3　言语处理入门

3.1　言语信号声学参数

3.1.1　言语信号及特征

语音是语言的物质外壳。语音信号是物理声学信号。语音信号具有如下特点：语音信号是时间依赖的连续媒体。因此处理的时序性要求很高，如果在时间上有 25 ms 的延迟，就会感到不连续。语音是人际交互的手段，因此语音起到了表事、表义、表情的作用。故对语音信号的处理，不仅是信号处理问题，还要抽取语义等其他信息。语音交互涉及言语链的全过程。语音处理应包括：语音信号特征表示、言语识别与理解、言语合成、语音编码、说话人识别等。

语音的波形展示出时域信号的形状。波形的变化会反映在时域特征、频域特征和感知的差异上。实验表明，人类对语音的感知与语音信号的频谱特性关系密切。人类的听觉对语音的频谱特性更为敏感，如共振峰频率和带宽等。

波形图是语音幅度随时间变化的二维图。由于语音产生方式不同，在波形图上有不同的形状。如图 3.1 所示是汉语音节 zhi1 语音波形起始的一部分，以此图为例来说明语音波形图中的几种主要形状。

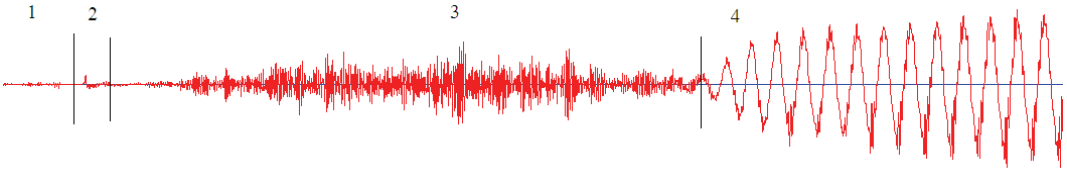


图 3.1　不同语音段的波形示意图

从波形图上的差异可以区分以下几类发音。

无音段或幅度较小的随机噪声段：特点是波形幅度明显小于发音段，波形无规则，如图 3.1 中竖直线所分开的第 1 段。

辅音：包括塞音（爆破音）、塞擦音和擦音等，擦音的特点是波形幅度略大于无音段，波形无规则，一般处于具有周期性波形的元音之前，如图 3.1 中竖直线所分开的第 3 段所示。爆破音：特点是时长很短，仅有一两个脉冲，幅度大于无音段，一般处于辅音前端，如图 3.1 中竖直线所分开的第 2 段所示。

元音：特点是波形幅度明显大于无音段，波形具有周期性，如图 3.1 中竖直线所分开的第 4 段所示。

要想根据波形分辨是什么音素或音节，是很困难的。但可以通过语音信号抽取其特

征以区分不同的音素或音节，如提取基频、样值幅度或能量、时长、语谱，以及这些特征的变化和模式。具体算法请参见“基础与资源”第 4 节言语特征分析。本节将介绍相关的概念。

3.1.2　时长

时长（duration，音长）是指音段（音素、音节等）在发音时的持续时间。通常时长使用时间的单位来表示，如毫秒、秒等，它表示了声学单元的绝对长度。如果是数字语音，有时也用抽样点的个数来表示。这时音长的持续时间等于抽样周期（抽样频率的倒数）乘以抽样点的个数。涉及时长问题的研究包括语速、音段时长、韵律特征分析等。

语音音段可小可大，如音素、声母、韵母、音节、词语等。通常音段越大，时长越长，这表示了音段的绝对长度。但音段相同，时长不一定相同。可以计算音段的平均长度，如计算音节的平均时长，以比较说话速度的快慢。或计算相同声母的平均时长，以分析发音方法对声母时长的影响，或区分不同的声母。

在语音分析中，应更关心在音段相同的情况下，音段的相对时长及其时长变化，如轻声音节、普通音节、重读音节的时长差异；音段在不同韵律结构的位置，时长的变化；语言的熟练程度导致辅音时长的差异等。

在普通话和多数汉语方言中，绝大多数情况下，时长对于区别字词的意义作用不大，但对韵律、语气、感情的表达有影响。而英语中元音的长短有区别意义的作用，例如 ship（船，短 i）和 sheep（羊，长 i）。

1．声母时长

声母的持续时间较短，而且不同声母的时长也不同。数据表明：送气音时长比不送气音长，塞擦音比塞音长，送气的塞擦音最长，不送气的塞音最短。本书统计了所附语料录音的声母时长，单音节中的声母时长分布如表 3.1 的第二列所示。声母在词语中的时长变化如表 3.1 的第三列所示（以单音节生母平均时长为 1）。通常词语中的声母时长比单音节的声母时长短。在连续语流中，韵律单元内部音节的声母时长会更短。

表 3.1　汉语声母时长

	单音节中声母	词语中声母	歌唱中声母
	时长/ms	变化比例	变化比例
不送气塞音	2～10	0.9	1.1
送气塞音	40～70	0.8	1.5
不送气塞擦音	20～70	0.7	1.5
送气塞擦音	75～140	0.8	1.5
鼻音	30～70	0.7	1.6
擦音	70～130	0.8	1.6
边音	40～100	0.6	1.6

表 3.1 中，送气的塞擦音（平均时长约 80ms），长于送气的塞音（平均时长约 70ms），

长于不送气的塞擦音（平均时长约 60ms），长于不送气的塞音（平均时长约 30ms）。鼻音和边音（平均时长约 50ms）长于不送气的塞音，同时短于其他的音。

在歌唱声音中，声母并不与韵母按同样的比例拉长，其变化的比例如表 3.1 的第 4 列所示（以单音节声母平均时长为 1）。声母的演唱时延长比例约为：送气的塞音、清擦音、送气的塞擦音、不送气的塞擦音延长比例约为 1.5 倍；不送气的塞擦音延长 1.1 倍；鼻音、边音延长 1.6 倍。

2．韵母时长

韵母时长是指韵母的持续时间。一般情况下，韵母时长大于声母时长。音节时长增加时，主要是韵母时长延长。以不送气塞音作声母的音节，韵母时长差不多等于音节时长。以其他辅音作声母的音节，韵母的时长比例变小，但也会超过音节时长的 50%。如音节 nang1 的时长为 0.59s，其中韵母时长约为 0.47s。表 3.2 给出单音节阴平调中不同韵母的时长示例。从表中看出，在单音节中，韵母不同或所在音节不同，韵母时长占音节时长比例有差别。

表 3.2　阴平调、单音节中韵母的时长　　单位：s

韵母/音节	a1	a/ba1	a/pa1	a/sa1	a/sha1	a/na1
韵母时长/音节时长/s	0.461	0.519/0.529	0.374/0.514	0.417/0.617	0.440/0.654	0.400/0.467
韵母/音节	yi1	i/di1	ing/ting1	ang/cang1	eng/cheng1	ong/long1
韵母时长/音节时长/s	0.545	0.573/0.593	0.447/0.557	0.530/0.720	0.508/0.628	0.400/0.465
韵母/音节	wu1	u/gu1	ua/kua1	uo/zuo1	ui/zhui1	ou/mou1
韵母时长/音节时长/s	0.482	0.584/0.624	0.456/0.556	0.453/0.543	0.548/0.618	0.405/0.480

3．音节时长

音节时长（syllable duration）是指按音节切分的音段时长。当音节单独发音时，单音节的时长与调类有关。统计表明，音节单独发音时，普通话上声音节最长，去声音节最短。通常更关注四声时长的相对关系。表 3.3 给出本音典录音数据的单音节时长和时长比例。平均时长反映的是发音速度。

在特定发音速度下，连续语流中的音节时长与声调的相关性较弱。主要与音节所在的位置、该音节重读程度、语调以及语言习惯等因素有关。人们在讲话时，根据语法结构、表达需求，不断变换声学单元的时长。让语音表现出不同的缓急、节奏、轻重。

表 3.3　音节时长

声调	阴平	阳平	上声	去声	轻声	平均
音节平均时长/ms	444	480	711	403	266	461
时长相对比例	1.0	1.08	1.60	0.91	0.6	1.04

3.1.3　基频

基音频率（fundamental frequency，pitch，基频）是物理声学术语。复合音中包含一系列谐音，其频率差又为一个不是太小的恒定值，这种复合音中最小谐音（基音）的频率称为基频。

语音信号中基频的高低取决于声带的松紧，是音高的声学物理参数。基频以“赫兹（Hz）”为计量单位。如男声的平均基频为 100～150Hz，女声的平均基频为 250～500Hz。童声的平均基频为 300～600Hz。

本音典正文给出很多音节的基频曲线。图中基频曲线已进行中值平滑。主要供读者观察基频曲线的走势，比较不同声调基频曲线的差异。详细的参数请查看随书的光盘。

汉语是声调语言，在语流中，基频数值实时变化。在语音数字信号处理中，有时会用基频、基频包络近似表示音高，并分析汉语的声调特性。

基音周期（fundamental period）是基频的倒数。基音周期反映的是声带振动周期。图 3.2 是音节 a1 中的一小段语音波形和相对应的声门波。声门波记录了声带的振动情况。从语音波形上来看，可以认为语音波形相邻两个峰值点或者谷值点的距离为一个基音周期。在语音处理中，通常采用短时分析的方法估计基音周期（短时自相关法就是最常用的算法），然后再转换成基频参数。

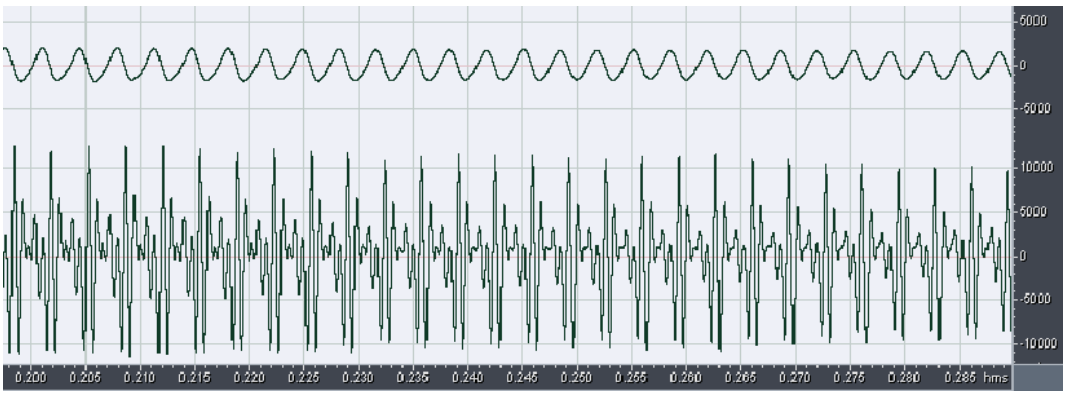


图 3.2　语音波形和声门波

3.1.4　音高

音高（pitch）是语音基频（在言语声中指声带的振动频率）的一种听觉主观心理量。当声音频率由小到大变化时，听觉便产生一种与此相应的由低到高的不同音高知觉。音高与声音的基频相关，但它们之间的关系是非线性的。只有在较高频段，主观感觉的音高大体上与客观测量频率的对数成正比。音高的单位是“美（Mel）”。响度为 40 方（F）、1000Hz 纯音的音高被定义为 1000Mel。音高与频率近似地满足以下方程：

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

图 3.3 中的曲线表示了纯音频率与音高的差异。300Hz 纯音的音高约为 379Mel，500Hz 纯音的音高约为 585Mel，2000Hz 纯音的音高约为 1585Mel。

一个稳态复合声，甚至一个瞬时声对人耳也有一定的音高感觉。如果复合音中包含一系列谐音，其频率差又为一个不太小的恒定值，这种复合音的音高等于最小谐音——基频的音高。

虽然基频的赫兹坐标与听觉的 Mel 坐标不是线性对应，但相差不大。言语工程中，常以基频的高低对音高进行度量。在言语处理中，术语“基频”与“音高”的使用也不尽严格。

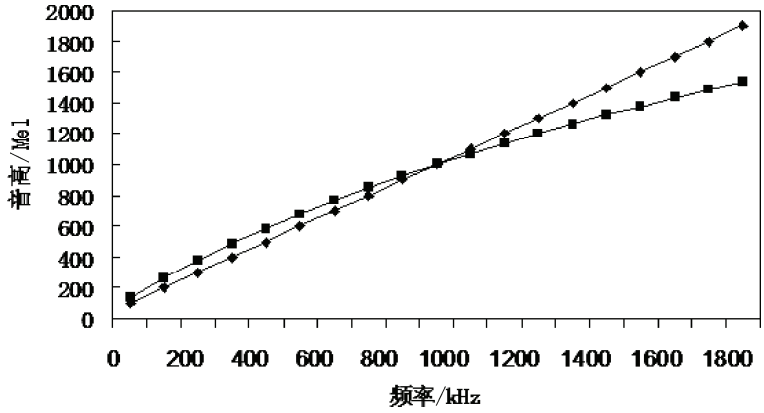


图 3.3　基频与音高

3.1.5　基频参数量化

在连续语流中，基频是随时间变化的。基频（音高）的变化形成了汉语音节的声调、词调、短语调、句调等。男声、女声、老人或儿童的基频绝对值不同，但变化模式相同，听者对声调的分类相同。因此专家提出了基频（音高）的多种相对表示方法，如 D 值、T 值、LZ 值、半音等。

（1）五度记调法给出了声调的数字标调法，而且可以用 1～3 个数字来标记声调的具体调值，参见本书“基础与资源”部分第 1 节汉语语音基础。五度记调法是一种相对表示，它把所有的音都规范成五度，用于比较声调的差异，但不给出具体的音高。在语言调查时，可采用五度记调法记录声调，简捷方便。专家提出的 D 值公式（沈炯，1985）、T 值公式（石峰；2008）可用于计算音高的五度值。D 值计算公式如下：

$$D = 5 \times \log_2 F / F_0$$

其中，F_0 为参考基频，F 为测量点基频。

T 值计算公式如下：

$$T = 5 \times (\lg x - \lg b) / (\lg a - \lg b)$$

其中，a 为调域上限基频，b 为调域下限基频，x 为测量点基频。

（2）半音：研究表明，赫兹坐标上的线性变化并不对应于听觉感知上的线性变化。如果把频率转换成乐律的半音程，那么频率的变化的半音程数与听感上的距离是比较一致的。这样就可以把以“赫兹”为计量单位的频率转换成音乐中的音阶。

在音乐上，将一个倍频程在对数域上等分为 12 份，每份为一个半音。设 F_l 为音阶，对任意测量点基频 F 都按下式量化：

$$F_l = 12\log_2 F/F_0$$

其中，F_0 为参考基频，量化结果 F_l 即 F 相对 F_0 的半音程数。表 3.4 是某数据库的语音基频和音高的量化结果。其中，$F_0 = 131\text{Hz}$，为标准低音 C 所对应的赫兹数。

表 3.4　某数据库基频参数的量化

	句首基频/Hz	音阶	量化	句尾基频/Hz	音阶	量化	平均基频/Hz	音阶	量化
阴平平均值	332	E4	16.1	322	E4	15.6	315	D#4	15.2
阳平最大平均值	327	E4	15.9	298	D4	14.3	271	C#4	12.6
阳平最小平均值	220	A3	9.0	199	G3	7.3	169	E3	4.5
阳平调域	107		6.9	99		7.0	104		7.5
上声最大平均值	288	D4	13.7	263	C4	12.1	242	B3	10.7
上声最小平均值	182	F#3	5.7	145	D3	1.8	124	B2	–0.9
上声调域	106		7.9	118		10.3	119		10.5
去声最大平均值	371	F#4	18.0	356	F4	17.3	335	E4	16.3
去声最小平均值	229	A#3	9.7	213	G#3	8.4	164	E3	3.9
去声调域	142		8.4	143		8.9	155		10.1

3.1.6　基频模式参数

1．基频规格化模式参数

基频包络（pitch contour）指基频随时间变化所形成的基频轮廓。对于音节来说，基频包络就是声调（字调）曲线。从本音典给出的普通话音节基频曲线可以看出，阴平音节的基频曲线高而平。阳平音节的基频曲线由低逐渐升高。上声音节的基频低起再向低，而后升高。去声音节的基频曲线由高逐渐降低。从而形成了汉语的声调模式（tone pattern）。在语音分析中，既要了解音与音之间参数数值上的差别，还要描述其随时间变化的规律（模式）。

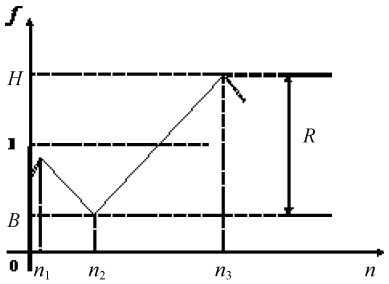


图 3.4　基频规格化模式参数

为了全面描述基频的变化特性，可以设置多个声学参数，如图 3.4 所示。包括 B（基频曲线的最小值）、H（基频曲线的最大值）、n_2（最小值位置）、n_3（最大值位置）、F（基频曲线

起始值）和 E（基频曲线终止值）。通过这些参数，结合时长、能量等参数，较好地描述基频模式，进一步分析汉语语音的韵律特征。

2．基于曲线拟合的基频模式参数

经过基频提取、平滑和归一化处理，可以得到每个音节的基频曲线。然后对基频曲线做三次曲线拟合，拟合曲线公式为：

$$g_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i$$

其中，g_i（x）表示 x 点基频值的基频拟合值。对于某音节来说，拟合得到的 16 个参数和阴平均值 $\overline{f_1}$ 共同组成汉语单音节音高模式。$[a_i, b_i, c_i, d_i]$ 表示模式参数，其中 i = 1,2,3,4 表示声调，反映了 4 个声调的相对变化关系，而 $\overline{f_1}$ 表示了发音人的音高水平。

基频模式参数处理过程如下：首先归一化时长，计算基频，并从中均匀抽取了 10 个基频样点值来代表整条基频曲线。然后计算基频的归一化值 F:

$$F = [\log_2 x - \log_2(\overline{f_1}/2)] \times 5$$

其中，$\overline{f_1}$ 为阴平的 10 个基频样点的均值，x 为测试点基频提取值，F 为归一化后基频值。本算法将一个八度归化到 0～5，阴平均值为 5。利用模式参数 $[a_i, b_i, c_i, d_i]$ 和阴平均值 $\overline{f_1}$ 就可以描述基频模式以及推导基频个性化特点。

对 4 个发音人的单音节语音处理结果如图 3.5 所示。反映了 4 个发音人基频模式的对比。可以看出 4 个声调的模式相似，但反映不同发音人的基频特点，如调域不同、去声的起点基频不同等。

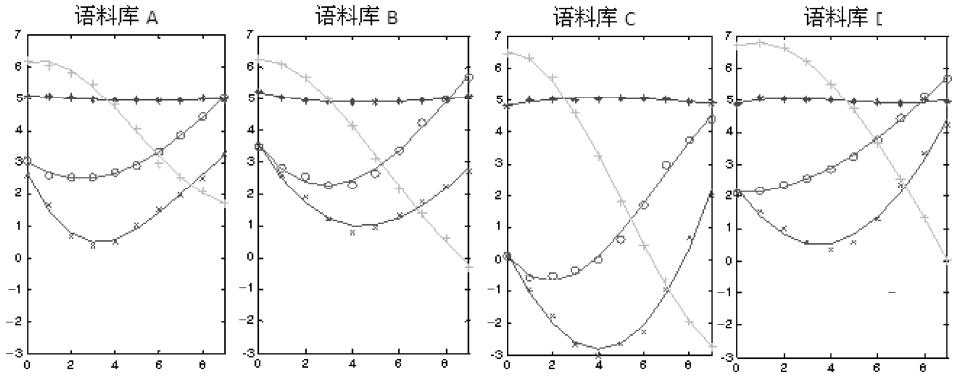


图 3.5　归一化基频曲线

3．韵律标音标准 ToBI

ToBI（Tones and Break Indices）是描述语言口语发音语调和韵律结构变化的标准框架，是一个韵律标音的业界标准，它可以被用于描述基频的高层表达模式。

ToBI 韵律标音系统包括平行的几层，反映韵律的多元性，每一层都定义了表示韵律事件的符号。在这些层中，最重要的是 Tone 层，用于语调的分析；Break-Index 层，指明了相邻两个词之间紧合的程度。另外还有两层：Orthographic（表音）层指明了语调中的词；Miscellaneous（混杂）层指明了不同的随机语音的效果。

社会科学院语言研究所提出了一种汉语韵律标音系统 C-ToBI，共分为 8 层，包括拼音层、声母/韵母层、声调和语调层、停顿层、重音层、语气层、口音层和话题转换层。

3.1.7 声音大小的度量

声音的本质是振动。声波在空气中传播导致某微小区域的气压变化是声压。振动幅度越大，声压越大。单位时间内流过单位面积的声波能量的平均值是声强。但人耳对不同频率的声音的辨别力是不一样的，响度级（loudness level）用来描述这种辨别灵敏度的物理量。通常声音越大，声压、声强、响度级越大。在听力测试、噪声评估时就必须校准声音的强度，如声强计、听力计。表 3.5 列出了声音大小的度量单位。

压电效应将声波转换成模拟电压信号，模拟/数字转换器（A/D）将模拟电压信号转换成数字序列，在对数字语音进行处理时，默认信号未失真，先将信号进行数字化，关注其通用的特性，以及语音幅度变化时对特性的影响。

表 3.5 声音大小的度量单位

	单位	符号	说 明
声压	帕（Pa）	P	某时刻的瞬时气压与标准气压的差。表示声波传播导致某微小区域的气压变化。听觉感知的声压范围约为 20μPa～20 Pa
声压级	分贝（dB） 贝尔（B）	L_p	$L_p=20\log_{10}P/P_0$，P_0是基准声压，$P_0=2\times10^{-5}$Pa。1B＝10dB。听觉感知的声压级范围为 0～120dB。0dB 相当于声压 20μPa。一般交谈时的声压级为 60dB
声强	瓦（W/m²）	I	单位时间内流过单位面积的声波能量的平均值。听觉感知的声强范围是 10^{-12}～1W/m²
声强级	分贝（dB） 贝尔（B）	L_I	$L_I=10\log_{10}L/L_0$，L_0是基准声强，$I_0=2\times10^{-12}$W/m² 1B＝10dB。听觉感知的声强范围约为 0～120dB
响度	宋（sone）		1 宋＝频率为 1000Hz，声压级为 40dB 的纯音的响度。主观度量单位
响度级	方（phon）		响度级等于与它等响的 1000Hz 纯音的声强级。主观度量单位

由于发音器官肌肉的紧张用力不同、气流的强弱不同，致使语音的强弱不同。语音的强弱与高低不同，它们有互补的作用，但不是必然的。通常音强提高时，音高有所提高。但也有强而低或弱而高的。

本音典中给出的图中，显示的波形幅度是相对缩小的抽样值。一般来说，音强大小与声调之间没有明显的对应关系。音强的增强一定伴随音高的提高。时长的加长或缩短，可以补偿音强的弱或强。重音可以用音高和时长的独立变化或协同变化来体现。普通话里的“孝子”和“儿子”里的“子”音强不同，前一个“子”音强比较强，后一个“子”音强比较弱。词语中的轻重音主要是音强的不同形成的，并且，声音的强弱在普通话中还有区别词义的作用，如“地道”中的“道”，分别读为轻声和非轻声时，所表示的意思是不一样的。

3.1.8　等响度曲线

人耳对不同频率的纯音的辨别力是不一样的，响度级（loudness level）正是被用来描述这种辨别灵敏度的物理量。响度级的单位为“方”（phon），方在数值上等于1 kHz 纯音的声强级。为了确定一个音的响度级，需要调节1 kHz 纯音的声强直到它听起来和目标音一样响，此时1 kHz 纯音的声强级数值上就等于该音的响度级。

人对声音响度的听觉感受与声压级不是线性关系。国际标准化组织发布了纯音正常等响度曲线标准，表示了声压级、响度与频率的关系。正常人耳的标准听阈曲线呈两旁高（低频和高频的听觉灵敏度低）、中间低的弧形。

图 3.6 是实验得出的等响度曲线（equal-loudness curves），已被采用为纯音正常等响度曲线标准。同一曲线上的点响度级相同，反映了频率、声强和响度级的关系，该曲线也被称作弗莱彻尔-蒙森曲线（Fletcher-Munson Curves）。观察图 3.6 可以发现：①在频率一定的前提下，声强越大响度越大，在声强一定的前提下，频率的变化也会影响响度的感知；②频率范围在3 ~ 4 kHz 附近，等响曲线彼此之间的距离最大，也就是说，此时人耳的分辨率最灵敏。相比之下，响度单位“宋”（sone）被用来刻画主观感受的声音响度及其变化，这种感觉与音强、频率和波形都有关系。定义一个高于听阈 40dB，频率为1 kHz 的纯音的响度为1 sone 。如果一个音被认为响度是该纯音的 k 倍，则其响度为 k sone 。响度 N和响度级 L 之间满足如下的转换关系：

$$N = 0.063 \times 10^{0.03L}$$

$$L = 33.33 \times \lg N + 40$$

上式表明，响度为 1 sone 的声音，其响度级为 40 phon，并且当响度 N的取值增加一倍，响度级 L 增加约 $33.33 \times \lg 2 \cong 10$ phon 。

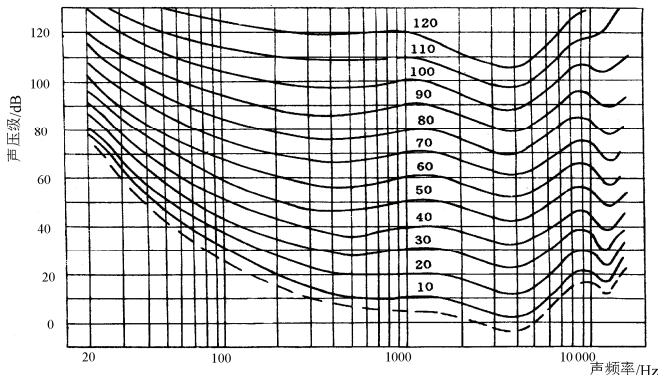


图 3.6　弗莱彻尔-蒙森曲线（等响曲线）

3.1.9　语音频谱

语音信号的频率分析在语音分析和感知研究中占有重要的地位。语音频谱（语谱，

语图）显示了语音信号的频率分布和各频率成分的幅度，反映了非常重要的语音特征。研究表明，人类感知语音的过程和语音频谱特性关系密切。

语音信号是一种非平稳随机信号。但由于人类发音器官运动缓慢，因此语音的特征不会瞬间变化。也就有理由认为语音特征的变化是缓慢的。在工程中，截取语音信号的一个短段（分析窗），并假设窗内语音特征不变。将抽取的特征作为该短段语音的特征。一般窗长为 10～30ms。

对语音信号进行频率分析就可得到语音频谱。如果只分析一个窗，可得到一个二维频谱。如果把整个语音分成很多个短段，逐个计算，就可得到语音的三维频谱。

由于气流在人的声道里发生共振，造成语音信号中某些频率成分（共振频率）的幅度大于周围频率的幅度。共振的中心频率称作共振频率。共振峰指共振频率、共振频率处的频谱值以及频谱曲线半功率点之间的频带宽度。共振峰的频率和带宽将随声道形状改变而变化。

频谱作为特征的相关表示如下：

频谱质心均值（SC），截止频率均值（SR），谱倾斜均值（SS），低-高频谱比均值（SLH），频谱变迁均值（SF），频带周期性均值（BP），频带周期性均值（BP2）（2000~4000Hz），MFCC 系数的均值（$MFCC_i$-$MFCC_p$），MFCC 系数的一阶差分（绝对值）（D*）等。

1．二维频谱

二维频谱可采用多个算法获得，如利用傅里叶变换，计算 FFT 谱。具体算法参见本音典“基础与资源”部分第 4 节言语特征分析。FFT 谱给出了频谱太多的细节，不易于在图上观察语音的共振峰。求解 LPC 系数，再计算频谱，可得到 LPC 谱。LPC 谱上的峰值点很好地描述了元音的共振峰，但它对零点的描述欠佳。求解 CEP 系数，再计算频谱，可得到 CEP 谱。CEP 谱对零点的描述比 LPC 谱好，但难以兼顾对共振峰描述。图 3.7 是单音节阴平 a 的 FFT 谱、LPC 谱。

二维频谱是一个短时（窗内）语音的频谱。如果要了解语音频谱随时间的变化，则需移动分析窗的位置，并计算每一个位置所对应的分析窗的频谱。

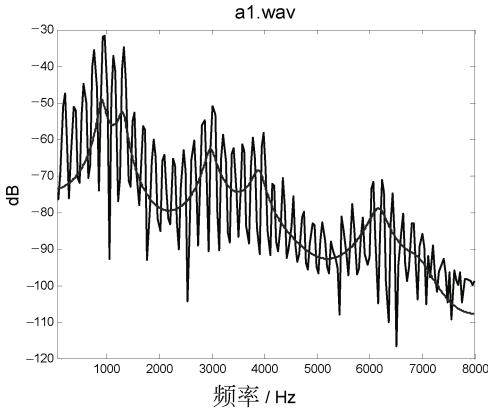


图 3.7　单音节阴平 a 的 LPC 谱

2．三维频谱

移动分析窗的位置而得到一个随时间变化的语音频谱，就可以得到一个三维语谱，并以二维图像表示它。它表示了语音的不同时刻、不同频率成分的幅度。

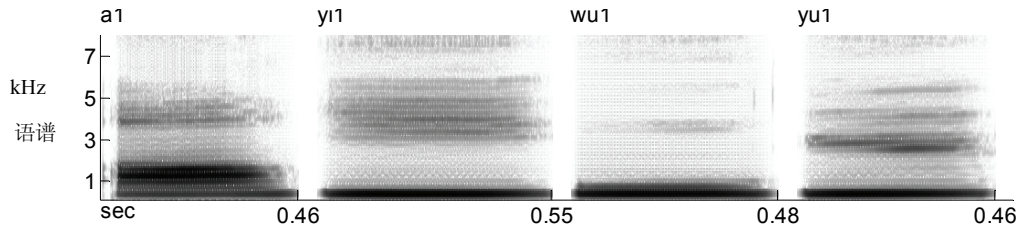
语谱图在 1941 年由贝尔实验室的研究人员发明，它在二维坐标中显示语谱表示三维信息。纵轴表示频率，横轴表示时间，颜色的深浅表示特定频带的能量大小。语音的许多特性直观地呈现在语谱图上。调整多通道滤波器的带宽，可以得到时间依赖的语谱的不同表示。当滤波器的带宽为 300Hz 时，得到的是宽带语谱图，主要用于共振峰研究；当滤波器的带宽为 45Hz 时，得到的是窄带语谱图，主要用于声调语调研究。

音典正文的每一节给出了 4 个音节的宽带语谱图。由于语谱图反映了语音信号的动态频谱特性，本身又反映了语谱随时间的变化（横轴为时间轴），因此在语音分析中有重要的使用价值。有经验的语音学家可以认谱知音，工程技术人员可根据语谱图估计语音的特征。元音、辅音的语谱图各有特点。

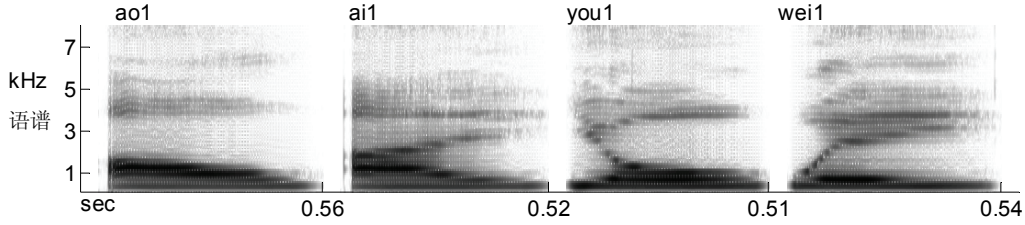
3．元音的语谱

在语谱图上，可以观察元音的频率分布，观察共振峰的位置、带宽以及共振峰的结构和轨迹。从图中可以看出：

（1）不同元语的共振峰的频率分布不同，带宽不同、幅度不同。对比宽带语谱图上以各深色横带（横杠，Bar）在纵轴上的分布可区分不同的元音，如图 3.8（a）单元音的语谱图所示。从图中可以看出，a 的第一共振峰明显高于其他元音，i 的第二共振峰最高。u 与 ü 的共振峰也有明显差异。不同人发音的共振峰位置会不相同，但其分布结构是相似的。



（a）



（b）

图 3.8　单元音和复元音的语谱图

（2）单元音（如图 3.8（a）所示）的共振峰频率平稳，在整个发音过程中基本不变。

双元音（如图 3.8（b）所示）的共振峰频率从一个音变化到另一个音的共振峰。

4．辅音的语谱

在语谱图上，可以观察辅音的频率分布，观察冲直条的位置、乱纹的分布以及浊音共振峰的结构和轨迹。不同辅音的频谱分布不同、幅度不同。图 3.9 是 4 个音节起始 0.25ms 的语谱图。它们的声母分别是塞音 d，塞擦音 ch，擦音 s 和浊音 m。从图中可以看出：

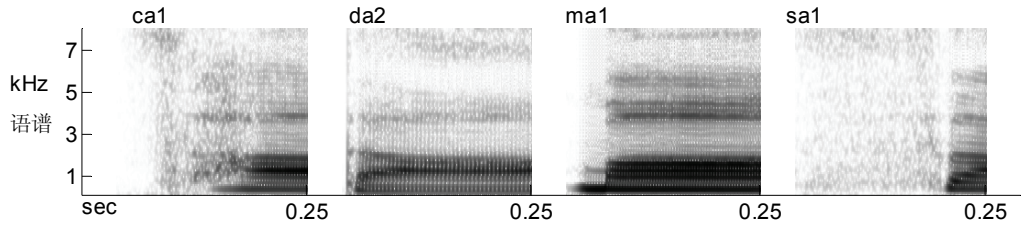


图 3.9　辅音的语谱

（1）乱纹：擦音是一段无规则的非周期性的波形。其频谱表现为无规则的乱纹（fill）。持续时间在几十毫秒，其长短与发音方法和发音速度有关。在频率轴上能量集中区的位置与发音部位有关。图 3.9 中音节 sa1 元音之前有一段乱纹，对应于擦音 s。

（2）冲直条：塞音、塞擦音在除阻的开始是一个爆破段，其波形是一两个脉冲。频谱表现为持续时间很短（约 5 ms）、与垂直频率轴平行的较窄的黑条，称为冲直条（spike）。在 p、g 和 k 的谱图中均明显地看到冲直条。b 和 d 的时长短，t 的幅度小，在谱图中的冲直条不太明显。冲直条在频率轴上能量集中区的位置与发音部位有关。图 3.9 中音节 da2 元音前面较窄的冲直条对应于塞音 d。

受发音方法的影响，塞音和塞擦音又分为送气和不送气。不送气塞音时长最短，送气塞擦音时长最长。与之相对应的频谱的长度也不同。

（3）横杠：浊音是一段准周期性的波形。其频谱是一些横杠，类似于元音的共振峰。如鼻音 m 和 n，边音 l，舌尖音 r 的谱图，而且主要表现在低频部分。参见图 3.9 中 ma1 的元音之前的一段频谱。

（4）空白：在塞音的持阻段，或辅音与元音之间会出现短暂的无声段，频谱看起来空白。参见图 3.9 中 sa1 的元音之前的频谱空白。

3.1.10　音色

音色：也称音质。指人在听觉上区别具有同样响度和音高的两个声音之所以不同的属性。主要由声音频谱的分布决定，也与波形、相位等有关。

对于乐音来说，发音乐器的发音体、发音方法、共鸣器的形状大小都会影响声音的音色，如大提琴与小提琴的音色不同。

发音体不同，音色不同。如胡琴和口琴、提琴与钢琴、弦乐与打击乐的音色不同，原因就在于发音体不同，或是琴弦，或簧片。发音方法不同，音色不同。如同一把小提

琴，用琴弓拉和在必要时用手指弹拨发出的音是不一样的。共鸣器不同，音色不同。如大、小提琴，二者的发音体都是弦，发音方法都是用弓拉，但是大提琴的共鸣器很大，小提琴的共鸣器很小，音色就不一样，大提琴浑厚、低沉，小提琴明亮、悠扬。描述音色的词语还有“深沉”、“明亮”、“锐利”、“纤细”、“浑浊”、“生硬”等。

在任何语言中，音色都是区别性的重要的要素。对于人类的语音来说，发音部位、发音方法、口腔共振腔的形状，以及不同人的发音器官参数的不同等，造成了声音的不同音色。每个音素都具有相对固定的共振峰分布，并且音素种类也会影响到频谱的特征。普通话中发 b 时，主要发音器官是上唇和下唇，发g时，主要发音器官是舌根与软腭，因而造成了声音的不同。同样，g和 h 这两个音，主要发音器官都是舌根与软腭，但g是用爆发方法发音，h 是用摩擦方法发音，发音方法不同，因而声音不同。再比如 u 和 o 的共鸣器都是口腔，但发 u 时口腔开度要比发 o 时小，因而声音不同。

在声学上，音色体现在频谱的分布上，频谱质心、截止频率、频谱下倾，以及周期性成分和非周期成分的强弱等。一般来说，频谱下倾与音色的明亮程度相关：频谱下倾越小，高频成分越强，则音色越明亮。频谱下倾越大，高频成分越弱，则音色越柔和。音色特征与说话人，以及情感状态等有关。不同的说话人由于发声器官的差别，其声带振动情况和共振峰分布都有自身的特点。

基于音色特征，可以区分不同的音、不同的发音人，以及发音人的语音情感。音色也是区别意义的最重要的要素。

3.1.11　言语情感

1．情感描述

人际进行言语交流时，除了语言直接表达的语义信息外，代表讲话者精神状态的情感（emotion，affect，mood，feeling）也传递了重要的信息。情感是未直接表达的精神状态，包括情绪（emotion）和情感（affect，feeling）。情感是人对客观事物的态度体验。

情感能力是人类智能的重要标志，情感在理性行为和理性决策中也起重要作用。随着科技的飞速发展，情感计算受到计算机科学、心理科学等学科的关注。涉及言语、脸像、图像和视频等，研究的内容包括基于情感的认知、分类、识别、合成等。

情感分类是情感分析的基本问题。中国古代名著《礼记•礼运》中就已有“七情”的分类法：“喜、怒、哀、惧、爱、恶、欲七者弗学而能。”中医理论则认为“喜、怒、忧、思、悲、恐、惊”是 7 种情志。人类的情感复杂多变，对情感的分类也没有一个公认的标准。实际上，汉语中有很多细致描述情感、情绪的词语。上述情感范畴分类符合人们的直觉和常识，但在情感计算的研究和应用中受到限制。除了这些基本情感外，更多情况下表现出的是复杂情感、细微情感。

对于情感表示还可以使用维度的连续表示方法。情感维度观给出了描述情感空间的理论构想。如果对情感实现测量，进而可以明确各种情绪范畴的定位和关系。因此，基于情感维度的研究可能处理丰富的细微和混合情感，为情感计算提供了量化的理论和方法基础。

科学心理学的创始人威廉·冯特（Wilhelm Maximilian Wundt, 1832—1920）最早明确地提出情感的三维说（Pleasure-Arousal-Dominance）。他认为，情绪情感由愉快-不愉快、激动-抑制、紧张-松弛三个维度组成，每个维度都在对立的两极之间变化，而且这种变化是连续的心理过程。基于 PAD 三维情感模型，对文本、言语情感、脸像表情、图像情感语义进行评估。

2．汉语情感词语

人类用来表达情感的符号，可能是词汇、图符、颜色甚至字体。最基本的情感符号是以语言文字形式表示情感的形容词，在心理学研究中叫作“情绪体验词”。情感符号都可以转写为特定的情绪体验词，因此，选取情感词语是情感研究的基本问题。

在综合现有多种词典和语义资源的基础上，构建了一个汉语情感词语表，详细内容请参考“基础与资源”第 12 节汉语常见情感词语。情感词语表构建的原则如下。

现代汉语的二字词在词汇中占绝对优势，表示情感的词也不例外，如悲伤、惊奇等。但这些词表示的情感实际上是复杂情感。在分析研究了汉语的情感表达后，我们先确定一些用单字词表示的核心情感类（约 44 类）。核心情感类选取的原则是覆盖基本情感范畴，且能覆盖绝大部分常见情感状态。应尽量保证情感的分布均衡。核心情感类如下：哀，爱，安，傲，悲，惨，怅，诚，烦，奋，愤，感，孤，恨，慌，悔，惑，急，惊，警，静，窘，愧，怜，凉，乱，美，闷，藐，漠，慕，怕，疲，誓，痛，颓，喜，羞，厌，抑，忧，犹，专，妒。

在核心情感类的基础上，增加词语，并扩充情感类到 184 类。扩充的词语包括二字词，三字词或多字词。显然此时所表达的情感是混合情感。

参照情感与认知的关系，以及词语的应用语境，将与汉语情感词语表中的词语分为情、态和评论三类。同时，我们选用 PAD 三维情感模型的量表评估了情感词语表中的词语。将某一特定的情感状态通过愉悦度、激活度、优势度三个维度构成的坐标来表示，将研究的情感对象扩充到情感空间中的复杂和细微情感。

除了表中选择的词语外，汉语中还有很多描述情感的词语。这些词语对情感的表达是有差别的，但有些差别并不是很大。我们参照语义词典和 Hownet 的语义相似度计算，将更多的词语（情感相近词）纳入情感类。

3．色彩词语与情感

在人类的活动中都离不开颜色，那么如何理解颜色呢？颜色是物理特征，语言能对颜色的心理形象进行表达和传递。本书“基础与资源”第 12 节还给出了一个色彩形象词表，这里给出关于色彩形象简单的介绍，更多内容请参考《色彩形象坐标》和《形象配色艺术》（小林重顺）。

1）色彩形象坐标

色彩是形象的代表。色彩词语不仅指颜色本身，还体现了人类神经生理基础和文化内涵。文字中的情感词语让语言更加生动多彩。研究形象的意义和语言的表示，具有重要的意义。

《色彩形象坐标》一书，从美学角度出发，将色彩在整体形象中进行定位，建立了

色彩形象坐标。色彩形象坐标将色彩、配色、语言、环境与人有机结合。书中根据人对色彩的心理体验选出 130 种代表色，构成“色相和色调体系”。同时又选出了 180 个体现人们感性价值的形容词，把色彩、形容词和三色配色群相对应，并加以分类，形成了色彩形象坐标。色彩的“语言形象坐标”如图 3.10 所示。

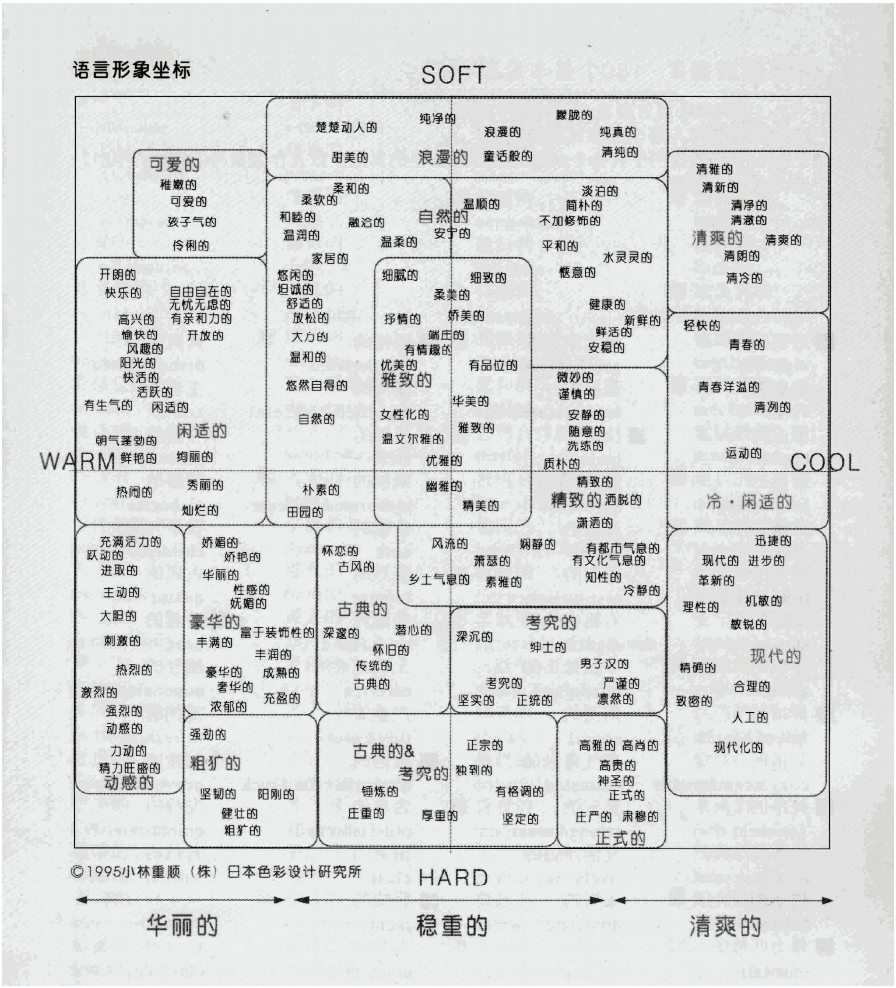


图 3.10　色彩的“语言形象坐标”（《色彩形象坐标》，小林重顺）

2）配色形象坐标

所有的色彩都不是孤立的，常常以配色的形式存在。三色配色是配色的基础。《色彩形象坐标》一书将三色配色形象和“语言形象坐标”相结合，并将各种各样的配色置于冷-暖、软-硬二维坐标系中，形成“配色形象坐标”。通过形象理解配色，使形象搭配理论具体化，进一步将色彩、配色、语言、环境以及人有机结合起来。

在《形象配色艺术》一书中，通过 360 个形容词描述五色配色形象，并将相似的配色归成一个配色群，选出概括性强的形容词作为配色群代表，进一步将其划分为 16 个形

象模式区域。

参考色彩形象坐标，我们选择了形容词表，抽取了大规模图像的颜色、结构等信息，实现了按形容词指定的情感的图像颜色变换，以及图像基于情感的分类，情感表达预测等。

4．情感语音特征

语音的情感表达与声学特征有着密切联系，语音情感的声学表现可以分为韵律和频谱两方面特征。传统分析以韵律特征为主，最常用的韵律参数为基频、时长和能量，而频谱方面则多采用 MFCC 参数进行分析。对语音的情感分析和识别采用的声学和生理特征如下。

声学特征：基频（F0）及其一阶差分（dF0）、音节时长（Dur）、短时能量（Ene）、F0 主分布比（F0DominantRatio）、F0 均值（MeanF0）、F0 最小值 （MinF0）、频谱质心参数（SC）、频谱变迁参数（SF）、频带周期性参数（BF）、频谱滚降系数（Rolloff）、频谱低高频比（LHRatio）。

生理特征：心电图的 R 波峰值（PRH）及其一阶差分（dPRH）的均值、心跳间期（HP）及其一阶差分（dHP）的均值、呼吸的最低（minR）和最高（minR）点、呼吸的平均值（meaR）和中位值（medR）、呼吸数据的一阶差分的平均值（dR）。

5．语音情感建模

早期的情感建模研究一般基于规则驱动，研究者将统计分析的结果转化为规则描述来确定不同情感语音在声学特征上的变化。主要方法是对韵律特征或者共振峰等参数统计分析所得到的结果进行模板化而形成的规则。

情感特征建模常常也采用基于语境分类的建模方法，分别建立统计分类模型或者概率模型。例如，基于统计分类的 KNN 模型、支持向量机（Support Vector Machine，SVM）、决策树、神经元网络、隐马尔可夫模型（HMM）等。此外对基频重置现象采用线性回归综合神经元网络进行了建模，达到了较高的预测精度。以上建模方法不仅可以用于韵律特征，也可以针对频谱特征进行建模。基于数据驱动，采用码本映射的方法实现了频谱特征的变换建模。使用语音的 Mel 频率倒谱参数（MFCC）构建了中性的特征模型，并通过实验发现它可以很好地应用于情感识别的前端。如使用了基于 GMM 和决策树的概率模型来对情感特征进行建模和预测。

3.2 语 流 音 变

3.2.1 语流音变

本节再次讨论语流音变（本书在 1.5.1 已有介绍）。汉语拼音字母代表的是一个个汉语音位。音素（segment）是与音位相对应的语音单位。在发音过程中，当用这些音素组成更大的音段单位——音节、词、短语、语句、篇章时，语音发生了变化。这就是语流音变，称这些变化了的音素为音位变体。虽然在给汉字注音时，没有改变拼音字母，

但可以引入更多的国际音标符号来描述这些变化了的语音。

影响语流音变的因素很多，如发音人、发音速度、表达要求、语流节奏、该音的语音环境等。还包括情感、语言、声学、语音、生理的影响等。这里仅介绍音段语境对语流音变的影响。

语流音变造成语音的变化很多。从声学上讲，语音的变化可能是音高、时长、能量和频谱的改变。从而导致声调、节奏、重音和语调的改变。这里主要介绍音段语境引起的语流音变所导致的语音频谱的变化。

在连续的发音中，发音器官连续性的动作，语音所处的环境不同，相邻的音互相影响而发生变化。语流音变造成语音频谱的变化表现为元音共振峰轨迹改变、辅音音轨改变、韵尾丢失等。本音典正文在**【区分辨正】**的标题下，给出一些语流音变的示例，请参考之。

3.2.2　辅音的协同音变

辅音的频谱主要表现为清音冲直条、乱纹的有无和强频区频谱，浊音共振峰的结构和轨迹。辅音所处语境不同，其频谱可能有所改变。

1．辅音受后接元音影响的协同音变

汉语有 21 个辅音可以作声母，与作为韵母的元音拼接成音节。辅音受后接元音影响发生音变。韵母按照韵头的不同，可分成开口呼、齐齿呼、合口呼、撮口呼 4 类，简称四呼。四呼是按照发音开始时的口型和舌位进行分类的。表 3.6 给出四呼分类表，并给出韵母首音素的共振峰和发音描述。共振峰的值是从本音典录音提取的。同一发音人在不同语言环境下的发音或不同发音人的参数会有差异。

表 3.6　汉语音节四呼的分类

四呼韵母分类	首音素典型共振峰	发音描述
a，ai，an，ao，ang	F1＝1.2kHz，F2＝1.5kHz	开口呼，舌位低，展唇
o，ou	F1＝0.6kHz，F2＝0.9kHz	开口呼，舌位半高，圆唇
e，ei，en，eng，er	F1＝0.6～1.2kHz， F2＝1.3kHz	开口呼，舌位半高，展唇
i，ia，ie，iao，iou，ian，in，iang，ing	F1＝0.6kHz，F2＝3.3kHz	齐齿呼，舌位高，展唇
u，ua，uo，uai，uei，uan，uen，uang，ueng，ong	F1＝0.3kHz，F2＝0.6kHz	合口呼，舌位高，舌面后，圆唇
ü，üe，üan，ün，iong	F1＝0.3kHz，F2＝2.4kHz	撮口呼，舌位高，舌面前，圆唇

辅音可能随后接元音不同而发生变化。图 3.11 给出以辅音 p[p^h]作声母的音节起始部分频谱。可以看出 p 在不同韵母前，其频谱分布稍有不同。p[p^h]的音轨与元音共振峰相呼应。在齐齿呼韵母前，高频成分增强。音典正文中给出了辅音声母与四呼韵母拼接时的短时频谱，请参考。

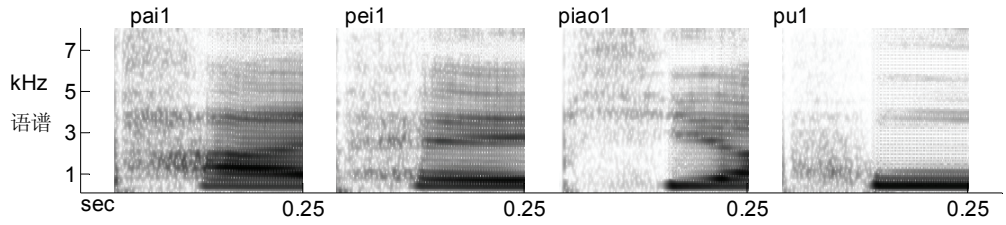


图 3.11　辅音受后接元音影响的协同音变

2．辅音受前音节韵尾影响的协同音变

音节之间的辅音音变，首先考虑声母辅音受前音节韵尾的影响所发生的音变，即 V1C2V2 中 C2 的音变。在分析 V1C2V2 中 C2 的音变时，前音节韵母按照韵尾（韵腹）的不同，韵母的口型和舌位可能对 C2 造成影响，发生音变。同时，还要考虑 C2 所在音节为轻声时 C2 的音变。韵母的韵尾（韵腹）分成 6 类，如表 3.7 所示。

表 3.7　韵母的韵尾（韵腹）分类

韵母的韵尾（韵腹）分类	韵尾音素共振峰	发音描述
a，ia，ua	F1＝1.2kHz，F2＝1.5kHz	舌位低，展唇
o，ao，iao，uo	F1＝0.6kHz，F2＝0.9kHz	舌位半高，圆唇
e，ie，üe，er	F1＝0.6～1.2kHz，F2＝1.3kHz	舌位半高，展唇
i，ai，ei，uai，uei	F1＝0.6kHz，F2＝3.3kHz	舌位高，展唇
u，iou，ou	F1＝0.3kHz，F2＝0.6kHz	舌位高，舌面后，圆唇
ü	F1＝0.3kHz，F2＝2.4kHz	舌位高，舌面前，圆唇
an，en，ian，in，uan，uen，üan，ün	F1＝0.3kHz，F2＝2.1kHz	前鼻音
ang，eng，iang，ing，iong，uang，ueng，ong	F1＝0.3kHz，F2＝1.2kHz	后鼻音

图 3.12 是词语“俏丽”，“魅力”，“概述”和“凑数”的语谱图。可以看出，声母 l 和 sh 受前音节韵尾和所在音节的韵母的影响，其频谱的强频区频率有所不同。

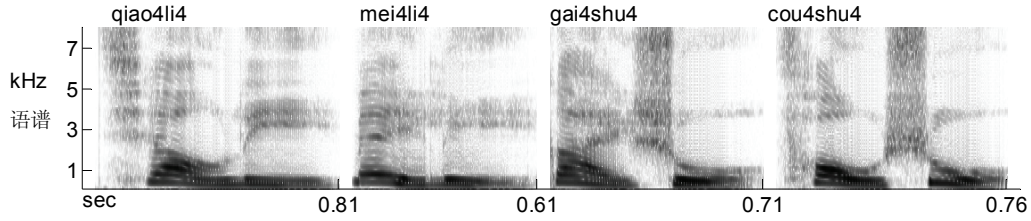


图 3.12　辅音受前音节韵尾影响的协同音变

3．韵尾辅音受后音节辅音影响的协同音变

普通话中的韵尾辅音包括-n 和-ng，如表 3.7 所示。在连续语流中，受后音节声母辅音的影响产生协同音变。即 V1C1C2 或 V1C1V2 中 C1 的音变。

图 3.13 是“艰难”，“允诺”，“人民”和“门面”的语谱图。可以看出，词首音节韵尾辅音-n 受后音节的影响，其共振峰轨迹有所改变。

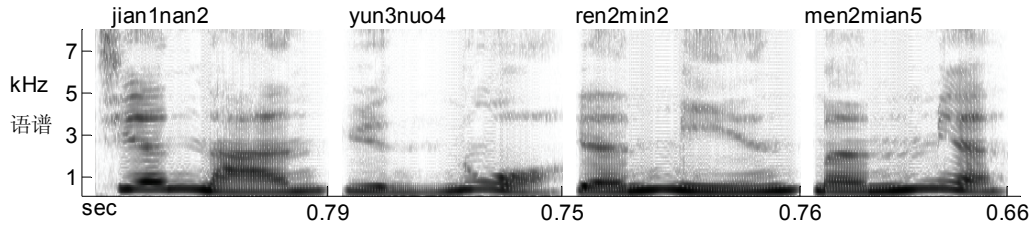


图 3.13　韵母中韵尾的音变

3.2.3　元音的协同音变

在音节中或在连续语流中，元音受到相邻音素的影响会发生音变。下面分别介绍。

1．韵母中首元音起始部分的音变

在非零声母的音节中，韵母的首元音受声母的影响会发生音变，改变共振峰轨迹。图 3.14 是 bai3，zha1，mo1 和 zhou1 音节起始 25ms 的语谱图。可以看出 bai3 中元音 a 起始处共振峰，受 b 的影响，F1 和 F2 的频率低。zha1 中元音 a 起始处共振峰，受 zh 的影响，F1 的频率低，F2 的频率高。其他韵母中首元音的音变可参看音典正文各节中的综合图形，观察其频谱和共振峰轨迹的变化。还可以查看【区分辨正】关于韵头 a（o，e，i，u，ü）的频谱中的内容。

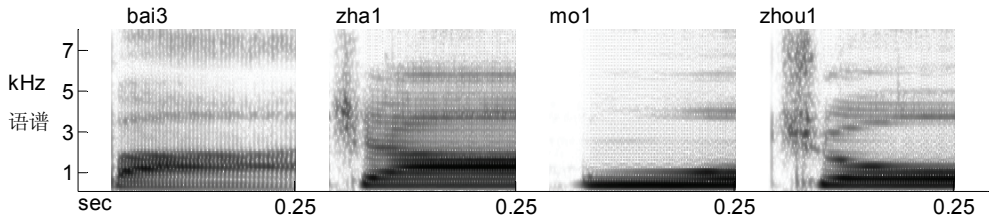


图 3.14　韵母中首元音的音变

2．复韵母中韵腹元音的音变

汉语复韵母可能包含 1～3 个元音，复韵母中韵腹元音受前后音素的影响会发生音变。图 3.15 是 wai1，yao1，yang1 和 yan1 音节的语谱图。可以看出，4 个元音 a 的共振峰受前后不同音素的影响，F2 的频率和轨迹不同。在 wai1 中元音 a 的 F2 没有稳定段，频率逐渐升高。在 yao1 中元音 a 的 F2 没有稳定段，频率逐渐降低。在 yang1 中元音 a 的 F2 有稳定段。在 yan1 中元音 a 的 F2 没有稳定段，频率逐渐降低，但维持较高频率。其他复韵母中韵腹元音的音变可参看音典正文中的综合图形，观察其频谱和共振峰轨迹的变化。

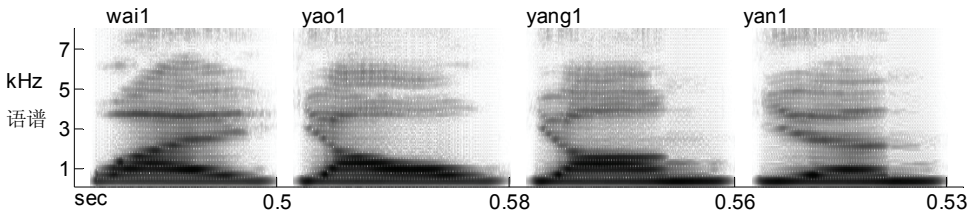


图 3.15　韵母中韵头的音变

3．韵母中韵尾的音变

韵母中韵尾受前后音素的影响会发生音变。图 3.16 是“辍学”，“苟且”，“西湖”和“女孩儿”的语谱图。可以看出，这些词的词首音节的韵尾共振峰轨迹发生变化。在“辍学”和“苟且”中，词首音节韵尾 o 和 u 的 F2 升高。在“西湖”和“女孩儿”中，词首音节韵尾 i 和 ü 的 F2 下降。

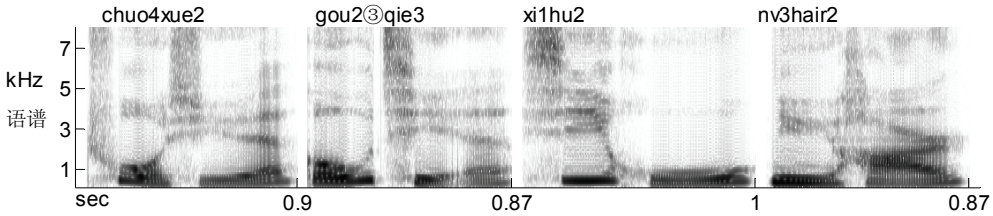


图 3.16　韵母中韵头的音变

3.3　韵 律 基 础

3.3.1　韵律

韵律（prosody）是超音段（supersegmental）的相关术语。韵律原用于对诗的韵律和作诗法的研究。在语音上，狭义地讲，韵律与语音的突显方面的区分有关，尤其表现为音高、时长和响度的变化。在音系学上表现为音调、音高重音或重音的使用。韵律具有层级结构（prosodic hierarchical structure），见 3.3.2 节。

韵律表达要满足表事、表情、表义的需求，在自然语言中起着非常重要的作用。韵律体现出不同的语气、态度、感情色彩等，可以帮助听者更好地理解说话人的语义。从语言表达来看，韵律表达特征包括节律（rhythm）、重音（stress，accent）和语调（intonation）等，见 3.3.3 和 3.3.4 节。

韵律的声学参数是指语音的基频（pitch）、幅度（amplitude）、时长（duration）、频谱（spectrum）以及它们随时间的变化模式。统计归纳这些声学参数的规律可以建立韵律模型，见 3.3.5～3.3.8 节。

韵律在感知上的知觉为语音的响度轻重、音高高低、语速快慢、停延长短、语段节奏、音色明暗等，以及它们随时间的变化，见 3.3.8 节。

韵律也受到语音发音、语法、语义和语用的约束。学者们从不同的方面研究韵律，如语言（文）学、音系学、语音学、言语工程，见 3.3.9 节。

3.3.2　韵律层级结构

每一个韵律元素称作韵律成分。韵律成分的大小不一。通常认为一个较小的韵律成分包含在一个更大的韵律成分中，由此形成了韵律的层级结构（prosodic hierarchical

structure)。音系学认为韵律成分从小到大依次是莫拉、音节、音步、音系词、附着语素词组、音系短语、语调短语和韵律语句。言语工程中，较多关注音节（sylleble)、韵律词（prosodic word)、韵律短语（prosodic phrase)、语调短语（intonational phrase)、语调以及语篇韵律（discours prosody)。有的专家提出了韵律架构说，阶层式架构由音节（SYL)、韵律词（PW)、呼吸句群（BG)、韵律句群（PG)、语篇韵律（DP）构成。

言语的韵律组织受到语义、语法、词法、语音、生理以及心理等不同层面约束和协调。其外在表现是将语音从长到短按多个层次划分成不同的韵律成分，韵律成分之间有长短不等的静音。各韵律成分的特征、变化模式共同形成了韵律系统。

（1）韵律词（prosodic word）也叫音系词（phonological word)，是韵律层级结构中的一个层级。一般短于韵律短语，是相当于音步的音系单位，是节律的基本单位。韵律层级理论认为，一组附着形式和相联系的词一起形成一个音系词。从韵律的角度来研究词，韵律词是语流中节律的基本单位。为了与语素词或句法词分开，叫作韵律词。音步可以构成韵律词。

韵律词的主要特点是：韵律词内部不能停顿，在韵律词边界的停顿不是必需的。汉语大多数的韵律词为两个音节，一般为三个音节以下的语法词或词组。单音节的韵律词一般由单音节词延长为一个音步构成。三音节音步是超音步。

韵律词是语音-语法界面上的基本节奏单位。它总是作为一个稳定音步出现。韵律词具有相对稳定的音节音高协同模式。内部不会出现韵律分界现象（停顿和音高重置等)。韵律词的音高和边界调负载了重音配置和韵律高层信息。

韵律词和语法词之间存在一定的联系，但二者也有差异。大多数情况下，一个双音节或多音节韵律词是一个语法词。单音节语法词也经常前附或后附到另一个语法词上，构成一个韵律词。少数时候，单音节词可以延长成为一个韵律单位，构成单音节韵律词。这些单音节词多数为功能词，例如连词、介词等，也有少数是动词、代词等。

（2）韵律短语（prosodic phrase）是韵律层级结构中的一个层级，对应于音系短语。由一个或多个韵律词组成。这个层级单元具有某种语调特性。通常认为韵律短语的长度是 7 个音节，变化长度为两个音节。这与呼吸群的长度相当。

韵律短语具有相对稳定的短语调模式，如音高音阶的逐步下倾（declination)。具有相对稳定的短语重音配置模式，即与句法结构相关的常规重音模式。例如，偏正结构一般是偏重，主谓结构一般是谓重，述宾结构一般为宾重，述补结构一般为补重。韵律短语内的韵律词之间存在停顿。在韵律短语边界的停顿是必需的，且停顿均值长于韵律词边界的停顿均值。韵律短语和语法短语之间存在一定的联系，但二者也有差异。

（3）语调短语（intonational phrase）是韵律层级结构中的一个层级。一般长于韵律短语。这个层级单元是较长的短语或包含在长句中的小句，具有某种语调特性。

在语法上，相当于较长的短语或较短的语句。语调短语有特定的语调模式。它可能通过一些方式与句法或篇章结构相联系。

语调短语（intonation phrase）相当于语法上的分句，是句子层面上的音系规则作用的辖域。由于它与韵律词及韵律短语具有一定的同构性，所以有时候一个语调短语可能

就是由一个韵律短语或者一个韵律词构成的。它的主要特点如下。

语调短语内部可能包含不止一个韵律短语调模式和韵律短语重音模式，因而会出现相关的节奏分界，主要表现为韵律短语末尾音节的延长（或伴有较短的无声停顿）和韵律短语之间的音高（包括音阈和调域）重置。具有取决于不同语气或句型的语调模式（例如与陈述、疑问和感叹相关的不同音阶走势）。

（4）句调（intonation，语调）是韵律层级结构中的一个层级。一般长于语调短语。这个层级单元是较长的语调短语或具有语调特性的语句，也称语调。语调在功能上用来表达不同的语气、语义、情感等。语调在感知上是语音的轻重缓急和抑扬顿挫的知觉。语调在韵律学上表现为节奏、重音、调阶的变化。语调在声学上表现为由音高、调域、时长，音强和停顿等的变化形成旋律模式。语调涵盖了韵律的诸多问题，下面将仔细介绍。

（5）语篇是韵律层级结构中的一个层级。一般长于句调，相当于文本中的段落或篇章。语篇韵律在功能上用于实现不同的表达风格。在韵律上，语篇韵律表现在语速、重音、调阶等的变化上。在声学特征上，由时长、调域等不同声学特征的变化实现。语篇表现为韵律对段落与篇章的整体约束，并约束了句调形成的模式。

3.3.3 语言表达与言语韵律

韵律是语言表达的手段。关注韵律问题的学科包括语言文字学、语音学、声学、言语工程等。文字表达上的节奏、凸现和语气，语音表现为节律（rhythm）、重音（stress，accent）和语调（intonation）。虽然它们不能等同，但它们的相关性是显而易见的。这里更关注韵律在语音上的表现。

1．节律

在言语中，由于声学特征变化表现出有规律的间断所产生的知觉模式，称之为节律（rhythm，节奏）。相邻韵律单元之间韵律特征差异、静音都可产生言语间断知觉。

对于文字表达，可以根据语法结构在文字间插入标点符号。对于语音表达，所表现出的音段间的停顿、音段的延缩以及音高的高低变化规律和模式。节律的声学征兆为停顿前音段延长、音节调域（或音高上下限）变化、停顿后音节音高重置、停顿等。

停顿可以细分为：能使意义产生变化的区分性停顿；语噎、哽咽、气喘吁吁、欲言又止等形成的生理性停顿；能形成语句结构的结构停顿；放在焦点词语后面的逻辑停顿，起到突出、强调的作用。

2．重音

在言语中，对那些念得比较重的音叫重音（stress, accent）。重音可表现为词重音或句重音。注意不能孤立地看待重音特征的变化，重音的实现以其他成分的对比存在为基础。重音的体现必须建立在它与其他相邻成分的相互关系上。

词重音指在大多数非单音节汉语词语中存在的重读和非重读音节的对比。赵元任指出，“实际上，在没有中间停顿的一连串的带正常重音的音节中，不论是一个短语还是复合词，其实际轻重程度不是完全相同的”。尽管轻重对比不一定会影响所表达的意义，

但轻重处理不当，会感到不自然。

句重音可以分为语法重音、逻辑重音。在具体的语句中，某些语法成分经常读重，这种重音叫语法重音。为了突出、强调句中某些词语而需要重读，这种重音叫逻辑重音。

语法重音也称一般重音或意群重音，出现在平常的说话里，只是在原来的重音节上再稍加重。所表达的语气是基本的、中性的。句子的一般重音在句末的最后一个短语中实现。

一般重音与语法的关系：通常谓语比主语读得重。在主语＋谓语（＋补语）＋宾语句型中，宾语读得重。如果是双宾语，后宾语读得重。在主语＋谓语＋补语句型中，补语读得重。

逻辑重音比词重音、一般重音强，可以重叠在它们上面，也可以加强“非重音音节”，改变原来的重音结构。强调重音的作用可能是夸张、肯定、特指、并比、对应等，以加重语气。

重音体现在每个音节上，以基频凸显（pitch prominence）为主要特征。基频凸显指在语流中，某重读音节的音高明显高于或低于周围音节而突出的现象。重音的主要声学特征为增加音强，延续时长，扩大调域，音色也比较清晰。

3．语调

语调（intonation）是韵律层级结构中的一个层级。本节重提语调，是把它看成语言表达的手段，介绍言语工程中关注的语调问题。

语调的分类：根据语句音高包络的形状可分为平语调、降语调、升语调、曲语调等。根据语调表达的语气可分为陈述语调、疑问语调、祈使语调、感叹语调等。根据语调实现的表达功能，分为中性语调、表情语调。如果按此分类，中性语调是指陈述事实时采用的语调，其他均可归属表情语调，如表达疑问、命令、高兴、愤怒等的语调。

赵元任先生认为英语语调的结构有两个层次。首先将语调分成“句首”和“句身”两部分。句首再分为“句首前/主句首”，句身分为“核心/句尾”。4 个成分线形排列。“核心”是调群几个重音音节的最末一个音节，其后的音节是“句尾”，其前为“句首”。在某个语句中，这 4 个成分不一定全都出现。

言语工程中，对语调的研究主要涉及语调模型、语调的韵律特征和语调的声学特征，以及语调预测和语调生成。

与语调感知相关的声学参数有基频、调域、能量、时长、停顿、频谱等，这些参数受到生理、心理、发音、语言，以及语境、场景、语用的影响，随时发生变化，而且它们之间互相关联。这些声学参数中，基频的变化对语调最为重要。它对传达陈述、疑问等不同的语义、重音的表达以及情感的表现都有比较重要的意义。能量、时长则与基频一起共同形成完整的语调表现。

语调模型通常基于语调生成的叠加观、平行观等建立，其中叠加观对语调模型影响最大。学者们提出了各种方法建立韵律模型，以产生这些声学参数，如基频（包络，pitch contour）模型、时长模型等，其主要的关注目标为语调音高的变化。

在语调的描述中，常使用声调的低音点和高音点所连接得到的低音线和高音线来表

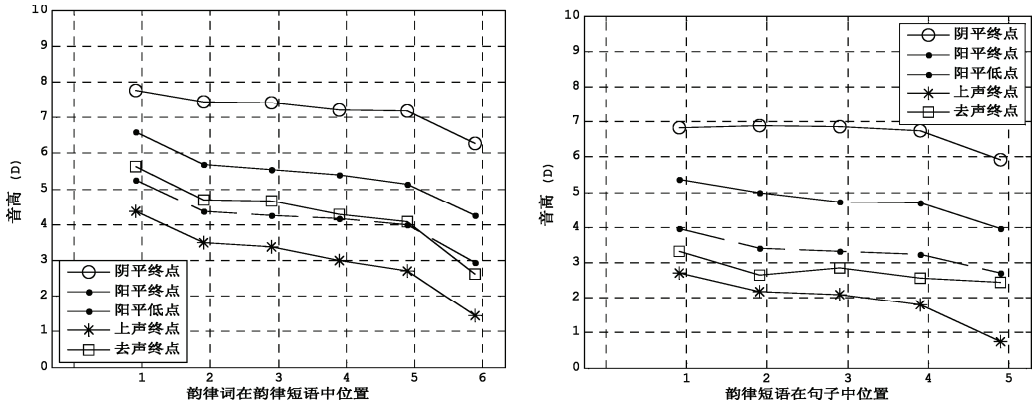
示语调中除去声调之外的短语调与句调成分。可采用以下特征对音高线的特点进一步描述：语调的基调、降阶（音高下倾）、突显等。基调的高低主要表现为音高线整体上移/下移。降阶主要表现为音高走势的上扬/下倾或平缓/曲折，局部音高的升/降等。基调和降阶的变化伴随着调域，能量、时长、停顿、频谱等的改变。

1）降阶

在语句中，当出现后续音节的音高低于前面音节的音高的趋势，称为音高下倾或降阶。在语调的描述中，常以连接各声调的低音点（基频最低点）所得到的低音线，来表示语调中除去声调之外的短语调与句调成分。低音点可能指上声音节基频的最低点、阳平音节基频的起点、去声音节基频的终点等代表声调低音特征的点。低音点连线的走势作为语调的特点来描述语调，如语调下倾，语调上扬。语调曲折可以看成语调局部的下倾或上扬。

赵元任先生认为汉语的中性语调，即不包含任何特殊的语义与情感的语调，呈现一种整体的下倾趋势。这种语调的变化主要体现在句子的低音线的变化上。在短语的内部，低音线呈现比较明显的下降趋势。在短语之间，可以看到基频的重置。但如果将各个短语结尾的低点连接起来，则可以看到句调在全句中的下降趋势。

我们统计分析了两万句新闻文本语料的录音。语句长度包含 12～28 音节，每个语句含 3~6 个韵律短语，韵律短语含 2~5 个韵律词。采用 D 值表示音高。在选取的语料库中对相同韵律位置的韵律成分进行分析，分析了音节的基频下限均值和分布。图 3.17（a）中曲线是韵律短语中韵律词的尾音节终点及阳平低点音高的连线。图 3.17（b）中曲线是语句中韵律短语的尾音节终点及阳平低点音高的连线。同一条曲线中音节的声调相同。从大规模语句音高的统计结果表明，汉语陈述句语句以及语句中的语调短语的低音线均呈音高下倾的趋势，且各条曲线的下降趋势一致。句首和句尾韵律成分下降稍大。



（a）韵律短语中韵律词的音高连线　　（b）语句中韵律短语的音高连线

图 3.17　汉语陈述短语语调模式

2）基调

基调的高低主要表现为音高线整体上移/下移。在下倾语调中，基调可以使用语句下倾的起点来表示。基调的变化反映了句子整体的基频平移。在情感语调中，可以观察到比较明显的基调变化，如在高兴语调中基调提高，在悲伤语调中基调降低。在篇章韵律中，篇章首、中、尾语句的基调可能不同。

3）起伏度

通常语句的语调走势不会是平坦的，而是处于起伏变化中。语法上的焦点，在语音上形成重音，导致语句的局部特征突显。各种不同的情感与重音的体现则集中表现在高音线上。重音将引起重音音节的高音点提高，同时还会使重音后的音节的高音点下降。重音对低音点也有一定的影响，但表现远没有高音点明显。

3.3.4　语调模型

1．语调生成的叠加观

语调是一个韵律层级，一般是较长的语调短语或具有语调特性的语句，其特征主要表现为音高在语调辖域内的变化。在语音中，字或词往往有自身固有的音高变化，如在重音语言中词重音的位置及在声调语言中的字调。语调的实现就表现为这种字、词固有的音高变化与语调对音高变化影响的叠加。这是语调生成的叠加观的基本观点。

汉语是一种声调语言，赵元任先生曾指出“汉语的语调实际是词的或固有的字调和语调本身的代数和”。汉语的语调是字调与语调对音高的控制相互叠加的结果。对汉语语调进行分析与描述，要将这些不同层级对音高变化的影响区分开。

字调与语调对音高影响的相互叠加并不是简单的基频相加的“代数和”，而是对调阶进行叠加的“代数和”。在汉语中，字调的调型是比较稳定的，并不会因为语调的影响而发生调类改变，如处在升调处的阴平音节并不会变成阳平。语调对字调音高的影响是对字调整体的调阶的影响，语调的音高的变化是通过字调整体调阶发生上浮或下沉实现的。字调与语调呈现出一种相互依存、相互制约的关系。

语调的生成过程是一个由整体到局部的规划过程。语调整体的变化被首先确定，如语句的基调、基频下倾等。语句中各个词语、音节的音高变化在语调所确定的调阶约束下实现。即语调与声调叠加是一个由上而下的叠加过程。

2．Fujisaki 模型

Fujisaki 认为人在说话时存在一条音高基线。当人说话时，对音高的控制会使得音高偏离这条基线，再逐渐回到这条基线。

在 Fujisaki 模型中，人对语句音高的控制可以分为两种，短语控制（phrase control）与重音控制（accent control），这两种控制分别通过短语命令（phrase command）与重音命令（accent command）来实现。短语控制用来表示句子整体的短语级的语调变化趋势，而重音控制用于字调、词调、重音等短时间的音高变化。短语控制与短语调、句调相对应，重音控制则与词调相对应，两者叠加生成语调的方式与语调生成的叠加观是一致的。Fujisaki 模型示意图如图 3.18 所示。

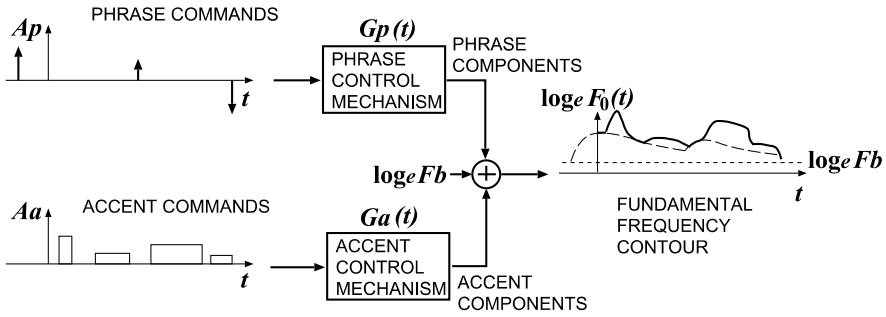


图 3.18 Fujisaki 模型示意图

短语命令为一系列脉冲，其通过一个临界阻尼二阶线性系统，形成语调中缓慢变化的音高曲线。这一音高变化与音高基线一同组成了语调变化的长时效果，确定了语句中字、词音高实现的基础。在汉语中，这一语调变化确定了每一个音节声调的调阶。通过短语命令可以观察到短语的整体音高变化。

重音命令为一系列阶跃函数，其通过另一个临界阻尼二阶线性系统，形成语调中短时的音高变化。这一变化在不同语言中对应的语言元素有所不同，如在英语中为单词重音，日语中为词调，汉语中为声调等。在汉语中，每个声调可以通过一个或两个重音命令来实现。

Fujisaki 模型从对音高曲线变化直接建模的目的出发，结合语调生成的叠加观与人对音高控制的生理特点，得到了由音高基线、短语控制、重音控制叠加生成语调的模型。这一模型符合自然语调生成的特点，已经成功地应用于世界上多种不同的语言。

Fujisaki 模型也提供了一种通过合成对语调进行分析的方式，即通过模型生成音高曲线的方式，得到语调的一系列的特征，如短语命令等。短语命令的大小、位置又可以应用于对语调的进一步分析，如分析不同语气、情感下语调的变化等。

3．PENTA 模型

PENTA 模型的全称是并行编码目标逼近模型（parallel encoding and target approximation），是许毅在量化目标逼近模型（qWA）的基础上提出的。它是一种针对汉语进行设计的韵律模型。PENTA 模型认为语音的韵律是语言的各种不同通信功能在语音的基频曲线上进行并行编码的过程。这些不同的通信功能包括文本、词汇、语句、焦点、话题、情感、意图等。共同编码的结果得到了每个音节声学特征实现的目标值，包括音高目标、音强目标、时长目标等。在 PENTA 模型中，各种通信功能的过程如图 3.19 所示。在语音实现过程中，语音的声学特征受到生理因素等的制约，以逐渐逼近目标声学特征的形式得到实现。

PENTA 模型中的音高目标可以表示线性函数 $x(t) = mt + b$，其中 m 与 b 分别为音高目标的斜率与起始音高。当线性函数斜率为 0 时，称为静态目标，即目标为一个固定的基频值；当 m 不为 0 时，称为动态目标。汉语中的阴平、上声的音节的音高目标分别为“高”与“低”的静态目标，而阳平、去声音节的音高目标分别为“升”与“降”的

动态目标。

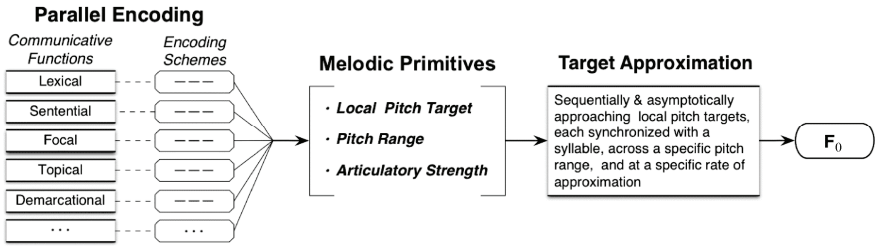


图 3.19　PENTA 模型示意图

根据人的生理发音特点与实验结果，PENTA 模型被设计为三阶临界阻尼系统，其响应包括强迫响应与自然响应两个部分。其中，强迫响应的部分即为音高目标本身，而自然响应部分为指数衰减的成分，表现为基频曲线不断向音高目标逼近的过程。PENTA 模型中的动态与静态音高目标与基频曲线向音高目标逼近的过程如图 3.20 所示。

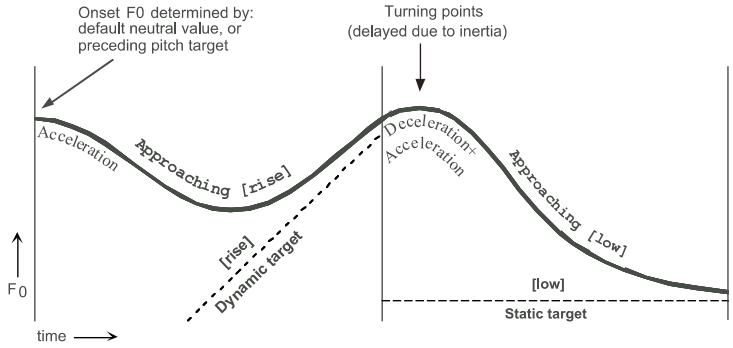


图 3.20　音高目标及基频曲线逼近过程

在 PENTA 模型中，句子的语调变化由句中各音节音高目标的变化表示，而音高目标则是在句中各种不同的通信功能并行编码的过程中生成。模型体现了不同的通信功能与句子韵律的关系。PENTA 模型专门针对汉语进行设计，音高目标的设计与汉语声调的感知相符合，可以描述汉语中不同声调组合形成的复杂基频变化模式。

3.3.5　韵律的声学特征——时长

音段时长（音长）指的是声学单元的长度，是韵律的重要特征之一，它描述了语流的时间结构。在语流中，声学单元的时长复杂多变。不同音素的时长不同；同一音素在不同语境下的时长不同；发音速度、轻重读法、韵律成分或位置等都会影响时长。在语音处理中，相对时长更被关注。

时长对区分不同韵律边界具有重要贡献，表现为韵律边界前音节延长、各音节时长比改变、停顿长度的变化等。分析表明：在语流中在非韵律边界处的音节缩短，在韵律边界前的音节延长。通常认为无声段（停顿）越长，边界的知觉等级也越高。

1．时长特征描述

（1）时长作为韵律特征的表示如下（以音节为例）。

音节时长均值（MeanDur），音节时长标准差（DDur），音节时长最大值（MaxDur），音节时长最小值（MinDur），音节时长范围（RangeF0），音节时长主分布比（DurDR），音节时长一阶差分的均值（绝对值），音节时长一阶差分的标准差（DDDur），音节时长的斜率与截距（DurS，DurIn）等。

（2）时长作为韵律特征的相关因素如下。

声母类型，韵母类型，音节在韵律词中的位置（首、中、尾），韵律词在韵律短语中的位置（首、中、尾），韵律短语在语句中的位置（首、中、尾），语调类型，音节的重音属性（重重、重、标准、轻读、轻声），音节到重音的位置（–2，–1，1，2）。

（3）语流中的静音段（停顿）也作为韵律特征，表示如下。

停顿时长均值（MeanPDur），停顿时长标注差（DPDur），停顿时长最大值（MaxPDur），停顿时长最小值（MinPDur），停顿时长范围（RangePDur），停顿时长主分布比（PDurDR），停顿时长一阶差分均值（DPDur），停顿时长一阶差分标准差（DDPDur），停顿时长的斜率和截距（PDurS，PDurIn）。

2．语句中音节时长变化的特点

在连续语流（语句）中，音节时长发生变化，与该音节单独发音、词语中音节的时长均不相同。分析表明，时长与该音节所处韵律结构位置相关，语调不同导致音节时长的分布不同等，如韵律边界前音节时长延长；重读音节（词）时长延长，其余音节相对缩短；疑问焦点在句尾的疑问句的句尾延长等。语句中的时长模式归纳如下。

（1）连续语流中，音节的平均时长短于音节单独发音时的时长。各音节的相对时长比例符合一定的规律。

（2）词语（二音节词、三音节词、四音节词）的时长模式是基础。

（3）音节处于韵律结构中的位置影响音节时长变化。通常韵律单元尾音节，时长延长；所属韵律单元在韵律结构中的层次越高，时延越长。

（4）音节（词语）为重音时，则时长增大。重音改变词语中各音节的时长模式，改变语句内部各短语的时长比例。如重读音节（词）时长延长，其余音节相对缩短；疑问焦点时长延长。

某语料库中，音节在不同韵律边界处的时长分布如表 3.8 所示。

表 3.8　音节在不同韵律边界处的时长分布

	音　节	韵律词边界	韵律短语边界	语调短语边界	语句边界	平　均
语料库 1	273 ms	325 ms	392 ms	388 ms		
语料库 2	216 ms	235 ms	297 ms	315 ms		
语料库 3	221 ms	210 ms	298 ms	325 ms	307 ms	245 ms

3.3.6　韵律的声学特征——基频

音高（基频）是韵律最重要的特征，它描述了语流的音高结构。在语流中，声学单元的音高复杂多变。不同音素的音高不同；同一音素在不同语境下的音高不同；发音速度、轻重读法、韵律成分或位置等都会影响音高。在语音处理中，相对音高更被关注。

音高对区分不同韵律边界具有重要贡献，表现为韵律边界之前的音节低音点下降，通常认为低音点下降越低，边界的知觉等级也越高。焦点音节或韵律边界之后的音节高音点提高（重置），通常认为高音点提升越高，语音的表现也越丰富。在语流中非韵律边界处的音节调域变窄。

1．基频相关的韵律特征

（1）基频均值。在无声调语言中，通常计算语音的基频平均值（均值），并以此来度量不同音段基频的差异或评估发音人的平均音高。如计算语句中音节、韵律词或韵律短语的基频均值，可以看出重读音节的音高比普通音节高，词调、短语调、语调的升降。如计算某发音人语音的平均值，根据男声的平均音高比女声低的规律，就可以推测发音人的性别。

（2）调域。调域（pitch range，基频范围）是指个人说话时音高的变化范围。有时也指一般言语中使用的音高范围，或用于副语言目的的扩展范围。有时会关注言语中音高的局部最高值与局部最低值的差别（也称音高跨度，pitch span）。描述调域的参数有调域、调域下限、调域上限等。吴宗济先生的研究表明语调变化在音节基频上的反映是基频中值的移动和调域的展缩。

调域下限（baseline，低音点，基频最小点）指基频范围下限的基频值。对于单音节来说，不同声调音节的调域下限不同，上声音节的调域下限最低。在语流中，调域下限的变化范围比较大。它受到音节在韵律单元中的位置、语调以及相邻音节特性等因素的影响。通常调域下限的降低是韵律边界的标志。有时会参照调域下限来分析基频的高低、变化，以及韵律结构等。

调域上限（topline，高音点，基频最大值）指基频范围上限的音高值。有时会参照调域上限来分析基频的高低、变化，以及重音等。对于单音节来说，不同声调音节的调域上限不同，去声音节的调域上限最高。在语流中，音节受到相邻音节特性、位置、重读、语调等因素的影响，调域上限不是稳定不变的。

2．基频特征描述

（1）基频作为韵律特征的表示如下。

F0 均值（MeanF0），F0 标准差（DF0），F0 最大值（MaxF0），F0 最小值（MinF0），F0 的中位值（MedF0），F0 音域（RangeF0），F0 主分布比（F0DR），F0 一阶差分的均值（绝对值）（MeanDF0），F0 一阶差分的标准差（DDF0），F0 斜率与截距（F0S，F0In），（MeanF0-MinF0）/（MedF0-MinF0），MeanF0/MedF0。

（2）基频作为韵律特征的相关因素如下。

声调类型，音节之前及之后的音节声调类型（阴平、阳平、上声、去声、轻声），

音节在韵律词中的位置（首、中、尾），韵律词在韵律短语中的位置（首、中、尾），韵律短语在语句中的位置（首、中、尾），语调类型，音节的重音属性（重重、重、标准、轻读、轻声），音节到重音的位置（–2，–1，1，2）。

3．语句中音节音高变化的特点

语流中，音高会发生各种各样的变化。下面结合几个例句进行说明。

声调相同，在语句中不同位置的音节，其音高不同。图 3.21 是语句：“粗树树粗树不秃，秃树树秃树不粗”的基频曲线。该语句中 6 个阴平音节的基频均值分别为 342Hz，297Hz，279Hz，339Hz，295Hz，290Hz。另外 8 个去声音节的基频也各不相同。

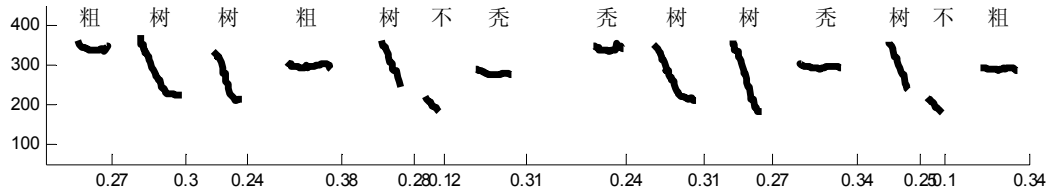


图 3.21　语句“粗树树粗树不秃，秃树树秃树不粗”的基频曲线

声调相同，在语句中不同位置的音节，其基频最小值不同。图 3.22 是语句：“刘华成名前，常常学弹琴”的基频。该语句中的音节都是阳平，但它们的基频包络不同。这些音节的基频最小值为：234Hz，215Hz，207Hz，232Hz，169Hz，223Hz，225Hz，225Hz，177Hz 和 181Hz。有的研究者以阳平基频最小值的连线研究陈述句语调下倾。

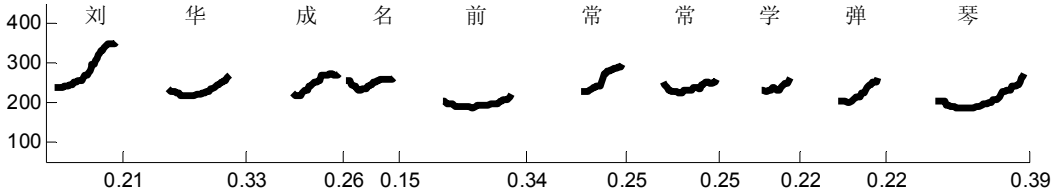


图 3.22　语句“刘华成名前，常常学弹琴”的基频曲线

声调相同，在语句中不同位置的音节，其调域不同。图 3.23 是语句：“玉不琢，不成器。人不学，不知义”（yu4 bu4 zhuo2，bu4 cheng2 qi4，ren2 bu4 xue2，bu4 zhi1 yi4）。的基频曲线。各音节的基频最大值、基频最小值如表 3.9 所示。在语流中，调域上限的变化范围也比较大。它受到音节重读的程度、在韵律单元中的位置、语调以及相邻音节特性等因素的影响。通常调域上限的增高是焦点的标志。

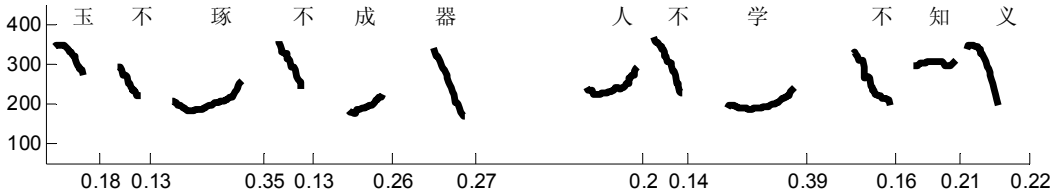


图 3.23　语句“玉不琢，不成器。人不学，不知义”的基频曲线

表 3.9　语句“玉不琢，不成器。人不学，不知义”的调域列表

汉字	玉	不	琢	不	成	器	人	不	学	不	知	义
拼音	yu4	bu4	zhuo2	bu4	cheng2	qi4	ren2	bu4	xue2	bu4	zhi1	yi4
最大值/Hz	358	297	257	358	222	341	292	367	239	336	312	361
最小值/Hz	272	215	158	237	175	167	216	227	185	179	284	196
调域/Hz	86	82	99	121	47	174	76	140	54	157	28	65

疑问句与陈述句的基频包络不同。图 3.24 是疑问语句“商店几点开门？”（shang1 dian4 ji2③ dian3 kai1 men2？）和陈述语句“商店十点开门。”（shang1 dian4 shi2 dian3 kai1 men2。）的基频曲线。可以看出疑问句的基调较高，句尾音节的基频也高，陈述句的句尾基频较低。

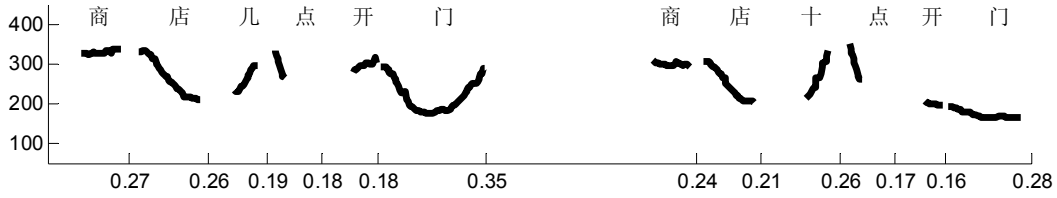


图 3.24　疑问句与陈述句的句尾基频对比

图 3.25 是不同情感下（悲伤、高兴）语句“他提拔哥哥”（ta1 ti2 ba2 ge1 ge5）的基频曲线。可以看出在悲伤情感下，语句的调域窄，基频低；在高兴情感下，语句的调域宽，基频高。

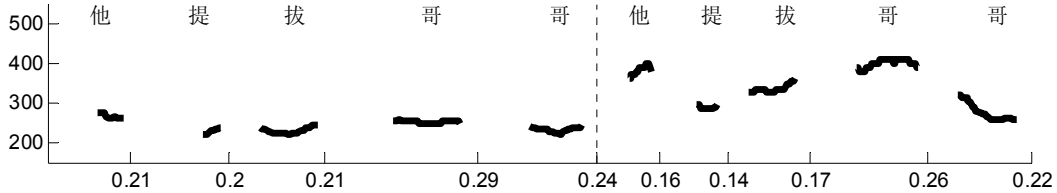


图 3.25　情感语句（悲伤、高兴）的基频包络对比

语调体现了由语音学、音系学、语言学、语用学综合决定的语音时域和频域特性。在音高上，主要表现为音高整体上移/下移，音高走势的上扬/下倾或平缓/曲折；局部音高的升/降；时长的延/缩；音强的高/低等。

3.3.7　韵律预测

言语工程中需要对韵律参数进行建模。韵律建模就是要计算言语中的韵律特性，建立韵律模型，以产生这些韵律参数。通常韵律模型分为两类：一种基于规则，另一种基于大规模语料库，采用数据驱动方法。当然它们也不是绝对分开的。

1．韵律预测的规则模型

在较早期的文语转换系统中，其韵律模型主要是基于规则的。规则模型认为人在讲话时，发音器官的运动是连续的，声道的形状不可能突变。因此连续语句不是孤立声学

单元的简单拼接。专家们研究各种语音现象，总结归纳出人类发音过程中的规律并将其写成规则。按照这些规则计算参数的变化轨迹，并合成出连续自然的语音。系统中的基本规则（knowledge）和参数值都是手工建立的。最有名的基于规则的模型是 Klatt 的参数模型，这是共振峰模型。源滤波器规则模型还有 LSP 模型、正弦模型等；在波形合成系统中建立的是非源滤波器模型。

2．韵律预测的统计模型

目前语音合成系统大多采用数据驱动技术，从而韵律建模也转向了统计模型。早在 1996 年，开始采用神经网络选择合成基元的办法，构造了一个数据驱动的语音合成系统。目前也有采用决策树、HMM、GMM、DBN 等训练韵律模型。

3.3.8　基于决策树的韵律建模中的特征集

语句中语音声学参数的数值和变化具有一定的规律性。在言语工程中常采用统计的方法来描述。如计算在某种条件下某种参数的均值、方差、变化率等。也可以采用决策树方法对条件不同的语音进行分类，然后再计算每类语音的参数，并求出均值。

在基于 HMM 模型的合成的训练过程中，首先明确影响语音参数的发音信息和语境信息，然后根据这些信息设计决策树问题集，就可以根据这些信息，将受不同影响的发音单元分类，然后分别训练 HMM 模型。此处不介绍 HMM 的训练过程，而是列出训练涉及的特征集示例。

在某基于 HMM 模型的汉语合成系统中，训练涉及的特征包括：汉字信息；发音信息；时长信息；前、后音节发音信息；音节个数、位置信息；韵律词的个数及位置信息；韵律短语的个数及位置信息等。

训练涉及的特征共有 108 个，如表 3.10 所示。这些特征都是基于文本标注的特征产生的。其中涉及的发音特征共有 24 个，包括拼音、声调、声母、韵母、声母发音部位、声母发音方法、韵头发音方法、韵尾发音方法等 8 类，考虑到当前拼音、前音和后音，共 24 个特征。涉及的韵律特征共有 84 个。

表 3.10　基于决策树的韵律建模中的特征集

文本标注类	特　征　类
汉字	后一音节声调（5 类）
拼音	后一音节声母
音调	后一音节韵母
声母	后一音节声母 Type
韵母	后一音节声母 Class
声母 Type（22 类）	后一音节韵母 Type
声母 Class①（10 类）	后一音节韵母 Class②（按韵头分为 5 类）
韵母 Type	后一音节韵母 EndType（按韵尾分为 6 类）
韵母 Class	当前词中的音节个数（w 类）
韵母 EndType	前一词中的音节个数（wf 类）

续表

文本标注类	特　征　类
音节开始时刻	后一词中的音节个数（wb 类）
音节结束时刻	当前短语中音节个数（p 类）
前一音节拼音	当前句子中音节个数（s 类）
前一音节声调（5 类）	当前音节在词中的正位置（1-w 类）
前一音节声母	当前音节在词中的倒位置（w-1 类）
前一音节韵母	当前短语中词个数（wp 类）
前一音节声母 Type（10 类）	当前句子中词个数（ws 类）
前一音节声母 Class（10 类）	当前词在短语中的正位置（1-wp 类）
前一音节韵母 Type	当前词在短语中的倒位置（wp-1 类）
前一音节韵母 Class（按韵头分为 5 类）	当前句子中短语个数（ps 类）
前一音节韵母 EndType（按韵尾分为 6 类）	当前短语在句子中的正位置（1-ps 类）
后一音节拼音	当前短语在句子中的倒位置（ps-1 类）

对于决策树的特征集和问题说明如下：

① 声母分类：按照声母发音方式的不同可以分为 7 类。在这个基础上，针对现有声学参数基频 F0 在处理清音和浊音时的不同，将擦音中的浊音声母 r、零声母单分一类。另外，对于语音中的停顿等静音段单独划分一类；这样，在该合成系统中，对声母的分类总数为 10。

② 韵母分类：按照韵头分类，根据四呼，加上特殊音节的韵母，韵头一共分为 5 类。儿化后的韵母分类略有不同，例如 zhi、zi 中的 i 在未儿化时分类为“开”，而儿化之后分类为“齐”。

按韵尾分类，除了上面提到的 4 类，还增加鼻尾音、儿化韵尾，因此可以分为 6 类。

③ 对于未知相关的特征，分别提问是否相等和是否小于等于。例如“当前韵律词在韵律短语中的正序位置是否等于 3？”和“当前韵律词在韵律短语中的正序位置是否小于等于 3？”。这里小于等于的问题有助于对韵律单元内部进行进一步的划分。

④ 问题集设计：在决策树聚类的过程中，需要相应的问题集来对模型空间进行划分，划分时主要考虑语境环境的影响，即对那些可能对发音产生不同影响的语境提问。

由于决策树是二叉树，问题集中的问题必须是真假型，例如“当前拼音是不是 yang？”，“前一音节的韵尾发音方法是不是‘开’？”，“后一韵律词的音节个数是不是小于等于 3？”等。

3.3.9　韵律、语法与语义

韵律组织在言语交流中起着非常重要的作用，它不仅是正确表达语义的关键，还能直接反映讲话人的态度、意向、情绪以及对听话人的期望等许多文字无法表达的信息。人在讲话时，总是先将这些意识层的信息转化成音系层的表达手段，如选择怎样的语调、轻重模式、节律模式等，最后再通过控制发音器官来实现相应的声学目标。韵律是语义、句法、语用等要素的综合表现。

韵律的正确表达可以弥补文本信息的不足，有效地分化歧义，提高表达效果。可见，韵律离不开句法，句法支配韵律，但韵律有时也会制约句法。

在语调生成中的短语和词语与语法结构中的语法短语和语法词并不完全相同，但它仍然受到语法结构的制约，如语调结构的短语边界一定出现在语法结构允许的位置。语调还会受到句子语法类别的控制。例如对陈述句与疑问句来说，它们的句子整体基频变化趋势并不相同。

1．句法支配韵律

在语言的文本表示中，语言符号线性排列。语法结构的分析表明，语言单位呈现出层次结构，语音中的语言意义单元通过语音韵律单元的切分来体现。语音分段的价值必须通过表达一定的语法意义的段落来体现。

韵律结构与语法结构有着非常紧密的联系。韵律受到句法结构的支配。语流中韵律边界必须以句法结构为基础，韵律边界总是出现在句法结构允许的位置，并在句法结构成分之间。但是句法结构的边界并不是设置韵律边界的必要条件。另外，韵律边界的层次倾向于同句法结构的层次保持某种程度的一致。但句法结构的层次高低并不是韵律边界层次的决定因素，因为韵律边界并不总出现在较高的句法层次上。

例如，韵律词的构成是以语法词为基础的。通常，如果前后没有依附成分或单音节语法词，双音节或三音节语法词就是韵律词；如果有，应当与依附成分或关系紧密的单音节语法词共同组成韵律词。对于单音节的语法词，一般只有当其处于重读地位,或者因为左邻右舍已经是完整的音步而无所依附时，才通过延长而构成独立的韵律词。其他情况下，它前附或后附到另一个语法词上，与其共同组成韵律词。句法对韵律分段的支配作用体现在以下几个方面。

（1）句法单位与韵律单位之间不存在一一对应关系。句法单位可能是语素、词、短语和句子等；韵律单位可能是莫拉、音节、音步、音系词、附着语素词组、音系短语、语调短语和韵律语句等。

（2）语流中韵律边界必须以句法结构为基础，韵律边界总是出现在句法结构允许的位置，并在句法结构成分之间。但是句法结构的边界并不是设置韵律边界的充分条件。

（3）韵律边界的层次总是倾向于同句法结构的层次保持某种程度的一致。但是句法结构的层次高低通常并不是韵律边界层次的决定性因素。因为在很多情况下韵律边界并不出现在较高的句法层次上。

2．韵律制约句法

韵律结构与句法语法结构不一定一致，也不存在一一对应关系。虽然韵律离不开句法，但有时也会制约句法，它们之间是相互影响和制约的关系。首先，二者的不一致性体现在，韵律上常常有“跨界”的现象，即“语法上有平级结构关系的成分，节律上离开平级的右接成分向左跨越上级语法边界，与没有直接语法关系的成分合并为一个节律单元”。主要原因之一，就是句法层级与韵律层级有着不同的生长规律。句法层级倾向于纵向生长。一个句子中每增加一个修饰成分，得到的句法结构树就会增加一个层级，句

子越长、结构越复杂，得到的树的深度就越大。而韵律的层级却是有限的，句子长度和复杂度的增加只能导致各层级上并列的韵律单元数目的增加，而非深度的增加，即韵律层级倾向于横向生长。句法也受韵律的影响与控制。如韵律边界的正确界定可以弥补文本信息的不足，有效地分化歧义，提高表达效果；韵律也可以改变句法，使一些非法句法合法化，或制约一些合法的句法变成非法等。如：

（1）语音分段首先服从于韵律分界，句法层级分界次之。自然语流中，由于韵律的基本单位是音步（韵律词），因此语素、单字词都将与相邻音节组合成韵律词（短语）。

（2）通常双音节动词与宾语形成动宾结构时，如果宾语为双音节词，则该动宾结构形成一个或两个韵律词，常用的结构为[2+2]，如“区分类别”；如果宾语为单音节词，则该动宾结构为一个韵律词，如“区分类，种植花”等，其结构为[2+1]。对于[2+1]结构，若重音落在动词上，则这样的结构不符合三音节韵律词的重音结构（抑-扬），此时有必要修改文字，以符合韵律需要。如“区分类别，种植花草”。也有些[2+1]结构可以成立，如“糊弄人，管管他”。这时动词的第二个音节轻读（轻声），该结构形成的是韵律词的重音结构（抑-轻-扬）。

（3）韵律可以改变句法。如使核心重音后移。边界并不出现在较高的句法层次上。然而在语音中，韵律边界是一个重要的句法结构的标志。韵律结构的改变可以引起对整句话语法结构理解的改变，如“上班时间，不得玩游戏、打私人电话”，如韵律短语边界仅出现在标点符号处，就会造成“不得玩游戏”与“打私人电话”相并列的印象，从而改变句子本身的语法结构。

3．基于语法信息的韵律标注

在文语转换系统中，要对输入的文本进行分析，预测韵律结构。即先建立起一个韵律模式，并以此来指导韵律相关声学参数的确定，并生成最终的波形。韵律结构预测有基于规则和基于统计的方法。基于统计学习方法预测韵律结构的步骤是：首先设计或收集一个包含大量同时带有语法标注（包括分词和词性标注）和韵律标注信息（韵律词和韵律短语标注）的语料的数据库， 然后用某种学习算法（包括 CART、决策树、Markov 模型、基于转换的学习方法 TBL 等）建立一个训练模型，并用从数据库中提取出的语法和韵律特征参数对模型进行训练，从而得到最终的韵律预测模型。用此模型作用于新的只有语法标注的文本，就可以自动生成韵律结构的标注。

为了训练韵律预测模型，我们研究了韵律与语法、语义之间的关系，撰写了“基于语法信息的韵律结构标注规范”，对大量的文本语料进行韵律标注。标注规范的主要内容如下。

1）韵律词标注规则

（1）韵律词的标准长度为两个音节，可以是 1~4 个音节。

韵律词的标准长度为两个音节，与两个汉字的语法词相对应，如“北京”。韵律词的长度可以是三个音节，与三个汉字的语法词的相对应，如“西红柿”。计入语法词中的介词、语素词等词，韵律词的长度可以大于三个音节，如“我的同学”、“小不点儿”。韵律词的长度可以是一个音节。当单音节语法词（主要是连词、介词等功能词或单音节动

词）处于重读地位，或者因为左邻右舍已经是完整的韵律词而无所依附时，也可以单独构成单音节韵律词，如“三”。

（2）韵律词边界一定是在语法词边界处。

鉴于在绝大多数情况下，一个韵律词由一个或多个语法词构成，因此，认为语法词边界才有可能成为韵律词边界。

（3）语法词可以与另一个相邻的语法词附和，构成一个韵律词。

单音词或者单音语素要成为韵律词，需与相邻词合并。虚词（代词、助词、介词等）在汉语中表现为轻声(且不携带重音)，常贴附于相邻的实词上，形成韵律词，如助词“了”、“过”、“着”。有些单字词不负载重音，常贴附于相邻的实词上，形成韵律词，如助动词“能”、“得”、“要”；否定词“不”；介词“跟”、“对”、“在”；指代成分“他”、“你”。动词后只能指派一个重音。动词后与其后较短语法词可以合并为韵律词，如“吃饭”。

2）韵律短语标注规则

（1）韵律短语边界一定是在韵律词边界处。

韵律层级结构的无递归性质决定，韵律词一定包含在韵律短语中。

（2）韵律短语的长度一般为 5~9 个音节，可以是 1~11 个音节。

韵律短语常常由表示指代/数量的定语＋中心语、助动词＋动词、4 字以上并列结构、三字以上述宾结构、三字以上介宾结构，以及中间带有“的”、“地”、“得”、“了”、“着”、“过”标记的定中、状中或述补结构构成，长度通常超过 4 个音节。又由于受到呼吸的限制，它的长度不会太长，因此一般小于 10 个音节，对应于一个小的呼吸群。

当韵律短语中含有“的”（助词）、“和”（连词）、“中”（方位副词）等功能词时，其长度可以略大于 9 个音节。这主要是因为在这些词之后，即使不插入韵律词边界，从听感上也会感觉到停顿，如“并恢复当地的法律秩序”。

当有些韵律词独立成句，或处于特别强调位置，或孤立于韵律短语之外时，也可能单独构成韵律短语，这时韵律短语的长度为 1~3 个音节。

（3）注意避免切分过于零散。

切分时，在可能的情况下不要太零散，即避免出现多个短韵律短语（1~4 个音节）相连的情况。例如在带“的”、“与”等的短语中，如果不是太长（小于 9），中间可以不插停顿。

（4）长句可能有几种不同的切分方案，在两可的情况下，应从以下几方面考虑。

最好以更符合句法结构的方式标注，即确保韵律层次和句法层次的对应。应该尽量与意群相符，即保证韵律短语边界不跨越意群板块的边界。尽可能保持前短后长的格局。

3）固定标注规则

标注固定规则主要考虑语法结构、固定停顿或者约定俗成的习惯。

（1）“的”、“了”、“着”、“得”、“地”等结构助词与前面的词结合成一个韵律词。

（2）单、双音节的名词、动词与后面的方位词结合成韵律词。

（3）在三音节限制内，介词可以和后面的名词、动名词、代词等结合，条件是后面的词是该介词的直接、单一宾语。

（4）单音节副词和后面的介词一般结合成韵律词。

（5）单/双音节的副词可以跟形容词/动词/动补结构组成韵律词；在三音节限制内，还可以前附单音节副词、代词、介词或向后与动词性、名词性语素等结合。

4）可选标注规则

标注时，要考虑语音、语法、语义的平衡。

（1）语句韵律平衡原则与句法结构相比，韵律短语应尽量更符合句法结构。比较下面语句的两种标注，第一种标注主要考虑语句中各韵律短语的长度，第二种标注首先考虑语句的语法结构，调整了各韵律短语的长度。第二种标注结果要优于第一种结果。

/世界|和平与・发展的|崇高|事业・是不可阻挡的。/
/世界|和平与|发展的・崇高|事业・是不可阻挡的。/

（2）语句韵律平衡原则与意群结构相比，韵律短语应尽量更符合意群结构。比较下面语句的两种标注，第二种标注首先考虑语句的意群结构，以便设置合适的语调和轻重模式。

/密切|党同・人民|群众|联系的|决心。/
/密切|党同|人民|群众・联系的|决心。/

标注应考虑语句的焦点。下面第二种标注可使焦点表现在“出现”和“临时”上，可以让听者更好地表达语义。

以往|出现・一个|大的|纠纷，・乡镇|就组成・一个|临时|工作组|进村，・俗称“|消防队”。
以往・出现|一个|大的|纠纷，・乡镇|就组成|一个・临时|工作组|进村，・俗称“|消防队”。

（3）连续两个单音节词，可合并为一个韵律词，如“也就”。连续三个单音节词，可合并为一个或两个韵律词，要看各词是否需要重读；重读时的标注如下：

“这|也是”，“而|让他”。

（4）顿号后不一定都作为韵律短语边界，可作韵律词边界。下面语句的第二种标注注意到修饰语修饰或限制的范围，可以让听者更好地表达语义。

上班|时间，・不得|玩游戏、・打私人|电话。
上班|时间，・不得・玩游戏、|打私人|电话。

3.4 言 语 知 觉

言语知觉涉及人耳、大脑加工和解释言语声的方法。人耳可以听到的声波的频率范围约在 20～20 000Hz。

语音信号处理区别于广义的数字信号处理的一个显著特点是在语音信号处理的过程中，对语音信号的分析必须与人对语音的感知特性联系起来考虑，孤立地将语音信号当作一般的数字信号往往事倍功半。对于言语感知的研究，主要集中于心理学和语言学领域。

3.4.1 听觉

（1）听觉（hearing）是声音作用于听觉系统引起的感觉。听力是听觉感受声音的能力，通常以听阈的高低表示。听力学是研究人的正常听力及在各种生理或病理情况下听力变化规律的学科。

听觉的过程如下：声波传递到达人耳，耳廓确定声音方位，声波沿耳道进入，引起鼓膜振动，经中耳、内耳共振、放大，并将振动转换成神经电信号，传输给大脑进行信息解码。人耳的内耳由耳蜗、前庭、半规管和听神经组成。耳蜗螺旋状的蜗管约 35mm 长，不同部位对应不同频率的声音共振，耳蜗底部对高频声音响应，耳蜗顶部对低频声音响应。

（2）听域是指感知的声音的范围。正常人耳能够感知的频率范围为 16.4 Hz ~ 16 kHz，年轻人可听到 20 kHz 的声音，老年后则下降到 10 kHz 左右。正常人能感知声音的强度范围是 0 ~ 128 dB SPL（Sound Power Level，声压级），这里基准声压级（0 dB SPL）的定义是 $10^{-16}\,W/cm^2$。

（3）听阈（auditory threshold）是足以引起听觉的最低声压级，通常以分贝数表示。听阈低表示听觉灵敏或听力好。

纯音的听阈的高低与刺激声的频率有关：1 kHz 纯音的听阈约为 4 dB，10 kHz 时听阈约为 15 dB，到 40 kHz 时达到 50 dB 左右。当声压级增大到一定强度时，人耳会感到不适（不适阈）或疼痛（痛阈），正常人的不适阈约为 120 dB，痛阈约为 140 dB，且均与频率无关，不适阈与痛阈均属于感觉阈，反映人耳对声压的容忍程度。

人耳对音强的分辨率称为强度差阈，对于不同频率的声音，强度差阈也是不同的。一般说来，在中频段，人耳能感受到约 325 级音强。

强度差阈是人耳对指定频率声音所能分辨的最小强度差。强度差阈在声音强度低，强度差阈高。声音强度增大时，强度差阈降低。

频率差阈是人耳对指定强度声音所能分辨的最小频率差。声音强度增大时，频率差阈降低。人耳对声音时长的感知能力与声音时长本身有关，还会影响到听阈。当声音短于 200ms 时，其声音越短，听阈升高越多。

频率差阈与音强和频率本身都有关系。在声强为 40 dB 左右时，在 100 ~ 500 Hz 范

围内，人耳的频率差阈约为1.8 Hz，在500 Hz～16 kHz，频率差阈 $\Delta f \approx f \times 0.035$ 。如果此时音强过高或者过低，人耳对频率的分辨能力都会下降。

此外，人耳对时间的分辨率大约为2 ms，也就是说人耳能够感知距离为2 ms的两个高低不同的音。

采用不同频率的纯音测试人的听阈，可得到听力图，用于描述听力损失。

3.4.2　心理声学

心理声学（psycholinguistics）是语言学的一个分支，研究语言行为与构成这种行为的基础心理过程之间的相关关系。研究人们可以用语言作为解释心理理论和过程的手段，如由于语言影响记忆、感知、注意和学习等所起的作用。或研究心理制约对语言使用的影响，如记忆限度如何影响言语发生和理解。

（1）听觉中枢系统主要由耳蜗核、上橄榄复合体、下丘、内侧膝状体以及听觉皮层组成。耳蜗核主要完成对听觉谱特征的提取，其中原生型神经元保持了时间-位置编码信息（平均局部同步率），建立型和斩波型神经元保持了发放率-位置编码信息（发放率），中止型和累积型神经元有很强的侧抑制现象，其数学模型为侧抑制网络。上橄榄复合体、下丘和内膝体都被当作听信息的中转站，其生理功能和数学模型没有公认的描述。初级听觉皮层对听觉谱按对数频率轴、局部对称性和局部谱带宽三个方向独立处理，即完成对频率转移方向和频率转移速度的检测，数学上等价为保持听觉谱幅度和相位的仿射小波变换。与外周模型结合起来，整个听觉的处理类似局部的倒谱分析。

（2）时间分辨率是指听觉系统对声音信号在时间上的快速变化加以反应的能力。在言语和其他声学交流方式中，信息的传递是通过信号时程的改变来实现的。经过长期进化，人的听觉系统具备了非常精细的时间分辨能力，而听觉系统时间敏感度的缺损也往往伴随着言语理解的障碍。

（3）掩蔽效应是指一个声音的听阈因另外一个声音存在而被提高的现象。前者称为被掩蔽声，后者称为掩蔽声。被掩蔽声刚能被觉察到时掩蔽声的强度称为被掩蔽声的掩蔽阈限。掩蔽效应已被成功地运用于语音编码领域来提高语音质量。

一个纯音可能被另一频率的纯音所掩蔽。当掩蔽声出现时，被掩蔽声就听不到了。要想听到被掩蔽声，就要提高它的听阈值。掩蔽声越强，被掩蔽声的听阈取值越大。噪声的存在也会影响纯音的接收，即对纯音产生掩蔽。一个纯音处于以它为中心频率，具有一定频带宽度的连续噪声中，如果在这一频带内，噪声功率等于纯音的功率，则此纯音可能刚好被掩蔽。这时，纯音处于能被听到的临界状态，称此连续频带的带宽为临界带宽，临界带宽的取值随着中心频率的不同而变化，如表3.11所示。

（4）GAP 阈值是指受试者刚刚能感觉到连续声音中断时该中断的最小时程。GAP阈值的大小可以反映出听觉系统对在时间上快速变化的声音响应的灵敏度。这是应用最广的一种检测听觉系统时间分辨率的心理物理学方法。GAP 阈值不受经验的影响。无经验受试者所测得的阈值和那些经过训练的被试相比结果非常接近。另外，该阈值与前后背景噪声信号的声强无关。

表 3.11 临界频带

序号	中心频率/Hz	带宽/Hz	序号	中心频率/Hz	带宽/Hz
1	50	100	13	1850	280
2	150	100	14	2150	320
3	250	100	15	2500	380
4	350	100	16	2900	450
5	450	110	17	3400	550
6	570	120	18	4000	700
7	700	140	19	4800	900
8	840	150	20	5800	1100
9	100	160	21	7000	1300
10	1170	190	22	8500	1800
11	1370	210	23	10 500	2500
12	1600	240	24	13 500	3500

3.4.3 听力测试

受到噪声、疾病、药物的影响，听觉系统受到损伤，或随着年龄的增长，听力下降，引起听觉障碍，称听力损失。听力损失在 25～40dB 定为轻度，40～70dB 定为中度，70dB 以上为重度。

（1）主观测听法（Subjective Audiometry）是听力检查方法的一种，指通过观察受试者主观判断后做出的反应，检查患者听力损伤情况的方法，如耳语检查、秒表检查、音叉检查、听力计检查等。但此方法容易因年龄过小、精神心理状态失常等多方面因素而影响正确的测听结论。

（2）客观测听法（Objective Audiometry）是听力检查方法的一种，指不需要患者对声刺激做出主观判断反应，客观测定听功能情况的方法，其结果较精确可靠。常用的客观听测法有以下几种：通过观察声刺激引起的非条件反射来了解听力水平（如瞬目、转头、肢体活动等）；通过建立条件反射或习惯反应来检查听力水平（如皮肤电阻测听、西洋镜测听等）；利用生物物理学方法检查听力水平（如声阻抗-导纳测听）；利用神经生物学方法检查听力水平（如耳蜗电图、听性脑干反应）等。

（3）纯音测听是以纯音信号为刺激声，通过观察受试者的反应，测定耳聋性质及程度的方法。纯音听力计是进行纯音测听所必需的设备。纯音听力计（Pure Tone Audiometer）利用电声学原理，通过电子振荡装置和放大线路产生各种不同频率和强度（Intensity）的纯音，经过耳机传输给受检者，分别测试受试者各频率的听阈强度。纯音测听可为耳聋的定性、定量和定位诊断提供依据。

声强以分贝（Decibel，dB）表示。用纯音听力计测出的纯音听阈均值为听力级（Hearing Level，HL）。听力计以正常人的平均听阈为标准零级（Standard Zero Level），即正常青年人的听阈在听力计上为 0dBHL。听力减退时需增加声音强度方能听到声音，所增加的强度即为听力损失的程度。

具体测听流程，气导和骨导略有不同，下面以气导为例进行阐述。气导测试纯音给音频率顺序通常为 1000Hz、2000Hz、3000Hz、4000Hz、6000Hz、（8000）Hz、（250）Hz、500Hz、1000Hz（括号中为可选频率）。若两次 1000Hz 阈值差别大于 10dB，则应重新测试。若两个倍频频率阈值差值＞20dB，则应作半倍频频率阈值测定。另外，给声的时间应持续 1～2s，且给声间隔不得低于给声时间。气导测试方法分为上升法和下降法两种，临床多采用上升法。上升法的具体步骤包括，在设定频率下开始给声，如受试者无反应，则以 5dB 为单位增加声强，直至受试者出现反应；然后，以 10dB 为单位降低声强至无反应，再以 5dB 为单位增加声强至做出反应，如此反复给声 5 次，若三次反应均在同一听级，则所得结果即为听阈。

（4）听力图（audiogram）指纯音测听检查结果的记录曲线（听力曲线）。听力图的横轴为频率，单位为 Hz。纵轴为听阈，单位为 dBHL。实际测试中，纯音测听采用测试若干个频率（例如 500，1000，2000）下的听阈，将这些点连接起来形成听力曲线，作为受试者的听力图，见图 3.26。

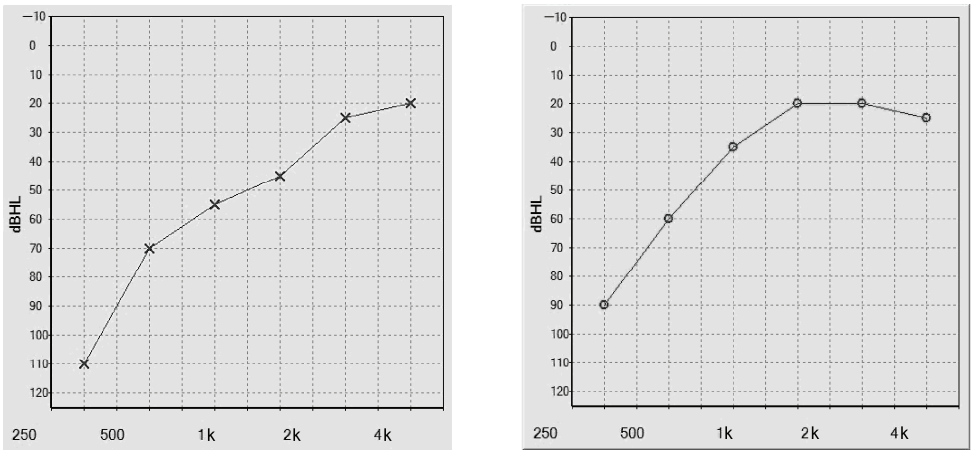


图 3.26　某受试者左耳（左图）和右耳（右图）的纯音听力图

（5）听力测试的听觉途径包括气导和骨导。气导是指声音凭借空气经过外耳、中耳传到内耳的过程。除以空气为介质外，额骨和牙齿等人体组织也可以传导声音振动到听觉神经，使人像通过耳朵一样听到声音，成为骨导。骨导通过骨导耳机来实现，骨导耳机的工作原理是利用高灵敏度的振动传感器，将电信号转化为人体组织（骨头等）的振动，从而听到声音，其最大的优点是具有非常好的抗噪声能力。

纯音气导平均听阈（Pure Tone Average，PTA）指在纯音测听中，听力图上 500Hz、1000Hz、2000Hz 三个频率下听力阈值的平均值。

纯音气导平均听阈与听力损失分级的对应关系如表 3.12 所示。

表 3.12　纯音气导平均听阈与听力损失分级对应关系

纯音气导平均听阈/dB	分　级
–10~15	正常
16~25	微小听力损失
26~40	轻度听力损失
41~55	中度听力损失
56~70	中重度听力损失
71~90	重度听力损失
91~	极度听力损失

3.4.4　言语清晰度

心理语言学（psycholinguistics）是语言学的一个分支，研究语言行为与构成这种行为的基础心理过程之间的相关关系。研究人们可以用语言作为解释心理理论和过程的手段，如由于语言影响记忆、感知、注意和学习等所起的作用，或研究心理制约对语言使用的影响，如记忆限度如何影响言语发生和理解。

言语清晰度表示无意义语言单位（如语音、音位或音节）的清晰度测试得分，等于听音人正确接收的语言单位数目与发送给听音人要求其做出响应的语言单位总数之比，以百分数表示。它可以在一定程度上反映人对于不同语言单位的信号正确接收的难易程度。

3.4.5　言语知觉

言语知觉是听音人从言语声信号中提取对应的语言学要素的过程，是对言语感知和理解的能力。人对言语的感知除了受言语信号的物理特性制约外，还受音位的影响。不同民族，不同地区的方言都有自己的音位系统。在特定语音里，属于不同音位的音素有区别意义的作用，即使它们之间的物理特性差别很小，使用这种语言的人也能辨别清楚。相反，属于同一音位的不同音素，不管差别多大，由于没有区别意义的作用，使用这种语言的人就往往难于辨别。例如，汉语普通话中“脑”和“老”，其声母分别为 n 和 l，它们属于两个音位，熟练掌握普通话的人对“脑”和“老”的发音区别得很清楚。可是在兰州方言里 n 和 l 不分，讲兰州方言的人，往往就不能区分普通话的“脑”和“老”的发音。

（1）元音感知指人耳对元音的感知与分辨能力。声谱分析显示元音是具有周期性的，而且一个元音有多个共振峰。元音的感知与其共振峰频率值和共振峰的相对关系有关。复合元音的感知与端点处（起始、结尾）的共振峰频率、共振峰的相对关系有关。

影响元音感知的主要因素包括：①共振峰，在语音学中认为元音的主要区分在于第一、二共振峰的不同，第一共振峰与发音时舌位的高低有关，第二共振峰与舌位的前后有关；②语言频域，语音信号的频谱分布在言语感知中具有关键作用，人耳对语音信号的频域信息十分敏感，而听力学上也认为语言频率对人耳的言语感知有重要作用；③时

序特征，单元音韵母发音过程平稳，整个发音过程中特征变化不大，而复合韵母由不同的音素分别构成其韵头、韵腹、韵尾，发音过程中时序变化较明显。

（2）辅音感知指人耳对辅音的感知与分辨能力。辅音是非周期性的，具有连续的频谱，其声学参量比元音更为复杂。它不能离开元音自成音节，总是同前后邻接的元音一起构成听辨上具有区别意义的信息。辅音除了本身固有的声谱特性外，还要靠它在时序上的声学特征起信息作用。因此，发音的轻重和延续时间的长短，以及邻近音的影响也是构成辅音知觉的重要因素。汉语普通话辅音的时长和强弱变化有一定的规律，如清擦音，都是先弱后强，如果把开头一段切去或减弱，就可能听成塞擦音，例如把 s 听成 z。

影响辅音感知的因素包括：辅音的强频区（或共振峰）、辅音到元音的过渡辅音后接元音的“嗓音起始时间”（VOT）、辅音的“过渡音征”（T）、声带是否振动等。

辅音随前后所接元音一起听辨。长/短、清/浊、辅音后接元音的“嗓音起始时间”（VOT）、辅音的“过渡音征”（T）。

（3）声调知觉（tone perception）指人耳对声调的感知与分辨能力。声调是指音节在发音过程中的高低抑扬性。声调信息主要存在于元音上。声调的感知依赖于音节的调型段，其声调知觉中心的位置位于韵母段中部。基频信息是声调感知的基础，对于一个音节来说，基频包络就是声调（字调）曲线。同一个人，他的不同声调的基频平均值和基频的变化率不同。此外，音节音高的变化，对时长和音强都可能产生影响。在声调感知过程中，当基频无法提供更多信息时，时长和音强将成为感知声调的依据。一般来讲，上声时长最长，阳平其次，去声最短。

影响声调感知的因素包括：基频曲线的均值、范围、最小值、最大值、最小值位置、最大值位置、初始值、终止值等。

（4）言语知觉加工指人在接收到输入的言语信号后，经过语言学知识与语境的加工，从而达到理解言语信号所要表达的意思的过程。加工过程的确切机制目前还未达到统一的认识。

① 词的知觉加工从接收语音输入开始，基本阶段包括：感觉输入与词的初始接合、激活（与输入匹配的词项状态的改变）、选择（从激活的词集中选择目标词）、识别（选择阶段的终点）和词汇信息提取。词的知觉加工还受到语境的制约。语境的影响可分成结构性与非结构性影响两种。结构性影响指系统各层次之间的相互制约，如词对音位加工的影响，句法对词加工的影响等。非结构性影响指由于词之间的语义和语用联系产生的词之间的相互影响。

② 语句的知觉加工指听者为了解语句的全部内容，在识别词的同时，对句子的韵律特征进行的知觉加工。在语句的知觉加工过程中，音段知觉加工不同于孤立音节和词加工的情况。言语信号中包含大量冗余信息，听者会寻找作为知觉决策基础的可靠信息，去提取各种知识源，包括重读音节、词的开始和结尾部分等。实验还证明， 在一些知觉分析水平上可能存在选择延迟。当输入信息不充分而需要等待另外的信息时，决策可能会延时。通过研究作语音决策的扫描窗口，人们发现在句子中，其后续 6 个音节（约 2s）以内的语境会对音位知觉加工产生影响。

③ 协同发音（Co-articulation）指在口语中，音素、音节或词等各种语音单元并不是简单机械地串列，而是按照一定的规则结合变化的有机体系。在连续语流中，这个体系的各个相邻语音单元之间会相互叠套，彼此渗透而产生协同发音现象。协同发音一般可分为先行与后续两种。前者是指前一音的发音期间，发音器官就已经开始作下一个音的发音，逐渐向着下一个音的部位和态势移动。后者是指在后一个音正式开始发音以后，由于惯性的作用，前一音的发音器官态势的残留过程会叠加在后一音的发音上，形成对它的音值的影响。

3.4.6 言语测听

言语测听法是用专门编制的测听词表来检查患耳的言语识别阈（Speech Reception Threshold）和言语识别率（Speech Discrimination Score）。有些病人的纯音听力较好，却听不懂语意。在这种情况时，纯音听力图并不足以反映病人的听功能状态，需要使用言语测听法来判定。言语测听的测听语料通常包括单音节字、扬扬格词、安静环境下句子与噪声环境下句子 4 种。

（1）言语识别阈指在言语测听时，受试者能正确复述一半测听词表中的语音时的最小声强级（dBHL）。

（2）言语识别率指在言语测听时，每次测试以固定的声强播放测听词表中的语料，受试者能正确复述的语料数量占词表语料总数的百分比。

（3）PI（Proportion-Intensity）曲线指言语识别率测试时，把不同声强级下的识别率值绘制在以声强为横轴、百分比为纵轴的坐标系中，形成的言语听力图（Speech Audiogram）。

PI 曲线的横轴代表刺激强度/信噪比，纵轴代表言语识别率。正常人的 PI 曲线形状类似侧向拉伸的 S 形，随着给音强度/信噪比提高，言语识别率逐渐增长，但强度/信噪比提高到一定程度之后得分会趋于稳定，即达到最大言语识别率。图 3.27 中，灰色曲线带所示即为正常人 PI 曲线的分布区域。由于测试言语识别率常常使用音位平衡的测试材料，所以最大言语识别率往往又称为 PBmax。临床研究表明：正常人的 PBmax 一般在 90%～100%；传导性听力损失患者的 PI 曲线比正常人向右平移（即气骨导差），但是 PBmax 仍可达到 80%～100%；血管瘤患者可能低至 60%；感音神经性听力损失的患者 PBmax 可能在 0%～100%范围内变化；蜗后病变患者的 PI 曲线可能会在达到 PBmax 之后，随着强度/信噪比的继续提高出现得分的异常降低，称为 PI 曲线的“回跌（Roll Over）”。PI 曲线的形状可以由阈值和斜率两个参数来描述，阈值描述了 50%言语得分的强度或信噪比，斜率（%/dB）描述了识别率受给声强度或信噪比影响的程度，PI 曲线斜率是与听力残障程度和致病原因相关联的重要指征参数。

（4）音位平衡指在言语测听中，一个测试材料中音位的组成与日常言语中的音位组成相同，即不同音位在测试材料中出现的频率应与其在日常言语中出现的频率相同。

音位是一个音系学概念，关乎词义。不同的语言、不同的方言都有其独特的音位归纳系统。讲话人略微改变语音参数（时长、共振峰频率、辅音特征）时聆听者仍能得出

正确的反应。而语音参数的较大变化则可使语音听起来像是另一个音位。音素则是音位在构音上的声学表现形式，是一个语音学概念，一个特定的音位可以表现为一组不同的音素，称为音位变体，但它们都被视为同一音位。

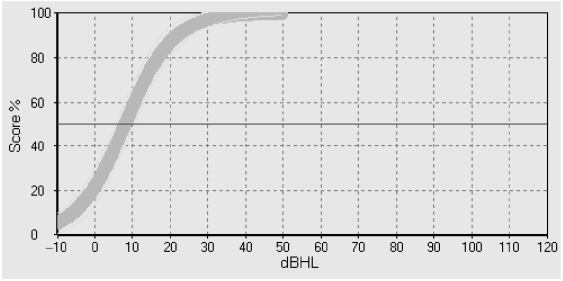


图 3.27　听力正常人 PI 曲线分布图（灰色曲线带）

（5）扬扬格词在汉语中指重音格式为“重重”的双音节词。扬扬格是英语诗的一种音步，指由连续两个重读音节组成的双音节词。汉语的重音系统由正常重音和轻声两类组成。正常重音都有声调，轻声没有声调。因此可以说，汉语双音节词有“重重”和“重轻”两种格式。

噪声下言语测听指使用带有背景噪声的测试语料，考察受试者在噪声环境下听力情况的方法。助听器使用者最常见的问题是在嘈杂环境中言语理解不佳。近年的研究表明，嘈杂环境下助听器满意程度的差异反映了患者在嘈杂环境中言语理解能力的差异。患者相对于正常人而言，为达到对字、关键字或句子 50% 的识别正确率，而比正常人所额外提升的信噪比称为“信噪比（Signal-Noise Ratio, SNR）损失”。许多研究都表明，听力图相似的患者，其信噪比的损失却可能有很大差异。常规的测听内容不能测试或预测患者在噪声中的言语理解能力。此外，了解信噪比损失有助于专业人士推荐相应的技术（如全方向麦克风、指向式麦克风、阵列麦克风或无线调频麦克风）以解决嘈杂环境中的种种问题。因此有关 SNR 损失的言语测试近来逐渐开始在临床中使用。

（6）听觉外周系统主要由外耳、中耳、耳蜗以及听觉神经（第 8 对神经）组成。外耳以及中耳的听小骨的作用主要是汇聚和传导声音，完成空气与耳蜗内液体间的阻抗匹配。声音通过听小骨的运动引起耳蜗内流体压强的变化，从而引起行波沿基底膜传播。不同频率的声音产生不同的行波，其峰值出现于基底膜的不同位置。声音频率沿基底膜的分布是对数型的，同时基底膜的振动幅度同声音强度有关。沿基底膜分布的内毛细胞随位置的不同具有不同的机械谐振特性和电谐振特性，基底膜的振动使得内毛细胞超极化或去极化，由此使得连到内毛细胞的听神经具有调谐特性听。听神经只对一定频率范围的声音响应，其最敏感的频率称为特征频率，其发放率随声刺激强度增加直至饱和。

3.4.7　听力补偿

指人的听觉系统对言语信号进行自动规范，减少语音变异的机制。由于讲话者的声

道长度存在差别，发音时的姿势和在不同语境下使用的发音方法也不尽相同，口语中使用的多数语音学区别性特征的声学语音学表现，在讲话者之间存在重大差别。通过知觉补偿和规范化过程，听者能够忽略和减小这种差异。

（1）助听声学技术指一切用于改善听力、提高与他人对话交际能力的工具、设备、装置和仪器等。现代的助听器主要是电声放大器，其使用麦克风接收声音信号并转化为电信号，借助放大器放大，而后有受话器将电信号还原为声音信号传送至人耳。

由于放大器利用不同的助听声学处理技术，从输出特性上，可以分为线性助听器和非线性助听器；从调节方式上，可以分为不可编程助听器和可编程助听器；从信号处理方式上，可以分为模拟助听器和数字助听器。

可编程助听器区别于不可编程助听器，在于前者使用数字调节器，可使得助听器的调节通过计算机实现，可以拥有更多的调节项而不借助多余的模拟元器件。

模拟助听器区别于数字助听器，在于前者放大器的信号处理器使用的都是模拟元器件，而随着数字信号处理能力的发展，目前很少有人使用模拟助听器。

现在使用的主要是数字助听器，先通过麦克风接收声音，将模拟信号转换为数字编码，而后经过滤波、压缩放大、限幅等数字化处理后还原成模拟信号输出，传入患者耳内。数字助听器比模拟助听器有更大的精确度、更小的内部噪声和失真、更长的寿命并能做更加复杂的计算。

助听器声学处理技术包括以下几种。

① 压缩技术。正常人听阈为 10dB 左右，不适阈为 100dB 左右，动态范围为 90dB。而听阈为 50dB 的感音神经性耳聋患者，其不适阈仍为 100dB，因而其动态范围就是 50dB。为了使患者能够听懂人的语音，将正常语音压缩到患者的动态范围内，即压缩技术。常用的算法包括：压缩限幅，宽动态压缩。

② 移频技术。将无高频残余听力或高频残余听力较差的助听器佩戴者无法通过普通助听器听到的且重要的高频言语信息转移至仍存有较好残余听力的低频阶段。常用的算法包括：线性移频压缩，非线性移频压缩。

③ 声反馈技术。由于助听器体积的限制，受话器与麦克风的距离往往很近，使得放大的声音信号由受话器输出时，其中的部分能量会通过助听器通过气孔或耳道壁间隙经空气泄漏，被麦克风获取，导致输入信号循环放大，引起某些频率的能量出现共振，发出啸叫声。消除这种声反馈所使用的技术称为声反馈技术。常用的消除算法有：消除反馈路径，增加麦克风与话筒距离，加强助听器外壳和耳模密封度；降低助听器增益。

④ 方向性麦克风技术。采用声学延时装置（阻尼器），使得特定或可变方向的声波在经过前后入声口时能量抵消，达到对特定方向的声波噪声屏蔽的作用。

麦克风降噪技术：将噪声从带噪声的语音信号中区分开，并去除噪声的技术。常用的算法包括：基于时域模型的参数或模型法；基于频域分析的谱减法、短时谱估计、对数谱估计等；基于时域频域分析的小波分析法等。

（2）人工耳蜗技术的生理基础。人们通过耳蜗毛细胞接收声音，当毛细胞出现重度损伤的时候，人们就会出现严重的耳聋。人工耳蜗是一种电子装置，其利用替代损伤的

毛细胞的方式，借助电流刺激听觉神经，使得听觉神经重新获得声音信号。

人工耳蜗由体外装置和体内装置两部分组成。体外部分包含麦克风、言语转换器、发射线圈，而体内部分含有接收线圈、处理器、刺激电极及参照电极。人工耳蜗利用不同部位与不同声音频率的对应关系，由体外麦克风接收声信号，通过言语处理器将模拟信号进行数字编码等处理，借助发射线圈传送到体内的接收线圈，把包含相应频率及电流强度的脉冲传递到多个刺激电极。多个刺激电极接收到信号后，信号通过听觉神经传送到听觉中枢进行辨别处理。

人工耳蜗技术包括以下几种。

① 言语处理器。将声音进行滤波分析并且数字化为编码信号，将编码信号送到传输线圈，传输线圈将编码信号以调频信号的形式传入位于皮下的植入体的接收或刺激器。

② 人工耳蜗植入体。将适量的电能传至耳蜗内部电极序列，沿着序列上分布的电极刺激耳蜗内的残余听神经纤维，电声信息继而沿听觉通路传至大脑进行编译。

③ 言语编码策略。控制着对环境声音及言语的数字化处理。根据人工耳蜗植入者个人需求，调整音调、响度和时相特征，设计相应言语编码策略，可以使其言语感知能力显著提高。

3.5　言 语 工 程

言语工程是基于神经学、生理学、心理学、语音学、声学、语言学，利用计算机科学的技术和成果，研究言语的发声、传送、感知、理解等科学理论、技术方法和应用问题。

3.5.1　言语链

言语链（speech chain）是指人类的言语过程，包括言语的产生、传播和接收的过程。在朗读时，说话人根据文字、词法和句法规则，朗读发声。在口语交流中，首先大脑产生说话的意向（intention），接着生成概念（concept），选择适当的词汇，按语法组织成语言。再根据语义、语用的需求，协调发音器官的动作，发出声音。言语过程可描述为以下几个阶段。

（1）大脑产生电信号，以表示说话的意向，并生成概念，选择适当的词汇，按语法组织成语言。这是语言产生（language generation）阶段。

（2）大脑的电信号沿运动神经传递，刺激发音器官。发音器官协调工作，发出声音（产生声波）。面部的肌肉、器官和体态与发音器官配合，送出多种信息，以便让听者更好地理解语音。与此同时讲话者的听觉系统接收到自己的声音，并随之修改。这是言语生成（speech production）阶段。

（3）传输：声波凭借质点运动而传播。

（4）接收：声波传递到人的听觉系统（外耳、中耳、内耳）。内耳的基底膜被声波刺激而振动，激发神经元产生脉冲，传递给大脑，从而感知到声音。这是言语识别（speech

recognition）阶段。

（5）理解：听者神经的语言层，通过一系列复杂的处理过程，辨认出说话人，理解其信息内容。这是言语理解（speech understand）阶段。

3.5.2 语音编码

语音编码原指用二进制数表示模拟信号的过程。这里是指利用语音信号中的冗余和人类听觉机理，通过语音处理技术，在语音质量或听感损失允许的情况下，尽量降低语音码率的处理过程。可用于减少语音传输占用的信道资源或语音存储占用的存储资源。

语音信号能进行压缩编码的基本依据是：第一，从语音信号的产生机理和它的结构性质表明，语音信号里存在很大的冗余度。语音压缩解决的本质问题是识别冗余度，研究去除冗余度的方法。研究表明，语音编码的极限速率估计为 80 ~100bps（Bit Per Second）。第二，利用人类听觉的功能特点——“掩蔽”现象，一个强的音能抑制一个同时存在的弱音的听觉。语音编码从信息论角度可以分为无损语音编码和有损语音编码两类。

1．无损语音编码

语音的无损编码指对语音信号进行编码后，不考虑传输损失的情况下，接收端能够毫无差错地译码。它在某些领域有着非常重要的应用，例如说话人身份识别，重要的声音数据保存等。一般信号的无损编码算法有 LWZ、Rice、Huffman、LPAC 等。这些算法的共同点在于首先利用一种预测算法，从一定程度上降低原始语音信号的自相关度，然后对预测残差进行熵编码。无损压缩率不能达到很高，编码码率通常会比较高。当语音信号出现强噪声时，无损语音编码压缩率会下降。

2．有损语音编码

相对于无损语音编码，有损情形下的语音编码的研究更为重要。有损语音编码的应用范围也比无损语音编码广得多，因此是语音编码研究的重点。

常用的压缩编码手段有两类：一类是降低量化每个语音样本的比特数，同时保持相对好的语音质量；另一类是先对数字信号进行分析，提取特征参数，例如线谱参数、线性预测器参数和反射参数，这些参数携带着语音信号的主要信息，因此可以实现很低的编码速率。解码器收到压缩后的参数进行信号重建，但这些信号的合成质量有欠缺。

语音编码一直沿着两个方向发展：波形编码和参数编码。波形编码算法力图使重建波形保持原来的波形形状，参数编码通过对语音信号特征的提取及编码，使重建的语音信号具有较高的可懂度。波形编码具有适应能力强、语音质量好等优点，缺点是编码速率高。参数编码具有编码速率低的优点，缺点是合成语音质量差，容易受噪声干扰。波形编码通过抽样和量化过程表示模拟信号，参数编码首先利用某种语音生成模型来表示语音信号的产生，然后用语音的特征提取方法提取必要的参数。波形编码的方法常常利用语音信号的统计特性和听觉特性对语音信号进行量化以达到压缩编码的目的。参数编码则是对语音特征的参数进行编码，包括声道特性参数、音调、清/浊音判决和基音信息。

（1）基于语音数据的统计特性进行编码，其典型技术是波形编码。其目标是使重建

语音波形保持原波形的形状。PCM（脉冲编码调制）是最简单最基本的编码方法。它直接赋予抽样点一个代码，没有进行压缩，因而所需的存储空间较大。为了减少存储空间，人们寻求压缩编码技术。利用音频抽样的幅度分布规律和相邻样值具有相关性的特点，提出了差值量化（DPCM）、自适应量化（APCM）和自适应差值量化（ADPCM）等算法，实现了数据的压缩。波形编码适应性强，音频质量好，但压缩比不大，因而数据率较高。

（2）基于语音的声学参数，进行参数编码，可进一步降低数据率。其目标是使重建音频保持原音频的特性。常用的音频参数有共振峰、线性预测系数、滤波器组等。这种编码技术的优点是数据率低，但还原信号的质量较差，清晰度低。

将上述两种编码算法很好地结合起来，采用混合编码的方法，这样就能在较低的码率上得到较高的音质，如码本激励线性预测编码（CELP）、多脉冲激励线性预测编码（MPLPC）等。

（3）基于人的听觉特性进行编码。从人的听觉系统出发，利用掩蔽效应，设计心理声学模型，从而实现更高效率的数字音频的压缩。其中以 MPEG 标准中的高频编码和 Dolby AC-3 最有影响。

3.5.3　言语合成方法

言语合成（speech synthesis，语音合成）是指用机械、电子、计算机等人为的方法生成汉语言语的过程与技术。17 世纪法国人研制了一个机械式的会说话的装置。自 19 世纪出现了电子合成器以后，言语合成研究得到了飞速的发展。从采用的合成分法，可分为发音参数合成、声道模型参数合成、波形编辑合成、基于 HHM 模型合成和基于 DBN 合成；从合成策略上，可分为频谱逼近和波形逼近。从技术发展来看，经历了从语音合成算法研究、文字-语音转换，到概念-语音转换研究的阶段。人们用语言进行交互时，用声音来表达自己的意向、情感。语音合成系统的输出语音不限于单个音节，而是连续语句。在这样的系统中，输入是文字（概念），输出的是语音流，言语合成的最终目标是让机器像人那样讲话。

言语合成的方法有声道模型参数、发音器官参数语音合成、参数合成方法、拼接合成、HMM 模型合成等。

（1）发音器官参数语音合成方法直接对人的发音过程进行模拟。它定义了唇、舌、声带的相关参数，如唇开口度、舌高度、舌位置、声带张力和肺气压等。由这些发音参数估计声道截面积函数，进而计算声波。这种方法有可能产生逼真的语音。但由于人发音生理过程的复杂性，理论计算与物理模拟之间的差异，合成语音的质量暂时还不理想。

（2）共振峰参数合成方法是基于声道共振原理，抽取语音难的共振峰（参照音典正文中韵母共振峰图），再合成出语音。它的优点是占用的存储空间小，与语音编码相结合时数码率较低，并能够较灵活地控制合成语音的音色，但合成辅音的清晰度低。共振峰参数提取困难。较为著名的共振峰合成器是 MIT 教授 D. Klatt 设计的串并联混合型共振

峰合成器。他用串联通道产生元音和浊辅音；并联通道产生轻辅音。还可以对声源做各种选择和调整，以模拟不同的嗓声。

（3）线性预测（LPC）合成是基于声道模型参数合成语音的方法。声道模型参数语音合成这种方法基于声道截面积函数或声道谐振特性合成语音，这类系统需要的存储量低，音质适中，易于实现韵律修改。

（4）波形拼接合成是对取自自然语流的语音基元进行拼接，来产生合成语音。这种语音合成技术用原始语音波形替代参数，而且这些语音波形取自自然语流中的音节、词或句子，它隐含了声调、重音等细微特性，合成的语音具有较高的清晰度。

然而，任何一个语音单元的声学特性，会随着语音环境的变化而改变。那么，语音基元选取、基元拼接、波形修改成为本合成方法的关键技术。

① 基元选取。波形拼接合成预先存入大量的语音单元基元。选取指在已知文本序列，从候选语音数据库中选择适当的语音基元。在选取基元之前，先选择韵律特征参数，如音段参数（当前音节的声母类型、韵母类型、声调类型，前音节的韵母类型、声调类型，后音节的声母类型、声调类型）；位置参数（当前音节在韵律词中的位置、在韵律短语中的位置、在语句中的位置、到前一个重音的位置、到后一个重音的位置、韵律词在韵律短语中位置、韵律词在句中的位置、韵律短语在句中的位置）；关联参数（与前音节的耦合度、与后音节的耦合度、前音节韵律关联参数、后音节韵律关联参数、所在韵律词的信息、所在韵律短语的信息、所在语句的信息等）；然后定义韵律匹配代价函数，并计算候选单元与相应的文本单元韵律特征参数信息的不匹配程度。选择使韵律匹配代价函数值小的语音基元作为拼接候选基元。

② 基元拼接。基元拼接时将选定的语音基元拼接起来，以拼接连续自然的语流。拼接过程中需要计算由拼接处的特征参数所决定拼接代价函数。特征参数包括时长、基频特征、幅度、重音属性、语调信息、语谱等。还可计算这些特征参数的一阶连续性。如果拼接代价函数值太大，可插入平滑、滤波等处理。

③ 波形修改。为了满足合成语音的音高、语速的需求，提高合成语音的自然度，波形修改是必需的。波形编辑合成中常用的波形修改算法是 PSOLA（基音同步叠加）。该算法按以下三步实施：它以基音周期的整数倍为窗长，对原始波形进行分析，产生中间的表示；然后对中间表示进行修改；将修改过的中间表示重新合成为语音信号。由于修改的参数不同，又分为时域 TD-PSOLA、频域 FD-PSOLA 和线性预测 LP-PSOLA。PLOLA 算法的优点是计算复杂度低，在保持音质不发生明显变化的前提下能够将基频修改 ±10% 左右。但当修改幅度超出这一范围时，PSOLA 算法在频谱结构和相位上都出现失真。

（5）HMM 合成。基于隐马尔可夫模型（Hidden Markov Model，HMM）的语音合成是一种可训练的参数合成方法。它分为训练与合成两个过程。在训练过程中，从语音中提取出基频和谱参数进行 HMM 参数化建模，并通过语境参数对模型进行聚类，得到一个语境相关的参数化模型。合成过程中，前段文本分析对输入文本进行语言学上的处

理得到每个训练单元的语境信息，之后从模型中选取出合适的 HMM 序列，进行状态时长估计和声学参数生成，得到基频和谱参数序列，经过参数化合成器生成合成语音。

3.5.4　文语转换

文语转换（Text To Speech，TTS）是把文字自动转换成言语（语音）的技术。TTS 系统中除了语音合成模块外，还包括文本分析、韵律生成模块。

文本分析是文语转换系统的前端。它对输入的文本进行分析理解，为后端语音合成器提供必要的信息，如读音、停顿、韵律等信息。不同的合成后端需要的信息也各不相同。对于简单的系统，可能文本分析只需要提供读音信息就够了；对于高自然度的任意文本的合成器，文本分析要给出更详尽的语言学或语音学信息，使得合成器合成的语音有更多的可调节的余地。结合自然语言处理和人工智能的研究成果，在充分“理解”文本的基础上，它的输出信息尽可能做到有正确的读音，有轻重缓急的标记，甚至包含不同的感情风格等。应该说，理想的文本分析器同时就是一个理想的自然语言理解程序。

TTS 系统要想取得高质量的声音，必须具备韵律处理和模拟的功能。语调、节奏和重音这些韵律特征是通过超音段特征——音高、时长、音强及频率分布的变化表现出来的。因此，这些超音段特征的修改成为韵律合成的基础。超音段特征的修改可通过多种方法实现，如修改基频模式、共振峰模式等。其中 PSOLA 算法获得了广泛的应用，它为波形合成方法实现 TTS 提供了有力的支持。

要实现韵律模拟，需解决韵律规则、韵律描述、计算模型和修改算法等问题。这要借助于语音学、语言学、心理学、信号处理等学科的成果。首先要研究韵律变化的特点，抽取韵律规则，找出韵律与声学参数的映射关系，给出定量的数学描述，建立计算模型。最后还要设计韵律修改算法。目前对自然语流中韵律现象的研究还远未达到人们的期望。

韵律建模就是要计算言语中的韵律特性，建立韵律模型，以产生这些韵律参数。通常韵律模型分为两类，一种基于规则，另一种基于大规模语料库，采用数据驱动方法。当然它们也不是绝对分开的。

3.5.5　声音转换

语音转换（Voice Conversion）利用语音处理技术，改变声音的某些特性。如改变时长、基音、音强、频谱等。可以实现语速的加快/减慢、音高变化、音色变化、男女声转换、发音者性别模拟等，可用于语言学习、声音伪装等。

从变换方法的角度，音色变换主要采用了三种方法：码本映射，基于 GMM 的线性映射和基于 HMM 的合成器。

码本映射的基本原理是对于源说话人的语音，先进行分帧和加窗处理。然后对各帧短时语音分别进行线性预测分析，从而得到一个 LSF 矢量集。利用预先训练的源码本，对 LSF 矢量集进行量化，得到源码字序列。对文本对齐的目标说话人的语音执行同样的过程，得到目标码字序列。再利用 DTW（Dynamic Time Warping）算法对两个码字序列进行时间对齐。整个源和目标说话人的语音库的训练过程结束后，得到若干具有映射关

系的码字序列。然后统计各目标码本中的码字，与各源码本中码字的对应次数，得到源和目标码字之间的概率关系。

语音修改通过对语音声学特征参数进行修改来获得具有特定听感属性的新的语音。当修改的声学特征与个性化感知相关时，得到的结果具有新的个性化特征。语音修改通常建立在语音模型的基础上。语音模型对语音生成或信号本身进行一定的假设，用数学建模的方法表示信号。模型参数代表一定的物理和声学意义，可以分别进行修改，再通过合成算法重建语音。常见的语音修改算法包括 PSOLA，正弦模型和 STRAIGHT。这些算法在可控参数，修改音质，计算复杂度方面存在很大差异。

3.5.6　言语识别

言语识别（speech recognition，语音识别）是指利用计算机自动识别语音的技术，有狭义和广义之分。狭义的语音识别特指利用计算机识别出语音信号所表达的内容，其目的是要准确地理解语音所蕴涵的含义，例如将语音转换成其所对应的文字。而广义的语音识别则泛指利用语音信号识别出其中所包含的“任何感兴趣”的内容的一种技术，例如利用语音信号中所包含的特定人的信息进行说话人身份辨认的说话人识别技术。

语音识别本质上是一种模式识别的过程，其基本结构原理框图如图 3.28 所示，主要包括语音信号预处理、特征提取、特征建模（建立参考模式库）、相似性度量和后处理等几个功能模块，其中后处理模块为可选部分。

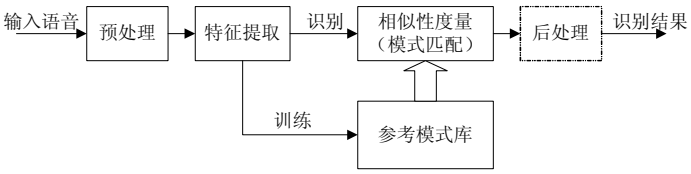


图 3.28　语音识别原理框图

一个语音识别系统主要包括训练和识别两个阶段。无论是训练还是识别，都需要首先对输入的原始语音进行预处理，并进行特征提取。下面具体说明各个模块的功能。

预处理模块，对输入的原始语音信号进行处理，滤除掉其中不重要的信息以及背景噪声等，并进行语音信号的端点检测，即判定语音有效范围的开始和结束位置，并进行语音分帧以及预加重等处理工作。

特征提取模块负责计算语音的声学参数，并进行特征的计算，以便提取出反映信号特征的关键特征参数，以降低维数并便于后续处理。语音识别系统常用的特征参数有幅度、能量、过零率、线性预测系数（LPC）、LPC 倒谱系数（LPCC）、线谱对参数（LSP）、短时频谱、共振峰频率、反映人耳听觉特征的 Mel 频率倒谱系数（MFCC）等。特征的选择和提取是系统构建的关键。

在训练阶段，用户输入若干次训练语音，系统经过上述预处理和特征提取后得到特征矢量参数（序列），然后通过特征建模模块建立训练语音的参考模式库（可能为参考模

板或者模型等），或者对已在模式库中的参考模式作适应性修正。

在识别阶段，将输入语音的特征矢量参数（序列）和参考模式库中的模式进行相似性度量比较，将相似度最高的模式所属的类别作为识别的中间候选结果输出。

而后处理模块则对上述得到的候选识别结果继续处理，通过更多的知识（例如：语言学的语言模型、词法、句法和语义信息等）的约束，得到最终的识别结果。语音识别分类：从不同的角度和要求出发，语音识别有不同的分类方法。

3.5.7 说话人识别

说话人识别（Speaker Recognition），又称声纹识别（Voiceprint Recognition），是利用语音波形中所包含的反映说话人生理和行为特征的语音参数自动识别说话人身份的技术。

说话人识别可以分为两个范畴，说话人辨认和说话人确认。说话人辨认（Speaker Identification）根据一段语音确定说话人是否已经注册，并判断其身份；说话人确认（Speaker Verification）则根据说话人的语音确定该说话人是否与其所声称的身份一致。说话人识别任务从对语音的要求上可以分为：与文本无关的说话人识别和与文本有关的说话人识别。与文本无关的说话人识别指模型训练语料不要求特定的语言和内容，而且训练语料与测试语料之间也不要求一致；与文本有关的说话人识别指模型的训练语料是由用户按照给定的文本朗读得到，测试语料应与训练语料相一致。

说话人识别与言语识别的区别在于，说话人识别希望从语音信号中提取出说话人的特征，以区分说话人是谁；而语音识别的抽取不同人的发音共性，以区分语音信号的内容。

说话人识别主要包括两大功能模块: 特征提取和模式匹配。特征提取是指从经过预处理的语音信号中提取出能唯一表现说话人身份的有效而稳定可靠的特征。目前说话人识别系统主要依靠语音的低层次声学特征进行识别，这些特征主要包括线性预测系数及其派生参数、由语音频谱导出的参数、反映听觉特性的参数以及上述参数的组合等。

模式匹配是将识别时的特征模板或模型与训练时得到的模板或模型作相似性匹配，计算它们之间的相似性（距离）。模式匹配的主要技术有动态时间规整（DTW）、向量量化（VQ）、隐马尔可夫模型（HMM）、人工神经网络（ANN）、高斯混合模型（GMM）、支持向量机（SVM）等，也可以将上述多种方法与不同特征进行有机组合来提高说话人识别的性能。

说话人识别被认为是最自然的生物特征识别方式，非接触式的交互易被人们接受。然而，语音基本上属于一种行为特征，身体状况、紧张程度等因素会引起语音的改变。另外，环境噪声对语音识别的影响非常大。所以基于语音的身份鉴别系统的性能在很大程度上依赖于采集、传输、存储等数字化设备的精度，更重要的是该项技术需要说话人本身的配合。从提高识别系统准确性和鲁棒性来说，说话人的语音识别与人脸识别、唇动识别等其他生物特征识别技术相结合是行之有效的方法。

3.5.8　可视语音

语音是人们日常交流的重要方式。但语音稍纵即逝，噪声环境会影响交流效果，听力障碍者的交流困难。实际上，也可以通过多种形式“看见”语音。

作为语音载体的声波通过耳朵由听觉来感觉到，而现代信号处理技术可以将声波进行处理，再借助一定的仪器用图表的形式表现出来，从而“看见”语音。另一方面，在日常生活的人与人交谈中，虽然声波看不到，但可以看到产生声波的发音器官在运动，它与语音有着必然的联系，这也是一种看得见的语音。

将“讲话的头”（Talking Head）与声音同步地结合起来（可视语音合成）是语音合成的重要进步。它提高了合成语音的可懂度，减轻听力障碍者的交流困难。

1．视位

20 世纪 60 年代，Fisher 提出了视位（viseme）的概念。视位指人们发某一音位时，可视发音器官所处的状态。描述一个视位可以有各种各样的表示方法，大致分为参数表示方法和非参数表示方法两大类。由于不同语言所包含的音位不同，不同语言的视位也有所不同。国际标准 MPEG-4 进一步明确了视位的定义，并把国际音标分成了 15 个基本静态视位，如表 3.13 所示。

表 3.13　MPEG-4 所定义的基本静态视位

视位编号	视　位	例　词
0	none	na
1	p, b, m	put, bed, mill
2	f, v	far, voice
3	T,D	think, that
4	t, d	tip, doll
5	k, g	call, gas
6	ts, dz, S	chair, join, she
7	s, z	sir, zeal
8	n, l	lot, not
9	R	red
10	A:	car
11	E	bed
12	I	tip
13	Q	top
14	U	book

2．汉语视位

从音位的角度来考虑，汉语音位有 32 个，包括 22 个辅音音位和 10 个元音音位；从汉语发音的基本组成单位来考虑，可以分为 21 个声母和 38 个韵母，其中韵母又可分为单韵母和复合韵母。针对汉语的发音习惯，从声韵母的角度来研究汉语视位的分类，

如表 3.14 所示为汉语基本视位集。

首先通过对汉语静态视位参数的相似度分析建立汉语视觉混淆树，再根据这一混淆树和建树过程中的误差变化曲线确定汉语静态视位分类。

表 3.14　汉语基本视位集

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		

3．可视韵律

可视韵律（Visual Prosody）指的是话语过程中，话音与文字、韵律结构紧密相关的面部和头部动作。如在强调某个词的时候，往往会伴随有下意识的点头动作；经常会在提高嗓门的时候扬起眉毛或者有轻微的抬头动作。这些肢体动作往往不受话语内容本身的影响，持续时间短暂，但起到了辅助言语的提示性作用。这类肢体动作与说话人发音口型之间的关系，类似于语音合成中韵律与发音音位的关系，称为可视韵律。在语音合成中，如果只是准确地按照文本内容合成语音，而不考虑自然语音中的声调、语调等韵律信息，那么合成的语音就如同机器人在说话。类似地，如果在虚拟说话人合成中，只是按照发音音位生成口型动作，缺少了与语音韵律相关的自然肢体动作，虽然在理解上不影响言语表达，但是合成效果会大打折扣，给人以呆板不自然的印象。由于可视韵律动作与话语语义相关度并不大，而与语音韵律十分相关，这一特点使得研究者可以直接在语音中提取与韵律相关的表现力特征，从而驱动说话人生成可视韵律动作。