《语音学概要》知识点汇编

1、 声波，是空气粒子的振动，是一种纵波，它的传播方向与振动方向一致。

2、 基频 （fundamental frequency）是复合波中最低且通常情况下最强的频率。语音中的基频就是声带振动的频率。男人和女人的语音的基频有很大的区别。基频的变化轨迹决定了声调、韵律、情感等。

3、 计算机内存储的声音采样是麦克风处的瞬时声压值，一般采用相对声压，单位是 dB，可以计算为：

$$L = 20\log_{10}(P/P_0) \qquad (\mathrm{dB})$$

P_0是闻域，是人耳所能感知的最小声压，为 2×10^{-5}Pa。所以闻域的相对声压是 0dB。

4、 元音通常有 3-5 个共振峰，不同元音的共振峰往往有差别。

5、 一个系统（滤波器）可以用它的单位冲激响应信号来刻画。对任意激励,其响应等于单位冲激响应和激励信号的卷积。证明过程见课件。

6、 在离散傅里叶变换（Discrete Fourier Transform，缩写为 DFT）中，首先将任意离散信号都视为一组正交基信号的新型组合，而变换后得到了各基信号的组合系数。正交基信号是复数，组合系数也是复数，其原因请参见文中解释：

https://www.zhihu.com/question/21314374/answer/542909849

语音时频域分析方法知识点

1） 时频分析环节在识别类应用中对应特征提取，在编码合成类应用中对应压缩表示机理

2） 语音是非平稳信号，但可以认为在一个较短的时间内是平稳的，因为在较短的时间内声管形状不会有大的改变。一般用短时窗截取一段信号进行分析，并在信号中以一定步长向前滑动。这一过程也被称为分帧，帧长一般取 10~30ms，滑动步长也被称为帧移，每次滑动前一窗与后一窗信号间重叠部分的长度被称为帧叠。帧叠是帧长的一半是一种常见的分帧设计，可以保证相邻帧间分析结果的连续性。时域分析常用方窗，保证采样值的真实性。频域分析常用 Hamming 窗和 Hanning 窗（窗函数见课件）。DFT 变换需要对信号做周期性延拓，但一帧信号通常不是周期性的，因而频率分析时很容易产生吉布斯效应，导致频谱失真。分帧时加 Hamming 窗或 Hanning 窗可以缓解这一现象。

3） 短时能量和短时过零率是两个常用的时域特征，计算公式见课件。短时能量刻画了声音的强弱，过零率是在时域上对声波的振动情况进行了粗略的刻画，反映的还是频率信息，为了避免直流噪声、背景噪声等影响过零率的计算结果，常采用门限过零率。在语音识别算法中，能量经常是声学特征向量中的一维特征，它对区分当前帧是语音还是背景音，是元音还是辅音有重要的作用。能量的变化轨迹也对识别语音的韵律模式有一定的参考意义。过零率统计特征对区分语音和音乐有帮助。

4） 在现实世界中背景声音持续存在，识别算法需要从连续的声音数据中找到语音（或其它目标声音）的起始点和结束点，这一任务被称为“端点检测”。也常被称为“语音活动检测”(Voice Activity Detection, VAD)。分析可知，元音的能量高于辅音，辅音的能量与背景音的能量没有太大区别。但辅音的过零率明显高于背景音。利用这一特性，可以设计基于短时能量和短时过零率的双门限端点检测方法。实际上算法中用了三个门限，用了一个较高的能量门限确保当前信号一定是语音，而不会是某种能量较大的环境声音。再用一个较低能量门限找到元音的起点和终点。最后通过一个过零率门限找到辅音的起始点和终

点。算法过程见课件。算法中的门限值一般根据当前信号的情况自适应的确定。还有很多其它的端点检测方法。

5） 掌握基频和基音周期的概念，掌握基于短时自相关函数的基音频率检测方法。元（浊）音有周期性，辅（清）音没有周期性。如果一个信号有周期性，其自相关函数一般也有周期性。元音的自相关函数会在基音周期的整数倍处出现峰值，设第一个峰值处为$\hat{k}$，它就对应基音周期，所以语音的基频等于采样频率除以$\hat{k}$。计算语音信号的自相关函数，然后在一定范围内搜索最大值，该最大值的位置就是$\hat{k}$，限定搜索范围是为了避免把倍频误当作基频，比如可以在 0~900 的范围内搜索基频。清音的自相关函数没有明显的周期性，表现为最大值处的自相关函数值$R(\hat{k})$小于等于能量的四分之一，也就是 $R(0)$的四分之一。基于短时自相关函数的基频检测算法见课件。还有很多其它的基频检测算法，如平均幅度差法。

6） 语音内容的可分性主要蕴含在其频域表示中。较长语音段的声学特性的可视化频域分析方法是语谱图。它是较早出现的基于专用硬件设备实现的频域分析方法，可以看做是由每帧频谱拼接而成。它是一种包含三个维度信息的可视化表示，横坐标是时间，纵坐标是频率，通过灰度值或颜色值表示强度（特定频带能量的大小）。窄带语谱图和宽带语谱图的帧长不同，窄带语谱的帧长值较大，从而傅里叶分析后频线较多，带宽较窄。窄带语谱图的频域分辨率高、语谱图上横的线条明显。宽带语谱图的时域分辨率高、可以看见语谱图上纵的线条。

7） 分析人类的发声器官和发声机理，声道对声门波进行了频率调制，从而产生了不同语音内容。研究者提出基于同态解卷来得到声道系统的单位冲激响应的前提是，假定声门脉冲已知，一般假定其为准周期三角波。在之后的研究工作中证明这一假定过于粗糙了。

8） 掌握倒谱域、复倒谱、倒谱的概念，倒谱域是一种伪时域，无实际的物理意义。对声音信号$x(n)$。

复倒谱计算为：

$$\hat{x}(n) = Z^{-1}[\log[Z[x(n)]]]$$

倒谱计算为：

$$c(n) = Z^{-1}[\log|Z[x(n)]|]$$

复倒谱是可逆变换，倒谱不是可逆变换。倒谱具有很好的压缩特性和鲁棒性， 因而在语音声学特征中常被采用。

9） 掌握语音信号产生的系统模型，见课件。将声道简化为若干节均匀界面无损声管的级联，给出了声道滤波器为全极点模型的假设。声门激励对元音和辅音不相同，元音是以基音周期为周期的冲激序列，再经过声门波整形得到的。辅音则是随机噪声。声道滤波器的控制参数就是全极点模型的系数。辐射模型则仿真了口唇的辐射作用。

10） 求取全极点声道滤波器系数的思路：假定其是 AR 模型这种特殊的全极点模型，然后通过求取最佳线性预测器来求 AR 模型系数。掌握求解线性预测系数的方程组的推导方法。在对分析窗进行假定的条

件下，将该方程组简化为 yule-walker 方程组，并给出快速求解算法：Levinson—Durbin 递推算法。掌握其推衍参数线性预测倒谱系数的快速计算方法。详细内容见课件。

11） 掌握人类的听觉感知机理，以及 Mel 尺度下的听觉感知模型，掌握 Mel 频域倒谱系数的快速计算方法。详细内容见课件。

《语音编码》知识点汇编

1、语音编码是语音相关各研究方向中最早获得成功应用的研究方向，它是支撑各种语音通讯方式发展（市话、长途电话、移动通讯、VoIP、即时通讯等）的最关键技术之一。也普遍存在于各种音视频多媒体、流媒体数据中。

2、语音数据不论是磁盘上存储，还是在网络信道上实时传输，都需要考虑数据规模的大小。每秒钟语音需要占据的磁盘空间可以计算为：采样频率◊量化 bit 数◊声道数。要达到实时通讯的效果，语音的传输速率也应该这样计算，单位为比特率，bps 或 bit/s。

3、根据采样定理，采样频率要大于信号最大频率的两倍，否则就会发生混叠失真。所以采样前应该用低通滤波器对信号进行滤波，只保留必要的频率成分。对语音而言，采样频率不应该低于 8KHz，即 4kHz 以内的频率成分如果损失，会影响语音的可懂度。采样频率越大，音质越好，但语音数据的规模也就越大。同样、量化 bit 数越高，量化误差越小，声道数越多，声音越立体，声音质感越好，但数据规模也会越大。为了节省磁盘空间，或降低对传输速率（网络带宽）的要求，需要对语音进行编码压缩。

4、语音编码方法可以粗略地分为“波形编码”、“参数编码” 和“混合编码”，其概念见课件。

5、信噪比的计算公式（见课件），其反映了信号的失真程度，单位为 dB，信噪比越大，失真程度越小。

6、脉冲编码调制（Pulse Code Modulation，PCM）是最简单的编码方式，就是声卡采样量化后的形式，其采样点按顺序存放。根据采样值中大幅值比较少的特性，可以对采样值进行对数变化，然后可以降低量化比特数，进行 8bit 量化，这种编码方式被称为压扩 PCM。要求会计算 PCM 和压扩 PCM 的比特率。

7、自适应差分脉冲编码 ADPCM 是一种应用广泛的波形编码，利用了相邻采样点间的强相关性了。

8、掌握 DPCM 的编码原理，内容见课件。掌握需要在编码器中包含一个解码器的道理。

9 各种波形编码算法的输入一般是 PCM 或者压扩 PCM。由于波形编码以采样点为单位来进行编码压缩，为了达到压缩的效果，一般需要减少压缩后信号的量化 bit 数，这通过量化器来完成，通常会带来量化误差。可以理解为新量化 bit 数没有足够的分辨率来存储压缩后的采样值。

10、掌握 ADPCM 的编码原理，内容见课件。要求能写出 ADPCM 编码的时域计算公式。边信息是以帧为单位的信息，有别于波形编码中其它以采样点为单位的信息。掌握 ADPCM 中的边信息有哪些。

11、掌握 LPC-10 的编码原理，内容见课件。其编码参数有哪些？其中线性预测系数采用了反射系数的形式，二者可以互相转换。反射系数插值特性好。解码端需要生成特定形状的声门波（以基音周期为周期）。

12、混合编码不生成特定形状的声门波，而是用“合成分析”和“闭环搜索”的方法来确定能最好生成语音信号的声门波，遍历所有的候选声门波，逐一合成语音，选择其中误差最小的候选声门波。为了使候选声门波更简洁，更容易编码表示，采用长时相关滤波器去除声门波中周期相关性，得到更稀疏的没有周期性的激励信号，因而语音生成时，也需要多一个长时综合滤波器。并在合成分析中，计算合成语音与真实语音间的

误差时，采用了感觉加权滤波器。其效果是计算误差时降低误差信号中被掩蔽频段的贡献，增强未被掩蔽误差频段的贡献。详细的计算公式见课件。

13 在频谱上 0dB 的等响度曲线也被称为掩蔽曲线，低于此掩蔽曲线的信号无法被人耳听见。语音信号的存在会改变此掩蔽曲线，某个频段信号强，掩蔽曲线也会被相应拉高。从而使之前可被人耳听见的信号，无法再被听见。具体图示见课件。

14 CELP 的候选激励信号从一个码本中挑选，最佳码本（合成分析后感觉加权误差最小）的编号被传输到解码端，从同样的码本中获得该激励，用于语音生成。其过程见课件。

15 解码语音的质量常用 MOS 评分来评估。

16 深度自编码器等深度学习技术可用于实现更高效的语音编码。

《语音识别技术》知识点汇编

1、语音识别任务的分类方法（见课件）。语音识别系统的输入是声学特征向量的序列，连续语音识别系统的输出是语言符号（基元）的序列，它是一个序列到序列的映射任务。

2、用模板匹配方法解决孤立词（命令词）识别的设想是，保存一次或多次命令词语音的实例作为模板，识别时计算待识语音的特征向量序列与所有命令词各模板间的距离，距离最小的模板所对应的命令词就是识别结果，因为距离越小，内容就越相似。即模板是声学特征向量的序列。模板和待识语音即使属于同一个命令词，其各声学基元的持续时间也存在差异，按序对齐各特征向量计算欧式距离再累加的方法缺乏合理性，因为会在不同基元间计算欧式距，累积距还是会比较大。一个想法是将它们对准了以后再计算累积欧式距作为序列间的距离。对准问题可以通过遍历所有的对齐关系解决，对准情况是所有对齐情况中累积欧式距最小的。然而，考虑到计算量，我们采用了一个动态规划算法，DTW 算法，详见课件。

3、由于语音的声学表示（声学特征向量）有较大的动态变化范围，尤其是在面向非特定人的语音识别系统中，即使采用了较多的模板，也无法保证能准确刻画其动态变化范围。因而，采用统计模型是一个合理的选择。语音承载了语言基元的序列，考察语音信号（语音特征向量序列），可以认为存在两种不确定性。第一是在每个语言基元上观察到特征向量具有不确定性，第二语言基元的持续时间，或者说随时间变化的语言基元序列也具有不确定性。我们采用了 HMM 模型来同时刻画者两种不确定性。

4、掌握马尔科夫链的数学定义。（见课件）

5、掌握 HMM 的数学定义。（见课件）

6、为了将 HMM 应用于语音识别系统，需要解决三个基本问题。（见课件）

7、掌握解决 HMM 三个基本问题的思路与算法（见课件）。会推导证明相关公式。

8、掌握如何利用 HMM 分别解决命令词、连接词和连续语音识别问题的技术框架（见课件）。

9、对课件中后向算法进行更正（感谢史云浩同学指出了课件中的错误），初值应为：$\beta_T(i)=1$。解释为 $P(O_{T+1}|q_T=i)=1$。因为 T+1 时刻没有语音特征向量，是空集，所以概率为 1。请大家自行推导$P(O|\lambda)$和 $\beta_1(i)$间的关系