**Applied Speech and Audio Processing 第三章读书报告**

**3.1语音的产生**

语音的发出大概有如下步骤：

1. 从肺部吹出空气，并用力度控制发出的音量；
2. 声门的开闭控制气流发出；
3. 声带的绷紧程度控制音调；
4. 口鼻的位置控制和影响发出的音色（这一部分侧重元音）；
5. 嘴唇和舌头位置影响发出的字节（这一部分侧重辅音）；

单音音节即单独发出一个音造成的音节。关于音节，分词是一大研究热点。

**3.2语音的特点**

3.2.1 语音分类

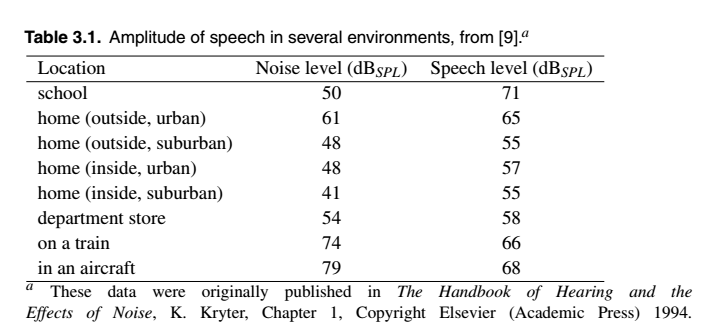
共振峰是语音的一个指征。对于识别语音来说，基音和辅音提供了绝大部分信息，其他共振峰只影响清晰度。

与英语相反，汉语普通话是音调化的（这一点上印欧语系的语言大多不是），所以在音调频率上普通话携带了非常多的信息。

3.2.2 语音的幅度分布

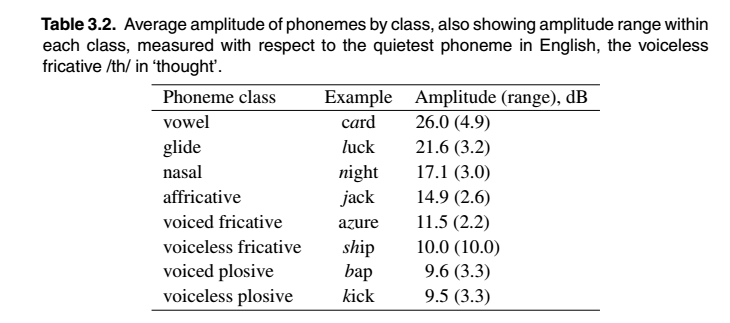
幅度分布就是描述不同环境下的音量大小，其中语音的清晰度受信噪比影响。平均噪声提高0.5db说话人音量大概会提高1db。

书中举了一些例子（大多是户外的情况）：



可以发现的规律是普遍来说户外的两项音量高于户内，有大型机械的场所高于教学、居民区。

3.2.3 语音的种类

不同类型的音节平均音量也不尽相同：

虽然平均起来元音的音量比较大，但实际上辅音对语音意义的识别贡献更大（英语）。英语中如果模糊掉所有辅音，要么听不懂意思，要么听起来千差万别；而指换掉元音相对来说会好很多。

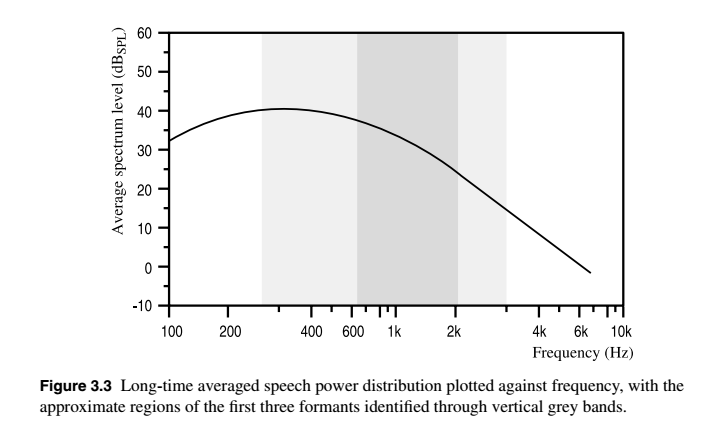
推测正是这个原因导致了连读和吞音的情况在汉语和英语间的差异。

3.2.4 频率分布

清晰度分布最多的频带与语言功率最高的频带并不统一，首先定义两个概念：

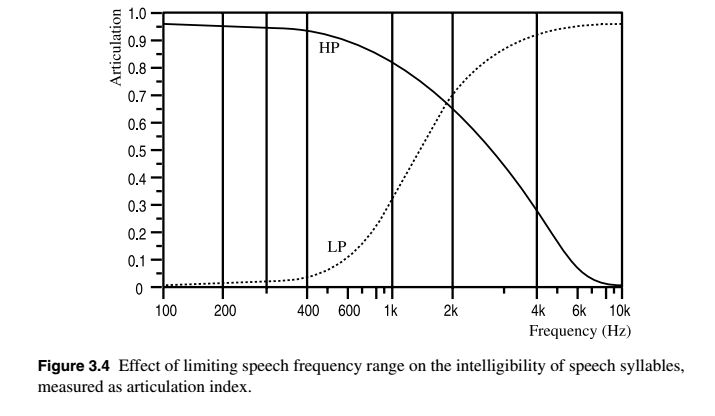
语言功率

男声的基频（大部分能量集中所在）大约集中在500hz，女声则为800hz。消除它们后并不影响听懂的过程，而影响识别说话人的能力（即识别是谁说出这句话的能力）。



清晰度

F2、F3集中的地方决定了信号的清晰度，移除它们（上图后两个灰块）声音会变得几不可辨。



上图的低通滤波只有25%的音频人能听懂，高通则有70%。

3.2.5 时域分布

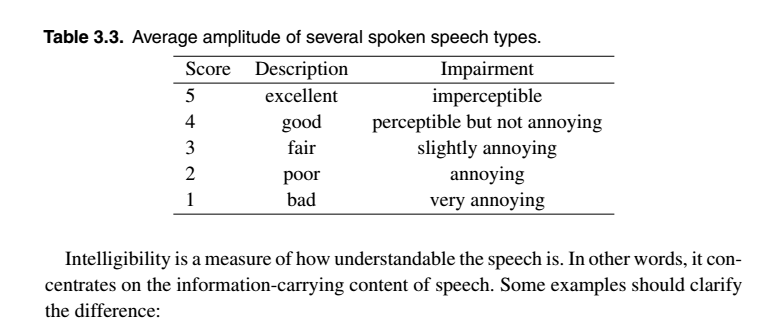
大多情况下说话人的发音间隔决定了语速，而单独音节的发音过程基本是稳定/半稳定的（在每20毫秒内），这是语音建模的理论基础。

每种语言的音节速率基本稳定。

重读的音节在时间上偏长。

**3.3 语音辨识**

3.3.1 清晰度与音质

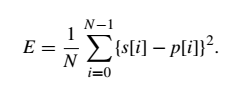
音质是对声音保真度的主观评价指标，衡量的是主观上它与原始声音的相似程度。

上表为一种典型的音质等级划分。

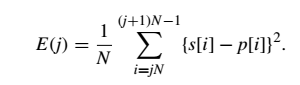
相比之下，清晰度则是语音所携带信息量的衡量标准。大多情况下二者直接相关，也有时它们会呈现相反的状态（如听起来很清晰但不包含语义的一个人的嗫嚅）。

3.3.2 音质的测量

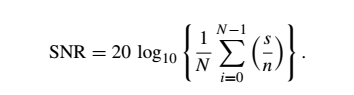
音质使用主观测量平均意见得分（MOS），被试在严格的监听环境下对前述标准就所听的素材打分。相对于较为麻烦的主观评测方式，随着技术的发展也出现了计算机自动识别音质的算法（大多数情况下管用）：PESQ、PSQM、MNB等。

其中有一种粗略算法：计算两个音频的均方误差：

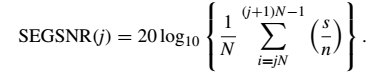
对应代码为：mse=mean((s-p).ˆ2)

同之前的很多分析一样，不做分割地分析长音频可能会丢失重要的细节。如果均方误差也是时变的，那对于长时间的信号做切割后再分析或许更适合，适合的分割长度通常为20-30ms：

代码：mse(j)=mean((s(j\*N+1:(j+1)\*N)-p(j\*N+1:(j+1)\*N)).ˆ2); %无交叠，N采样/帧。

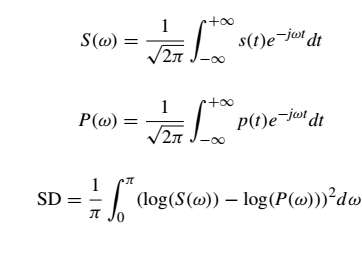
如用来求信噪比，则采用等式

代码：snr=10\*log10 (s./n)

如为分割后则：

代码：segsnr(j)=10\*log10(s(j\*N+1:(j+1)\*N)./n(j\*N+1:(j+1)\*N)) %单位为db，无交叠，每帧时长20-30ms

在频域，以上手段则不起作用，此时可用傅里叶变换计算频谱失真（单位db的平方）：



Matlab中的分析是有限长的，长度为N：

S=fft(s.\*hamming(N));  
S=20\*log10(abs(S(1:N/2)));  
P=fft(p.\*hamming(N));  
P=20\*log10(abs(P(1:N/2)));

SD=mean((S-P).ˆ2); %这里是先求对数再相减，等同于相除后再求对数

这个方法求的是频域的信噪比再平方，是谓频谱失真（spectral distortion/SD）。它可用于不同频谱加权来增加清晰度。

3.3.3 清晰度的测量

清晰的意思是听者能明白语音不同音节的意义。清晰度有主观评测方法若干，但目前缺乏确定测试其中语句辨别结果是否有效的手段。

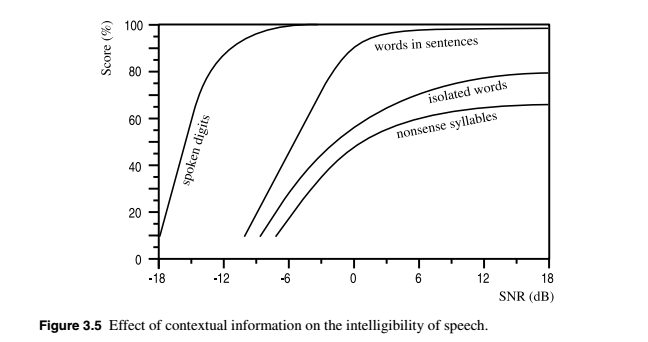
3.3.4 上下文、冗余和词汇量

上下文可用来补全丢失的信息。重要的词丢失也许也可以恢复，这个情况比较随机。

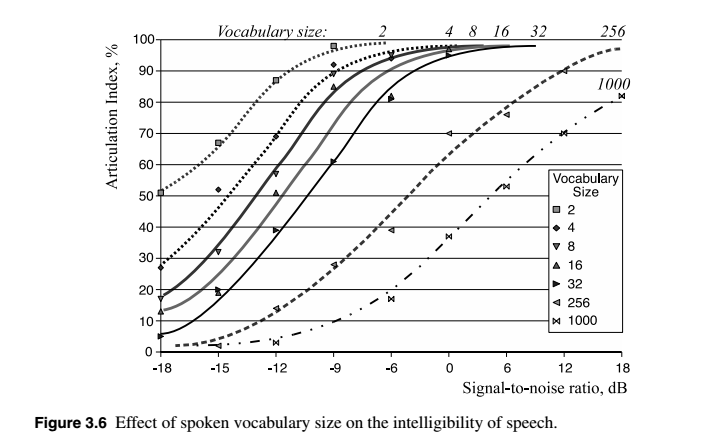
不合适的交流环境里词汇量会大幅下降。

冗余能增加对话的信息量，增大特定词汇丢失的时候依然保有关键信息的概率。

上下文信息的评价主观且复杂，有图如下：



横坐标为信噪比，纵坐标为正确的概率，可见相对来说识别的是个别词汇缺失。



由上图，词汇量对识别率的影响（音节的识别率，非词汇），横坐标为信噪比。可见词汇量越低越好识别。

**总结**

依旧是一些概念性的陈述和介绍，在语音的生理和识别层面。有几个代码，但是暂时没有找到适合用来实验的素材。