第一章 绪 论

1.1 概 述

语音信号数字处理是一门涉及面很广的交叉科学，虽然从事这一领域研究的人员主要来自计算机和通信等学科，但是它与语音学、语言学、数理统计学以及神经生理学等学科，也有非常密切的关系。作为一本为信息处理、通信和计算机科学等领域的高年级大学生、研究生、科研工作者和工程技术人员所写的基础教材，这本书着重从数字信号处理的角度来讨论这个课题。众所周知，语言是人类进行相互通信和交流的最方便快捷的手段。在高度发达的信息社会中用数字化的方法进行语音的传送、储存、识别、合成、增强……是整个数字化通信网中最重要、最基本的组成部分之一。

计算机的高速发展既对语音信号数字处理提出了越来越高的迫切要求(如用语音输入代替键盘输入以实现直接的人机对话)，同时也提供了效率不断提高的软、硬件实现手段。另一方面，语音也是人类赖以进行思维的主要工具，因此，这一学科与认知科学和人工智能等研究领域，必然有千丝万缕的联系。近年来，人工神经网络的研究有了飞速发展，语音信号数字处理的各项课题是促使其发展的重要动力之一，同时，它的很多研究成果，也体现在有关语音的各项应用之中。目前，世界科技界正在蓬勃开展的其它一些新研究课题，诸如模糊理论、混沌理论和子波(Wavelet)信号处理等，也都能够在语音信号处理的研究中找到用武之地。

语音信号数字处理涉及一系列前沿科研课题，是目前发展最迅速的信息科学研究诸领域中的一个。正如其它数字信号处理研究课题，语音处理的研究涉及三方面互相密切配合的任务和课题，这就是：应用、算法(包括基础理论和软件)和硬件系统，三者缺一不可。由于这一领域的发展非常迅速，发表的有关文献浩若烟海，进行面面具到的介绍既不可能也不必要。这里只介绍对当前研究工作有关的且最重要的基础理论和算法，并且迅即将读者引入当前最重要的研究课题(包括应用和系统)，而不把精力放在一些支流或已成为历史陈迹的内容上。

1.2 语音信号数字处理的应用

如上所述，几乎语音信号处理的所有研究课题都是受到应用驱动的。以语音编码为例，由于数字化的语音传输和存储，无论在可靠性、抗干扰、速交换、易保密和廉价格等方面都远胜于模拟语音。从 50 代以来，在通信系统中数字化语音所占百分比不断增加。现在已非常清楚，在未来的 ISDN(综合业务数字通信网)、卫星通信、移动通信、微波接力通信和信息高速公路等系统中将无一例外地都采用数字化语音传输和存储。在不到 50 年的时间里，语音编码已有了惊人的发展。最早的标准化语音编码系统是速率为 64kb/s 的 PCM 波形编码器，到 90 年代中期，速率为 4～8kb/s 的波形与参数混合编码器，在语音质量上已逼近前者的水平，且已达到实

用化阶段。

据预测，速率为 2.4kb/s 左右的语音编码器，在未来几年中将在性能和实用化两方面都接近于 64kb/s 的标准 PCM 编码器。语音识别的研究起步较晚，大规模的研究开始于 70 年代初期，近年来已取得了长足的进展。一些中、小词表的孤立或连续语音识别系统已进入市场。目前，研究的重点是实现大词表、非特定人的连续语音识别系统。它可以用于人机直接对话、语音打字机以及两种语言之间的直接通信等一系列重要场合。这是一个难度相当大的高科技课题。在当前，学术界的普遍看法是：在信号处理、计算机、语言学、语音学和人工神经网络等各界学者的通力合作下，这一难题很有可能在本世纪末取得突破性的进展。语音合成是人机对话的另一个重要环节，让机器将文本语言转换成具有人声特点、抑扬顿错自然流利的口头语言绝非易事，这一研究课题也正日益受到重视。其它一些重要的应用领域还包括语音增强（在强背景噪声或干扰中恢复“干净”的语音）和说话人识别及确认等。以上各个方面都是这本书所要讨论的内容。

1.3 语音信号数字处理的基础理论和算法

对于语音信号处理的基础理论和各种算法的研究包括紧密结合的两个方面。

一方面是从语音的产生和语音的感知来对其进行研究，前者涉及大脑中枢的言语活动如何转换成人发声器官的运动，从而造成声波的传播；后者涉及耳对声波的搜集并经过初步处理后转换成神经元的活动，然后逐级传递到大脑皮层的语言中枢。这一研究与语音学、语言学、认知学、心理学和神经生理学等密不可分。目前，对于这整个语言链的底层（或称为物理层），其中包括发声器官和耳的功能已经研究得比较透彻，但是对于其上层（即神经元的活动和大脑语言中枢的工作原理）则可以说还很不清楚。

另一方面，是将语音作为一种信号来进行处理。60 年代中期形成的一系列数字信号处理方法和算法：数字滤波器、快速傅里叶变换（FFT）、……与语音信号处理的要求分不开的。嗣后，在 70 年代初期产了线性预测编码（LPC）和同态信号处理的算法，它们已成为进行语音信号处理最强有力的工具，且广泛应用于语音信号的分析、合成及各个应用领域。80 年代以后，出现了一系列更重要的方法和算法，其中包括语音编码中采用的分析合成方法，简称为 ABS（Analysis By Synthesis）以及各种自适应处理方法和变换方法。在语音识别方面最重要的是与隐含马尔可夫模型 HMM（Hidden Markov Model）有关的一系列算法以及语言的概率模型。在编码和识别两个方面都非常重要的是与矢量量化（VQ）有关的各种算法。

应该注意，在研究各种算法时科研工作者通常采用两种方法，一种是用概率统计的方法，另一种是用规则的方法（或者说专家系统的方法）。虽然这两种方法互相渗透，不可能截然分开，但是仍能按照其主要观点和方法大致加以划分。从 80 年代至 90 年代中期的发展趋势看，前一种方法略占优势。而后一种方法则逐渐为人工神经网络的方法所取代或与之相结合。可以预期，在本世纪最后几年中，在前一方法继续平稳发展的同时，后一种方法将更加蓬勃地发展。

1.4 语音信号数字处理的硬件和实用系统

绝大多数语音信号数字处理系统需要按照实时方式或称为在线方式工作，这时对于系统的硬件环境要求很高(这里主要指系统的运算速度和内存容量的要求)。

随着语音处理算法的日益复杂，许多语音处理器的运算速度需要达到 10～20MIPS(Million Instructions per Second)，在未来几年中这个速度甚至要达到 50MIPS。而在语音识别与合成等领域中对于处理系统的内存容量往往要求达到若干 MB。实用的实时语音信号数字处理系统通常以两种方式实现：第一种是用一台计算机作为主机(微型机、小型机或工作站)插上一块或若干块数字信号处理板来构成，后者由通用或专用的数字信号处理芯片(DSP 芯片)及相应的存储芯片、接口芯片和 A/D、D/A 芯片组成。第二种则由专用或通用的 DSP 芯片及其它辅助芯片构成一个独立工作的系统。前者通常称为非脱机工作系统，用于识别、合成、增强或模拟实验中。后者称为脱机工作系统，用于编码、小词汇表识别与合成等场合。

通用 DSP 芯片的出现及其性能价格比的迅速提高为各种实用化语音信号处理系统的实现铺平了道路。美国 TI 公司在 80 年代中期研制出的第一代 DSP 芯片 TMS32010 和 TMS32020 完成一次乘/累加运算(16 位、定点)需要 200ns，第二代 DSP 芯片 TMS320C25 完成一次乘/累加(16 位、定点)运算需要 100ns，第三代 DSP 芯片 TMS320C30 完成一次乘/累加(32 位、浮点)运算只需要 50ns 且片内的 ROM 和 RAM 和片外可扩充的 RAM 容量都大大增加。此外，美国 AT&T 公司研制出的 DSP-16C 和 DSP-32C，美国 AD 公司研制出的 ADSP21010 和 ADSP21020 等芯片系列与上述 TI 公司的第二代和第三代 DSP 芯片大致处在相似的水平上。第三代 DSP 芯片及更高一代 DSP 芯片的出现将使语音信号数字处理技术的发展和实用化登上一更高的新台阶。

1.5 全书的组织

这本书重点介绍语音信号数字处理的基础理论、算法和应用，其中第二至五章是各种具体应用领域的共同基础部分。为了使做更深入研究的人员得以获得彻底了解，对很多算法都做了详尽的推导。那些仅对这些算法的使用感兴趣的读者，则可以把冗长的推导略去不读。对语音编码感兴趣者，可以进一步阅读第七和第八章。对于语音识别和说话人识别感兴趣的读者，阅读第六、第十和第十二章。对语音合成或增强感趣者则可阅读第九或第十一章。为了区分主次，对每一章中的有些节，凡是加上“*”号，表明这是一些内容较深的节次，可以在初学或学习时间不够的情况下将其略去。从每一章所附的文献中，有兴趣的读者还可以找到更多的学习内容。

第二章　语音信号的数字表示、基本组成单位、产生模型和短时分析技术

2.1　概　述

在研究各种语音信号数字处理技术及其应用之前，首先需要了解语音信号的一些重要特点，应知道它是如何由一些最基本的单位组成的，发声器官是如何发出这些音的，在此基础上可以建立一个既实用又便于分析的语音产生模型，这些问题可以归于声学语音学的范畴。通过对于语音信号发声过程的研究以及观察记录的各种语音波形，便可知道语音信号的频谱分量主要集中在 300～3400Hz 的范围内。如果用一个防混叠的带通滤波器将此范围内的语音信号频谱分量取出，然后按 8kHz 采样率对语音信号进行采样，就可以得到离散时域的语音信号。下面将讨论离散时域语音信号或称为数字语音信号。应该注意，为了实现更高质量的语音编译码器或者使语音识别系统得到更高的识别率，某些近代语音系统将此频率范围高端扩展到 7～9kHz，相应的采样率也提高到 15～20kHz。语音信号的另一个重要特点是它的“短时性”。在某些短时段中它呈现出随机噪声的特性，另一些短时段则呈现出周期信号的特征，其它一些是二者的混合。简而言之，语音信号的特征是随时间而变化的。只有在一短段时间间隔中，语音信号才保持相对稳定一致的特征，这短段时间一般可取为 5～50ms。因此，对于语音信号的分析和处理必须建立在“短时”的基础上。最重要的语音信号“短时特征”和“短时参数”包括它的“短时能量”、“短时过零率”、“短时相关函数”、“短时频谱”等。

语音信号的最基本组成单位是音素。音素可分成“浊音”和“清音”两大类。如果将不存在语音而只有背景噪声的情况称为“无声”，那么音素可分成“无声”，“浊音”和“清音”三类。在短时分析的基础上可判断一短段语音属于哪一类。如果是浊语音段，还可测定它的另一些重要参数，如基音和共振峰等。这里将讨论语音信号数字处理的这些基本知识、术语和分析技术。

2.2　语音信号的时域波形

在进行语音信号数字处理时，最先接触到并且也是最直观的是它的时域波形。为了获取一段语音信号的时域波形，首先将语音用话筒转换成电信号，再用 A/D 变换器将其转换为离散的数字化采样信号后存入计算机的内存中，最后将此信号取出，用绘图仪绘成时域波形。图 2-1 所示是一个男青年说的“欢迎你到深圳特区”这段话的语音时域波形。语音是在安静的环境下录取的。采样前经过频带为 0.1～3.4kHz 的带通滤波器进行滤波，采样率为 8kHz。每个采样信号用 12 位进行量化。这段语音的持续时间为 4 秒，图中横轴为时间，纵轴为语音信号的幅度。由于时间轴压缩得很短，从图 2-1 中无法辨别语音波形的细节，但是可以看到语音能量的

起伏，还可以大致分辨出话语中每一个字(音节)在此波形中的位置。为了仔细辨识语音波形，可以把时间轴拉宽。图 2-2(a)和(b)显示了这一段语音的波形细节，其中每一段横线伸展的

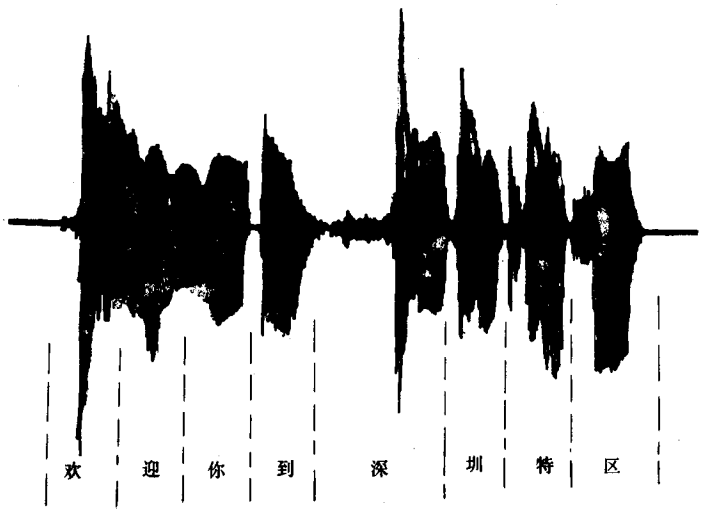
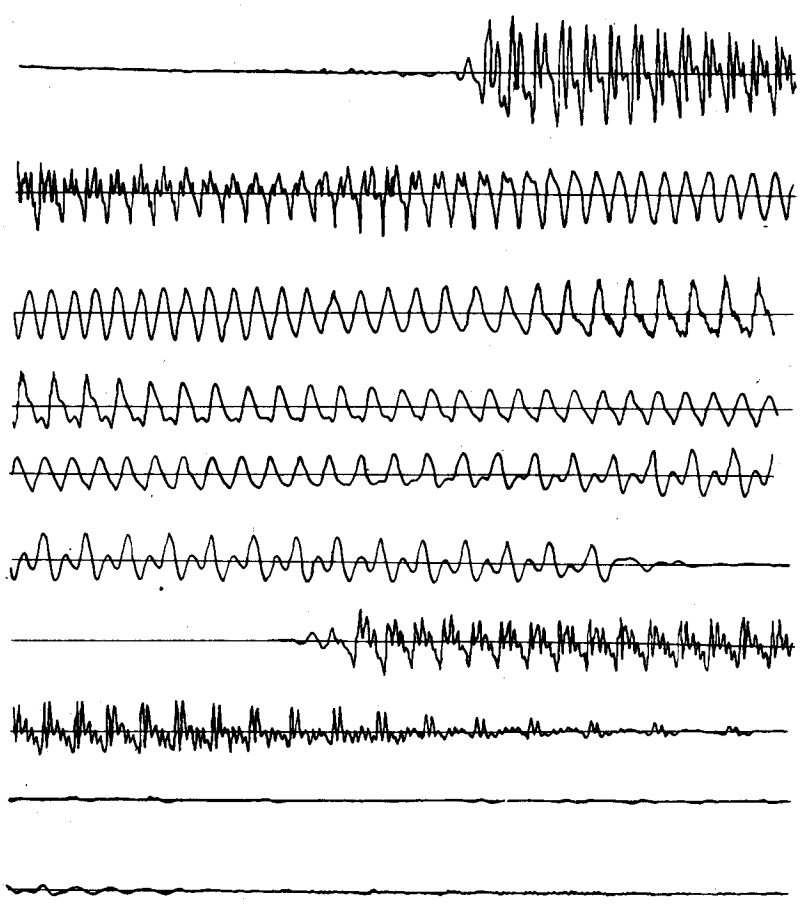


图 2-1　一段语音信号的时域波形



(a)

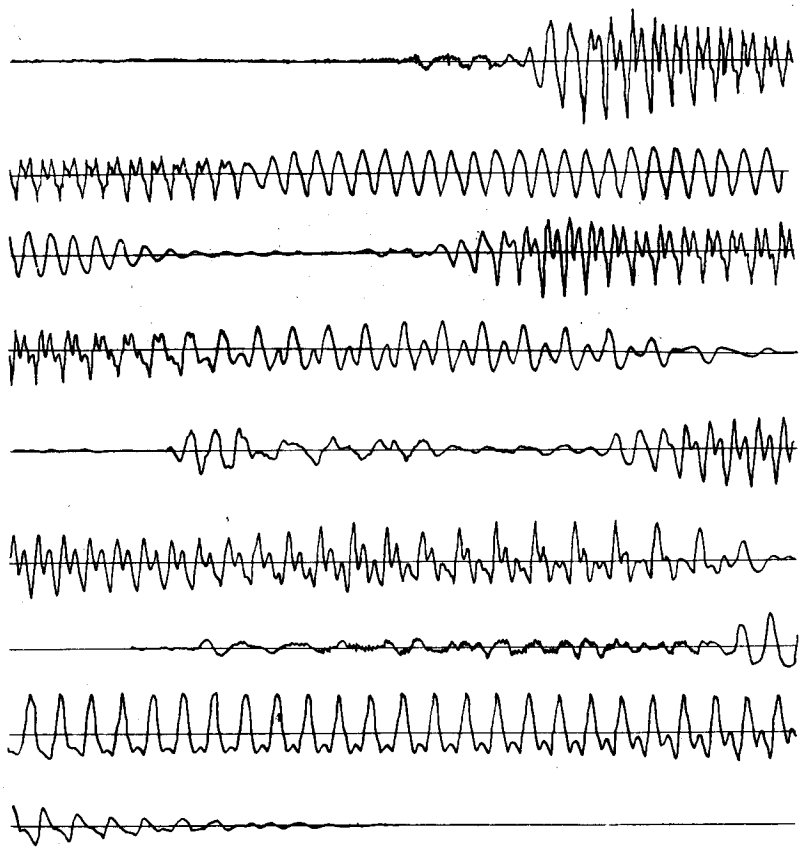


图 2-2　图 2-1 时域波形的展开图

范围是 200ms，上段的末尾与下一段的起点相连接，(a)和(b)相连接。由此图看出，语音信号具有很强的“时变特性”。在有些段落中它具有很强的周期性，有些段落中又具有噪声特性，而且周期性语音和噪声语音的特征也在不断变化之中，只有在较短的时间间隔中(例如 20～200ms)才可认为语音信号的特征基本保持不变。这一特点是语音信号数字处理的一个重要出发点。

2.3　发声器官

发声器官由三部分组成：喉、声道和嘴。下面分别介绍它们的结构和功能。

2.3.1　喉

喉位于气管的上端，其顶视解剖结构如图 2-3 所示。实际上它是气管末端的一圈软骨构成的一个框架，前方稍高处的软骨称为甲状软骨，前后方环成一圈的称为喉部环形软骨。喉中有两片肌肉，称为声带，它们的一侧由甲状软骨支撑，另一侧则由两块杓状软骨支撑和控制。后者又与环形软骨连接。当它们分开时声带是张开的，空气可自由地流过喉和气管(见图 2-3(a))，

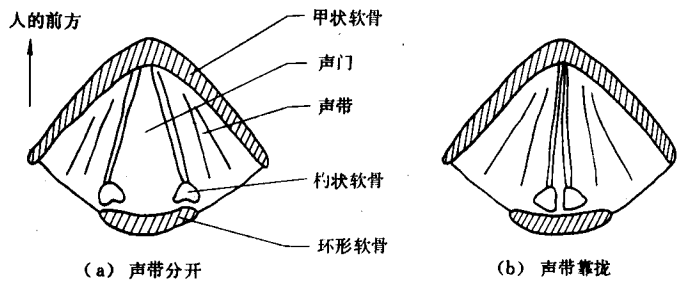


图 2-3 喉的解剖结构

正常呼吸时就处于这种情况。当它们合拢时，声带闭合将喉封住，在吃东西时食物就不会落入气管。两片声带之间的空隙称为声门。说话时两片声带在杓状软骨的作用下相互靠近但不完全封闭，这样声门变成一条窄缝(见图 2-3(b))。当气流通过这个窄缝隙时其间的压力减小，从而两片声带完全合拢使气流不能通过。在气流阻断时压力恢复正常，因此声带间的空隙再次形成，气流再次通过。这一过程周而复始的进行，就形成了一串周期性的脉冲气流送入声道。它的典型波形如图 2-4 所示。这一周期气流脉冲串的周期称为“基音周期”，用 T_p 表示，其倒数称

图 2-4 典型的声门脉冲串波形

为“基音频率”，用 f_p 表示。f_p 值取决于声带的尺寸和特性，也决定于它所受的张力。男性说话者的 f_p 值大致分布在 60～200Hz 范围内，女性说话者和小孩的 f_p 值在 200～450Hz 之间。用上面所述的方式发出的语音是“浊音”(Voice)。

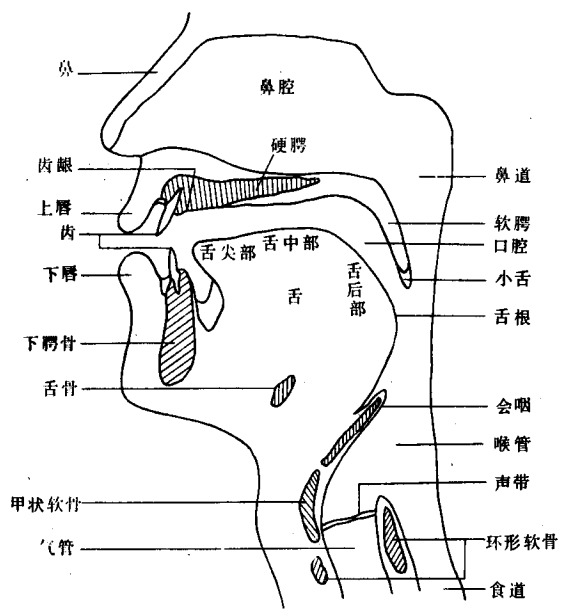


图 2-5 头的纵剖面，其中显示了各主要发声器官

2.3.2 声道

气流从喉向上经过口腔或鼻腔后从嘴或鼻孔向外辐射，其间的传输通道称为声道。声道的解剖结构(纵剖面)如图 2-5 所示。口腔的上顶分成两部分。前部是一块称为硬腭的骨头，它的作用是将口腔和鼻腔分开，并且支撑上排牙齿。后部由肌肉和连接组织构成，称为软腭，软腭的终端是小舌。当软腭在肌肉的作用下卷起贴在鼻道的后壁上时，鼻腔和口腔相互隔开；反之，二者连通在一起。硬腭前部的骨头较厚，其中固定着牙齿，沿

着牙齿的一圈边缘称为齿龈。口腔下部是由肌肉构成的舌头，它的前部与下腭相连，后部和喉部的骨头及头部其它骨头相连。

气流流过声道时犹如通过一个具有某种谐振特性的腔体。输出气流的频率特性既取决于声门脉冲串的特性。又取决于声道的特性。为了便于分析，可以把声道当作一段无损声管如图2-6所示，其中鼻腔和口腔是否耦合，取决于软腭的位置。有耦合时发出的语音称为“鼻音”，否则为非鼻音。对成年男性而言，声道的口腔段的长度为17cm左右，而鼻腔段的长度约为13cm。实际上声道的横截面积并非常数，所以声道模型中的声管应是一变截面积声管，而声道的频率特性主要取决于声道截面的最小值（一般称之为收紧点）出现的位置。这一收紧点的位置又主要由舌的位置来控制。

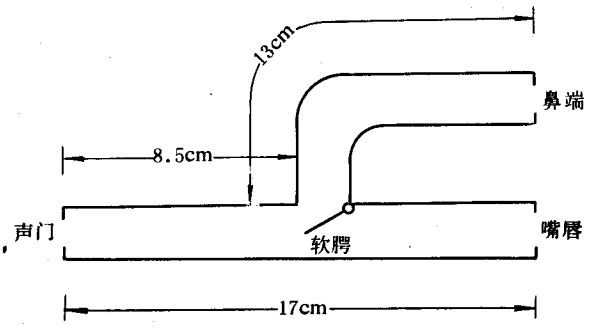


图 2-6　声道的无损声管模型

语音的另一种产生方式是声门完全封闭，这时声道不是受声门周期脉冲气流的激励而是利用口腔内存有的空气释放出来而发声。由于该气流通过一个狭通道时在口腔中形成湍流，因而明显地具有随机噪声的特点。相应的语音称为“清音”(Unvoice)。汉语发音中的韵母如[a]，[i]，[u]，[o]等均为浊音，某些声母如[s]，[sh]，[h]，[x]，[f]等为清音，另一些声母如[z]，[zh]，[J]等兼具二者的特点。[n]和[ng]是鼻音韵母。[m]，[n]，[l]是鼻音声母。

2.3.3　嘴

嘴的作用是完成声道的气流向外辐射。嘴的张开形状会影响语音频谱的形状，但是其作用较之声道而言是次要的。粗略而言，可以根据发音时嘴唇张开的圆形程度将一个音划归“圆唇音”或“非圆唇音”。

2.4　音素与音节

语音流由音素结合而成的最小单位，同时也是发声的最小单位是“音节”(Syllable)，音节可以结合成更大的单位——“词”。词进一步可结合成“节奏群”、“句子”等等。音素的英语对应词是phoneme，可以认为它是语音的最基本组成单位。事实上，同一音素与不同音素结合时，发音是有差异的。例如，[sh]这个音素在发“诗”([shi])这个音与发“书”([shu])这个音时，发音方式不完全一致，前者是非圆唇音，而后者是圆唇音。对于同一音素，它的各种不同发音方式称为“音素变体”(Allophone)。一个音节由元音(Vowel)和辅音(Consonant)构成。元音构成一个音节的主干，无论从长度看还是能量看，元音在音节中都占主要部分。所有元音都是浊音。辅音则出现在音节的前端或后端或前后两端。在汉语普通话中，每个音节都是由“辅音-元音”构成的（其中包括只有元音而没有辅音的纯元音音节，例如“啊”，这种情况称为“零辅音”），这种结构称为“C-V结构”。在其它语系中还可以出现“V-C结构”或“C-V-C”结构。在汉语中辅音也称为声母，元音也称为韵母。

单独发声的一个音节或是语音流中的任何一个音节都可能由9个部分组成，如图2-7所

示。其中1～4段属于声母(辅音)段,6～9段属于韵母(元音)段,第5段是二者的过渡段。对一个具体指定的音节而言,有可能只包括其中的某几段,但是第7段(主要元音段)是每一个音节都具有的。各段的特点及其发音机制将结合各个声母和韵母进行解释。

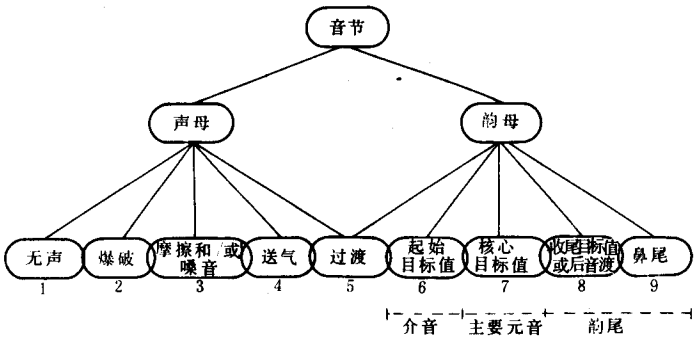


图 2-7 汉语普通话的音节结构框架(引自[1])

2.5 语音信号的"短时谱"、"语谱图"以及元音和辅音的特点

语音信号的最重要特征表现在它的"短时频谱"(简称为"短时谱")上。如果从语音流中利用加窗的方法取出其中的一个短段,再对其进行傅里叶变换,就可以得到该段语音的短时谱。采用数字信号处理(DSP)的手段,可以在数字计算机上非常方便、快捷地完成这一任务。图2-8所示是一段浊音和一段清音的时域波形及其短时谱,语音的采样率是10kHz,窗长为50ms(相应的样点数为500),窗形为哈明窗(在2.10中将较详细地讨论与此有关的短时分析问题)。浊音的短时谱有两个特点:第一,有明显的周期性起伏结构,这是因为浊音的激励源为周期脉冲气流。第二,频谱中明显地具有几个凸起点,它们的出现频率与声道的谐振频率相对应。这些凸起点称为"共振峰"(Formant),其频率称为共振峰频率。共振峰按频率由低到高排列为第一共振峰、第二共振峰,……,相应的频率用 F_1、F_2、……来表示。一般浊音中可以辨别的共振峰有5个,其中前3个(尤其是前2个)对于区别不同语音是至关重要的。清音的短时谱则没有这两个特点,它十分类似于一段随机噪声的频谱。

在DSP技术发展起来以前很久,人们早就用一种特殊仪器——语谱仪来分析和记录语音信号的短时谱。它将语音信号(经话筒变成了电信号)送进一排频率依次相接的窄带滤波器,各窄带滤波器的输出记录在一卷按一定速度旋转的记录纸上(各滤波器的由低到高按频率排列),信号强则记录得浓黑一些,反之则浅淡一些。由此得到的即是语音信号的语谱图,此图的水平方向是时间轴,垂直方向是频率轴,图上或深或浅的黑色条纹表征各个时刻的短时谱。图2-9给出了[i],[æ],[ə],[ɔ],[a],[u]这六个美国英语元音单独发声时的时域波形和语谱图,其中与时间轴平行的几条深黑色带纹称为"横杠"(Bar),它们相应于短时谱中的几个凸出点,也就是共振峰。由横杠的频率及宽度可以确定相应共振的频率和带宽。在一个语音段的语谱图中,有没有横杠存在是判断它是否为浊音的重要标志。

图2-10给出了若干辅音配以元音[a]发音时产生的时域波形图和语谱图,它们的花纹比较复杂,其中比较典型的花纹是横杠、乱纹和冲直条(语谱图中出现与时间轴垂直的一条窄黑条)。每一种辅音包括上面几种典型花纹中的一种或几种,它们与该辅音发音的特点有密切关系,这将在2.7中进行详细讨论。

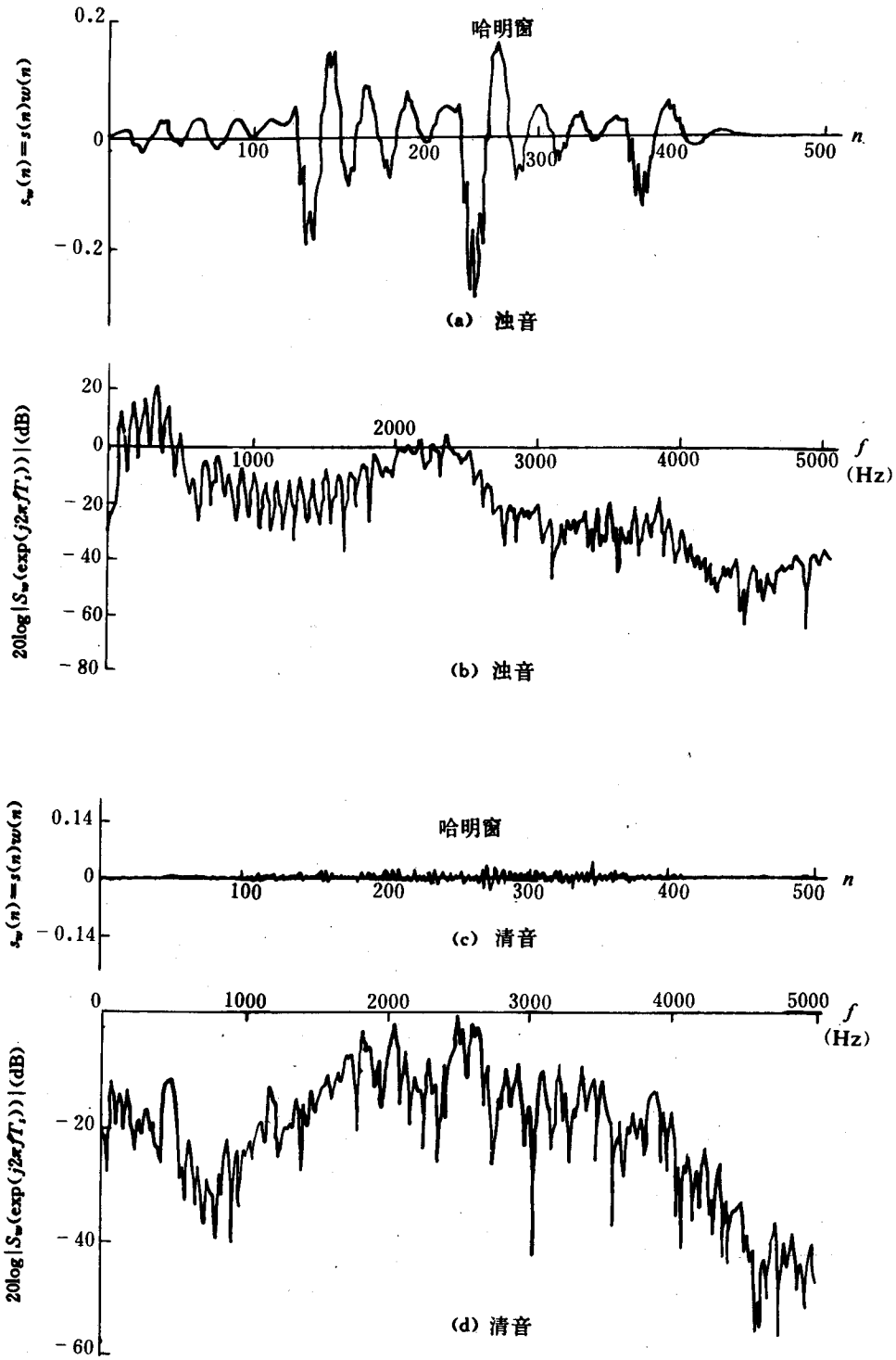


图 2-8　浊音和清音的时域波形和短时谱(引自[1])

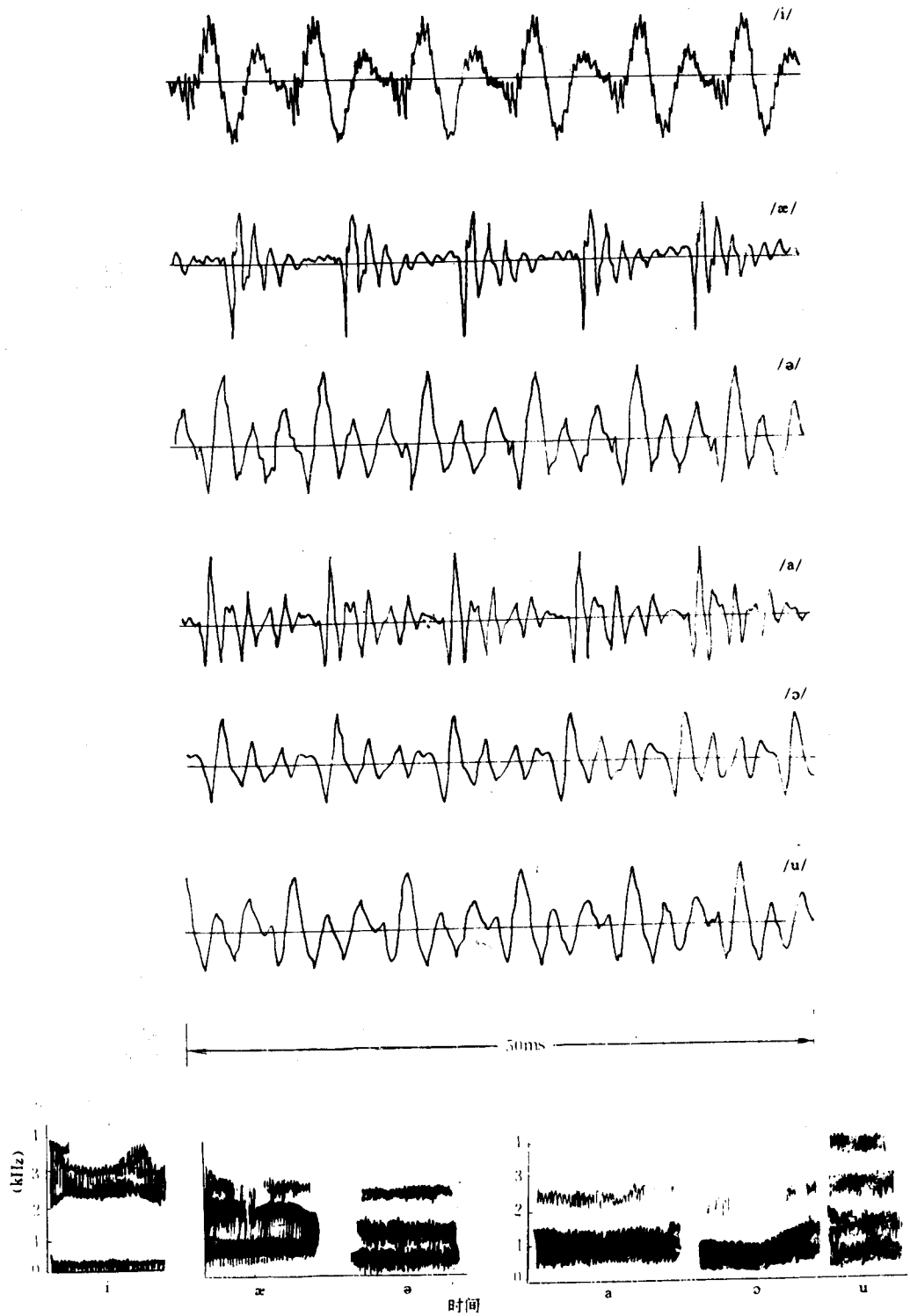


图 2-9　若干元音的时域波形图及语谱图(引自[2])

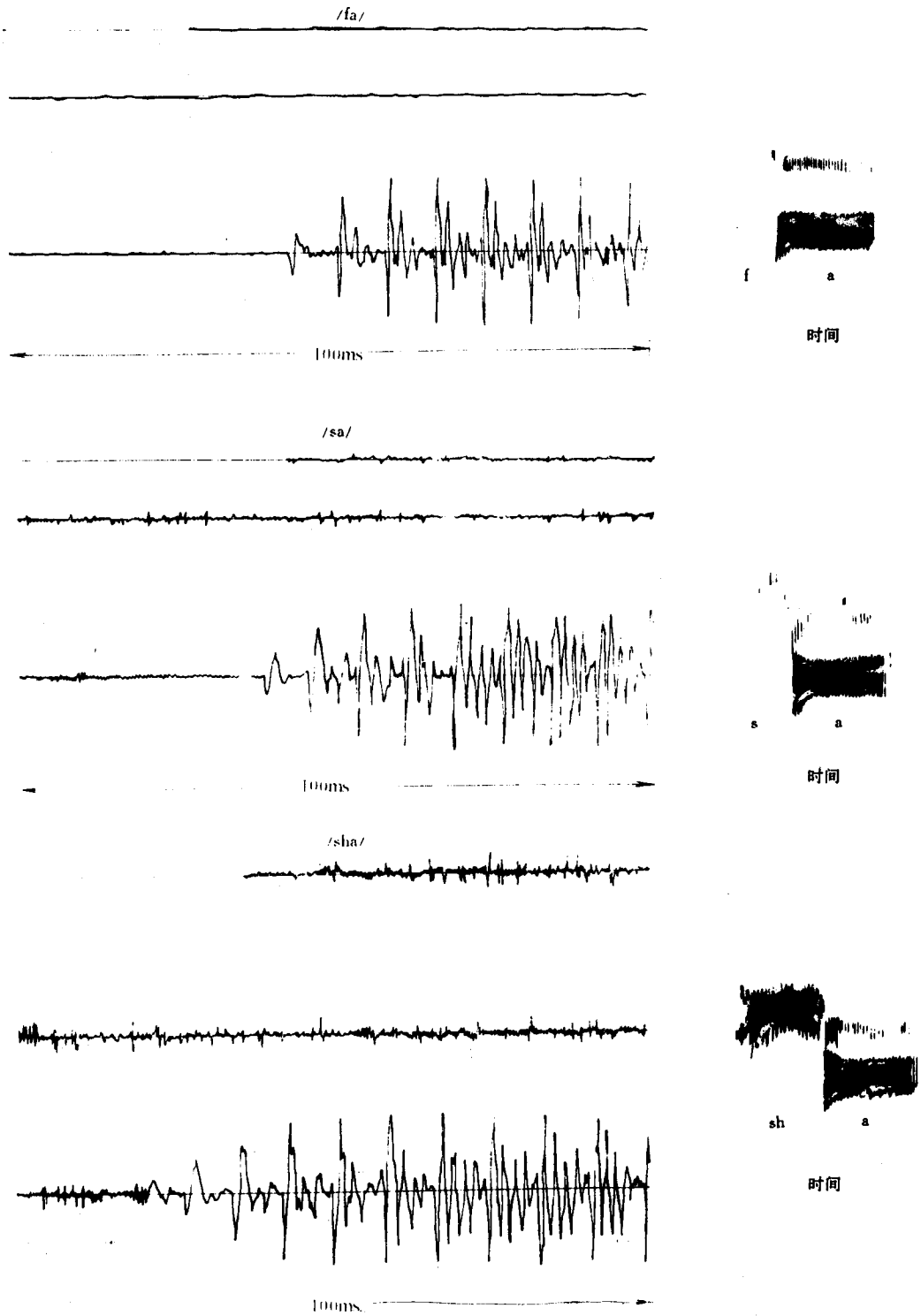


图 2-10　若干辅音的时域波形图及语谱图(引自[2])

2.6 韵 母

汉语普通话中每一个音节都包括一个元音，或称为韵母。韵母总共有 38 个，其中 8 个是单韵母，14 个是复韵母，16 个是鼻韵母。下面分别对它们进行介绍。

2.6.1 单韵母

8 个单韵母是[a]，[i]，[u]，[ü]，[l]，[ɿ]，[e]，[o]，其中前 6 个是稳定元音，这就是说，在单独发这些音时发声器官的状态基本不变，因而这些音的语谱图中共振峰的位置是基本不变的，后 2 个则有些变化。按照浊音的发声机制每一个韵母的产生过程如图 2-11 所示。声门气流脉冲串 $g(t)$的频谱为 $G(j2\pi f)$，由于声门脉冲串是周期性的，它的频谱具有谱线结构，谱线之间的间隔为 f_p。实测数据表明，20log $|G(j2\pi f)|$以每倍频程 12dB 的速度随着 f 的增高而下降。声道的频率响应 20log $|V(j2\pi f)|$中有若干共振峰，用 F_1，F_2 等来表示。唇的辐射所形成的频率响应用 20log $|R(j2\pi f)|$表示，它是一个随频率 f 递增的函数，幅射频率响应因张嘴的圆度不同而有所变化。语音信号 $s(t)$的频谱 20log $|S(j2\pi f)|$等于上列三个对数频谱之和。

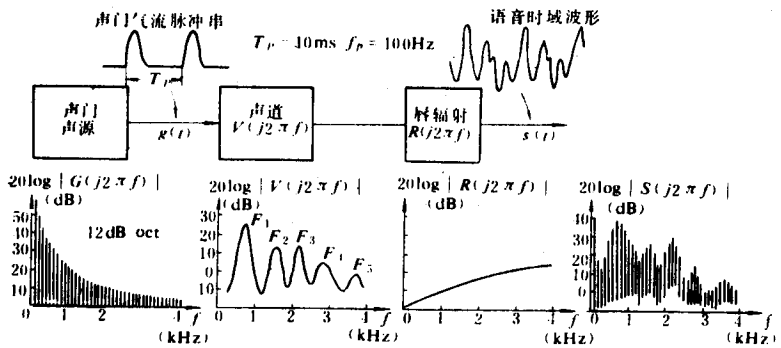


图 2-11 韵母的产生过程

各个韵母的区别可以按照发声的特点来描述，也可以按照对听觉起作用的频谱特点，尤其是共振峰的特点来加以描述，这两方面是紧密相关的。从发声的角度看，不同的韵母是由于声道形状的不同所造成的，而声道又可以用一段变截面积的声管来表示。如果给出声管的截面积随其轴向长度 l 的变化就能得到它的面积函数 $A(l)$。图 2-12 是一个面积函数的示例，为了分

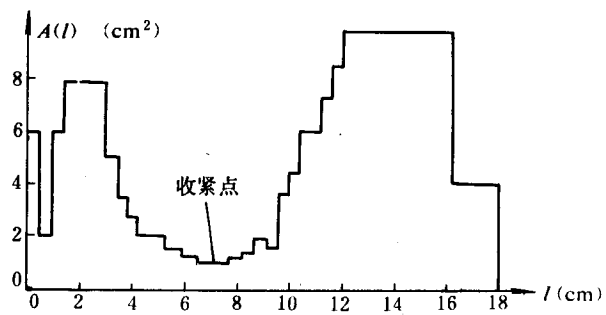


图 2-12 声道面积函数示例

析方便，一般把连续变化的面积函数表示成阶梯形状，其中 l 值为 0cm 表示喉部，l 值为 18cm 表示嘴唇处。采用流体力学的方法，可以计算出不同面积函数声管的频率响应及其共振峰。这里不打算详细介绍这些计算细节，因为可以用更有特征性的方法来描述不同韵母所对应的声道形状。声道形状主要取决于三个方面：第一是舌在口腔中的前后位置不同，造成声道中的收紧点（即声管中的最小面积点）位置不同。第二是舌位的高低，舌位越低嘴张得越大，所以也可以称为开口度大。反之，舌位越高开口度越小。第三是唇的圆展程度。其中前两个因素影响更大一些。舌位前后主要影响 F_2，收紧点越靠前（靠近嘴唇）则 F_2 越高，反之则越低。但是 F_2 的最高值不是出现在收紧点最前（$l=17$cm）的发音[1]，而是出现在舌尖紧贴上腭时次前的发音[i]。F_2 的最低值也不出现在收紧点最后（$l=0$cm）的发音[a]，而出现在次后的发音[u]。舌位上下，即开口度，主要影响 F_1。开口度越小，F_1 越低，开口度越大则 F_1 越高。唇的圆展程度则对 F_1 和 F_2 都有影响。F_1 和 F_2 受这三个因素的综合影响如图 2-13 所示。

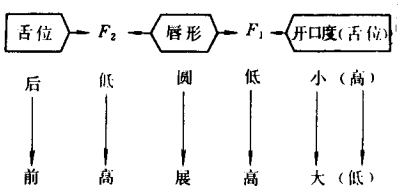


图 2-13　舌位前后，开口度大小和唇形圆展度对第一和第二共振峰 F_1 和 F_2 的影响（引自[1]）

在表 2-1 中列出了 8 个单韵母的发音特点、频谱特点和前 4 个共振峰的典型值（成年男性）。对于女性而言，各共振峰值大约较男性高 25%，而小孩大约高 35%。如果按照各个韵母的前两个共振峰的典型值，把它们标注在一张以 F_1 为横轴，F_2 为纵轴的平面图中，就可以得到一个元音三角形。这个三角形以[i]，[a]，[u]为顶点，其它元音都在三角形中。应注意，不同的人对同一韵母发声的共振峰值有很大差异，这表现在 F_1-F_2 平面上的同一元音占据的不是一个点，而是一个区域。不同元音所占的区域甚至有部分重叠，但是对于人们听音辨意不会造成任何困难。元音三角形如图 2-14 所示。

表 2-1　8 个单韵母的发音及频谱特点，前 4 个共振峰值

韵母	典型字的韵母	收紧点	开口度	F_1	F_2	F_3	F_4	频谱特点
[a]	巴，大	后	大	850	1300	2600	3700	整体强度高，F_1 特别高
[i]	一，希	前	小	300	2300	3000	3500	F_2 弱，F_3，F_4 近，故形成一个强区，频谱质心高
[u]	乌，路	后	小	350	650	2500	3300	F_1，F_2 形成强区，F_3，F_4 很弱，频谱质心低
[ü]	玉，居	前	小	300	2000	2500	3500	圆唇音，F_2，F_3 形成一个强区
[1]	兹，此，丝	前	小	400	1300	2700	—	舌尖前元音，频谱均匀无明显的质心
[ɿ]	知，吃，施	前	小	400	1600	2000	—	舌后元音（卷舌音）F_3 极低，F_2，F_3 形成质心
[e]	特，哥	中	中	520	1200	2400	—	开口由小到大，所以 F_1 由低到高
[o]	迫，魔	中	中	570	840	2400	—	开口由小到大，F_1，F_2 皆由低到高

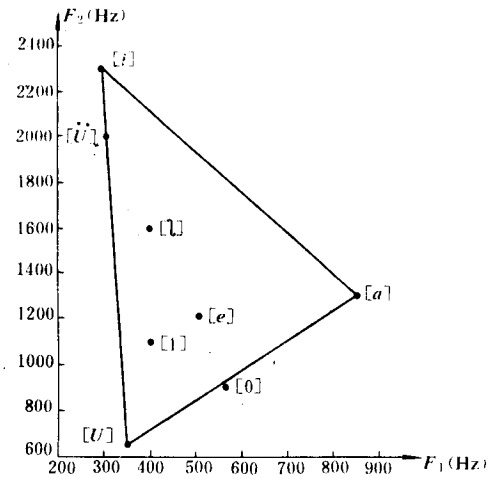


图 2-14　元音三角形

2.6.2　复韵母

其中包括 9 个二合元音:[ai],[ei],[au],[ou],[ia],[ie],[ua],[uo],[üe],4 个三合元音:[iao],[uai],[uei],还有一个[ər](这是而,二等字的韵母)也可以归入二合元音中。二合元音中有 5 个称为后响元音,它们是[ia],[ua],[uo],[ie],[üe]。这些韵母的开口从小到大,因而响度也从小到大。后响二合元音中第二个音素相当于图 2-7 音节框架中的第 7 段,即核心目标值,它是主要元音。它的第一个音素相当于框架中的第 6 段(起始目标值)通常称为介音。这里所说的第一和第二音素在一些文献中也称为第一和第二音位。[ai],[ao],[ei],[ou],[er]等 5 个二合元音称为前响元音,其中第一个音素是主要元音,相当于音节框架中的第 7 段,而第二个音素则相当于第 8 段(收尾目标值)也可称为元音性韵尾。后响二合元音的两个音素互相影响较小,它们的共振峰相当接近各音素单独发音时的目标值。前响二合元音的两个音素互相影响较大,其共振峰偏离单独发音时的目标值较远。

三合元音的第一音素是介音,第二音素是主要元音,第三音素是元音性韵位;它们分别与音节框架的第 6、7、8 段对应。第一音素与后两个音素的相互影响较小,而后两个音素之间的影响较大。

从语谱图看,复韵母的共振峰不像单韵母那样比较稳定,而呈现连续变化的动态特性。在元音三角形中,复韵母表现为一段从出发点到终止点的轨迹。

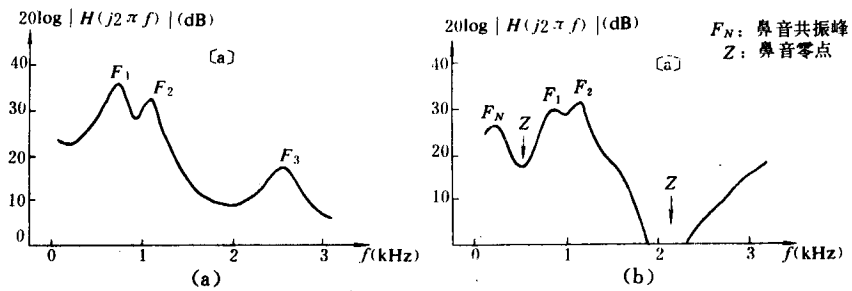
2.6.3　鼻韵母

鼻韵母是以[n]或[ng]收尾的韵母,前者称为齿龈鼻音,后者称为软腭鼻音。16 个鼻韵母是:

[an]	[ian]	[uan]	[üan]	[en]	[in]	[uen]	[ün]
[ang]	[iang]	[uang]	[eng]	[ing]	[ueng]	[ong]	[iong]

发鼻音时鼻腔与口腔相互耦合,这使得相应的声道传输函数不仅有极点而且有零点。对于鼻韵母而言,鼻音出现在尾部,这相当于音节框架的第 9 段,称为鼻尾。在鼻尾和主元音(第 7 段)之间有一个过渡段,称为后音渡,它在框架中是第 8 段。主元音之前还可能存在一个介音

(第 6 段),例如[ian]存在此介音。鼻韵母的重要特点之一是主元音既受介音较大影响又受鼻尾较大影响,后者称为元音鼻化。有些情况下,例如在发四声中的去声(即第四声)时,鼻尾有可能不发音(这称为鼻尾脱落),而由于元音鼻化,对该韵母的鼻化感觉不会改变。元音鼻化主要表现在两方面。第一,在原来的元音频谱中增加了多对新的"极-零点"。其中一对出现在低频(极点为 250Hz 左右。零点为 300Hz 左右),这就在 250Hz 附近造成一个很强的鼻音共振峰,可以用 F_N 表示。另一对则一般出现在各元音的 F_2 和 F_3 之间,这使得该段频谱出现一个深坑。第二个影响是元音各共振峰的宽度和强度都有较大变化。图 2-15 给出了一个非鼻化元音及相应鼻化元音的频谱对比。鼻韵母的第二重要特点表现在它的后音渡。对于[n],主元音的 F_1 和 F_2 在后音渡趋向于分离;而对于[ng],则趋向于平行。这是区分二者的重要特征。此外,鼻尾段常常表现出与 2.7 中将要讨论的鼻辅音相似的特点,因此在连续语音中如果某个音节具有鼻韵母而后续音节具有鼻辅音,那么这两个音节的首尾连成一片,很难将二者的分界定出来。



(a)为元音[a]的声道频谱　(b)为鼻化元音[ã]的声道频谱

图 2-15　元音[a]鼻化对声道频谱的影响

2.7 声　母

声母是一种辅音。所有辅音的共同特点是发音时声道处于某种受阻挡的状态。辅音的另一个重要特点在于这是一种动态特性很强的音,这就是说,发辅音时发声器官的状态变化较大。与之相应,辅音的短时频谱也随着时间而有很大变化。元音与辅音相反,发音时声道不受明显的阻挡 ,它的频谱结构相对稳定。这样,从辅音到元音必然有一个过渡阶段。在图 2-7 给出的音节框架中,声母包括其中的第 1 段至第 5 段或这 5 段中的某几段。由于声母的发音比较复杂多样,对它的发声过程的描述远比韵母复杂。粗略而言,可以按照发声方法和发声部位两个方面来区别及描述不同的声母。但是要具体描述各个声母发音过程各阶段的特点,则需要将这两方面结合起来进行细致的讨论。

众所周知,发辅音时声道的某个部位(发声部位)必然设置了某种障碍。障碍可以分成两大类。第一类,声道中的受障部位完全封闭,空气不能流通。这种情况称为阻塞。第二类,声道的受障部位没有完全封闭,尚留有一条缝隙,空气可以在受阻的情况下流动。这种情况称为阻碍。当辅音与后续元音相连结时必然有一个将阻挡撤除的过程,这称为除阻。除阻阶段就是辅音向元音过渡的阶段。图 2-16 给出了阻塞的横截面及几种不同阻碍的横截面示意图。

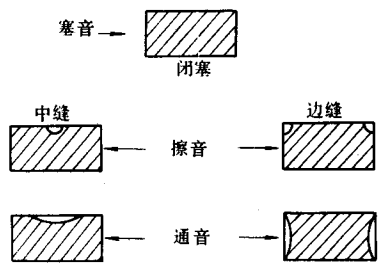


图 2-16　阻塞与阻碍的示意图

在声道阻塞情况下发的音称为塞音。如果声道阻碍的缝隙面积很小，所发的音是擦音。如果阻碍的缝隙面积大一些，所发的音则是通音。

为了研究的方便，可以把发音方法接近的若干个辅音划为一类。汉语普通话的辅音划分为下列各类。

(1)不送气塞音：[b]、[d]、[g]

(2)送气塞音：[p]、[t]、[k]

(3)清擦音：[s]、[sh]、[x]、[f]、[h]

()不送气塞擦音：[z]、[zh]、[J]

(5)送气塞擦音：[c]、[ch]、[Q]

(6)鼻音：[n]、[m]

(7)边音：[l]

(8)卷舌音：[r]

(9)半元音：[i]、[u]、[ü][注①] 和零辅音。

现在将对各类辅音的发音过程及主要特征分别进行介绍。

2.7.1　清擦音

在发擦音时，气流在肺部的压力下流过阻碍处的狭窄缝隙时形成湍流，这就造成了一个噪声源。这一噪声经过阻碍点到嘴唇之间的声道传输后向外辐射，形成了人耳可以听到的摩擦噪声。摩擦噪声的频谱形状取决于阻碍点的位置以及阻碍点到嘴唇之间的声道长度和形状。从语谱图看，摩擦噪声表现为一片乱纹，乱纹的深浅及上下限反映了该噪声的能量在频域中的分布状况。清擦音声母只包含音节框架中的第 3 段(摩擦段)和第 5 段(过渡段)。清擦音的一个明显特点是有持续时间较长的摩擦噪声频谱段(即乱纹段)，其长度为 200ms 左右。不同的清擦音有各自特定的噪声频谱，这是它们的主要特征。过渡段是阻碍去除并向后续元音过渡的发音过程，其间共振峰的走向也表现了不同擦音的特点，这称为过渡音征。对于清擦音而言，过渡音征对于各擦音的听辨不起主要作用。下面介绍各个擦音。

(1)[f]　这是“法”字的声母，是一个唇齿擦音。它的频谱特点是能量在频域内的分布很宽(1000～12000Hz)，而且十分平坦，没有明显的峰起。由于能量分散，谱图上显得十分浅淡。

(2)[s]　这是“思”字和“苏”字的声母，是一个舌尖齿龈擦音。[s]有两种音素变体，在思字

① 在这里采用汉语拼音的声母标注法，当一个音节以[i]、[u]、[ü]为声母(即以这些音为首)时应把它们写成[Y]、[W]、[Y].