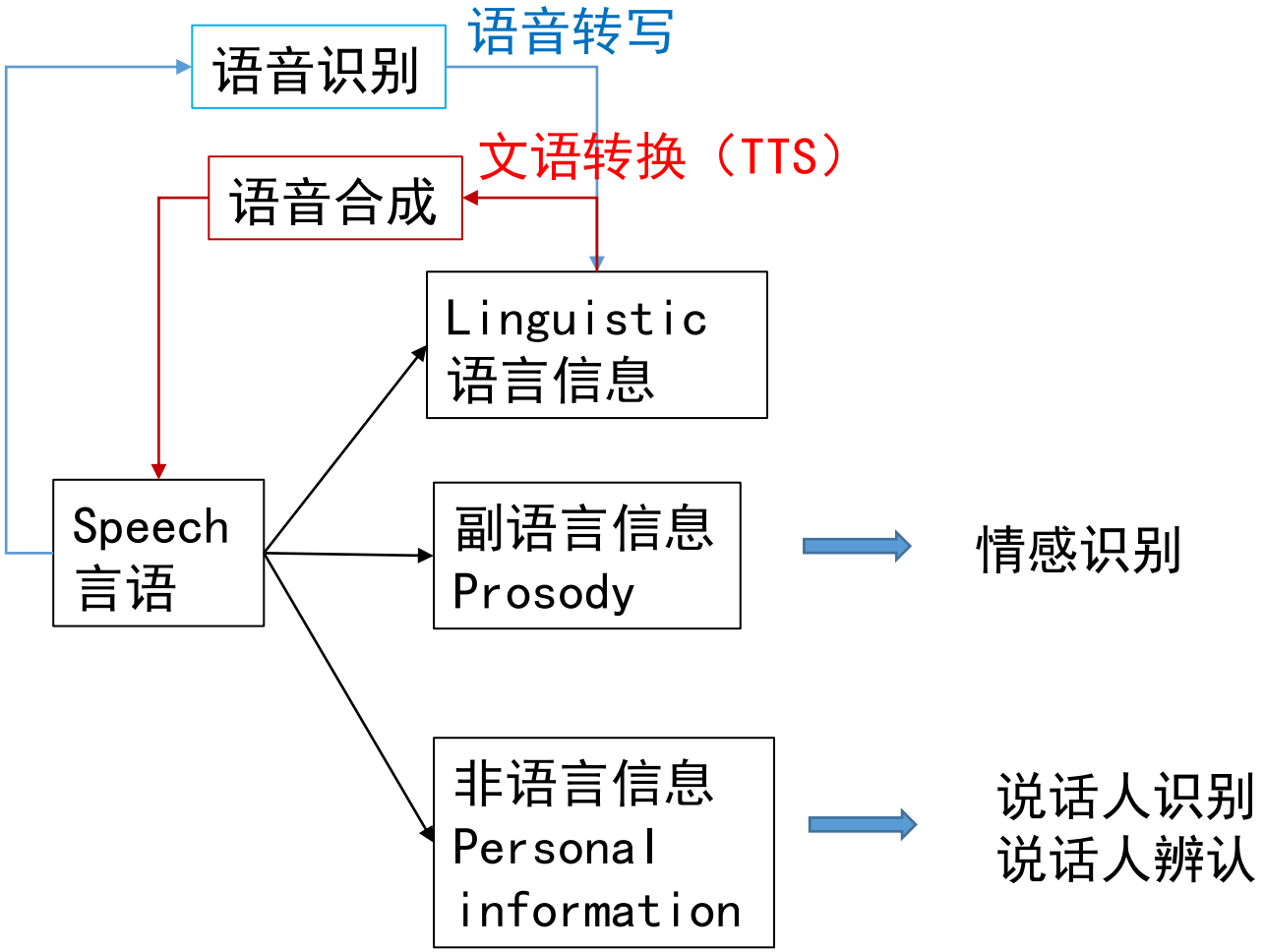
痛阈：声音大到人耳刚刚感觉痛时的大小

音调的主观（基本概念）

- 音调
  - 听觉分辨声音高低时，用于描述这种感觉的一种特性
  - 声带振动的频率（即基音）决定了声音频率的高低，频率快则音调高，否则音调低
  - 客观上用频率来表示音调，主观上感觉音调的单位时梅尔（Mel）尺度。梅尔频率与客观音频率 f 的转换关系为：

$$\mathrm{Mel}(f) = 2595 \log_{10}(1 + \frac{f}{700})$$

语音中语言、副语言和非语言信息的感知（基本概念）



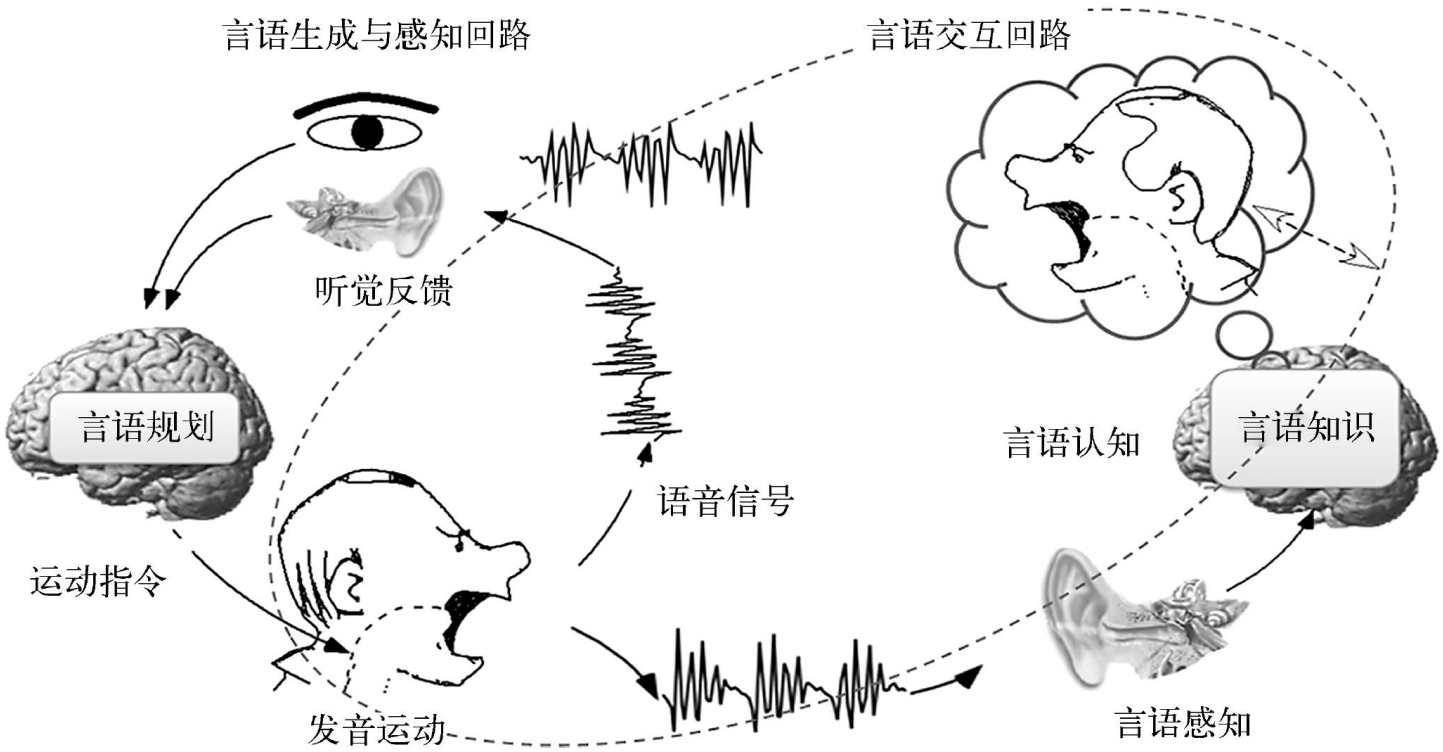
第一章　语音产生与感知机理

- 1.1 语音产生的机理
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用

第一章　语音产生与感知机理

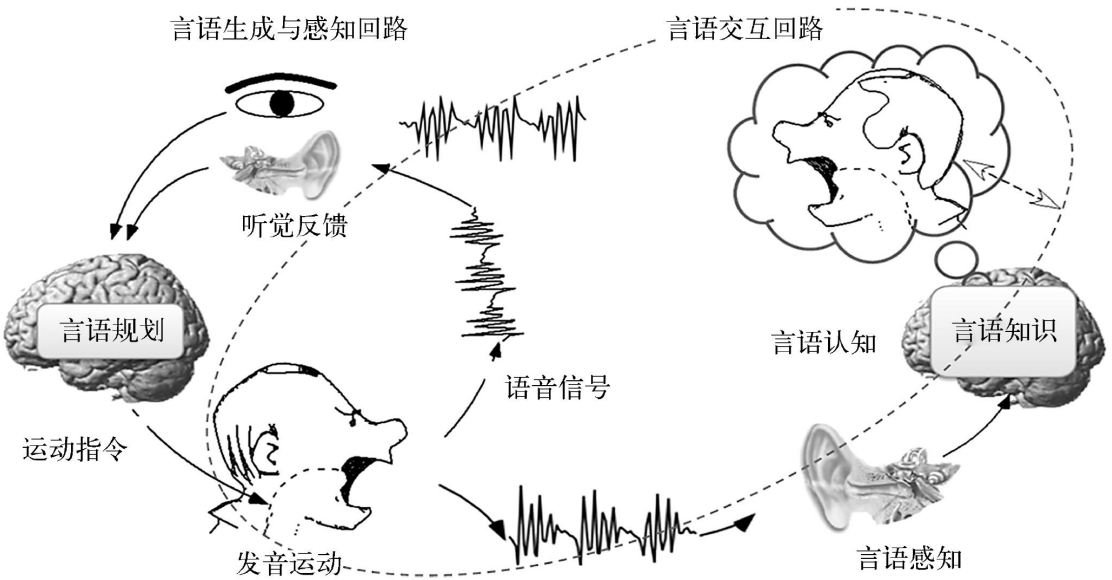
- 1.1 语音产生的机理
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用
  - 1.3.1 言语链（**掌握**）
  - 1.3.2 言语感知运动理论（**了解**）

言语链



- 生成感知认知回路和言语交互理解回路在同一知识语境和语言语境下工作。
- 反映了说话人的意图如何以言语的形式编码并传递给听话人。
- 反映了如何在听话人的大脑中解码、再现说话人的意图。

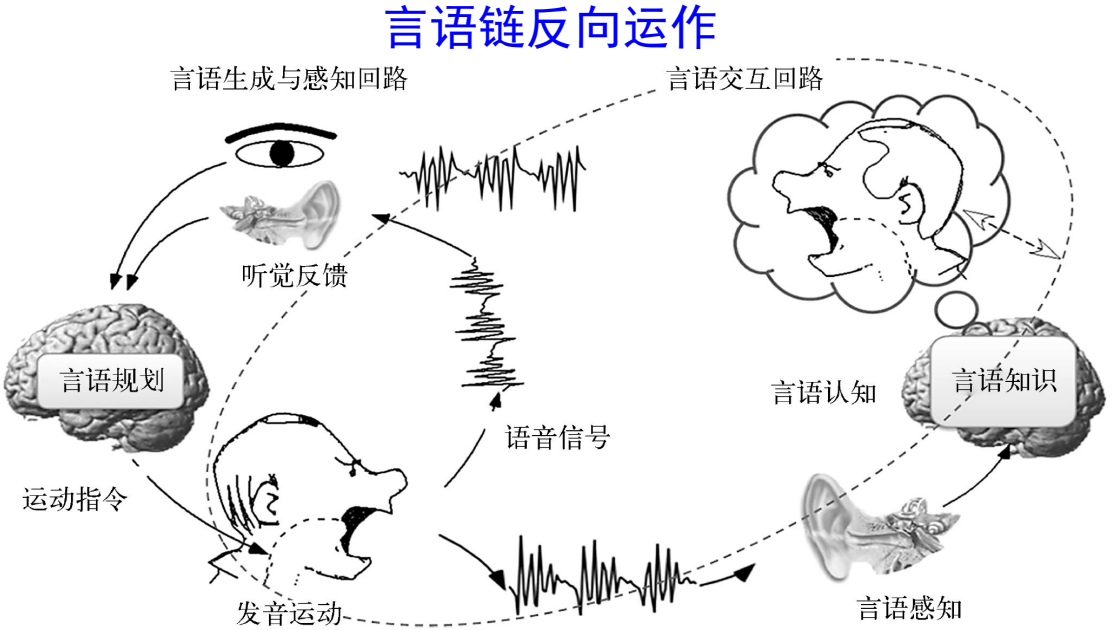
言语链



言语（语音）产生：

- 基于意图来决定说话的内容-》言语中枢进行言语规划（将抽象的概念映射到词汇表征，并以正确的顺序给出词汇的音素、语调和时长。）
- 大脑的言语运动中心编排对应的发音运动程序，并传递到后续神经系统，按照时序来执行。下一级神经水平运动指令又将动作信息传递给负责语音产生的所有肌肉。
- 由于肌肉收缩，肺部提供的空气流穿过声带，产生不同类型的声源，并经过口腔和鼻腔的调制，由唇部、鼻孔的声波放射以及声道壁振动辐射，产生语音信号。

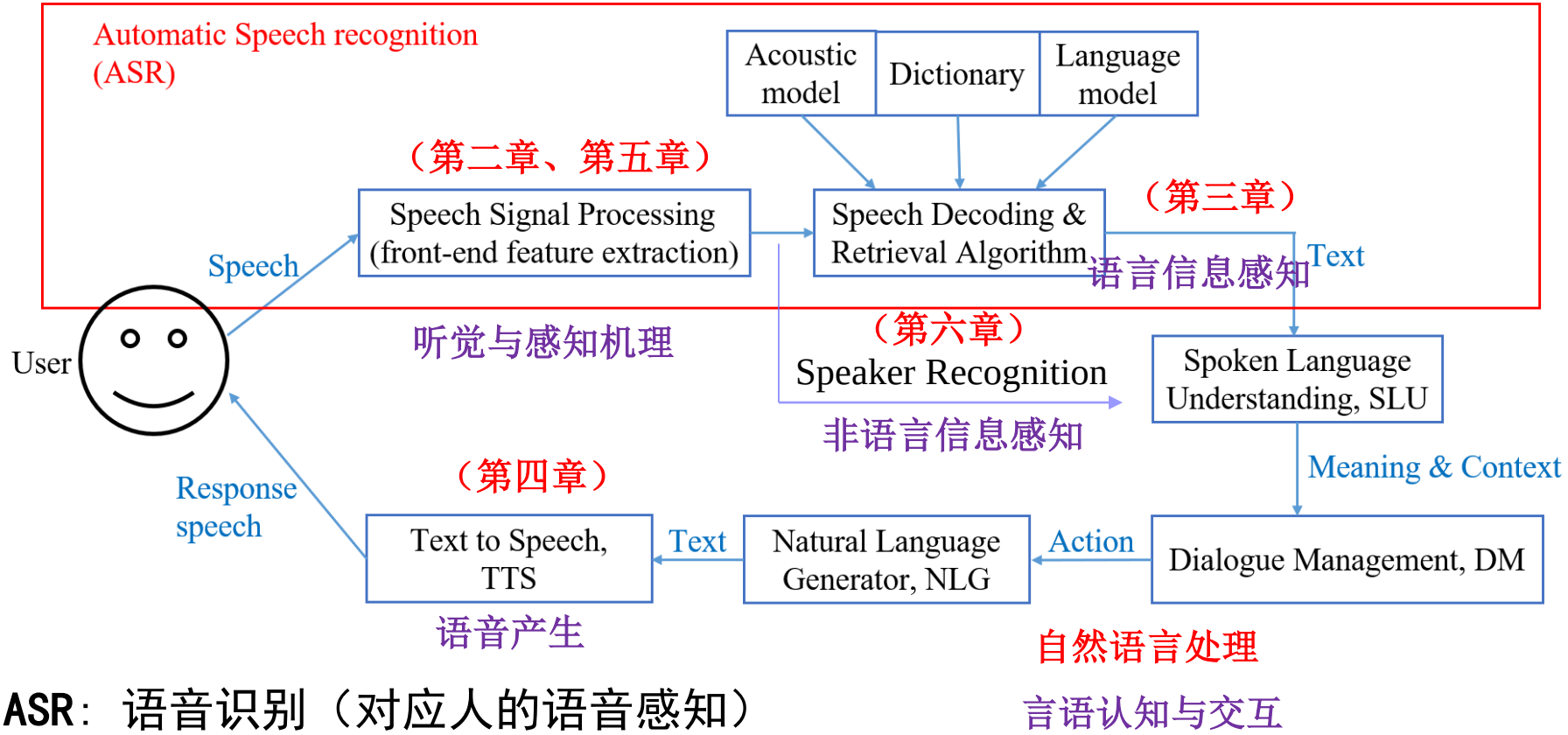
言语链



言语（语音）感知：

- 机械振动-》流体的行波-》神经脉冲，经由听觉神经传递给脑干、丘脑和听觉皮层。大脑的语言中枢通过整合语义、语调、时长和能量包络等信息识别传达意义的音素。
- 通过音质等额外信息辨识说话者、了解说话者的健康、情绪状态。
- 听话人的大脑高级中枢会有意识或潜意识地将这些输入的声学特征和语言信息与以前的记忆和当前的语境相结合，解读说话人的思维，“再现”说话人意图的某种程度的“复制品”。

言语链与人机语音交互（重点掌握）



ASR：语音识别（对应人的语音感知）

TTS：语音合成（对应人的语音生成）

Speaker Recognition：说话人识别/声纹识别

第一章　语音产生与感知机理

- 1.1 语音产生的机理
- 1.2 语音感知的机理
- 1.3 语音产生与感知的相互作用
  - 1.3.1 言语链
  - 1.3.2 言语感知运动理论

言语感知运动理论

- 言语感知运动理论(The motor theory of speech perception)：是将言语产生和感知过程融为一体进行研究考察的，主张在语音感知过程中，语音生成的运动系统的作用是必不可少的，强调音素感知不变的原因在于发音的语音姿势（弱运动假说：语音姿势是说话人语言的层面控制的，而不是实际的运动）。

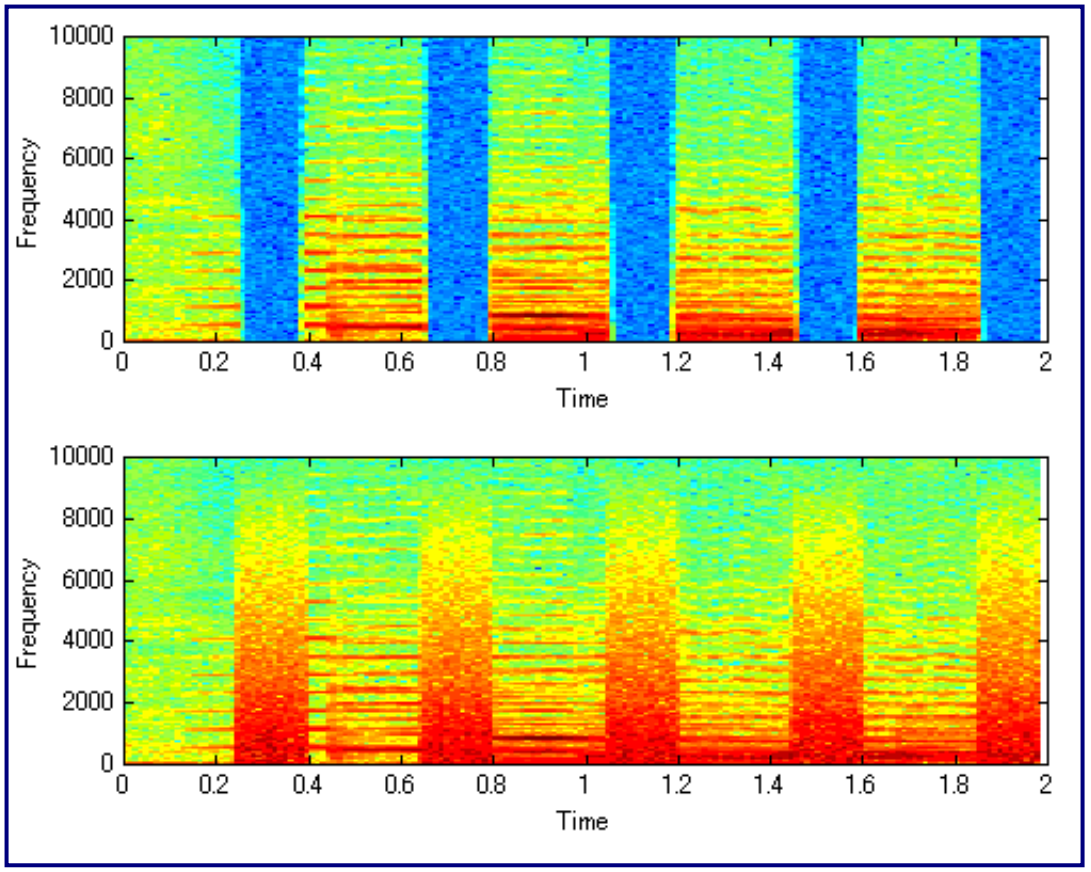
- 该理论主张，① 语音感知的对象是反映说话人意图的发音运动和驱动发音器官的运动指令所形成的脑内表征；② 语音生成和感知通过共享具有不变性的同一特征（语音姿势）而密切相关；③ 这个相关特性是在进化过程中获得而不是后天学习获得的，将语音的生成和感知自动关联在一起的发音感知的模块存在大脑中。

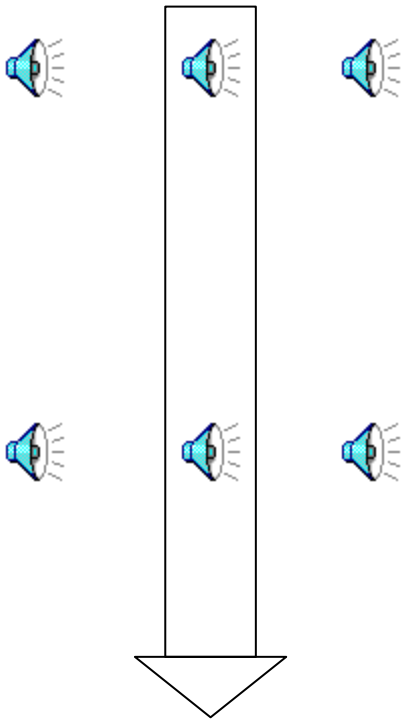
McGurk效应的工作机理

视听者融合发音的先验知识综合判断听觉内容

当给受试者播放听觉刺激“ba”同时呈现出视觉“ga”的嘴部动作时，部分受试者会将其识别为音节“da”。

你听到了什么？





通过人的音韵修复功能
修复了不存在的声音

第一章　小结

- 语音的发音器官
- 语音发音机理
- 语音发音运动及其范畴化
- 听觉器官
- 听觉感知机理
- 听觉特性
- 言语链
- 言语感知运动理论