语音合成技术的原理和应用

吴志勇　蔡莲红
清华大学计算机科学与技术系
智能技术与系统国家重点实验室
wuzy@tts.cs.tsinghua.edu.cn

一、语音合成在 TTS 系统中的地位

当今，“语音合成”的含义已远远超出“综合”的内涵，它往往是“语音合成”这一研究领域的统称。就合成器而言，它经历了机械式、电子式和数字式的发展历程。当前，语音合成的研究已经进入文字-语音转换（TTS）阶段。按照功能模块可以分为三大部分：文本分析、韵律建模和语音合成。语音合成是 TTS 系统中最基本而重要的模块。

对于 TTS 系统而言，其主要目的和功能就是要从有限的原始语音库单元出发，合成出具有无限词汇的连续语句。而汉语是一种韵律特征非常复杂的语言，对于同一个音节，出现在不同的语境下时，往往会表现出不同的韵律特性。因此必须在一定的韵律特性的要求下对原始语音基元进行韵律参数的控制和调整，以期得到符合当前语境情况的音变单元。而这些就是汉语语音合成模块所要完成的功能。

概括起来来说，语音合成的主要功能是：根据韵律建模的结果，从原始语音库中取出相应的语音基元，利用特定的语音合成技术对语音基元进行韵律特性的调整和修改，最终合成出符合要求的语音。

二、几种语音合成技术

语音合成技术经历了一个逐步发展的过程，从参数合成到拼接合成再到两者的逐步结合，其不断发展主要是人们认知水平以及要求的不断提高的结果。目前，常用的语音合成技术主要有：共振峰合成技术、LPC 合成技术、PSOLA 拼接合成和 LMA 声道模型技术。各种合成技术各有自己的优缺点，人们在应用的过程中往往将多种技术有机的结合在一起，或者将一种技术的优点运用到另一种技术上，以克服另一种技术的不足。这些方面，都已经有了不少成功的应用范例。

1、共振峰合成

语音合成的理论基础是语音生成的数学模型.该模型语音生成的过程是在激励信号的激励下，声波经谐振腔（声道），由嘴或鼻辐射声波。因此，声道参数、声道谐振特性一直是研究的重点。如图 1 所示的某一语音的频率响应图中，标有 Fp1、Fp2、Fp3……处为频率响应的极点，此时声道的传输频率响应有极大值。习惯上，把声道传输频率响应上的极点称之为共振峰，按照频率从小到大，依次记为 F1、F2、F3……。而语音的共振峰频率（极点频率）的分布特性决定着该语音的音色。

音色各异的语音具有不同的共振峰模式，因此，以每个共振峰频率及其带宽作为参数，可以构成共振峰滤波器。再用若干个这种滤波器的组合来模拟声道的传输特性（频率响应），对激励源发出的信号进行调制，再经过辐射模型就可以得到合成语音。这就是共振峰合成技

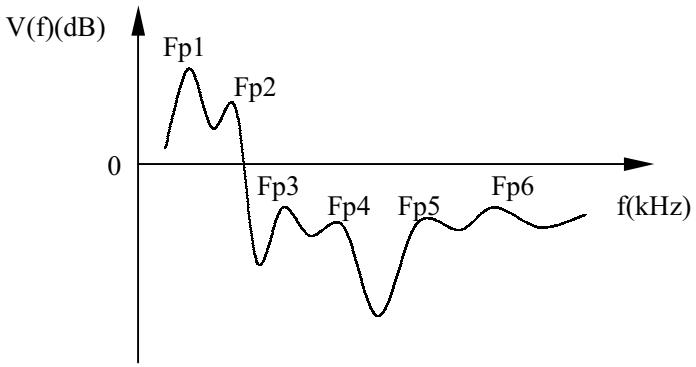

图 1、声道频域特性 （频率响应图）

(Fp1、Fp2、Fp3…对应于共振峰 F1、F2、F3…)

术的基本原理。

基于共振峰的理论出现了三种实用模型：级联型，并联型和混合型。这三种模型在实际系统中都得到了成功的应用。

1、 级联型共振峰模型

在级联型共振峰模型中，声道被认为是一组串联的二阶谐振器。V1、V2、V3、V4、V5 对应于具有不同共振峰频率及带宽的共振峰滤波器。级联型共振峰模型主要用于绝大部分元音的合成。

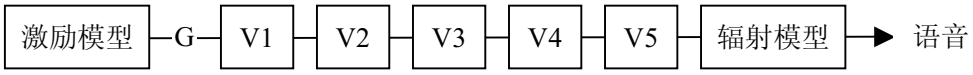

图 2、级联型共振峰模型

2、 并联型共振峰模型

许多研究者认为对于鼻化元音等非一般元音，以及大部分辅音，上述级联型模型不能很好的加以描述和模拟，因此构筑和产生了并联型共振峰模型。

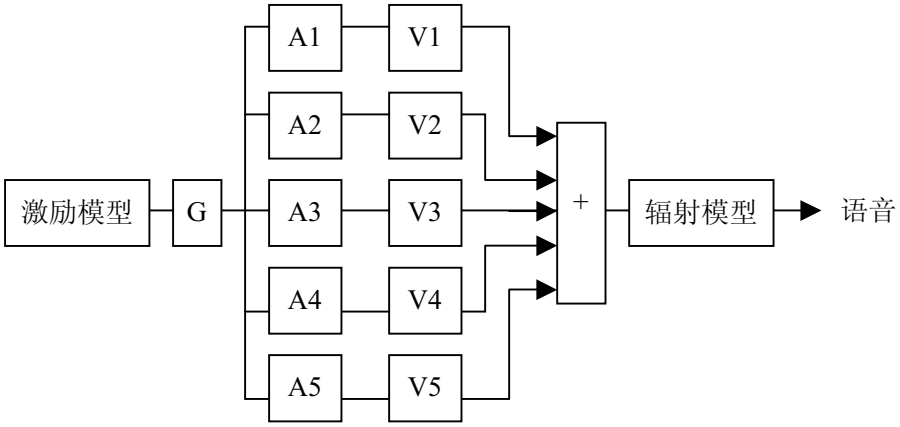

图 3、并联型共振峰模型

3、 混合型共振峰模型

级联型共振峰合成模型中，共振峰滤波器首尾相接；而在并联型模型中，输入信号首先分别通过幅度调节，再加到每一个共振峰滤波器上，各路的输出再叠加起来。两者比较对于

合成声源位于声道末端的语音（大多数的元音），级联型合乎语音产生的声学理论，并且无需为每一个滤波器分设幅度调节；而对于合成声源位于声道中间的语音（大多数清擦音和塞音），并联型则比较合适，但是其幅度调节很复杂。基于此种考虑，人们将两者结合在一起，提出了混和型共振峰模型，如图 4 所示。

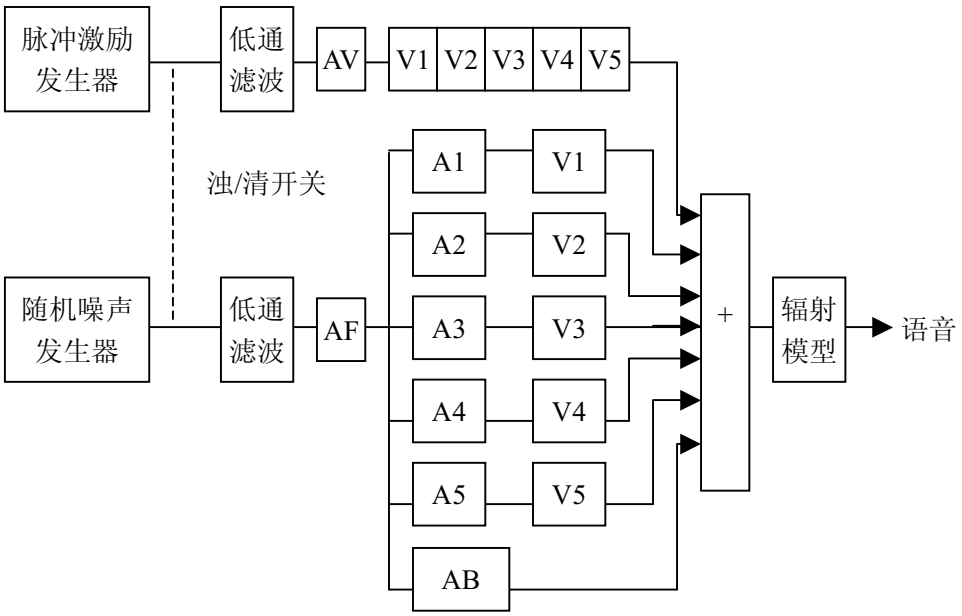

图 4、混合型共振峰模型

事实上，上述三种共振峰模型在实际中都得到了成功的应用。例如：Fant 的 OVE 系统就采用了级联型的共振峰模型[11]，Holmes 合成器采用的是并联型的共振峰模型[12]，而最为典型也是最为成功的 Klatt 合成器则构筑在混合型共振峰模型的基础之上[13]。

而在汉语语音合成方面，基于共振峰的模型也研制出了一些成功的应用系统。比如：社科院语言所 SIFS 合成器、中科院声学所研制的 KX-1 系统中基于 Holmes 的并联型共振峰合成器模型的[8]，而同样由中科院声学所开发的第二代共振峰合成器 KX-FSS 则是基于 Klatt 合成器的[9]。

共振峰模型是基于对声道的一种比较准确的模拟，因而可以合成出自然度比较高的语音，另外由于共振峰参数有着明确的物理意义，直接对应于声道参数，因此可以容易利用共振峰描述自然语流中的各种现象，并且总结声学规则，最终用于共振峰合成系统。

但是，人们同时也发现该技术有明显的弱点。首先由于它是建立在对声道的模拟上，因此对于声道模型的不精确势必会影响其合成质量。另外实际表明，共振峰模型虽然描述了语音中最基本最主要的部分，但是并不能表征影响语音自然度的其他许多细微的语音成分，从而影响了合成语音的自然度。另外共振峰合成器控制十分复杂，对于一个好的合成器来说，其控制参数往往达到几十个，实现起来十分困难。

基于这些原因，人们继续寻求和发现其他新的合成技术。从波形的直接录制和播放得到启发，人们提出了基于波形拼接的合成技术。其中 LPC 合成技术和 PSOLA 合成技术可以说是其代表。

与共振峰合成技术不同，波形拼接合成基于对录制的合成基元的波形进行拼接，而不是基于对发声过程的模拟。

2、LPC 参数合成

波形拼接技术的发展是和语音的编码、解码技术的发展密不可分的，其中 LPC 技术（线性预测编码技术）的发展对波形拼接技术产生了巨大的影响。

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。

对于利用 LPC 合成技术来进行汉语语音合成和汉语文语转换的研究，中科院声学所在这方面进行了大量的工作。1987 年崔成林、李昌立、莫福源等人引进了多脉冲激励 LPC 技术，1989 年莫福源等引入矢量量化，1993 年倪宏、李昌立等引入码激励技术，他们的这些工作对于 LPC 合成技术在汉语合成方面的运用作出了重要的贡献。

LPC 合成技术的优点是简单直观。其合成过程实质上只是一种简单的解码和拼接的过程。另外，由于波形拼接技术的合成基元是语音的波形数据，保存了语音的全部信息，因而对于单个合成基元来说能够获得很高的自然度。

但是由于自然语流中的语音和孤立状况下的语音有着极大的区别，如果只是简单的把各个孤立的语音生硬的拼接在一起，其整个语流的质量势必是不太理想的。而 LPC 技术从本质上来说只是一种录音＋重放，对于合成整个连续语流 LPC 合成技术的效果是不理想的。因此，LPC 合成技术必须和其他技术结合才能够明显改善 LPC 合成的质量。

一种典型的基于 LPC 合成技术的文语转换系统原理图[10]，如图 5 所示。

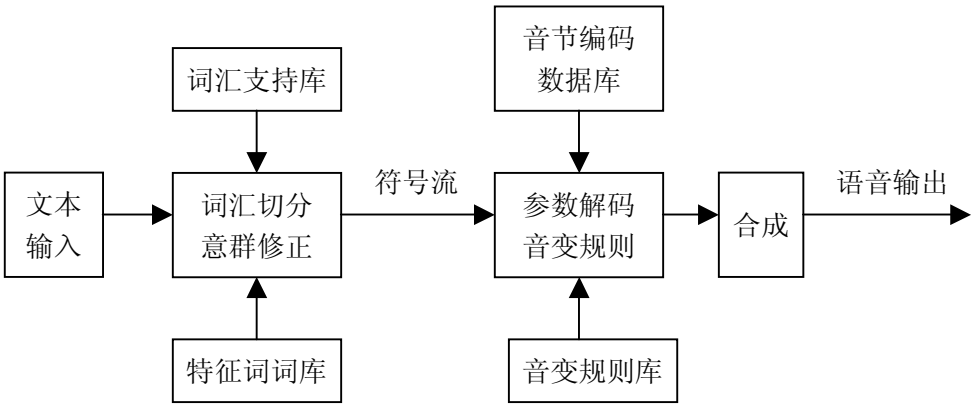

图 5、基于单音节和 VQLPC 技术的文语转换系统原理图[10]

(VQLPC：矢量量化的 LPC 技术)

3、PSOLA 合成技术

于 80 年代末提出的 PSOLA 合成技术（基音同步叠加技术）给波形拼接合成技术注入了新的活力。PSOLA 技术着眼于对语音信号超时段特征的控制，比如：基频、时长、音强等的控制。而这些参数对于语音的韵律控制以及修改是至关重要的，因此 PSOLA 技术比 LPC 技术具有可修改性更强的优点，可以合成高自然度的语音。

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。使得合成波形既保持了原始发音的主要音段特征，而且能够使得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。

如何将 PSOLA 技术应用于汉语文语转换系统，国内许多学校和科研单位进行了大量广泛深入的研究。清华大学、北方科大、中科院声学所等在对 PSOLA 技术研究的基础上，先后开发出了基于波形拼接的汉语文语转换系统，并且对于如何进一步完善该技术、如何进一步改善合成语音的自然度等都提出了一些措施。

PSOLA 技术保持了传统波形拼接技术的优点，简单直观，运算量小。而且它还能够方便的控制语音信号的韵律参数，具有合成自然连续语流的条件，受到了广泛的应用。

但是，PSOLA 技术也有其缺点。首先，PSOLA 技术是一种基音同步的语音分析/合成技术，首先需要准确的基因周期以及其起始点的判定。基音周期或其起始点的判定误差将会影响 PSOLA 技术的效果。其次，PSOLA 技术是一种简单的波形映射拼接合成，这种拼接是否能够保持平稳过渡以及它对频域参数有什么影响等并没有得到解决，因此在合成时会产生不理想的结果。

4、LMA 声道模型

随着人们对语音合成的自然度和音质的要求越来越高，PSOLA 算法表现出对韵律参数调整能力较弱和难以处理协同发音的缺陷，人们又提出了一种基于 LMA 声道模型的语音合成方法。这种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。

三、语音合成技术展望

目前主要的语音合成技术是共振峰合成技术和基于 PSOLA 算法的波形拼接合成技术。这两种技术各有所长，共振峰技术比较成熟，有大量的研究成果可以利用，而 PSOLA 技术则是比较新的技术，具有良好的发展前景。

以前两种技术基本上是互相独立发展的，现在许多学者开始研究两者之间的关系，试图将两者有效的结合起来，从而合成出更加自然的语流。例如清华大学刘灏、蔡莲红等进行了将共振峰修改技术应用于 PSOLA 算法[1]的研究，并用于 SONIC 系统的改进，研制出了具有更高自然度的汉语文语转换系统。

随着人们研究的不断深入和知识的不断积累，语音合成技术一定会取得新的进展，而让计算机像人类一样说出自然流畅的语言，一定会成为现实。

参考文献：

1. 刘灏（1997），语音信号的共振峰修改及其在汉语文语转换系统中的应用，硕士论文
2. 蔡莲红、魏华武（1994），汉语文语转换系统的研究与实现，应用声学，Vol.13，No.6，1994
3. 杨行峻等，语音信号数字处理，电子工业出版社，1995 年 8 月
4. 崔成林、李昌立、莫福源（1987），9600 比特/秒的多脉冲激励的线性预测语音编码的研究与实现，第三届语音通信与图象处理论文集，PP.46~50
5. 莫福源、刘清波、李昌立（1989），矢量量化合成器，第三届语音图象通讯信号处理论文集，PP.88~91
6. 倪宏等（1992），汉语合成规则的混合激励源研究，第二届全国人机语音通讯学术会议论

文集，PP.277~280
7. 刘庆峰、王仁华（1998），基于 LMA 声道模型的语声合成新方法，《声学学报》，1998 年第 3 期
8. 张家录，吕士楠等（1989），汉语文语转换的研究，信号处理，1989 年第 1 期
9. 吕士楠，周同春，谢咏圭（1992），用 KLATT 合成器合成汉语的初步研究，第二届全国人机语音通讯学术会议论文集，PP.281~286
10. 李彤，莫福源，李昌立（1994），基于单音节的汉语文语转换系统及其应用，第五届全国语音图象通讯信号处理学术会议论文集，PP.415~420
11. Fant,G.(1972) et. al., Speech Analysis, Synthesis and Perception, Springer, New York
12. Holmes,J.N.(1983), Research report: fromant synthesizers: Cascade or Parallel? Speech Communication, Vol.2, No.4, pp.251~273
13. Klatt,D.H.(1980), Software for a cascase/parallel formant synthesizer, JASA, Vol.67, No.3, P.971~995