第15卷 第10期 小型微型计算机系统 Vol. 15, No. 10
1994年10月 MINI-MICRO SYSTEMS Oct., 1994

# 波形编辑语音合成技术及在汉语TTS中的应用

蔡 莲 红
(清华大学，北京 100084)

**摘 要** 本文介绍了近几年迅速发展的基于波形编辑的语音合成技术。文中简介了它的研究内容、基本算法，还介绍了用波形编辑方法实现的汉语文一语转换系统。

**关键词** 语音合成，波形编辑，汉语文一语转换，PSOLA

## 1 前 言

言语是人类最习惯最方便的通信方式。随着人工智能和计算机技术的发展，人们期待着以语音方式进行人机交流。人机语音交互已成为人机交互(Human Computer Interactive)中的重要课题。语音是计算机、多媒体、通信中不可缺少的组成部分。语音合成、文一语转换、音频编解码的研究受到各国的重视。人们基于语音生成模型，研究合成语音的算法，如LPC、共振峰参数合成。它们的参数规范化、存储量少、理论基础坚实，但有限个参数很难适应语音的细微变化，语音质量也有待于提高。近几年国外对时域语音合成技术给予了极大的关注。日、法、德都有相应的文一语转换系统问市。基于波形编辑的语音合成用直观的波形替代参数，对波形进行灵活多变的修改，以适应所期望的韵律。目前，基音同步叠加技术(PSOLA)日趋成熟，使波形编辑合成法大受欢迎。本文重点介绍波形编辑合成技术及在汉语文一语转换(TTS)中的应用。

## 2 波形编辑语音合成技术

80年代末E. Moulines和F. Charpentier提出基于时域波形修改的语音合成技术[4]，在PSOLA(Pitch Synchronous Overlap Add)方法的推动下，此技术得到很大发展和广泛应用。

波形编辑语音合成技术是直接把语音波形数据库中的波形级联起来，输出连续语流。这种语音合成技术用原始语音波形替代参数，而且这些语音波形取自自然语音的词或句子，它隐含了声调、重音、发音速度的影响，合成的语音清晰自然。其质量普遍高于参数合成。

这种语音波形编辑技术多用于文一语转换系统中。如日本NTT基于波形文件实现日语规则合成系统。日本ATR的γ-TALK语音合成系统，使用了大小不规则的语音单元，采用

收稿日期：1994-08-10。本课题受国家自然科学重点基金资助。蔡莲红，副教授，主要从事计算机语音识别和合成、多媒体技术的研究。

单元集自动生成和快速构造算法，自动音调控制规则。法国CNET以双音素作语音基元，用基于HMM的语音匹配法进行特性标注，实现了法语文－语转换系统。德国波恩大学的语音合成系统接收有重音标注的音素串，以半音节类似的时域音元拼接，输出语声流。

波形编辑语音合成主要研究的内容有：

(1) 语音波形数据库中语音基元的选取。可供选择的基元有音素、双音素、多音素、词汇或句子。语音基元的大小与算法的复杂度和变化的灵活度成反比，与数据库的大小成正比。时域语音单元越小，拼接过程可能越复杂，但修改灵活度也就越大。

(2) 语音波形拼接过程中的平滑滤波。波形或频谱的不连续都会产生或大或小的噪声，而协同调音现象又使过渡段成为不可避免的问题。因此拼接过程中的平滑是必不可少的。

(3) 韵律修改。韵律在时域波形上的主要体现是时长、音高、音强及语音波形的形状之别。它们反映了语音在基频、共振峰、能量及谱分布特性上的差异。波形编辑语音合成能对韵律进行灵活方便的修改，使语气、语调、重音达到我们所要求的效果。

PSOLA(Pitch Synchronous Overlap Add)是一种韵律修改算法。它以基音周期(而不是传统的定长的帧)为单位进行波形的修改。算法直接作用于语音波形的数据，实现语音基元的拼接、韵律的修改。

(4) 语音基元的自动分割。其中包括语音基元的挑选、标记、剪切，以及语音数据库的设计和建立。

(5) 语言学分析和处理。波形编辑语音合成多用于文－语转换系统中，因此应包括语言学分析、语音规则处理、声学参数标注。

当前，越来越多的人研究波形编辑语音合成技术，并设计了相应的算法和系统。如法国CNET已实现了多语种文－语转换系统。他们设计了一种新的体系结构，保证了实时、多通道和交互功能。该系统中只采用一个合成器，就可实现多语种文－语转换，只须提供高层次语言处理模块。该系统已在电话网中，用于公共电话服务。又如日本NTT提出的共振峰修改算法。这个算法以频率、带宽和谱密度为参数，灵活地变化共振峰频率，克服由语音库引起的语音多样性的限制，合成出不同音质的声音(男、女、小孩或沙哑的声音)。

## 3 PSOLA算法简介

PSOLA就是基音同步叠加，其算法核心是基音同步。它把基音周期的完整性作为保证波形及频谱的平滑连续的基本前提。因此，为原始语音段加基音标记是算法执行的基础。浊音有基音周期，标记有效。对于清音，为保持算法的一致性，设标记为一适当的常数。

PSOLA算法按以下三个步骤实施：对原始波形进行分析，产生非参数的中间表示；对中间表示进行修改；将修改过的中间表示重新合成为语音信号。由于修改所针对的侧面不同，已提出TD(Time Domain)－PSOLA、FD(Frequency Domain)－PSOLA、LP(Linear Predictive)－PSOLA几种不同方法。下面先简介PSOLA的基本处理，再介绍几种方法的区别。

基音同步分析：数字语音波形$x(n)$的中间表示是由短时信号$x_m(n)$组成。$x_m(n)$是由基音同步分析窗$h_m(n)$对$x(n)$加权而得到的：

$$x_m(n)=h_m(t_m-n)x(n) \tag{1}$$

其中 $t_m$ 称基音标记。$h_m(n)$通常为 Hanning 窗，总是长于一个基音周期。因此相邻短时信号有重叠。一般窗长取本地基音周期 P 的某个倍数 $\mu$，则窗长的比例规则可表示成：

$$h_m(n)=h(n/\mu p) \tag{2}$$

h(n)为标准化窗，$\mu=2\sim4$ 比例因子

基音同步修改：把分析短时数据流 $x_m(n)$转换为修改过的合成短时数据流 $\tilde{x}_q(n)$，$t_m$ 也相应地改为$\tilde{t}_q$。这种转换包括三个基本操作：

- 修改短时信号的数量。
- 修改短时信号间的延时。
- 修改每个短时信号的波形。

合成语音的基音标记数量$\tilde{t}_q(n)$取决于基音和时间的修改因子 $\beta$ 和 $\gamma$。两个连续基音周期之间的延时必须等于合成语音的基音周期。由算法解决$\tilde{t}_q$到 $t_m$ 之间的映射。某个特定的分析短时信号 $x_m(n)$被选作产生指定的合成短时信号 $\tilde{x}_q(n)$。在 TD－PSOLA 方法中，选择某些分析短时信号，按延时序列 $\delta_q=\tilde{t}_q-tm$ 转换成 $\tilde{x}_q(n)$：

$$\tilde{x}_q(n)=x_m(n-\delta_q)=x_m(n+t_m-\tilde{t}_q) \tag{3}$$

基音同步叠加法合成：有多种叠加合成法可供选择。例如，合成信号 $\tilde{x}(n)$可由最小平方叠加合成中得到：

$$\tilde{x}(n)=\frac{\sum_q \alpha_q \tilde{x}_q(n)\tilde{h}_q(\tilde{t}_q-n)}{\sum_q \tilde{h}_q^2(\tilde{t}_q-n)} \tag{4}$$

其中 $\tilde{h}_q(n)$表示合成窗序列，$\alpha_q$ 为附加归一化因子，是为了补偿基音修改时能量的损失而增设的。上式可以简化为：

$$\tilde{x}(n)=\frac{\sum_q \alpha_q \tilde{x}_q(n)}{\sum_q \tilde{h}_q(\tilde{t}_q-n)} \tag{5}$$

式(5)的分母是一个时变归一化因子：补偿相邻窗口叠加部分的能量损失。该因子，在窄带条件下是常数；在宽带条件下，若合成窗长为合成基音周期的两倍时，该因子仍为常数。此时，若设 $\alpha_q=1$，则有：

$$\tilde{x}(n)=\sum_q \tilde{x}_q(n) \tag{6}$$

时间标尺的修改：时间标尺的修改可以与基音标尺修改同时进行，也可以独立地变换。

在后一种情况下，不需要频域的运算，也与TD－PSOLA的分析窗大小无关。

最简单的情况是时间标尺的修改因子γ为常数。此时，$\tilde{t}_q \rightarrow t_m$ 基音标记的映射简化为寻找最接近 $\gamma \tilde{t}_q$ 的 $t_m$。当需要减慢语速时，基音标记的映射为几个短时分析信号的重复；反之，愈增加语速，则删去短时分析信号中的某些波形段。如图1所示。

音高标尺的修改：音高标尺的修改总是与时间标尺修改相关交叉的，相对复杂一些。最简单的情况是音高、时间的修改因子相同：β=γ。这样，合成基音标记和分析标记是一一对应的关系，如图2中原始时间轴和第二条时间轴的关系。但是一般情况下，时间和音高是不相关的，这就需要对短时分析信号进行复制或删除，如图2所示。它可看成两个转换过程的结合：其一，用相同的因子β修改音高标尺和时间标尺；其二，用因子γ/β对时间量进行补偿。而实际中，这两步映射结合为一个映射，时间标尺和音高标尺的修改在一步内同时完成。

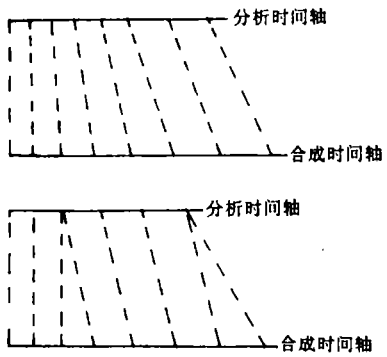


图1 使用TD－PSOLA算法进行时间标尺修改

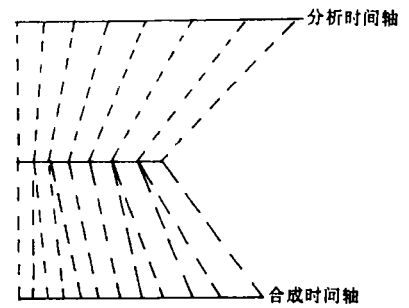


图2 使用PSOLA算法进行音高标尺修改

TD－PSOLA算法只作用于时域波形，算法简单，但有时会导致频谱的不连续。LP－PSOLA、FD－PSOLA是与LPC编码技术相结合的产物。LP－PSOLA算法的本质是在LPC滤波器应用之前使用TD－PSOLA算法，可认为是TD－PSOLA剪贴法的扩展。FD－PSOLA算法是在合成之前对每个短时浊音信号进行频谱分析，并在频域上进行修改。该算法可解决在窄带TD－PSOLA处理中因共振峰带宽的加大而引入的回音噪声。

## 4 PSOLA在汉语TTS中的应用

一般来讲，基于波形编辑方法合成语音的质量优于参数合成。近几年，国内也开展了此技术的研究，并将其应用于汉语文－语转换系统中，如声学所、清华大学和西安电子科技大学等。

我们经过多年研究，基于波形编辑算法，在微机平台上实现了汉语文－语转换系统[1,3] TH－Speech。它综合了语言学、语音学和语音信号处理技术。它可将计算机内的文本，经断句、分词、音变调变处理后，转换成连续自然的语声流。文－语转换是语音合成技术的延伸，它涉及到多个相关学科。本文重点介绍TH－Speech系统中语音合成基元的选取、重音韵律参数模型的抽取及韵律的修改。

### 4.1 合成基元的选取

汉语是声调语言。汉语有调音节是有意义的最小语言单位。其结构的基本类型分为V、

CV、VC和CVC(C表示辅音，V表示元音)。汉语有丰富的复元音，V可扩展成VV、VVV。汉语没有复辅音，且CVC的第二个辅音只能是鼻音(N)和塞音，为此可以把汉语音节结构归纳成以下框架：

$$(C)+(V)V(V)+(N,P)$$

括号表示结构中可有可无的成分，N和P同处一个括号内表示二者互相排斥。

在语音分析中，把音节开头的辅音称为“声母”，声母后的部分统称为“韵母”。普通话有22个声母、37个韵母。按照声母和韵母拼接的规律性，组成了410多个音节。再加上声调共有1200多个有调音节。

我们又研究了语音的音联现象。在音节内部，声母和韵母在时间上连接紧密，在特性上相互影响。而在相邻音节、词汇之间的影响就逐级变小。音素是发出各不相同音的最小单位，选音素作为合成基元，所需的存储容量小，但难于表达语音复杂多变的细微韵律特性。而汉语的音节特征明显，音节音联远小于音素音联，故TH－Speech选音节作合成基元。

汉语音节的声学特性有如下特点：

(1) 声调有辨义作用。且当声调不同时，不但基频频率、模式、调域不同，音色也不同。

(2) 轻声也有辨义作用，通常把它作为0声调。

(3) 除er自成音节外，全部音节都可能儿化。儿化卷舌的作用大多从韵腹开始，直到韵尾。

(4) 受语言、语义的影响，有时孤立音节读音与在连续语流中该音节的读音不同。

(5) 音节音联主要表现在词内部，如词内前音节的韵尾受后音节影响，发生截尾和韵律改变；词内后音节的声母受前音节影响，声母特性改变。

综上所述，TH－Speech中以音节为单位建立了语音数据库，其结构如下：

$$A=\{a_{ij}\}$$

i＝0～417，表示汉语拼音在音库中的序号。j＝0－9，是声调特征值，包括四声、轻声、儿化或特殊调值。

### 4.2　重音的韵律参数模型与修改

人们用语言表明事实、表达感情，传递消息。就语音的声学特性来说，在时域波形上表现为时长、基频、幅度和语音波形的变化。它们之间互相影响、相互依存。按照常识，语音幅度增大，主观感觉声音变强；基频提高，主观感觉音调提高。然而，人对幅度相同，但基频不同，或时长不同的声音，感觉强度也不同。

在TTS研究中，即要遵重语音表现的物理规律，又要注意人类的语言习惯，还要符合人的主观感知结果。我们重点分析了时长、基频和幅度之间的关系[2]，结论如下：

(1) 随着幅度的增大(重读)，发音的基频和幅度都是提高的。

(2) 随着幅度的增大(重读)，音节的时长可能增长，也可能不增长，但基音周期数一定增多。

(3) 随着幅度的增大(重读)，韵母平稳段加长。在归一化时长表示中，若把一条幅度曲线分为上升、平稳和下降三个阶段。可以看出，幅度上升阶段稍有缩短。

我们把上述实验结果应用于重音修改中。当把文本转换成声音时，按语义或语法规则指出某音需重读时，TTS改变相应的幅度(V)、基频(P)和基音周期数(N)，修改算法如下：

$$Vm=\alpha V \quad (\alpha=0.5\sim2)$$
$$Pm=\beta p \quad (\beta=0.8\sim1.5)$$
$$Nm=\frac{\alpha}{\beta}N$$

式中α、β分别为幅度和基频修改因子。根据需求V和P可独立修改，也可以让β与α建立某种相对关系。

实验结果表明，上述修改得到了重音感知效果。

## 5 结束语

国内开展波形编辑语音合成的研究中才刚刚开始，已取得可喜的成果，但要想使合成语言达到较高的自然度，还有许多问题有待研究。

## 参 考 文 献


〔1〕 Cai Lianhong, Wei Huawu. Research and Implementation of Text－to－Speech for Chinese, ISSIPNN'94, Hong kong P 583－586 94.4.

〔2〕 周俏峰，蔡莲红. 汉语句子重音的韵律参数模型的研究，第三届全国人机语音通讯学术会议，94.10，四川.

〔3〕 蔡莲红，魏华武. 一种基于语音学的汉语自动分词算法，计算机语言学研究与应用，北京语言学院出版社，1993.

〔4〕 E. Moulines, F. charpentier. Pitch－Synchronous Waveform Processing Techniques for Text－to－Speech Synthesis Using Diphones, Speech Communication 9, 1990, PP453－467.


# SPEECH SYNTHESIS BASED ON WAVEFORM COMPILATION AND APPLICATION ON CHINESE TTS


Cai Lianhong

(*Tsinghua University, Beijing* 100084)



**Abstract** This paper describes speech synthesis technology based on waveform compilation. That is rapidly developement. It's research content and basic algorithm PSOLA are introduced. In this paper, also introduces a Chinese Text－to－Speech system by waveform compilation.



**Key words** Speech synthesis, Waveform compilation, Chinese TTS, PSOLA