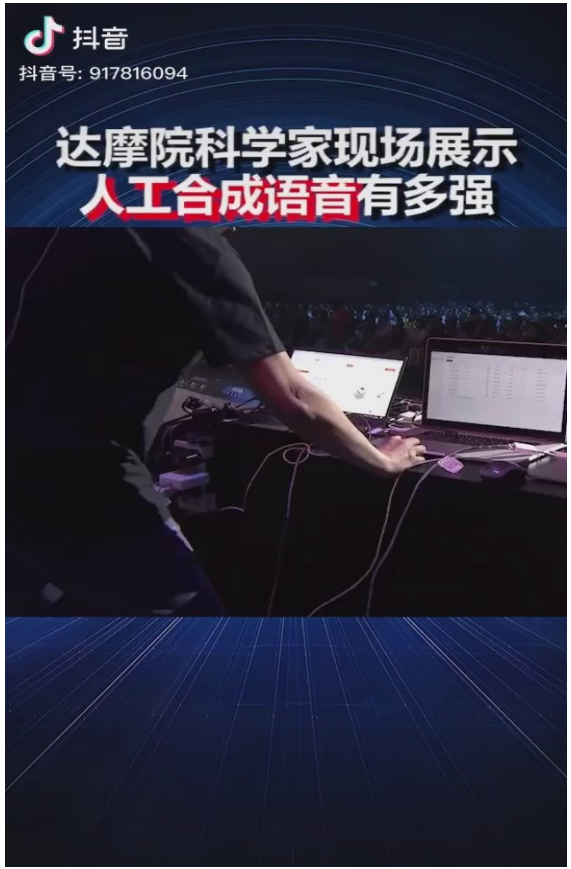
优必选
UBTECH

智能语音信号处理及应用

项目 4
智能小义之学会说话

CONTENT 目 录

课程导入



从视频当中，我们可以体会到语音合成技术的强大和它的应用广泛性。

语音合成的定义及相关技术

（一）语音合成的定义

语音合成，又称文语转换（ Text to Speech ， TTS ）技术，是通过机械的、电子的方法产生人造语音的技术。它是将计算机自己产生的、或外部输入的文字信息转变为可以听得懂的、流利的汉语口语输出的技术，相当于给机器装上了人工嘴巴。

语音合成的定义及相关技术

（一）语音合成的定义

语音合成技术以文本输入为起点，通过文本分析及声学建模进行预处理，转化为发音符号描述，进而通过声码器转化为语音波形，进而合成能听到的语音进行输出。

合成前端

输入文本——>发音符号化描述

合成后端

发音符号化描述——>合成语音

（二）语音合成技术的基本术语

合成单位

合成单元也称为合成单位，是语音合成系统处理的最小的语音基本单位。按由小到大顺序可分为音素、双音素、半音节、音节、词、短语和句子。合成单元越大，合成音质越好，但合成语音的数据量也越大。

合成参数

合成参数为参数合成和规则合成方式中，控制语音合成器以输出所需语音的一组参数，分为音色参数和韵律参数。

（二）语音合成技术的基本术语

合成语音库

合成语音库指在语音合成系统里合成单元编码或合成参数数据的集合

合成语音器

合成语音器是在语音合成系统中将合成参数转变为语音波形的软件和硬件系统。

合成音质

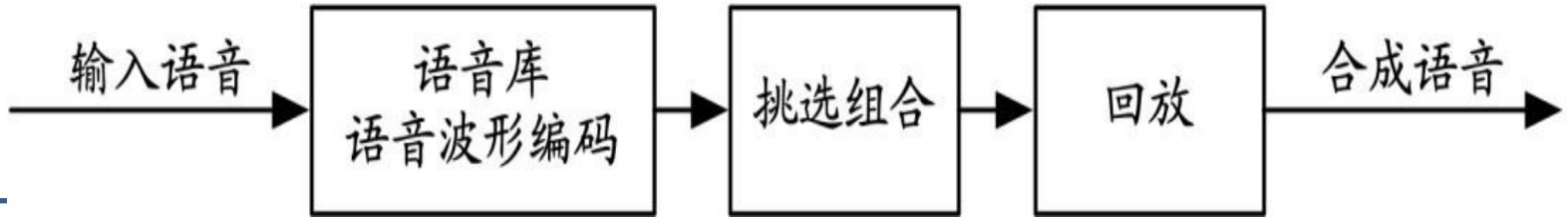
合成音质指语音合成系统输出语音质量，主要用清晰度、自然度和连贯性等指标评价。

（三）语音合成技术的分类

语音合成技术按合成单元的处理方法进行划分时，可以分为波形合成、参数合成和规则合成三类；按合成模型的划分方法进行划分时可以分为基于声学模型、发音模型和自然语音编码模型三类。

1. 基于合成单元的分类

（1）波形合成法



（三）语音合成技术的分类

（2）参数合成法

参数合成法，又称分析合成法，该方法的合成单元由音节、半音节或音素组成。为了节约存储容量，首先要对语音信号进行分析，提取语音参数以压缩存储量。

（3）规则合成法

规则合成法是一种以语音学规则为基础产生语音的高级合成方式，其主要是在系统中存储声学参数和音素组成音节、控制音调等韵律的各种规则，不必进行词汇表的确定。

（三）语音合成技术的分类

三种语音合成方式的比较：

项目		波形合成方式	参数合成方式	按规则合成方式
语音质量	可懂度	高	高	中
	自然度	高	中	低
词汇量		小（500 字以下）	大（数千字）	无限
合成方法		PCM,ADPCM	LPC,LSP, 共振峰	LPC,LSP, 复倒谱
数码率		9.6~64kbit/s	2.4~9.6kbit/s	50~75kbit/s
1 兆比特可合成的语音长度		15 秒 ~100 秒	100 秒 ~7 分	无限
合成基元		音节、词组、句子	音节、词组、句子	音素、双音素、音节
装置		简单	比较复杂	复杂
硬件主体		存储器	存储器和处理器	处理器

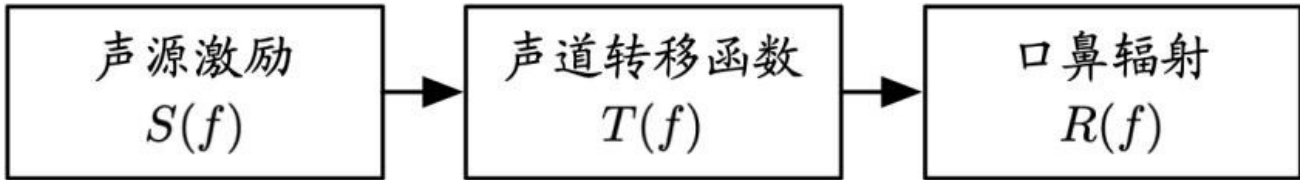
2. 基于合成模型的分类

（1）发音模型合成技术

发音模型合成技术是一种以人发音机理为基础的语音合成方法。

（2）声学模型合成技术

共振峰合成是一种重要的声学模型合成技术，该合成技术以滤波器理论为基础，其合成过程如下图所示：



2. 基于合成模型的分类

（ 3 ）波形编码和拼接

波形拼接的方法通过将语音波形本身信息自然的保留在合成的语音当中来合成质量高的语音。其中波形编码和拼接的主要合成技术为线性预测合成技术和 PSOLA 合成技术（基音同步叠加技术）。

（四）主要语音合成技术

1.LPC（线性预测编码）

LPC 合成技术本质上是一种时间波形的编码技术，目的是为了降低时间域信号的传输速率。其合成过程实质上只是一种简单的解码和拼接过程。

2.PSOLA（基音同步叠加技术）

PSOLA 技术的主要特点是：在拼接语音波形片断之前，首先根据上下文的要求，用 PSOLA 算法对拼接单元的韵律特征进行调整，使合成波形既保持了原始发音的主要音段特征，又能使拼接单元的韵律特征符合上下文的要求，从而获得很高的清晰度和自然度。

3.LMA（对数振幅近似）声道模型

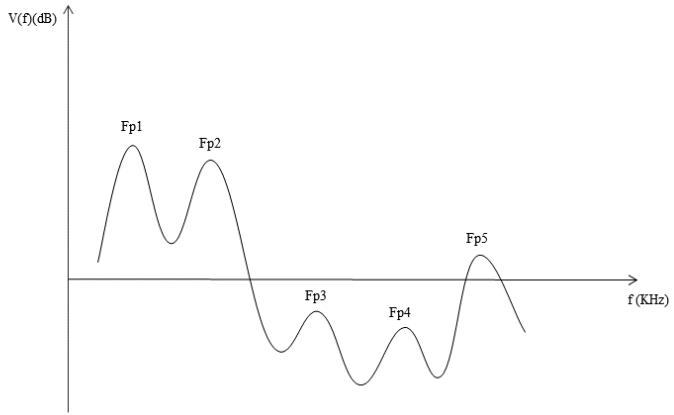
基于 LMA 声道模型的语音合成方法具有传统的参数合成可以灵活调节韵律参数的优点，同时又具有比 PSOLA 算法更高的合成音质。

语音合成技术实现原理

（一）共振峰合成

共振峰是指在声音频谱中能量相对集中的区域，反映声道的物理特征。

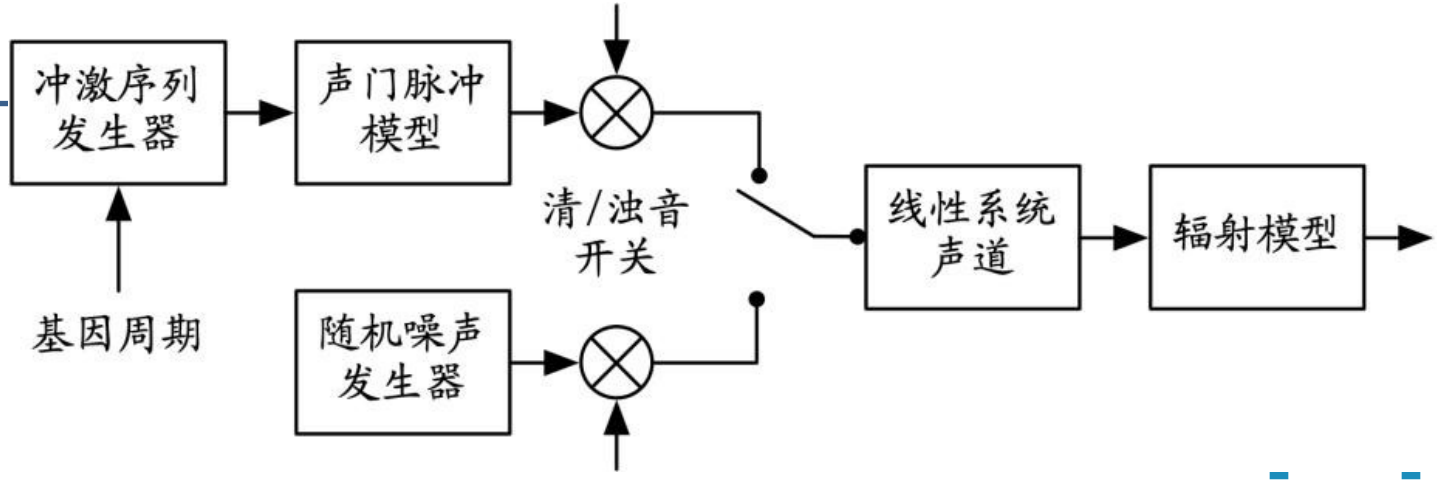
下图声音频率响应图中的 Fp1, Fp2, Fp3, …… 处为频率响应的极点，即为共振峰，其频率分布特性决定语音的音色。



（一）共振峰合成

1. 共振峰合成原理

语音合成时，可以通过多个共振峰滤波器，进行组合模拟声道的传输，然后将激励源处的信号进行调制，最后通过辐射模型输出合成语音。



（一）共振峰合成

2. 实用模型

（1）级联型共振峰模型

声道在级联型共振峰模型被看作是一组串联的二阶谐振器。在一般元音中，当零点可以由多个极点模拟时，可以使用全极点模型进行模拟，所以在大多数情况下声道模型的传输函数是一个全极点函数：

$$V(z)=\frac{G}{1-\sum_{k=1}^{N}a_k z^{-k}}$$

其中， N 是极点个数， G 是幅值因子， ak 是常系数。

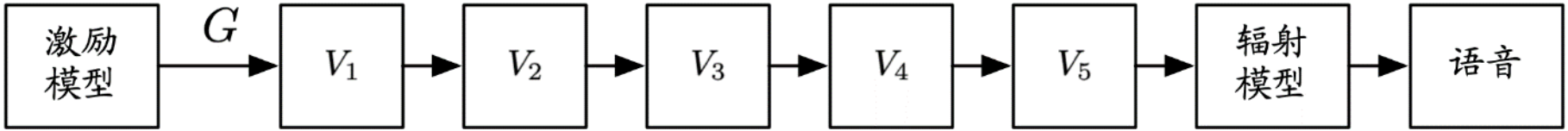
（一）共振峰合成

2. 实用模型

则上式可以分解成多个二阶极点的网络串联，公式如下：

$$V(z)=\prod_{k=1}^{M}\frac{1-2e^{-B_kT}\cos(2\Pi F_kT)+e^{-2B_kT}}{1-2e^{-B_kT}\cos(2\Pi F_kT)z^{-1}+e^{-2B_kT}z^{-2}}$$

式中 M 是小于 (N+1)/2 的整数。假如 N 取 10 ，则 M=5 ，该模型可以表示成下图：



2. 实用模型

（2）并联型共振峰模型

使用并联型共振峰模型进行描述和模拟，声道模型的传输函数可以表示为：

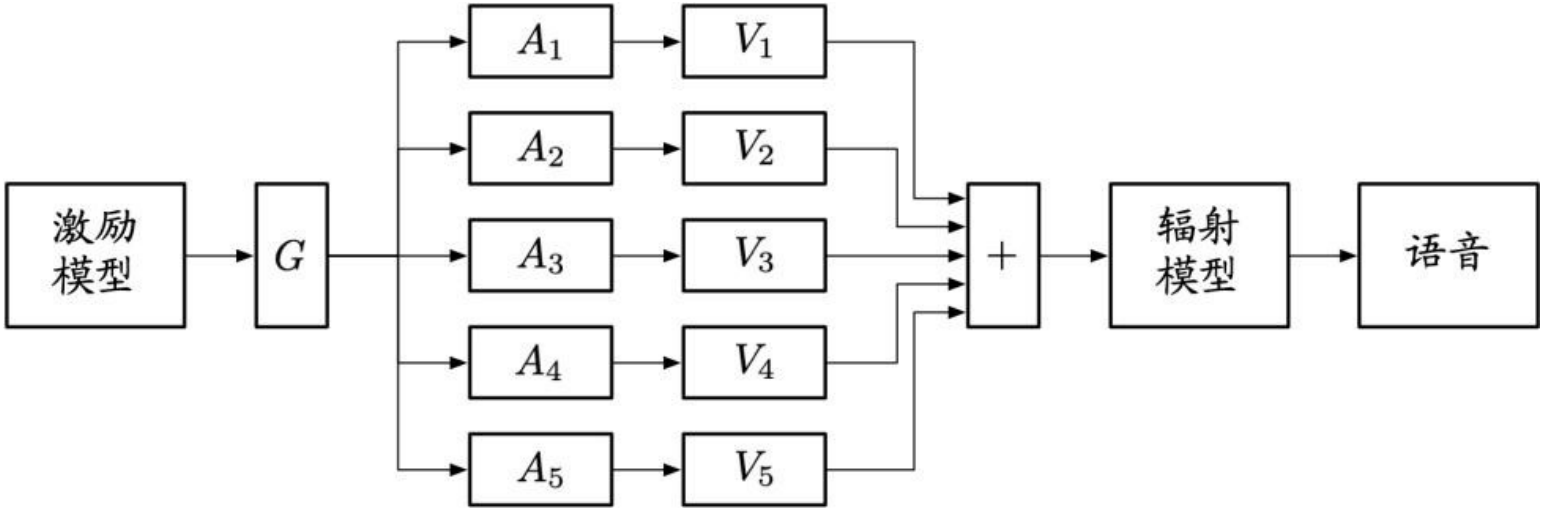
$$V(z)=\frac{\sum_{r=0}^{R} b_r z^{-r}}{1-\sum_{k=1}^{N} a_k z^{-k}}$$

通常，且设分子与分母无公因子及分母无重根，则上式可分解为如下部分分式之和的形式：

$$V(z)=\sum_{i=1}^{M}\frac{A_i}{1-B_i z^{-1}-C_i z^{-2}}$$

（一）共振峰合成

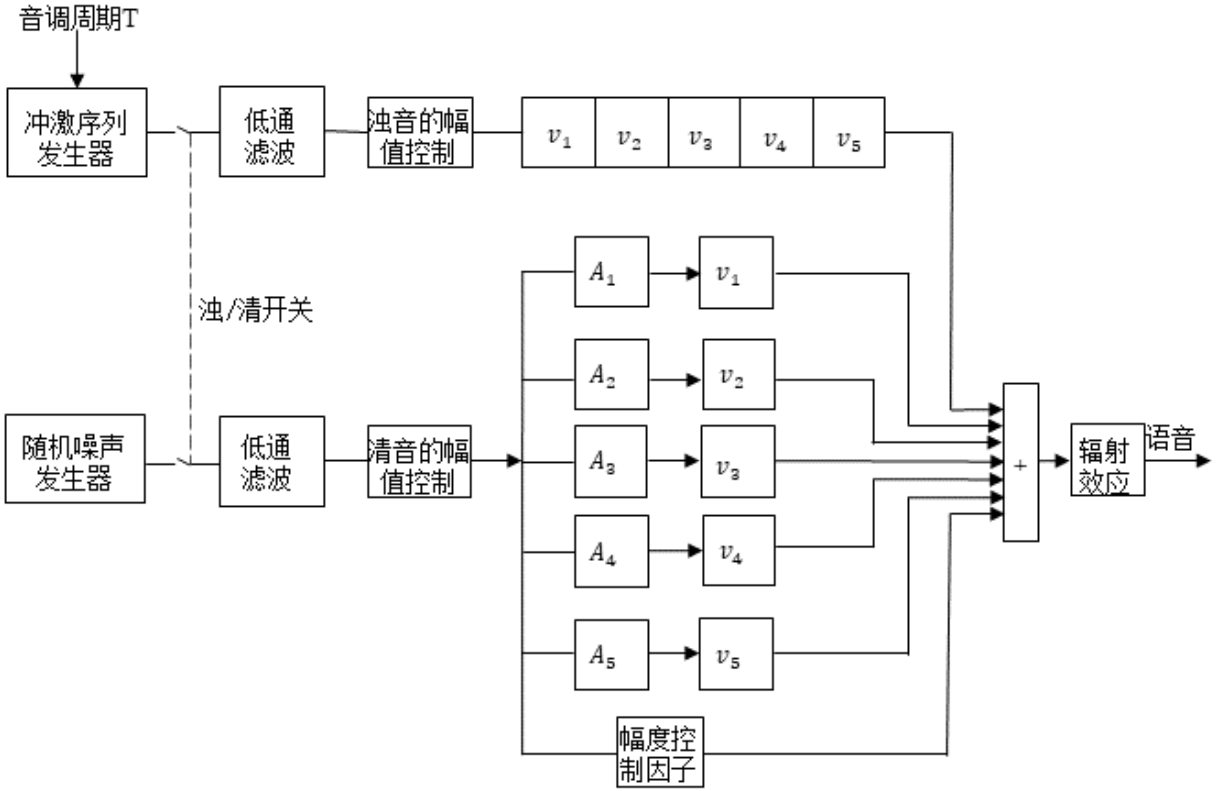
M=5 的并联型共振峰模型如下图所示。



语音合成技术实现原理

（一）共振峰合成

（ 3 ）混合型共振峰模型



3. 共振峰模型的优缺点

优点：声道模拟比较准确，合成语音自然度较高；

缺点：声道模型的合成质量受其精确度影响；共振峰模型不能对语音自然度的其他细微的语音成分进行精准表述，合成语音的自然度不够；共振峰合成器控制方式复杂，较难实现。

（二）线性预测合成

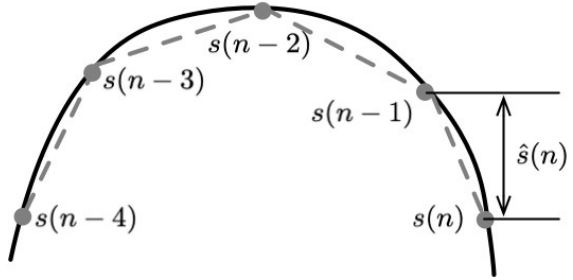
1. 线性预测分析的基本原理

线性预测分析的基本思想是由于语音样本之间的相关性，通过过去的样本数据可以预测现在或未来的样本数据。也就是说，一个语音的取样能够用过去若干个语音取样或它们的线性组合来逼近，在一定的准则下，将实际语音采样或线性预测采样之间的误差最小化，从而确定唯一的预测系数。

（二）线性预测合成

2. 线性预测分析的基本原理

线性预测参数合成的原理如下图所示。



上图中，是实际语音抽样值 s(n) 的线性预测抽样，可以有前 p 个过去的样点值线性表示，即

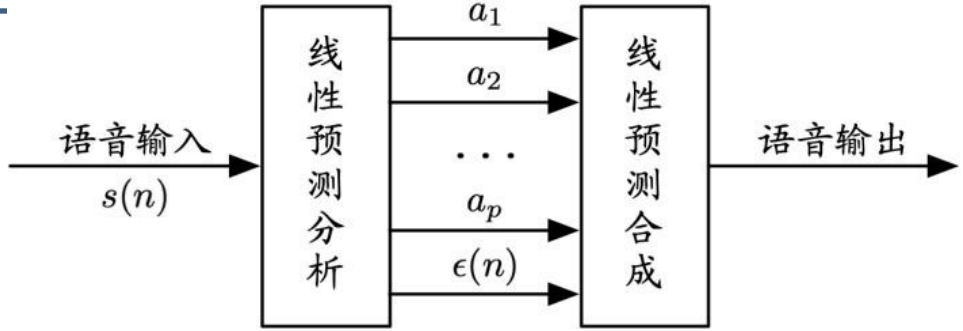
$$\hat{s}(n) = -a_i \sum_{i=1}^{P} s(n-i)$$

（二）线性预测合成

其中， ai 是线性预测系数。线性预测抽样写实际语音抽样之间必定存在预测误差 ε(n)，可以表示为

$$\varepsilon(n) = s(n) - \hat{s}(n) = s(n) + \sum_{i=1}^{p} a_i s(n-i) = \sum_{i=0}^{p} a_i s(n-i)$$

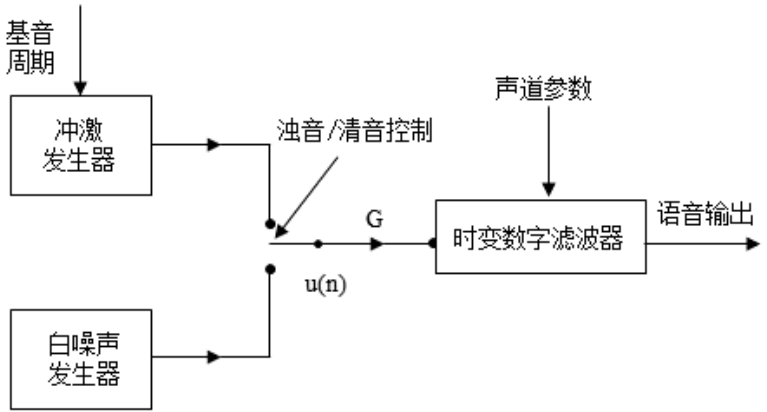
原始语音信号由预测误差和线性预测系数合成，且这个合成系统是一个线性时不变的全极点系统。线性预测分析和合成系统框图如下图所示。



语音合成技术实现原理

（二）线性预测合成

线性预测合成模型是一种由白噪声序列和周期冲击序列构成的激励信号，经过选通、放大并通过由语音参数控制的声道模型 - 时变数字滤波，就可以获得合成的语音信号的一种“源滤波器”模型。这种语音合成器的框图如下图所示。



（二）线性预测合成

线性预测合成的形式有两种：一种是直接用预测器系数构成的递归型合成滤波器，这种结构简单而直观，只需要定期地改变激励参数 u(n) 和预测系数，就能合成出语音。它合成的语音样本由下式决定：

$$s(n)=\sum_{i=1}^{p} a_i s(n-i)+Gu(n)$$

其中， ai 为预测系数； G 为模型增益； u(n) 为激励；合成样本为 s(n) ； p 为预测器阶数。

（二）线性预测合成

另一种合成的形式是采用反射系数构成的格型合成滤波器。它的合成语音样本由下式决定：

$$s(n) = Gu(n) + \sum_{i=1}^{p} k_i b_{i-1}(n-1)$$

其中 G 为模型增益， u(n) 为激励， ki 为反射系数， bi(n) 为后向预测误差， p 为预测器阶数。

（三）按规则合成

1. 韵律规则

韵律规则是合成规则中的一个重要组成。汉语语音节奏性很强，因此单个音节除具有自身韵律特征外，由于语流中音节所处的位置不同，会发生声调、音强、音长变化等协同发音现象，这就要求必须对自然语言中的韵律特征进行统计，得出规则。韵律规则有词调规则、句调规则、音长规则以及音强规则等。

（三）按规则合成

1. 韵律规则

（1）转接与音渡规则

汉语发音中存在“转接”和“音渡”，分别是在发音中声母和韵母的拼接过程中辅音与元音组合以及复合韵母内部元音与元音组合这两种基本情况。

（2）变调规则

一个字发音的声调在连续语流中会与这个字单独发音时的声调有所不同，在合成的连续语流中，只有通过这种声调变化的方式，才能使合成的语音具有较好的可懂度。

（三）按规则合成

1. 韵律规则

（ 3 ）语调规则

汉语的语调是由一系列音节调域组织起来的音高调节形式，是句子结构的一部分，也就是语气部分。换句话说，语调包括了音高、音长、音强以及重音、停顿等，是语言节律的总和。

（ 4 ）音长规则

汉语中音长主要体现在韵母的调型段长度，其中调长和调型段是密切相关的。通常情况下，上升（三声）音节最长，阴平（一声）、阳平（二声）次之，去声（四声）最短。

（三）按规则合成

优必选 UBTECH

1. 韵律规则

（ 5 ）重音规则

平时说话或朗读时读的比较重的音节或词语叫做汉语的重音。

汉语重音可分为词重音和句重音两大类。词重音是指词的某个音节可分为重轻等级，重轻等级可通过音长特征与声调域进行区分。汉语重音的声学特征表现在，音域加宽、音程加大，气流加强。

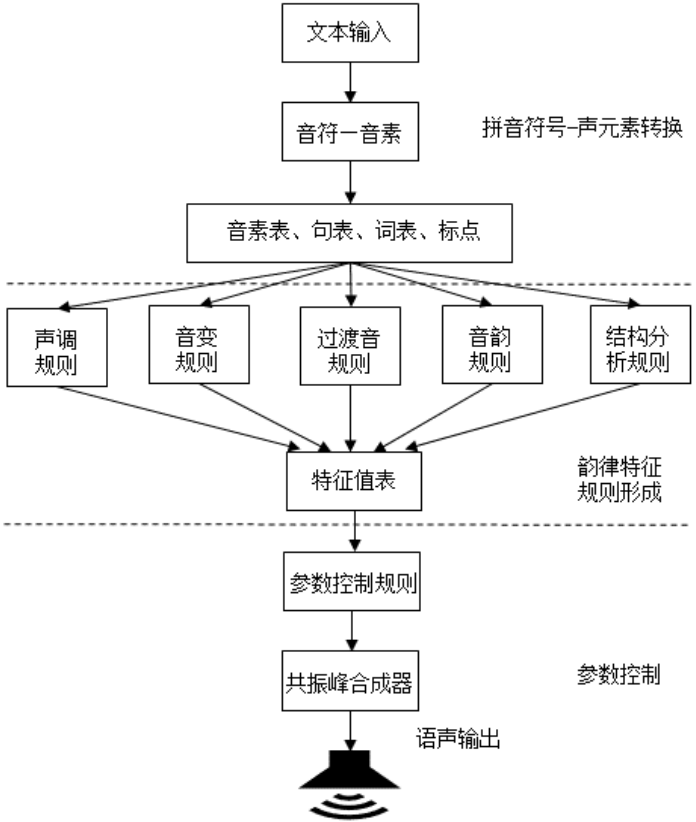
汉语的语句重音是指根据语句特点将一句话的某些部分重读，或是某个词，又或是某句话，取决于句子的内容。

语音合成技术实现原理

（三）按规则合成

2. 汉语按规则合成系统

右图所示是某语声文本材料以汉语拼音方式输入，在控制共振峰合成器产生语声时，使用文本材料规则产生各种控制参数参与控制。



（一）文本分析

文本分析是语音合成系统的前端。根据发音字典对文本进行处理，将输入的文本字符串分解为属性词和拼音。

文本分析的主要功能是计算机根据上下文识别文本，了解文本的范围、发音的类型、发音的方式，并告诉计算机发音方式，使计算机能够识别文本中的单词和短语，了解停顿位置和时间等。

（一）文本分析

文本分析分为三个步骤：

（1）标准化输入文本；

（2）词语划分和语法分析；

（3）根据文本在不同位置的结构、构图和标点，确定语音过程中不同声音的语气变化和轻重。最后，将输入文本转换为可通过计算处理的参数。

（二）韵律处理

自然语音中感情和语气的变化通过音高、音强和响度等的变化表现出来，这些特征称为韵律。韵律参数包括影响这些特征的声学参数，如基频、音长、音强等。语音中的停顿也是韵律的重要成分。

1. 韵律的特点

（1）说话方式

不同的语言表达方式和语言的特点会决定韵律的特点，讲话者在不同环境和不同心情下其说话的韵律也会不同。

（二）韵律处理

1. 韵律的特点

- 个性：指说话人长期稳定的特征；
- 情绪：与前者所述的个性不同的是，情绪与个性无关，个性是稳定的，而情绪是波动的，在进行建模时，需要将情绪考虑在内，只有这样才能将讲话者的信息完全表达出来。

2. 韵律规律

（1）字的声调规则

（2）词的声调规则

① 轻重音规则　②变调规则

3. 韵律生成和处理

韵律的生成离不开文字信息，因此韵律的输入采用文本分析的文字信息，通过文本信息来得知韵律所需发音。同时文本信息中不同的文字在不同的语境下发音不同，读音的轻重也不同，所以韵律的生成需要基于文本规则和大数据驱动来综合考虑。

基于文本规则生成韵律的方法主要应用在早期，但是该方法要求大量的音律数据、音色和声学数据，当规则非常复杂时，处理数据将会变得非常耗时，最终韵律结果的自然度也会大打折扣。

在当前的技术背景下，韵律生成已经可以使用神经网络和统计模型来实现，达到的效果也非常好。用该方法得到的韵律模型灵活性较强，更利于对指定人的韵律特征进行模拟，为多种语言的整合创造了条件。

总结

课程导入：感受日常生活中语音合成的魅力

语音合成的定义：一种通过机械的、电子的方法产生人造语音的技术

语音合成技术的分类：波形合成、参数合成、规则合成；声学模型、发音模型合成技术和自然语音编码模型

主要语音合成技术： LPC 、 PSOLA 、 LMA

语音合成技术实现原理：共振峰合成、线性预测合成、按规则合成

理解文本分析、韵律处理等语音合成相关知识

谢谢大家！