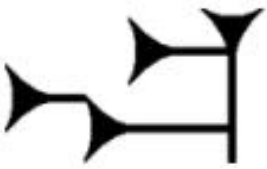


25. 语音学

组成我们在这本书中讨论的文本的字符，并不是随机的符号。它们也是一项惊人的科学发明：一种构成人类语言元素的理论模型。

我们知道的最早的书写系统(苏美尔语、汉语、玛雅语)主要是图形标识(logographic)：一个符号代表整个单词。但是从最早的阶段我们可以发现，一些符号也被用来代表构成单词的声音。右边的楔形符号叫 **ba**，在苏美尔语中的意思是“定量”，也可以纯粹用作声音/ **ba** /。我们拥有的最早的汉字刻在骨头上用于占卜，同样包含语音元素。纯粹的基于声音的书写系统，无论是音节(例如日语平假名)，字母(例如罗马字母)还是辅音(例如 **Semitic** 书写系统)，都可以追溯到早期的徽标音节系统，通常是



两种文化的结合。因此，阿拉伯文、阿拉姆文、希伯来文、希腊文和罗马文系统都源自西方闪族人(**Semitic**)的文字，推测该文字已被西方闪族佣兵从草书形式的埃及象形文字中加以修改。日本的音节是由汉字的草书音标演变而来的，在唐代传入中国的佛经(**Buddhist scriptures**)中，汉字被用来代替梵语(**Sanskrit**)。

口语是由较小的语音单位组成的，这是一个隐含的想法，是语音识别(将波形转换为文本)和文本到语音(将文本转换为波形)算法的基础。在本章中，我们将提供有关**语音学(phonetics)**的计算观点，对世界语言使用的语音的研究，它们如何在人的声道中产生，如何在声音上实现，如何数字化和处理。

25.1. 语音和语音转写

像“p”或“a”这样的字母已经是人类语音的有用模型，我们将在第 26 章看到如何将字母和波形映射在一起。尽管如此，更抽象地表示声音还是有幫助的。我们将用一串**音子(phone)**来表示一个单词的发音，这些音子是语音，每个音子都用取自罗马字母的符号来表示。

ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription	ARPAbet Symbol	IPA Symbol	Word	ARPAbet Transcription
[p]	[p]	parsley	[p aa r s l iy]	[iy]	[i]	lily	[l ih l iy]
[t]	[t]	tea	[t iy]	[ih]	[I]	lily	[l ih l iy]
[k]	[k]	cook	[k uh k]	[ey]	[eɪ]	daisy	[d ey z iy]
[b]	[b]	bay	[b ey]	[eh]	[ɛ]	pen	[p eh n]
[d]	[d]	dill	[d ih l]	[ae]	[æ]	aster	[ae s t axr]
[g]	[g]	garlic	[g aa r l ix k]	[aa]	[ɑ]	poppy	[p aa p iy]
[m]	[m]	mint	[m ih n t]	[ao]	[ɔ]	orchid	[ao r k ix d]
[n]	[n]	nutmeg	[n ah t m eh g]	[uh]	[ʊ]	wood	[w uh d]
[ng]	[ŋ]	baking	[b ey k ix ng]	[ow]	[oʊ]	lotus	[l ow dx ax s]
[f]	[f]	flour	[f l aw axr]	[uw]	[u]	tulip	[t uw l ix p]
[v]	[v]	clove	[k l ow v]	[ah]	[ʌ]	butter	[b ah dx axr]
[th]	[θ]	thick	[th ih k]	[er]	[ɜ~]	bird	[b er d]
[dh]	[ð]	those	[dh ow z]	[ay]	[aɪ]	iris	[ay r ix s]
[s]	[s]	soup	[s uw p]	[aw]	[aʊ]	flower	[f l aw axr]
[z]	[z]	eggs	[eh g z]	[oy]	[oɪ]	soil	[s oy l]
[sh]	[ʃ]	squash	[s k w aa sh]				
[zh]	[ʒ]	ambrosia	[ae m b r ow zh ax]				
[ch]	[tʃ]	cherry	[ch eh r iy]				
[jh]	[dʒ]	jar	[jh aa r]				
[l]	[l]	licorice	[l ih k axr ix sh]				
[w]	[w]	kiwi	[k iy w iy]				
[r]	[r]	rice	[r ay s]				
[y]	[j]	yellow	[y eh l ow]				
[h]	[h]	honey	[h ah n iy]				

图 25-1：英语辅音(左)和元音(右)的 ARPAbet 和 IPA 符号

用于转录世界语言的标准语音表示法是**国际音标(International Phonetic Alphabet, IPA)**，它是 1888 年首先开发的不断发展的标准。但是在本章中，我们将用 **ARPAbet(Shoup, 1980)**来表示音子，这是一个

简单的语音字母(图 25.1), 它方便地使用 ASCII 符号表示 IPA 的美式英语子集。

IPA 和 ARPabet 的许多符号都等同于熟悉的罗马字母。因此, 例如, ARPabet 音子[p]代表在鸭嘴兽(platypus), 美洲狮(puma)和车前草(plantain)的开始, 豹子(leopard)的中间或羚羊(antelope)的末尾的辅音。但是, 总的来说, 英语拼写字母和音子之间的映射相对不透明: 一个单个字母可以在不同的上下文中代表非常不同的声音。英文字母 c 对应于美洲狮(cougar)[k u w g a x r]中的音子[k], 但对应于细胞(cell)[s e h l]中的音子[s]。音子[k]除了显示为 c 和 k 外, 还可以显示为 x(狐狸 fox [f a a k s]), ck(豺狼 jackal [j h a e k e l])和 cc(浣熊 raccoon[r a e k u w n])的一部分。许多其他语言(例如西班牙语)在声音拼写法映射中比英语更加透明。

25.2. 发音语音学

发音语音学(articulatory phonetics)研究的是口腔、喉咙和鼻子中的各种器官如何调节肺部气流, 从而产生音子。

25.2.1. 发声器官

图 25.2 显示了语音器官。空气的快速运动会产生声音。人类通过将空气从肺部通过气管(从技术上讲是 trachea)排出口腔然后从嘴或鼻子中排出, 从而产生大多数口头语言的声音。空气通过气管时, 会通过喉, 通常称为亚当的苹果或语音盒。喉头包含两个小褶皱的肌肉, 即声带(通常在非技术上称为 vocal cords), 可以一起移动或分开移动。这两个褶皱之间的空间称为**声门(glottis)**。如果褶皱靠在一起(但不是紧密闭合), 则当空气通过时它们会振动。如果它们相距很远, 它们将不会振动。由声带褶皱产生并振动的声音称为**浊音(voiced)**; 没有声带振动而发出的声音称为**清音(unvoiced)**或无声。浊音包括[b], [d], [g], [v], [z]和所有英语元音等。清音包括[p], [t], [k], [f], [s]等。

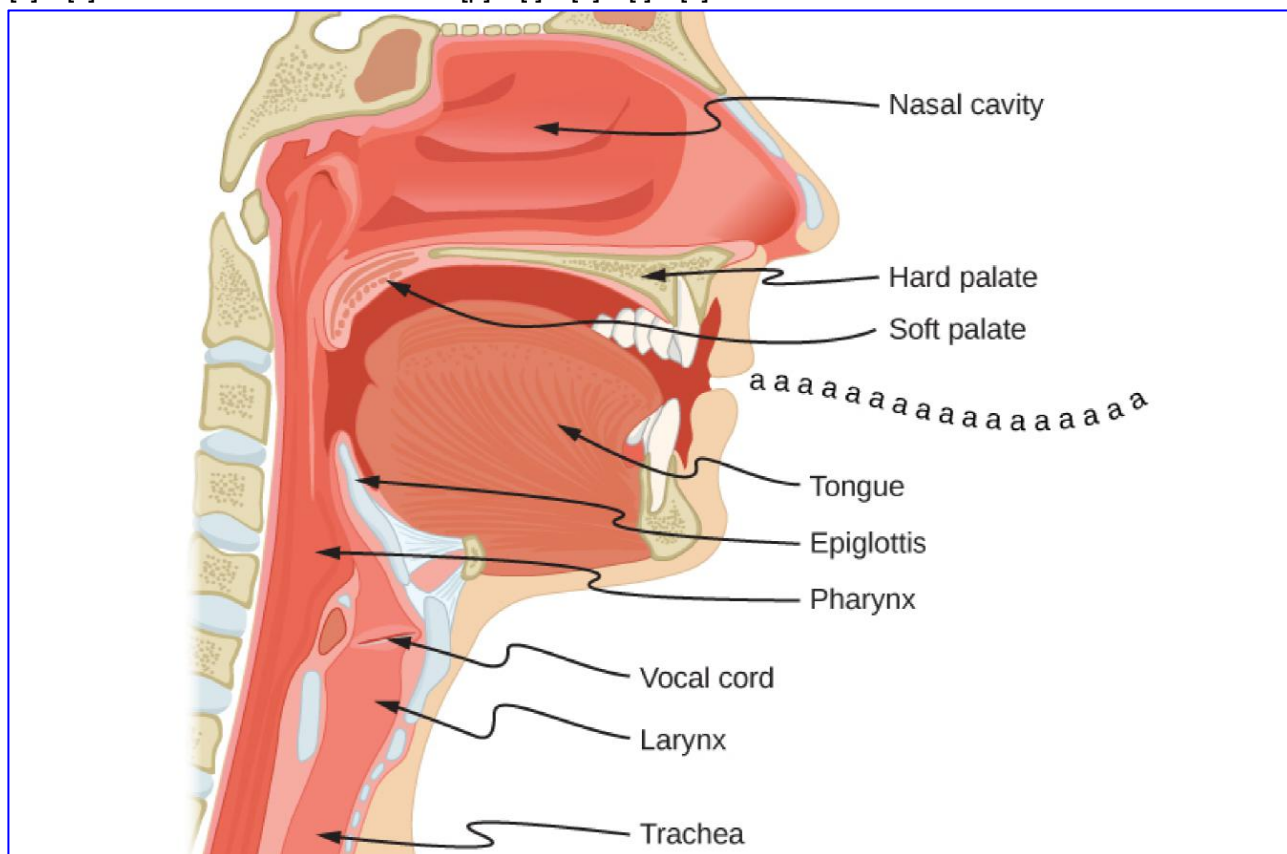


图 25-2: 发声器官的侧视图

图注: 来自 OpenStax University Physics, CC BY 4.0。

气管上方的区域称为**声腔(vocal tract)**。它由**口腔(oral tract)**和**鼻腔(nasal tract)**组成。空气离开气管后, 它可以通过嘴或鼻子离开身体。大多数声音是由空气通过嘴发出的。空气通过鼻子发出的声音称为鼻

音；鼻音(例如英语[m], [n]和[ŋ])使用口腔和鼻腔作为**共鸣腔(resonating cavities)**。

音子分为两大类:**辅音(consonants)**和**元音(vowels)**。这两种声音都是由空气通过嘴、喉咙或鼻子的运动形成的。辅音是由于气流受到阻碍而产生的,可以发声也可以不发声。元音的阻力较小,通常是发声的,而且通常比辅音响亮和持久。这些术语的技术用法很像普通用法:[p], [b], [t], [d], [k], [g], [f], [v], [s], [z], [r], [l],等等,都是辅音;(aa), (ae), (ao), (ih) (aw),(ow),(uw),等等,都是元音。半元音(semivowels)(如[y]和[w])具有两者的某些属性;它们像元音一样发声,但它们像辅音一样短而且音节少。

辅音:发音部位

由于辅音是通过阻挡气流产生的,我们可以根据它们的最大阻挡点、它们的**发音部位(place of articulation)**来将它们分类(图 25.3)。

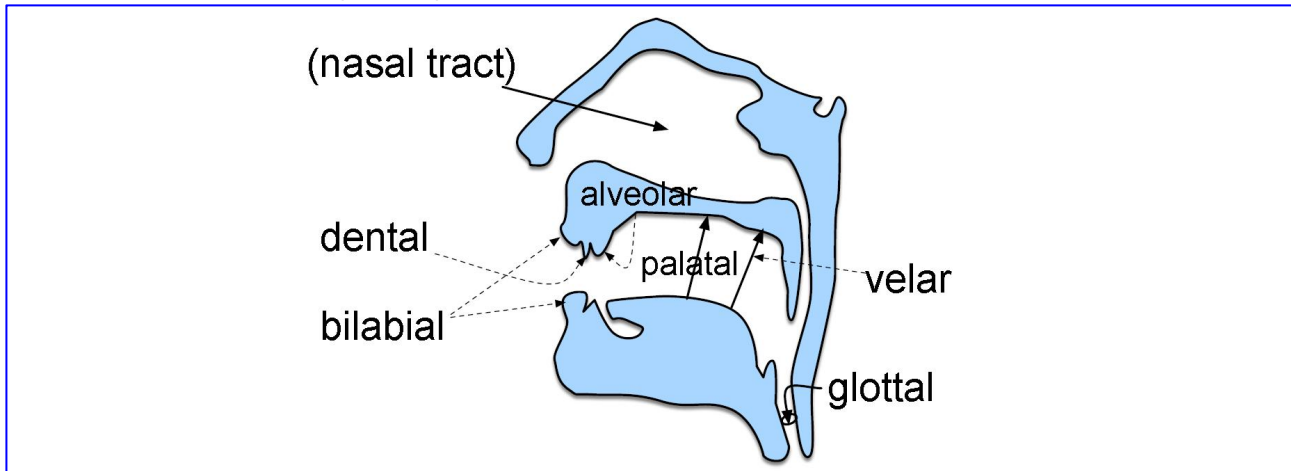


图 25-3: 英语主要的发音部位

唇音(Labial):辅音的主要阻挡是由两个嘴唇凑在一起形成的,具有**双唇音(bilabial)**发音部位。在英语中,这些词包括 possum(负鼠)中的[p], bear(熊)中的[b]和 marmot(土拨鼠)中的[m]。英语的**唇齿音(labiodental)**的辅音[v]和[f]是由下唇紧贴上排牙齿,让空气从上排牙齿的空隙中流过而成的。

齿音(Dental):舌头抵住牙齿发出的音是齿音。英语中主要的齿音是 thing(事物)中的[θ]和 though(尽管)中的[dh],它们是由舌头放在牙齿后面,舌尖在牙齿之间形成的。

齿龈音(Alveolar):齿龈是上排牙齿后面的口腔上颚的一部分。大多数说美式英语的人把舌尖靠在齿龈上而产生音子[s]、[z]、[t]和[d]。**舌面前音(coronal)**这个词通常用来同时指称齿音和齿龈音两者。

上腭音(Palatal):口腔的顶部(上腭)从上腭的齿龈脊后部急剧上升。龈腭音(palato-alveolar)[sh](shrimp 小虾), [ch](china 中国), [zh](Asian 亚洲)和[jh](jar 广口瓶)是由舌尖对着上升的齿龈脊后部发出的。牦牛 yak 的[y]发音是舌头的前部靠近上腭发出的。

软腭音(Velar):软腭是位于上腭最后部的一个可移动的肌肉瓣。声音[k](cuckoo 布谷鸟), [g](goose 鹅)和[N](kingfisher 翠鸟)是把舌头后部贴在软腭上发出的。

喉音(Glottal):喉塞音[q]是通过关闭喉头(将声带连接在一起)产生的。

辅音:发音方式

辅音的不同之处还在于如何限制气流,例如通过完全停止空气或部分阻塞来限制气流。此特征称为辅音的**发音方式(manner of articulation)**。发音部位和发音方式的组合通常足以唯一地识别辅音。以下是英语辅音发音的主要方式:

塞音(stop)是一个辅音,其中气流在短时间内被完全阻塞。随着空气的释放,这种阻塞之后会伴随着爆破声。阻塞的阶段称为**成阻(closure)**,爆破的阶段称为**除阻(release)**。英语有浊塞音诸如[b]、[d]和[g],也有清塞音诸如[p]、[t]和[k]。塞音也称为**爆破音(plosives)**。

鼻音(nasal sounds)[n]、[m]和[ŋ]是通过降低软腭让空气进入鼻腔而发出的。

在**摩擦音(摩擦音(摩擦音))**中,气流受到压缩,但不会完全被切断。由于压缩而产生的湍流会产生一种特有

的“嘶嘶”声。英语的唇齿摩擦音[f]和[v]是通过将下唇压在上齿上,使上齿之间的气流受限而产生的。齿擦音[th]和[dh]让空气在牙齿之间的舌头周围流动。龈擦音[s]和[z]是舌头在齿龈脊上产生的,迫使空气流过牙齿的边缘。在腭龈擦音[sh]和[zh]中,舌头位于齿龈脊的后部,迫使空气通过舌头形成的凹槽。更高音调的摩擦音(英语[s], [z], [sh]和[zh])被称为**咝音(sibilants)**。先摩擦而后阻塞的塞音叫做**塞擦音(affricates)**,这包括英语[ch](chicken 鸡)和[jh](giraffe 长颈鹿)。

在**半元音(approximant)**中,两个发音部位靠得很近,但还没有近到造成强气流的程度。在英语[y](yellow 黄色)中,舌头靠近口腔顶部,但又没有接近到足够产生摩擦音的强气流流。在英语[w](wood 木头)中,舌的后部靠近软腭。美式[r]的形成至少有两种方式:舌尖伸展并靠近上腭,或者整个舌面隆起接近上腭。[l]是指舌尖向上抵住齿龈脊或牙齿,舌头一侧或两侧降低,让空气流过。[l]之所以被称为**边音(lateral sound)**,是因为舌的两侧下降。

颤音(tap)或**(flap)[dx]**是舌对齿龈脊的快速运动。单词 lotus ([l ow dx ax s])中间的辅音是大多数美式英语方言中的一个颤音;说许多英国方言的人会用[t]来代替。

元音

像辅音一样,元音也可以通过发音时的位置来区分。元音的最相关的参数有三个:第一个参数是元音高度,大致相当于舌头最高部分所处的位置;第二个参数是元音前(frontness)或后(backness),表示这个高点朝向口腔的正面或后面;第三个参数是元音的圆状,嘴唇的形状是圆的或不是圆的。图 25.4 显示了不同元音时舌头的位置。

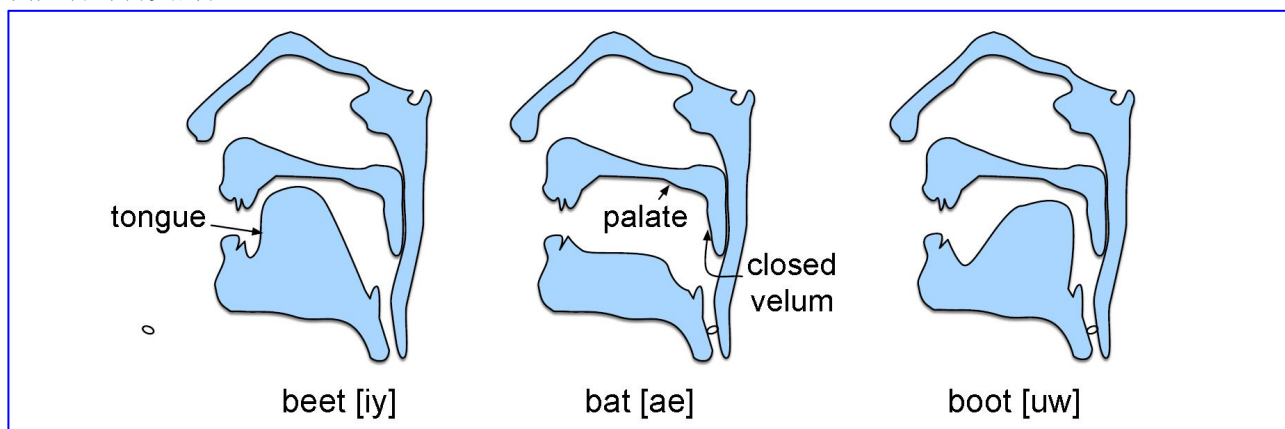


图 25-4: 英语三个元音的舌位: 前高[iy], 前低[ae]和后高[uw]

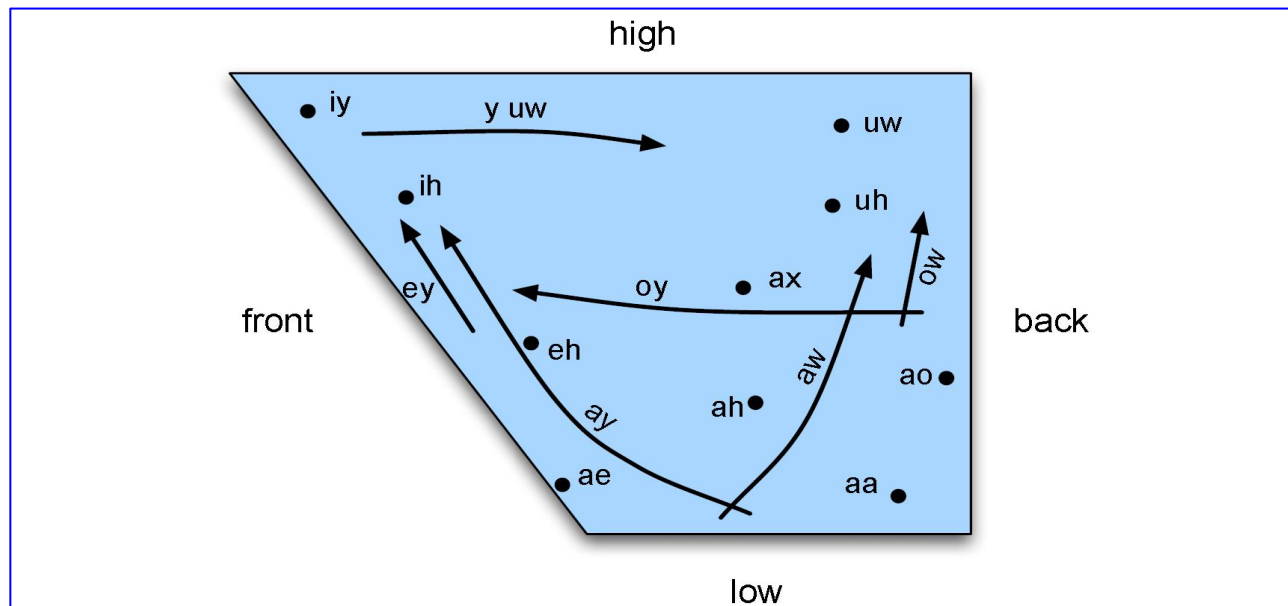


图 25-5: 英语元音的“元音空间”示意图

例如,在元音[iy]中,舌头的最高点对着嘴的前部。相比之下,在元音[uw]中,舌头的最高点对着嘴的后部。舌头朝前伸的元音叫做**前元音(front vowels)**;那些舌朝后翘的元音叫做**后元音(back vowels)**。注意,虽然[ih]和[eh]都是前元音,但[ih]的舌头比[eh]要高。舌头的最高点相对较高的元音被称为**高元音(high vowels)**;舌头最大高度为中等或较低的元音分别称为中元音或低元音。

图 25.5 显示了不同元音高度的示意图。这是一个示意图,因为高度这一抽象属性仅仅与舌头的实际位置大致相关;事实上,它更准确地反映了声学事实。注意,这个图表有两种元音:一种是用点来表示舌高,另一种是用路径表示舌高。在产生元音时舌头位置发生显著变化的一个元音是**双元音(diphthong)**。英语中双元音特别丰富。

第二个重要的元音发音维度是嘴唇的形状。有些元音发音时嘴唇是圆的(和吹口哨时嘴唇的形状一样)。这些**圆唇元音(rounded vowels)**包括[uw]、[ao]和[ow]。

音节

辅音和元音结合在一起构成一个**音节(syllable)**。音节是由一个类似元音的声音(或响音[sonorant])和它周围的最紧密相关的辅音结合而成的。狗(dog)这个词有一个音节,[d aa g](在我们的方言中);猫薄荷(catnip)这个词有两个音节,[kaet]和[nhp]。我们把位于音节核心部分的元音称为**音节核(nucleus)**。初始辅音(如果有的话)称为**音节头(onset)**。具有多个辅音的音节头(如 strike[s t r ay k])被称为复杂音节头。**音节尾(coda)**是跟随在音节核后面的可选辅音或辅音序列。因此,对于 dog 单词,[d]是音节头,[g]是音节尾。**韵(rime)**,或韵脚(rhyme),是音节核加音节尾。图 25.6 显示了一些音节结构例子。

自动将一个单词分解成音节的任务称为**音节切分(syllabification)**。音节结构也与语言的**音子配列(phonotactics)**密切相关。音子配列这个术语指的是音子在一种语言中相互跟随的限制。例如,英语对什么辅音可以一起出现在音节头有严格的限制;例如,序列[zdr]不能是一个合法的英语音节头。音子配列可以用语言模型或音子序列的有限状态模型来表示。

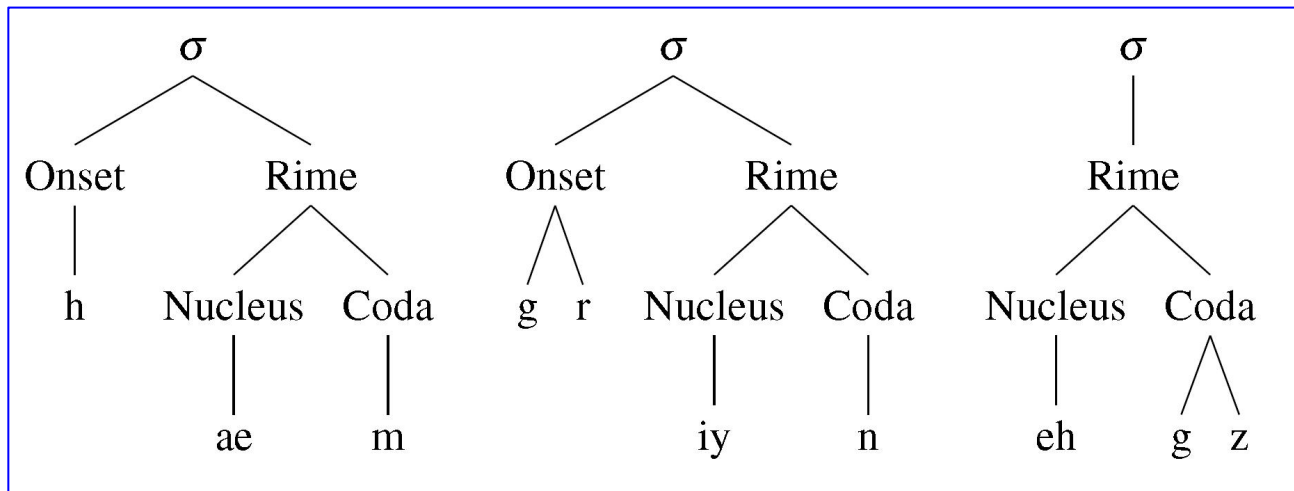


图 25-6: 火腿(ham)、绿色(green)、鸡蛋(eggs)的音节结构(σ =音节)

25.3. 韵律学

韵律学(prosody)研究的是语言的语调和节奏方面,特别是使用 **F0**、**强度**和**持续时间**来传达语用的、情感的、或对话-互动的意义¹。韵律可以用来标记话语结构,比如陈述句和疑问句之间的区别,或者对话的结构方式。韵律是用来标记一个特定的词或短语的显著性。韵律被大量用于副语言功能,如传达情感含义,如快乐、惊讶、或愤怒。而韵律在会话中管理依次发言起着重要的作用。

¹ 这个词在诗歌中有不同但相关的用法,意思是对诗歌韵律结构的研究。

25.3.1. 韵律突显度：重读、重音和中元音

在美式英语的自然发音中，有些词听起来比其他词更突显，这些词中的某些音节也比其他音节更突显。我们所说的**突显度(prominence)**是指这些单词或音节在听话者的感知上更加突显。在英语中，为了让一个单词或音节更加突出，说话者会提高音量，放慢语速(这样就会有更长的音长)，或者在发音过程中改变 F0，使其更高或更具变化性。

重读(accent) 我们通过一种称为**音高重读(pitch accent)**的语言标记来表示突出。突出的音高或音节被认为**负荷者(bearer)**(与之相关的)音高重读。因此，可以通过重读下划线的单词来表达这种语段：

(25.1) I'm a little surprised to hear it characterized as happy.

词汇重音(lexical stress) 带有音高重读的音节称为重读音节。并非单词的每个音节都可以重读：音高重读必须在具有词汇**重音(stress)**的音节上实现。词汇重音是单词在词典中的发音的属性；如果单词是重读的，则带有词汇重音的音节会更响亮或更悠长。例如，**surprised**(惊讶)一词的重音在其第二个音节上而不是第一个音节上(尝试通过说出 **SURprised** 来发重音在第一个音节上，希望这对您来说是错误的)。因此，如果单词 **surprised** 在句子中接收到音高重读，则第二个音节将更强。下面的例子显示了下划线的重读单词，其中带重音音节的重读用黑体表示(更响、更长的音节)²：

(25.2) I'm a little surprised to hear it characterized as happy.

重音被标在字典里。例如，CMU 词典(CMU, 1993)在 counter: [K AW1 N T ER0]或 table: [T EY1 B AH0 L]的词条中，用 0(非重音)或 1(重音)标记元音。词汇重音的差异会影响单词意义；对于单词 **content**，作为名词读音为[K AA1 N T EH0 N T](意思是“内容”)，作为形容词读音为[K AA0 N T EH1 N T](意思是“满意的”)。

弱元音和中元音 非重音的元音可以进一步被减弱为**弱元音(reduced vowels)**，其中最常见的是**中元音(schwa)**([ax])，如在 **parakeet**(长尾小鹦鹉)的第二个元音中一样：读为[p ae r ax k iy t]。在弱元音中，发音姿势不如完整的元音完整。并非所有非重音的元音都减弱了；任何元音，尤其是双元音，即使在非重音的位置也可以保持其完整的质量。非重读元音可以被进一步弱化为重读元音，其中最常见的是 **Schwa** ([ax])，如长尾鹦鹉的第二个元音：[p ae r ax k iy t]。在简化元音中，发音姿势不像在全元音中那样完整。并不是所有的非重读元音都减少了；任何元音，特别是双元音，即使在非重读位置也能保持其全部音质。例如，元音[iy]可以出现在 **eat** [iy t]中的重音位置，也可以出现在 **carry** [k ae r iy]中的非重音位置。

总之，有一个连续的韵律突显度，它通常是有用的代表级别，如重读、重音、全元音、弱元音。

25.3.2. 韵律结构

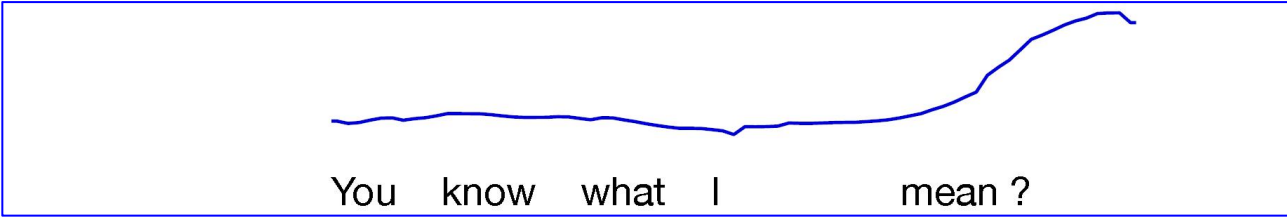
口语句子有韵律结构：有些词似乎自然地组合在一起，而有些词之间似乎有明显的中断或脱节。韵律结构通常用**韵律短语(prosodic phrasing)**来描述，意思是一个语段具有韵律短语结构，其方式与具有句法短语结构相似。例如，I wanted to go to London, but could only get tickets for France(我想去伦敦，但只能买到去法国的票)，这句话似乎有两个主要的**语调短语(intonation phrases)**，它们的边界出现在逗号处。此外，在第一个短语中，似乎还有另一组较小的韵律短语边界，通常称为**中间短语(intermediate phrases)**，它们把单词分开，就像 I wanted | to go | to London.一样。这些语调短语通常与句法结构成分相关(Price 等人 1991, Bennett 和 Elfner 2019)。

自动预测韵律边界对于 TTS 这样的任务很重要。现代方法使用序列模型，将原始文本或带有解析树等特征注释的文本作为输入，并在每个单词边界做出中断/不中断决策。它们可以在标签为韵律结构的数据上训练，如波士顿大学广播新闻语料库(Ostendorf 等人，1995)。

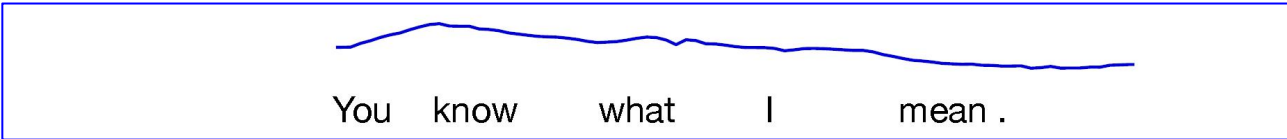
² 译者注：英语单词重音常被称为**重读(accent)**而不是**重音(stress)**，以区别于重音位置规则。但是为了简便，下面我们仍然使用“重音”这个比较熟悉的词语。---胡壮麟. 语言学教程[M]第五版中文本.北京:北京大学出版社, 2019: 38.

25.3.3. 音调

两个具有相同的突显度和短语模式的语段，由于有不同的**音调(tune)**，在韵律上也会有所不同。语段的音调是 F0 随时间的起伏。一个非常明显的关于音调的例子是英语中陈述句和**是非问句**之间的区别。同样的单词也可以用 F0 升调来表示是非问句(称为**问句升调(question rise)**):



或者 F0 的最后降低(称为**最后降调(final fall)**)来表示陈述句语调:



语言广泛使用音调来表达意义(Xu, 2005)。例如，在英语中，除了众所周知的是非问句的升调外，一个包含用逗号分隔的一系列名词的短语，在每个名词后面通常有一个称为**连续升调(continuation rise)**的短暂升调。其他的例子包括英语特有的表达矛盾和惊奇的轮廓。

突显与音调的联系

音高重读有不同的种类，它们与音调有关;例如，高的音高重读与低的音高重读有不同的功能。在不同的语言中有许多口音类的类型。其中一种类型是语调的**声调和间断索引(Tone and Break Indices, ToBI)**理论的一部分(Silverman 等人, 1992)。ToBI 中的每个单词都可以与图 25.7 中所示的五种音高重读类型中的一种相关联。ToBI 中的每个语段由一系列的语调短语组成，每个短语以图 25.7 所示的四种**边界声调(boundary tones)**中的一种结尾，这四种边界声调代表了语段音调的最终体。ToBI 有多种语言版本。

Pitch Accents		Boundary Tones	
H*	peak accent	L-L%	“final fall”: “declarative contour” of American English
L*	low accent	L-H%	continuation rise
L*+H	scooped accent	H-H%	“question rise”: cantonical yes-no question contour
L+H*	rising peak accent	H-L%	final level plateau
H+!H*	step down		

图 25- 7: ToBI 重读和边界声调标签

图注：美国英语语调的 ToBI 转录系统的重读和边界声调标签 (Beckman 和 Ayers 1997, Beckman 和 Hirschberg 1994)。

25.4. 声学语音和信号

首先，我们简要介绍了声波波形及其数字化和频率分析；鼓励感兴趣的读者查阅本章末尾的参考资料。

25.4.1. 波

声学分析是基于正弦和余弦函数的。图 25.8 显示了一个正弦波图，特别是函数：

$$y = A*\sin(2\pi ft)$$
 (25.3)

其中，我们把振幅 A 设为 1，频率 f 设为每秒 10 周。

回想一下基础数学，波的两个重要特征是它的**频率(frequency)**和**振幅(amplitude)**。频率是波在一秒钟内自我重复的次数，也就是循环的周数(cycles)。我们通常以每秒的周数来测量频率。图 25.8 中的信号在 0.5 秒内重复 5 次，即每秒 10 个周期。每秒钟的周期通常称为赫兹(hertz, 缩短为 Hz)，因此图 25.8 中的频率将被描述为 10Hz。正弦波的振幅 A 是 Y 轴上的最大值。波的**周期(period)**T 是完成一个周期的振动所需要的时间，定义为：

$$T = 1/f$$

(25.4)

图 25.8 中的每一周持续十分之一秒，因此 $T=0.1$ 秒。

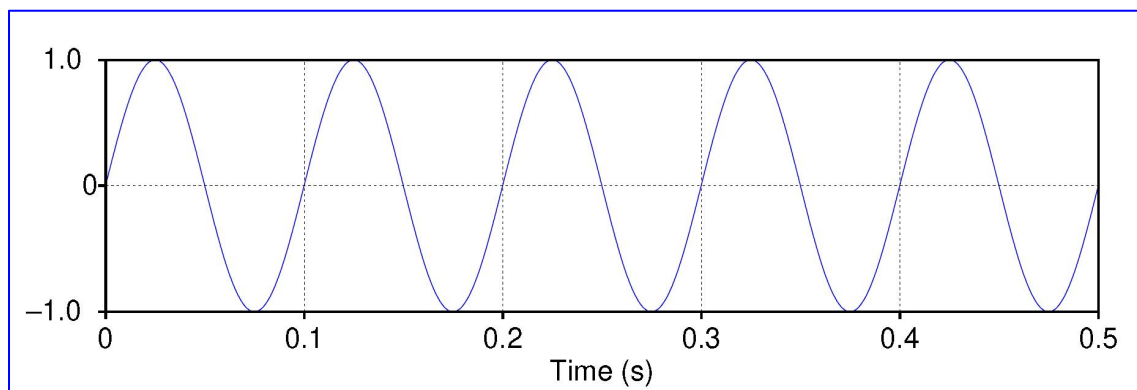


图 25-8：频率为 10Hz，振幅为 1 的正弦波

25.4.2. 语音声波

让我们从假设的波转向声波。语音识别器的输入就像人耳的输入一样，是一系列复杂的气压变化。气压的这些变化显然是由说话者引起的，并且是由空气通过声门进入口腔或鼻腔的特定方式引起的。我们通过绘制气压随时间的变化来表示声波。有时可以帮助您理解这些图表的一个比喻是一块垂直板，该板阻挡了气压波(也许在扬声器口前的麦克风中，或者在听众耳中的鼓膜中)。该图测量了该板上空气分子的压缩量或稀薄量(解压量)。图 25.9 显示了取自 Switchboard 语料库的一段短波形，其中元音[iy]来自某人说“she just had a baby(她刚生了一个孩子)”。

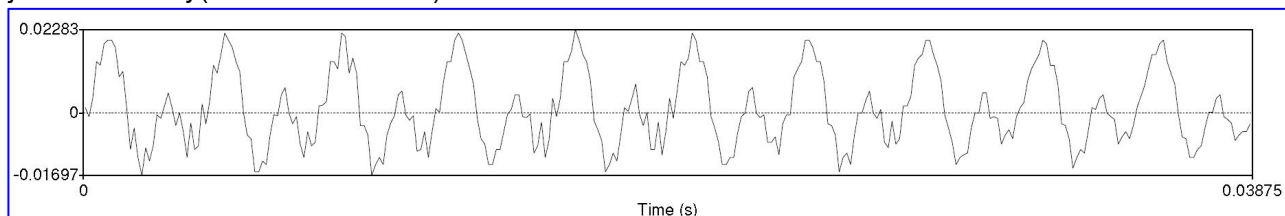


图 25-9：某语段中元音[iy]的波形图

图注：来自图 25.13 中的元音[iy]的波形图。y 轴表示空气压力高于和低于正常大气压的水平。x 轴表示时间。注意这个波有规律地重复。

将声波数字化(如图 25.9 所示)的第一步是将模拟表示(首先是气压，然后是麦克风中的模拟电信号)转换为数字信号。这种模数转换有两个步骤：**采样(sampling)**和**量化(quantization)**。为了对信号进行采样，我们在特定时间测量其幅度。采样率是每秒采样的数量。为了准确地测量波，每个周期中至少必须有两个样本：一个测量波的正侧部分，另一个测量波负侧部分。每个周期多于两个采样可提高幅度精度，但少于两个采样会使波的频率完全丢失。因此，可以测量的最大频率波是其频率为采样率一半的波(因为每个周期需要两次采样)。给定采样率的最大频率称为**奈奎斯特频率(Nyquist frequency)**。人类语音中的大多数信息的频率都低于 10,000 Hz。因此，要获得完全的准确性，就需要 20,000 Hz 的采样率。但是电话语音是由交换网络过滤的，电话只能发送小于 4,000 Hz 的频率。因此，像 Switchboard 语料库一样，8,000 Hz 的采样率足以满足电话带宽语音的需求，而麦克风语音通常使用 16,000 Hz 的采样率。

即使是 8000 赫兹的采样率，每秒的语音也需要进行 8000 次振幅测量，因此有效地存储振幅测量值是很重要的。它们通常以整数形式存储，要么 8 位(-128-127 的值)，要么 16 位(-32768-32767 的值)。这个将实值数字表示为整数的过程称为**量化(quantization)**，因为两个整数之间的差异作为最小粒度(量化范围)，所有接近于这个量化范围的值都以相同的方式表示。

一旦数据被量化，它将被存储在各种格式中。这些格式的一个参数是上面讨论的采样率和样本量。电话语音通常以 8 kHz 采样并存储为 8 位样本，而麦克风数据通常以 16 kHz 采样并存储为 16 位样本。另一个参数是**通道(channel)**数。对于立体声数据或两方对话，我们可以将两个通道存储在同一文件中，也可

以将它们分别存储在单独的文件中。最后一个参数是线性或压缩的单个样本存储。用于电话语音的一种常见压缩格式是 μ -law(通常写为 u-law, 但仍发音为 mu-law)。对数压缩算法(如 μ -law)的直觉是, 人的听力在小强度下比大强度下更敏感。对数以较小的忠诚度表示较小的值, 但以较大的错误为代价。线性(非对数)值通常称为线性脉冲编码调制(Pulse Code Modulation, PCM)值。这是将线性 PCM 样本值 x 压缩为 8 位 μ -law(其中 8 位的 $\mu = 255$)的公式:

$$F(x) = \frac{\text{sgn}(x)\log(1 + \mu|x|)}{\log(1 + \mu)} , \quad -1 \leq x \leq 1 \quad (25.5)$$

有很多标准的文件格式可以存储数字化的波文件, 比如微软的 .wav 和苹果的 AIFF, 它们都有特殊的文件头; 也使用简单的无头“原始”文件。例如, .wav 格式是微软多媒体文件 RIFF 格式的子集; RIFF 是一种通用格式, 可以表示一系列嵌套的数据块和控制信息。图 25.10 显示了一个简单的 .wav 文件, 其中包含单个数据块及其格式块。

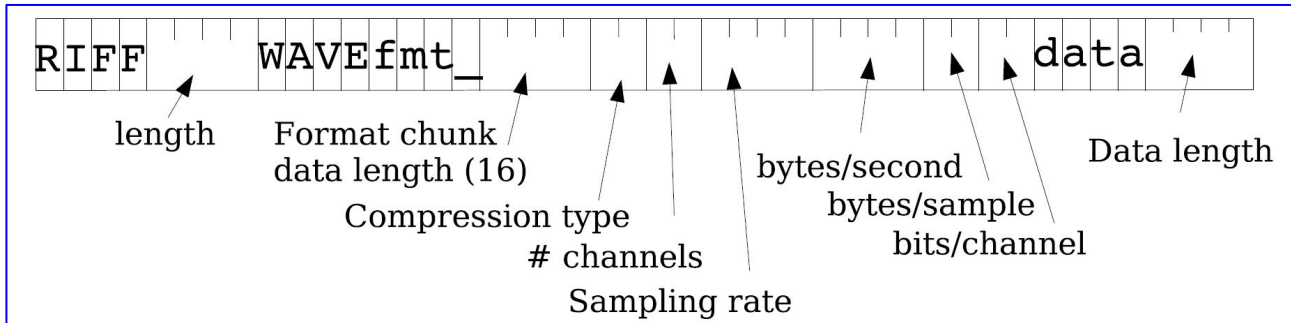


图 25-10: Microsoft wavefile 头文件格式

图注: Microsoft wavefile 头文件格式, 假设文件很简单, 只有一个块(chunk)。紧随这个 44 字节文件头的是数据块。

25.4.3. 频率和振幅;音调和响度

声波, 像所有的波一样, 可以用频率、振幅和我们前面介绍的纯正的正弦波的其他特性来描述。在声波中, 它们不像正弦波那样容易测量。让我们考虑频率。在图 25.9 中注意到, 虽然不是精确的正弦波, 但是波是周期性的, 图中捕获了在 38.75 毫秒内重复 10 次(0.03875 秒)。因此, 这段波的频率是 $10/0.03875$ 或 258Hz。

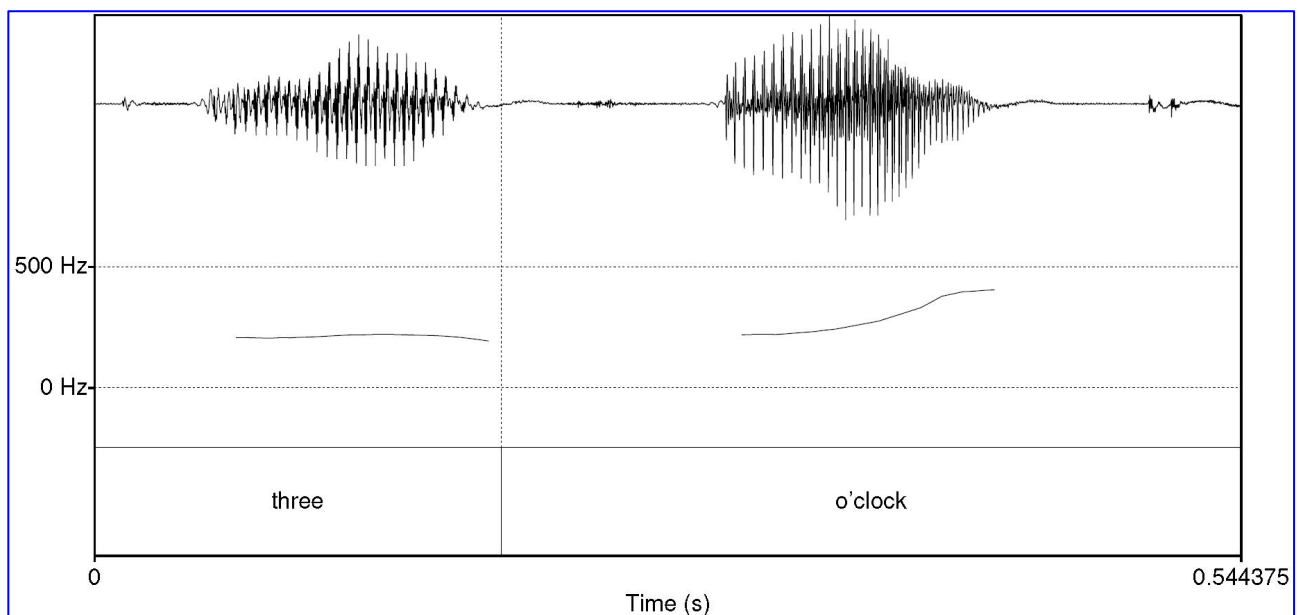


图 25-11: 问句“Three o'clock?”的音高轨迹

图注: 音高轨迹显示在波文件下面。注意问句终端 F0 的升调。注意在非常平静的部分 (“o’clock” 中的 “o”) 没有音高轨迹; 自动音高跟踪是基于浊音区域中的脉冲计数。如果没有浊音(或声音不足), 则不起作用。

这个 258 赫兹的周期性波是从哪里来的?它来自声带振动的速度;由于图 25.9 中的波形来自元音[iy], 所以它是浊音。请记住, 发浊声是由声带的定期开启和闭合引起的。当声带打开时, 空气通过肺部向上推动, 形成一个高压区域。当声带闭合时, 就不会有来自肺部的压力。因此, 当声带振动时, 我们期望看到图 25.9 所示的振幅有规律的峰值, 每个主要峰值对应声带的开口。声带振动的频率, 或复杂波的频率, 称为波形的**基频(fundamental frequency, F0)**。我们可以在**音高轨迹(pitch track)**中绘制 F0 随着时间的变化。图 25.11 显示了一个简短问句 “Three o'clock ?” 在波形下面表示。注意问句终端 F0 的升调。

图 25.9 中的纵轴测量的是空气压强的变量;压强是单位面积上的压力, 用帕斯卡(Pa)度量。纵轴数值越高(振幅越高)表示该点空气压强越大, 零值表示正常(大气)气压, 负值表示低于正常气压(稀薄度)。

除了这个在任何时间点的振幅值之外, 我们还经常需要知道在某个时间范围内的平均振幅, 给我们一些关于气压的平均变化有多大的概念。但是我们不能只对一个范围内的振幅值取平均值;正数和负数(大部分)会抵消掉, 留给我们一个接近于零的数字。相反, 我们通常使用 **RMS(root-mean-square, 均方根)** 振幅, 它在平均(使其为正)之前对每个数字进行平方, 最后在结束时取平方根。

$$\text{RMS amplitude}_{i=1}^N = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad (25.6)$$

信号的**强度(power)**与振幅的平方有关。如果一个声音的采样数为 N, 则其强度为:

$$\text{Power} = \frac{1}{N} \sum_{i=1}^N x_i^2 \quad (25.7)$$

而不是强度, 我们更多地是指声音的**音强(intensity)**, 它将强度归一化为人类听觉阈值, 并以 dB 为单位进行测量。如果 P_0 是听阈压力 = 2×10^{-5} Pa, 则音强定义如下:

$$\text{Intensity} = 10 \log_{10} \frac{1}{N \cdot P_0} \sum_{i=1}^N x_i^2 \quad (25.8)$$

图 25.12 显示了句子 Is it a long movie? 的音强图, 来自 CallHome 语料库, 再在下方显示波形图。

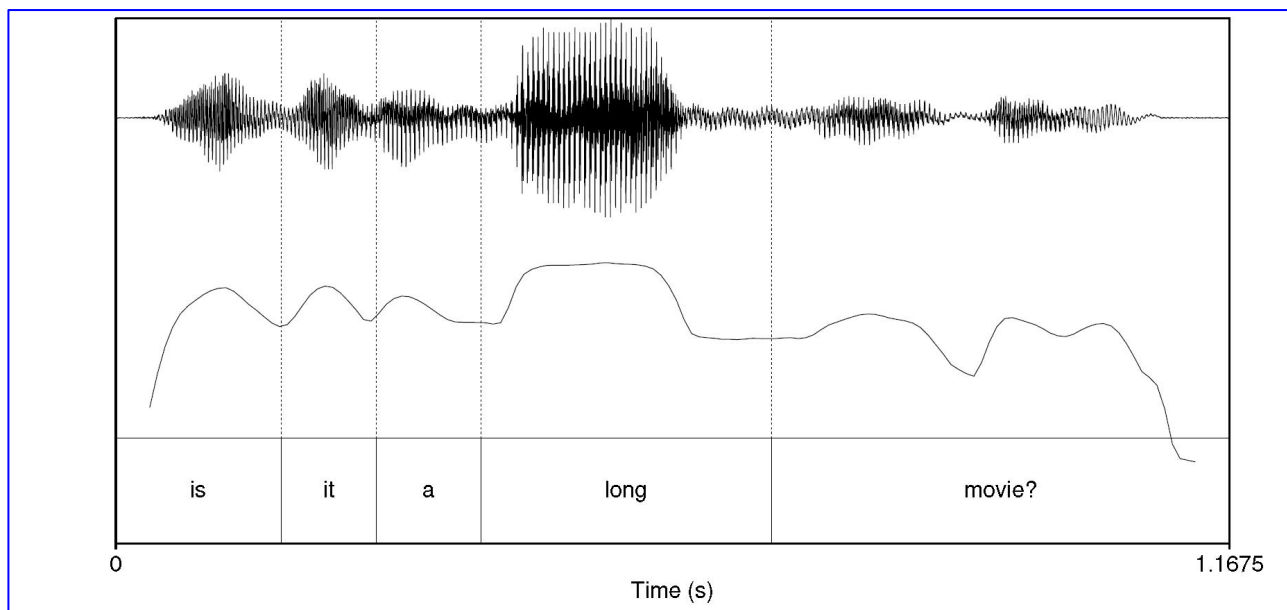


图 25-12: 句子 Is it a long movie? 的音强图

图注: 注意图中每个元音的音强峰值, 以及单词 long 的特别高的峰值。

音高(pitch)和**响度(loudness)**是两个重要的感知特性, 与频率和音强有关。声音的音高是基本频率的心理感觉或感知相关性; 通常, 如果声音的基频更高, 我们会认为它具有更高的音高。我们之所以说“通常”, 是因为这种关系不是线性的, 因为人类的听力对于不同的频率具有不同的敏锐度。粗略地说, 人的音高感知在 100 Hz 至 1000 Hz 之间最为准确, 在此范围内, 音高与频率呈线性关系。人耳听觉不能准确地表示高于 1000 Hz 的频率, 而高于该范围时, 音高与频率成对数关系。对数表示意味着高频之间的差异被压缩, 因此无法准确感知。音高感知尺度(scale)有多种心理声学模型。一种常见的模型是**梅尔(mel)**尺度(Stevens

等人 1937, Stevens 和 Volkmann 1940)。mel 是音高的一个单位, 定义为使在音高上等距的成对的声音被相等数量的 mels 分开。mel 频率 m 可以从原始声频计算如下:

$$m = 1127 \ln(1 + f/700) \quad (25.9)$$

我们将在第 26 章中看到, mel 尺度在语音识别中扮演着重要的角色。

声音的响度是一种强度的感知关联。因此, 高振幅的声音会被认为声音更响, 但这种关系也不是线性的。首先, 正如我们在上面定义 μ -law 压缩时所提到的, 人类在低强度范围内具有更大的分辨率; 耳朵对微小的力强度差异更敏感。第二, 已经证明, 在功率、频率和感知到的响度之间存在着复杂的关系; 某些频率范围内的声音被认为比其他频率范围内的声音更响。

有各种算法可以自动提取 F0。稍微滥用一下术语, 这些被称为 **音高提取(pitch extraction)** 算法。例如, 音高提取的自相关方法, 就是在不同的偏移量处将信号与自身相关联。具有最高相关性的偏移量给出信号的周期。有各种公开可用的音高提取工具包; 例如, Praat(软件包)提供了一个增强的自相关音高跟踪器(Boersma 和 Weenink, 2005)。

25.4.4. 从波形来解释音子

从波形的目视检查中可以学到很多东西。例如, 元音很容易识别。回想一下元音是浊音: 元音的另一个特性是它们倾向于较长且相对较响(如我们在图 25.12 的音强图中所看到的)。时间长度直接在 x 轴上显示出来, 响度与 y 轴上的振幅(平方)有关。我们在上一节图 25.9 中看到, 浊音是通过振幅的有规则峰值来实现, 每个主要峰值对应于声带的开口。图 25.13 显示了简短的句子 she just had a baby 的波形。我们已经用单词和音子标签来标签该波形。请注意, 图 25.13 中的六个元音[iy], [ax], [ae], [ax], [ey], [iy] 中的每个都有规则的振幅峰值代表浊音。

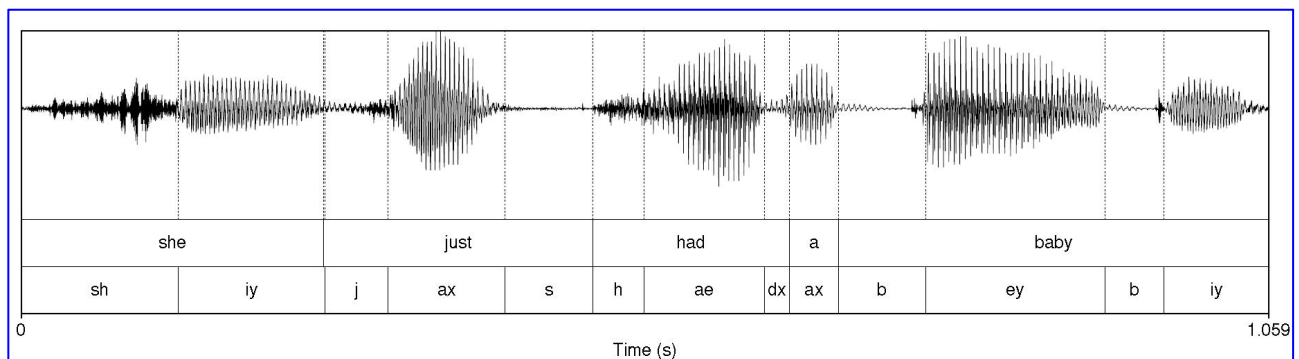


图 25-13: 句子 She just had a baby 的波形图

图注: Switchboard 语料库中 “She just had a baby” 这句话的波形图(会话 4325)。说话的人是女性, 1991 年 20 岁, 大概就是录音录制的时间, 她说的是美国南部中部地区的方言。

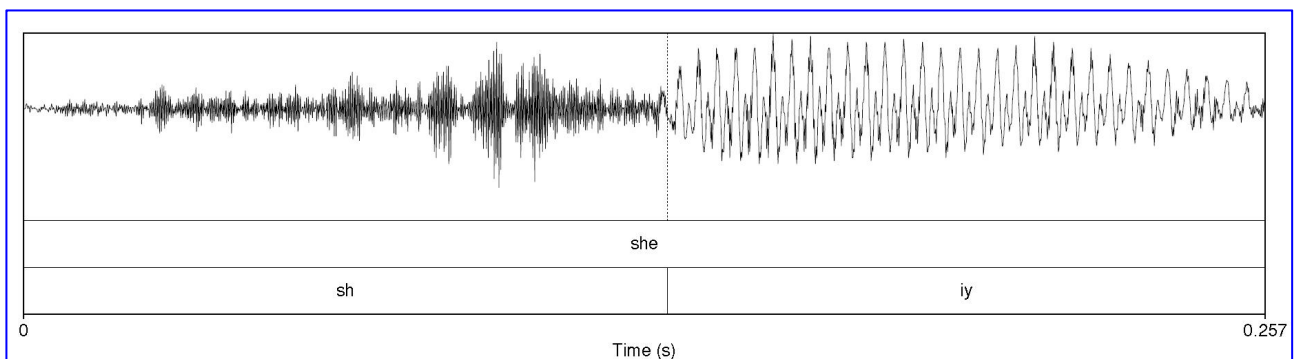


图 25-14: 单词“she”的详细视图

图注: 从图 25.13 的波形文件中提取的第一个单词 “she” 的详细视图。注意摩擦音 [sh] 的随机噪声和元音 [iy] 的有规则浊音之间的区别。

塞辅音包含成阻和紧随其后的除阻，我们经常可以看到一段静默或近乎静默的时期，随后是一个轻微的振幅爆破。我们可以在图 25.13 中看到这两个[b]在 baby 中的情况。

经常在波形中很容易辨认的另一个音子是摩擦音。回想一下，当狭窄的气流通道导致嘈杂的湍流空气时，就会产生摩擦音，尤其是像[sh]这样刺耳的摩擦音。产生的嘶嘶声具有嘈杂的不规则波形。在图 25.13 中可以看到这一点。在图 25.14 中甚至更清楚了，在这里我们只放大了她的第一个单词。

25.4.5. 声谱和频谱

虽然一些宽泛的语音特征(如强度、音高、浊音的出现、塞音的成阻、或摩擦音)可以直接从波形上来解释,大多数计算应用,如语音识别(以及人类听觉处理)是基于不同的声音表示频率的成分。傅里叶分析的精髓在于,每个复杂波都可以表示为不同频率的许多正弦波之和。考虑图 25.15 中的波形。这个波形(在 Praat 中)是由两个正弦波形(频率为 10 赫兹和 100 赫兹)相加而产生的。

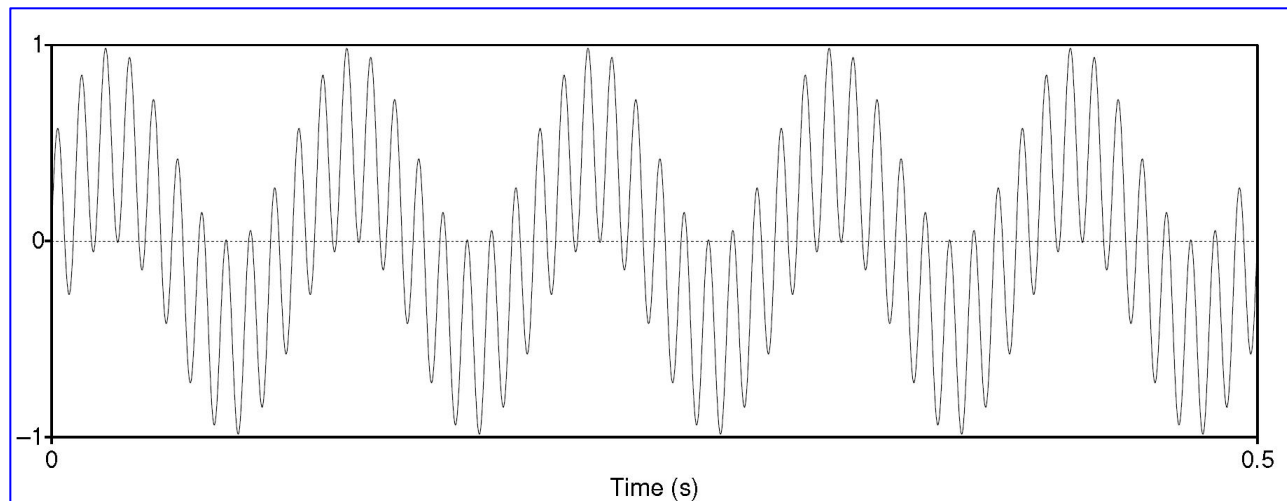


图 25-15: 一个波形是两个正弦波之和

图注: 图中一个频率为 10hz(注意半秒窗口内重复 5 次), 另一个频率为 100hz, 振幅均为 1。

我们可以用**声谱(spectrum)**表示这两个成分的频率。信号的声谱表示其每个频率成分及其幅度。图 25.16 显示了图 25.15 的声谱。以 Hz 为单位的频率在 x 轴上, 振幅在 y 轴上。注意图中的两个尖峰, 一个尖峰为 10 Hz, 另一个尖峰为 100 Hz。因此, 声谱是原始波形的另一种表示, 我们将声谱用作研究特定时间点声波频率成分的工具。

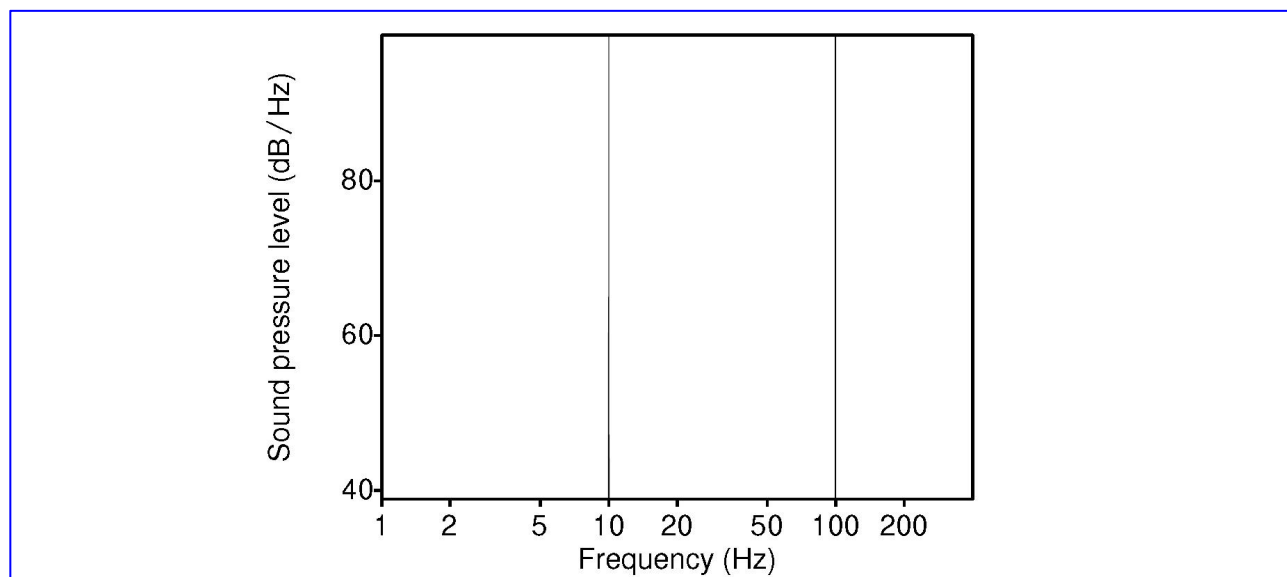


图 25-16: 来自图 25.15 中波形的声谱

现在让我们看一下语音波形的频率成分。图 25.17 显示了单词 **has** 的元音[ae]的一部分波形，该波形是从图 25.13 所示的句子中切出的。请注意，图中有一个复杂的波重复大约十次。但是还有一个较小的重复波，它在每个较大的模式中都会重复四次(注意每个重复波中的四个小峰)。复杂波的频率约为 234 Hz(我们可以弄清楚这一点，因为它在 0.0427 秒内大约重复 10 次，而 10 个周期/0.0427 秒= 234 Hz)。

微波(smaller wave)的频率应该是大波频率的四倍，大约是 936hz。然后，如果你仔细观察，你会看到 936hz 波中的两个小波(**little wave**)。这种**最小波(tiniest wave)**的频率大约是 936hz 波的两倍，也就是 1872hz。图 25.18 显示了图 25.17 中波形平滑的声谱，通过离散傅里叶变换(DFT)计算得到。

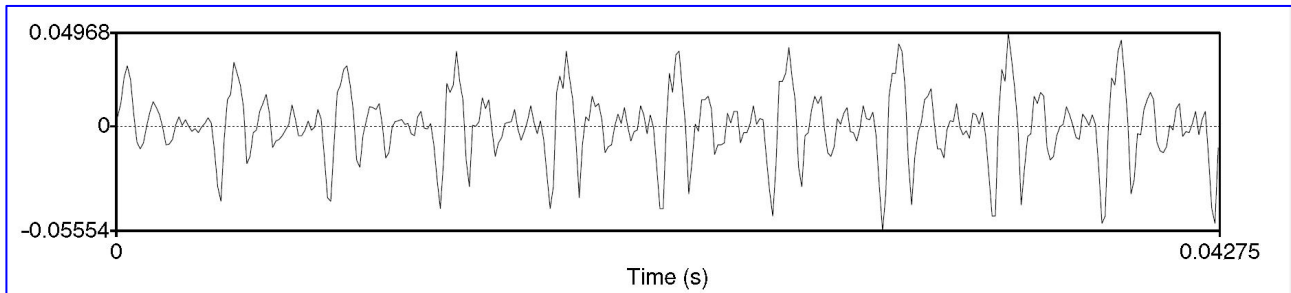


图 25-17: 从图 25.13 中截取了 had 单词中元音[ae]部分的波形

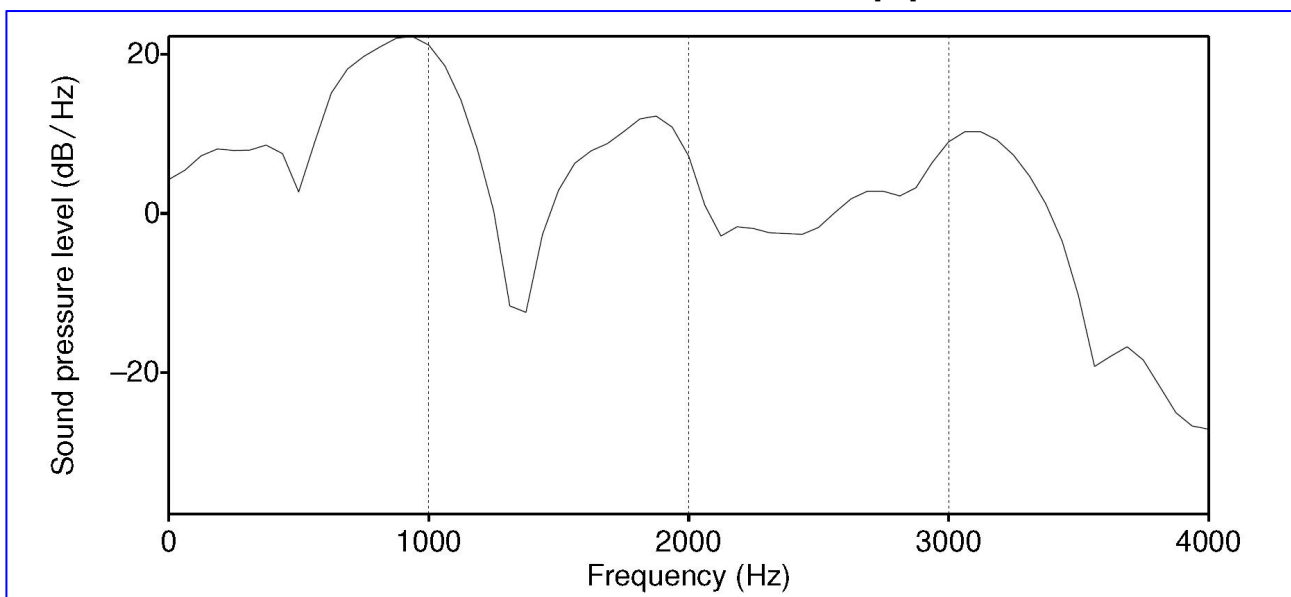


图 25-18: 来自图 25.13 中单词 had 的元音[ae]的声谱

声谱的 **x** 轴表示频率，**y** 轴表示每个频率分量的大小的某种量纲(以分贝(dB)为单位，这是我们之前看到的幅度的对数量度)。因此，图 25.18 显示了大约 930 Hz，1860 Hz 和 3020 Hz 处的重要频率分量，以及许多其他较低幅度的频率分量。通过查看图 25.17 中的波形，这是我们在时域中注意到的前两个成分！

声谱为什么有用?事实证明，这些很容易在声谱中看到的谱峰是不同音子的特色;音子具有声谱特色的“签名”。就像化学元素在燃烧时会发出不同波长的光一样，它使我们可以通过观察光的光谱来检测星体中的元素，我们可以通过观察波形的声谱来检测不同音子的特色“签名”。声谱信息的这种使用对于人类和机器语音识别都是必不可少的。在人类听觉中，**耳蜗(cochlea)**或内耳的功能是计算输入波形的声谱。

让我们看一下不同元音的声谱。由于某些元音会随着时间变化，因此我们将使用另一种称为**频谱(spectrogram)**的图形。声谱显示某个时间点的波的频率分量，而频谱则是一种构想方式：组成波形的不同频率是如何随时间而变化。**x** 轴显示时间，就像波形的表示一样，但是 **y** 轴现在显示频率(以赫兹为单位)。频谱上一个点的暗度对应于频率分量的幅度。非常暗的点具有高振幅，亮的点具有低振幅。因此，频谱是三个维度(时间 **x** 频率 **x** 振幅)可视化的有用方法。

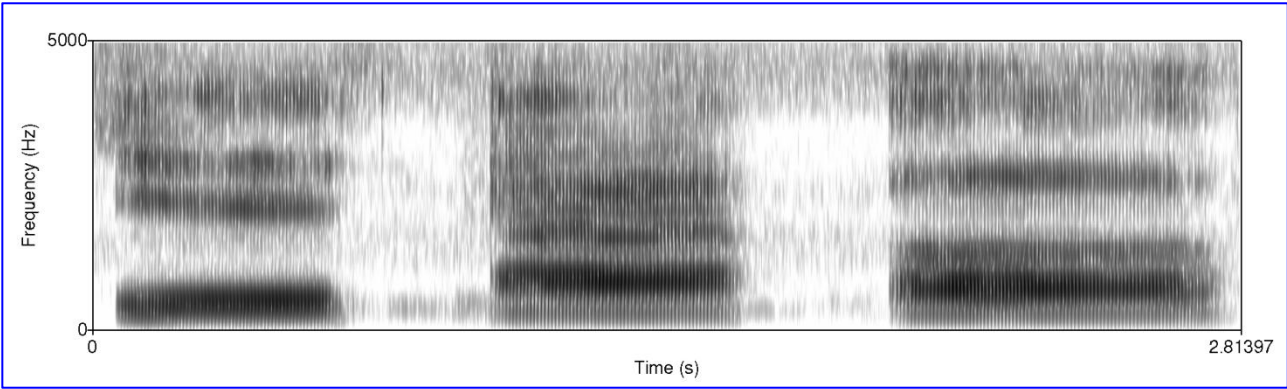


图 25-19：美国英语[ih]、[ae]、[uh]三个元音的频谱

图 25.19 显示了三个美式英语元音[ih]、[ae]和[ah]的频谱。注意，每个元音在不同的频带上都有一组暗条，每个元音的频带略有不同。每一个都代表了我们在图 25.17 中看到的同一种声谱峰。

每个暗条(或声谱峰)称为**共振峰(formant)**。正如我们下面所讨论的，共振峰是一种被声腔特别放大的频带。由于声腔在不同的位置产生不同的元音，它们会产生不同种类的放大或共鸣。我们来看看前两个共振峰，叫做 F1 和 F2。请注意，最底部的暗条 F1 对于三个元音的位置有所不同;它对于[ih]的位置低(集中在大约 470 Hz)，对于[ae]和[ah]的位置稍高(大约 800 Hz)。相比之下，从底部开始的第二个暗条 F2，对于[ih]的位置最高，对于[ae]的位置居中，对于[ah]的位置最低。

我们可以在**连续语流(running speech)**中看到相同的共振峰，尽管弱化和协同发音过程使它们有些难以看到。图 25.20 显示了“she just had a baby”的频谱，它的波形如图 25.13 所示。对于 just 的[ax]、had 的[ae]和 baby 的[ey]，F1 和 F2(还有 F3)都很清楚。

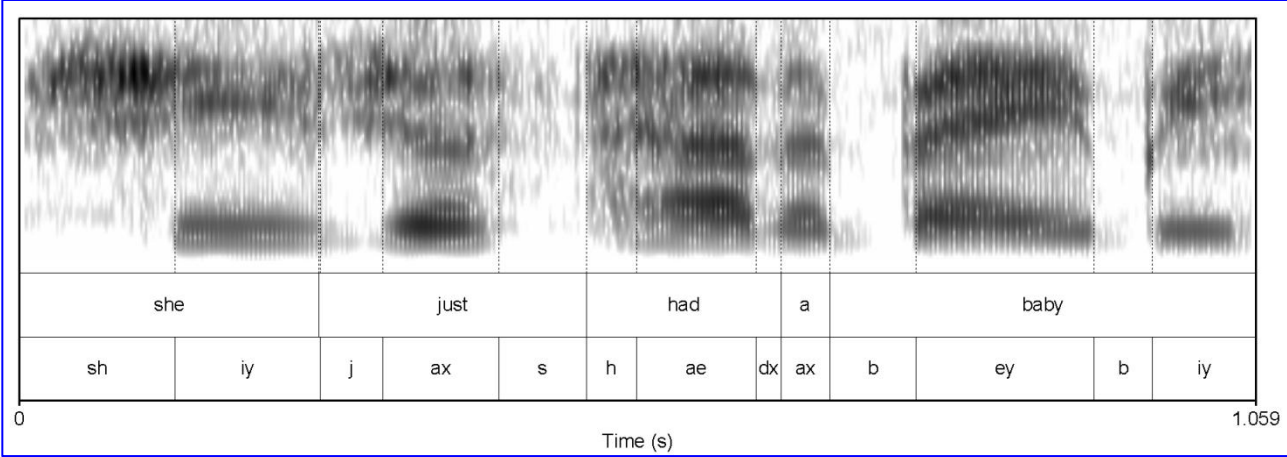


图 25-20：句子“she just had a baby”的频谱

图注：其波形如图 25.13 所示。我们可以把频谱看作声谱的集合(时间切片)，如图 25.18 所示按照端到端放置。

声谱表示可以为音子识别提供哪些特别的线索？首先，由于不同的元音在特征位置处有共振峰，因此频谱可以将元音彼此区分开。我们已经看到，样本波形中的[ae]具有 930 Hz, 1860 Hz 和 3020 Hz 的共振峰。考虑图 25.13 中语段开始时的元音[iy]。该元音的频谱如图 25.21 所示。[iy]的第一共振峰为 540 Hz，远低于[ae]的第一共振峰，第二共振峰(2581 Hz)远高于[ae]的第二共振峰。如果仔细看，大约 0.5 秒处，这些共振峰就会在图 25.20 中显示为暗条。

前两个共振峰(称为 F1 和 F2)的位置在决定元音身份方面起着很大的作用，尽管共振峰仍然因不同的说话者而不同。较高的共振峰更多的是由说话者声腔的普遍特征而不是单个的元音引起的。共振峰也可用于鉴别鼻音音子[n]、[m]和[ng]以及边音(liquids)音子[l]和[r]。

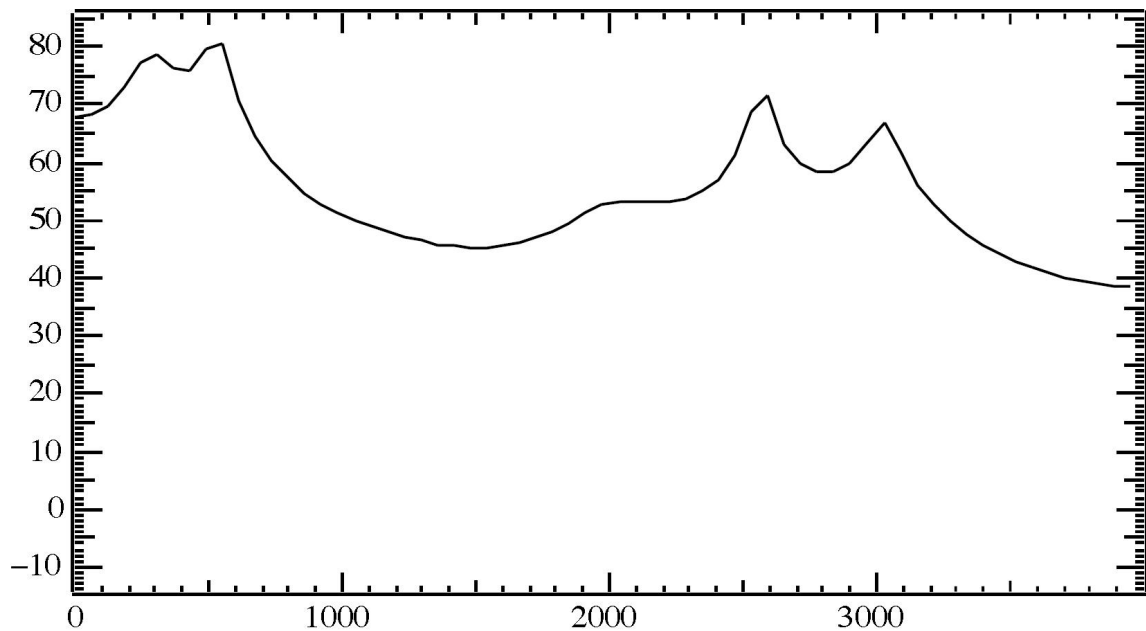


图 25-21: 元音[iy]的平滑(LPC)频谱

图注: 句子 She just had a baby 开始处的元音[iy]的平滑(LPC)频谱。注意, 第一共振峰(540 Hz)远低于图 25.18 所示的[ae]的第一共振峰, 第二共振峰(2581 Hz)远高于[ae]的第二共振峰。

25.4.6. 声源滤波器模型

为何不同的元音有不同的声谱特征? 前述, 共振峰是由口腔共鸣所致。**声源-滤波器模型(source-filter model)**是通过建模来解释声音的声学方式: 声门(声源)产生的脉冲如何由声腔(滤波器)成型。

让我们看它是如何工作的。每当我们有一个波, 如由声门脉冲引起的空气振动, 这个波也就会有**谐波(harmonics)**。谐波是另一种波, 它的频率是基波的数倍。例如, 一个 115 赫兹的声门褶振动导致 230hz、345hz、460hz 的谐波(其他波), 等等。一般来说, 这些波中的每一个都会更弱, 也就是说, 它们的振幅会比基频处的波小得多。

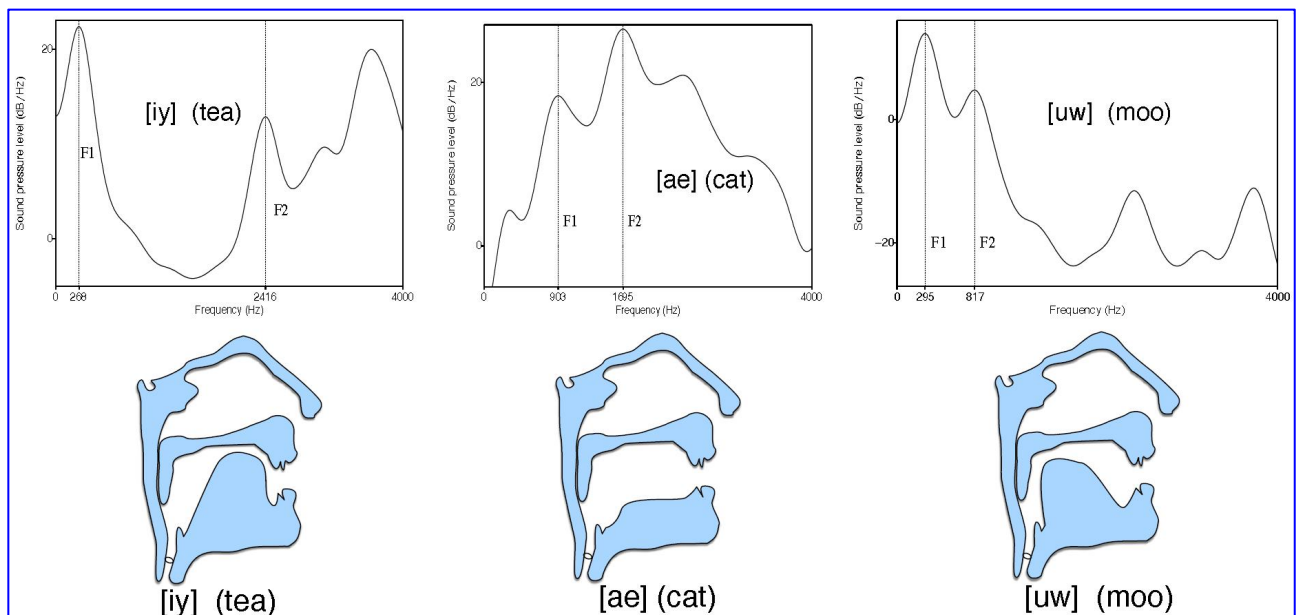


图 25-22: 作为过滤器的声腔位置的可视化

图注: 图中显示了三个英语元音的舌头位置以及生成的平滑声谱, 包括 F1 和 F2。

然而，事实证明，声腔起到了一种过滤器或放大器的作用；事实上，任何空腔，比如管子，都会导致某些频率的波被放大，而其他频率的波被衰减。这种放大过程是由空腔的形状引起的；一个特定的形状会引起特定频率的声音产生共鸣，从而被放大。因此，通过改变空腔的形状，我们可以放大不同的频率。

当我们发出特定的元音时，我们实际上是在通过把舌头和其他发音器官放在特定的位置来改变声腔的形状。结果是不同的元音导致不同的谐波被放大。所以相同基频的波通过不同的声腔位置会导致不同的谐波被放大。

我们可以通过观察声腔形状和相应声谱之间的关系来看到放大的结果。图 25.22 显示了三个元音的声腔位置和一个典型的声谱结果。共振峰是声腔在声谱中放大特定谐波频率的地方。

25.5. 语音资源

各种各样的语音资源可以用于计算工作。在线**发音词典(pronunciation dictionaries)**为每个单词提供语音转写。LDC 发布埃及阿拉伯语、荷兰语、英语、德语、日语、韩语、汉语和西班牙语的发音词汇。对于英语，《CELEX 词典》(Baayen 等人 1995)有 160595 个单词形式的发音，包括音节切分、重音、形态和词类信息。开源的 CMU 发音字典(CMU, 1993)有大约 134,000 个单词的发音。而细粒度的 110,000 个单词的 UNISYN 字典(Fitt, 2002)，为研究目的免费提供，给出了音节切分、重音、以及数十种英语方言的发音。

另一个有用的资源是语音标注语料库(phonetically annotated corpus)，其中一组波形用相应的音子字符串由手工标签的。TIMIT 语料库(NIST, 1990)，最初是德州仪器(TI)、麻省理工学院和 SRI 的联合项目，是一个包含 6300 个阅读句子的语料库，由 630 个说话者每人说 10 个句子。这 6300 个句子是从 2342 个句子中选出来的，其中一些被选为具有特定方言颤音的句子，另一些则是为了最大限度地扩大语音双音子范围。语料库中的每个句子都被语音手工标签，音子序列自动与句子的波形文件对齐，然后对已经自动标签过的音子边界进行手动校正(Seneff 和 Zue,1988)。结果是**时间对齐转写(time-aligned transcription)**：它是一个转写，其中每个音子与波形中的开始和结束时间相关联，如图 25.23 中的示例所示。

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

图 25-23：来自 TIMIT 语料的语音转写例子

图注：利用 ARPAbet 特有的特征进行严式(narrow)转写，如单词 had 中的[d]的上腭音化，单词 dark 中没有除阻的最后一个塞音，单词 suit 中的最后一个[t]喉音化为[q]，单词 water 中[t]读作颤音。TIMIT 语料库还包括时间对齐(未显示)。

Switchboard 转写项目的语音标注语料库包括从 Switchboard 语料库中提取的 3.5 小时的句子(Greenberg 等人 1996)，以及在音节水平上按照时间对齐的转写。图 25.24 显示了一个示例。

0.470	0.640	0.720	0.900	0.953	1.279	1.410	1.630
dh er	k aa	n ax	v ih m	b ix	t w iy n	r ay	n aw

图 25-24：Switchboard 中的短语的语音转写

图注：Switchboard 语料库中的短语 they're kind of in between right now 的语音转写。注意 they're 和 of 中的元音弱化，kind 和 right 中音节尾的删除，以及再音节化现象(of 由于与 in 的音节头相连接而变为[v])。时间是以秒为单位，从句子的开始到每个音节的开头。

Buckeye 语料库(Pitt 等人 2007, 2005)是自发的美国语音的语音转写语料库，包含来自 40 个谈话者的 300,000 个单词。语音转写的语料库也可用于其他语言，包括德语 Kiel 语料库、由中国社会科学院(Li 等人 2000)转写的普通话语料库。

除了像词典和语料库等资源外，还有许多有用的语音软件工具。本书中的许多图形都是由 Praat 软件包生成的(Boersma 和 Weenink, 2005)，它包括音高、声谱和共振峰分析、以及一种脚本语言。

25.6. 总结

本章介绍了语音学和计算语音学的许多重要概念。

- 我们可以用称为**音子(phone)**的单位来表示单词的发音。代表音子的标准系统是国际音标, 即 **IPA**。最常见的英语转写的计算系统是 **ARPAbet**, 它方便地使用 **ASCII** 符号来表示英语语音。

- 音子可以通过发音器官如何发声来描述;辅音的定义是根据其**发音部位(place)**、**发音方式(manner)**、以及声带是否**振动(voicing)**来确定;元音的定义是根据其舌位的**高低(height)**、**前后(backness)**、以及嘴唇是否**圆形(roundness)**来确定。

- 语音也可以用声学来描述。声波可以用**频率(frequency)**、**振幅(amplitude)**、或它们的感知上相应的**音高(pitch)**和**响度(loudness)**来描述。

- 声音的**声谱(spectrum)**描述其不同的频率成分。虽然一些语音属性可以从波形中识别出来, 但人类和机器两者都依赖声谱分析来检测音子。

- **频谱(spectrogram)**是一种随时间变化的声谱。元音是由称为**共振峰(formant)**的特征**谐波(harmonics)**来描述的。

25.7. 文献和历史说明

语音发音的主要见解可以追溯到公元前 800-150 年印度的语言学家。他们发明了发音部位和发音方式的概念, 弄清了发声的声门机制, 并理解了语音同化的概念。直到 2000 年后的 19 世纪末, 欧洲科学才赶上印度的语音学家。希腊人确实有一些基本的语音知识。举例来说, 到柏拉图(Plato)的 Theaetetus 和 Cratylus 时代, 他们将元音与辅音区分开来, 将停顿辅音与连续辅音区分开来。Stoics(斯多葛)提出了音节的概念, 并且意识到对可能单词的音位限制。十二世纪一位不知名的冰岛学者利用**音素(phoneme)**的概念, 并为冰岛人提出了一种音素书写系统, 其中包括音长和鼻音的变音符号。但是直到 1818 年, 他的著作才得以出版, 直到那时, 在 Scandinavia 半岛以外的地区还是鲜为人知(Robins, 1967)。通常认为现代语音学是从 Sweet 开始的, Sweet 在他的《语音学手册》(Handbook of Phonetics, 1877 年)中提出了实质上的音素。他还设计了用于转写的字母, 并区分了**宽式(broad)**和**严式(narrow)**的转写, 提出了许多最终被纳入 IPA 的想法。斯威特(Sweet)被认为是他那个时代最优秀的语音专家。他制作了第一个用于语音目的的语言科学记录, 并发展了发音描述的最新技术水平。他也名声不好很难与人相处, 亨利·希金斯(Henry Higgins)很好地抓住了这一特质, 乔治·萧伯纳(George Bernard Shaw)模仿他的舞台角色。音素是由波兰学者鲍杜因·德考特尼(Baudouin de Courtenay)首次命名的, 他于 1894 年发表了他的理论。

语音入门教材包括 Ladefoged(1993)和 Clark and Yallop(1995)。威尔斯(Wells 1982)是有关英语方言的确定性三卷本资料。

1950 年代末或 1960 年代初, 许多关于声学语音学的经典见解得到了发展。仅有的几个亮点包括诸如声音频谱之类的技术(Koenig 等人 1946), 理论洞察力诸如声源滤波器理论的锻造以及在发声与声学之间的映射中的其他问题(Fant 1960, Stevens 等人 1953, Stevens 和 House 1955, Heinz 和 Stevens 1961, Stevens 和 House 1961), 元音共振峰的 F1xF2 空间(Peterson 和 Barney 1952), 对重音的语音本质的理解, 对音长和音强的线索的使用(Fry, 1955), 对音子感知问题的基本理解(Miller 和 Nicely 1955, Liberman 等人 1952)。Lehiste(1967)是有关声学语音学的经典论文的集合。Gunnar Fant 的许多开创性论文已被 Fant(2004)收集。

优秀的声学语音学教科书包括 Johnson(2003)和 Ladefoged(1996)。Coleman(2005)从语言学的角度介绍了声学和语音的计算处理。Stevens(1998)提出了一个有影响力的语音产生理论。有许多的软件包用于声学语音分析。可能使用最广泛的是 Praat(Boersma 和 Weenink, 2005)。

25.8. 练习

25.1 Find the mistakes in the ARPAbet transcriptions of the following words:

- a. “three” [dh r i]
- d. “study” [s t uh d i]
- g. “slight” [s l iy t]
- b. “sing” [s ih n g]
- e. “though” [th ow]
- c. “eyes” [ay s]
- f. “planning” [p pl aa n ih ng]

25.2 Ira Gershwin’s lyric for *Let’s Call the Whole Thing Off* talks about two pronunciations (each) of the words “tomato”, “potato”, and “either”. Transcribe into the ARPAbet both pronunciations of each of these three words.

25.3 Transcribe the following words in the ARPAbet:

- 1. dark
- 2. suit
- 3. greasy
- 4. wash
- 5. water

25.4 Take a wavefile of your choice. Some examples are on the textbook website. Download the Praat software, and use it to transcribe the wavefiles at the word level and into ARPAbet phones, using Praat to help you play pieces of each wavefile and to look at the wavefile and the spectrogram.

25.5 Record yourself saying five of the English vowels: [aa], [eh], [ae], [iy], [uw]. Find F1 and F2 for each of your vowels.