杨顺安 著

面向声学语音学的
普通话语音合成技术

THE CHINESE SPEECH SYNTHESIS TECHNIQUE
ORIENTED ACOUSTIC—PHONETICS

社会科学文献出版社

面向声学语音学的
普通话语音合成技术

杨顺安 著

THE CHINESE SPEECH SYNTHESIS TECHNIQUE
ORIENTED ACOUSTIC—PHONETICS

Yang, Shun—an

社会科学文献出版社

（京）新登字 028 号

面向声学语音学的普通话语音合成技术
杨顺安 著

社会科学文献出版社出版发行
（北京建国门内大街 5 号 邮政编码：100732）
新华书店经销 北京管庄印刷厂印刷

850×1168 1/32 开本 4.875 印张 124 千字
印数 0001—1000
1994 年 3 月第一版 1994 年 3 月第一次印刷

ISBN 7—80050—514—6/H·24 定价：4.80 元

序

杨顺安同志的遗著《面向声学语音学的普通话语言合成技术》即将副印了，有关同志要我为此书作序。我作为顺安导师之一，加以共事科研十来年，谊兼师友，非比寻常，自然有许多话可说、该说。关于顺安的苦学及著书经过，他的女儿杨晶在《后记》中已说得很详细，现在只就我个人感受写几点。

顺安以一位在电子专业上有了十多年工龄，并屡有发明贡献的工程师，为了决心要攻克言语处理的难关，从根本做起；竟毅然把专业搁起，放弃了原单位的既得待遇，改弦更张，到我们这个社会科学部门，从工程师变为研究生，从头学起。他的立志之坚、牺牲之大，是少有的。回忆他初来不久时，在研究生院一面学习必修学科，一面在我所实验室实习。那时我们正在装一套新引进的计算机（当时微机尚未普及），让他参加调试。有人在旁问他是不是冲着这新机器来的？他笑摇着头，用手指指实验室的同仁，意思是说，是为参加这个队伍而来的。他此后在工作中也是这样表现、重人不重物的。他毕业后在我室工作的十年中，他的研究方向、方法，固然吸取了许多前人及本室的成果，但他的敏捷构思和不懈的实验，有了不少创新，给普通话语音的规则合成打下了坚实的基础。这在国内外同类课题的研究中，不夸张的说是领先的。因此这本著作在这个课题上是起着承先启后的作用的。

语音合成技术在目前已大有进展，但汉语语音自有其特点，特别是根据语音基本规则来合成，为人机对话应用，则是不能生搬硬套西方的成果的。顺安此书，不但详述语音产生的声学原理和

模型，更重要的是按照汉语特点，以声母、韵母、音节和多音节的连读作全面的分析，建立合成程序，为“文语合成”的应用作了先锋，其成果已获国家特奖，为同行所引用。此书不但反映了他的勤奋的脑力劳动，还蕴含着许多辛苦的体力劳动。因为当时他的家离工作单位，一在极东，一在极西，又无直接的通车路线，他是不问寒暑风雨，都从不迟到早退的。单位为解决他的困难，配给了离所很近的宿舍，原想是从此可以使他能多事休息，谁知结果竟成了他加班加点的好条件。我有时晚些下班，总见他晚饭后又来坐在机旁编他的程序了。他这样不顾劳逸结合的干，以致带病之身彻底垮了下来。他的研究成果正如日之方中时，而最重要的身体本钱却江河日下了。凡是认识他的或读过他的论著的，无不为之痛惜。

对顺安的成就及其不幸，我的笔墨是难于表达的。想到古代有两件事倒或可比拟：一是孔子之探问他的得意弟子冉伯牛的重病，再三叹息说：“斯人也而有斯疾也！斯人也而有斯疾也！……”。一是唐代杜牧为苦吟而早逝的天才诗人李贺的遗集作序，评价他的诗写道：“使贺且未死，稍加以理，奴仆命骚可也”。今天我对顺安其人其书的感想，此二言尽之矣。是为序。

1994年正月，吴宗济于语言研究所，时年八十有五。

目　录

一、引　言

1.1　人—机—人的言语通信

言语（speech）是人类特有的、最迅速、最方便和最自然的一种通信系统。在当今的世界上，能力非凡的电脑已迈出试验室，来到了工厂、办公室、乃至家庭。如果我们能教会电脑说话和听话，赋予它言语功能，实现人—机—人的言语通信，那么，将会给我们的工作和生活带来多么大的便利和乐趣啊！

语音合成（speech synthesis）技术，就是一种教会电脑说话的技术，它泛指利用电脑技术或数字信号处理技术重新产生人类言语声音的技术。以合成语音作为电脑的一种信息输出手段，则称为电脑的语音输出（computer speech output system）。语音合成与叫电脑“听话”的“语音识别（speech recongnition）技术或电脑的语音输入系统（computer speech input system）相结合，是实现人—机—人语音通信的重要环节。与利用磁带或唱片录音技术重放语音相比，利用语音合成技术来产生语音，具有更大的灵活性和可靠性。

电脑有了说话的本领，人们就可以给它们派上各种用途。最简单的是会说话的玩具，它们只须鹦鹉学舌似的说上几句简短的话，就能令儿童爱不释手。如果用来教学，就可以造出“电脑发音词典”，看书时遇到生词，只要键入这个词，电脑就会发出该词的标准读音，显示并朗读释文。人们可以制成供哑人使用的“说话机”，供盲人使用的“读书机”，可以用于电脑排版的校对工序。如果语音合成系统再与语音识别系统有机地结合起来，人们就无

须叩击令人眼花缭乱的键盘，直接用话音向电脑发号施令；也无须目不转睛地盯着屏幕，电脑会用清晰的话音及时向你报告各种信息。

可以毫不夸张地说，电脑是无所不能的，但是，它的一切功能都是人赋予的。要想叫电脑说话和听懂话，我们首先要教会它。言语功能对正常人来说简直是轻而易举的，脱口而出，一听就懂。而且，人的言语功能是非常高明的。我们可以一边咀嚼食物，一边说出让人能懂地话语；我们能在喧闹声中专听某人的话音；我们能闻其声知其人……。眼下，这些三岁儿童就有的本事，电脑却难于学会。婴幼儿天生地具备了言语中枢和言语器官，在言语环境中逐步习得了言语功能。但电脑只有运算器和存储器一类硬件，它的所有功能都是人们专门赋予的。要想真正地赋予电脑言语功能，我们必须首先明白，在言语产生和言语感知的过程中，人的大脑是如何工作的？图 1.1 以框图的形式表示出人的言语通信的几个过程，其中的细节，特别是涉及神经系统的过程的细节，人们的认识是非常粗浅的。现阶段，我们只能从形式上教会电脑说话和听话，而且，电脑的言语能力也是极为初步的，电脑所能说的话，有的词语很有限，有的很不自然；电脑所能听懂的话，只是有限的、断断续续的词语，而且只懂主人的话。

尽管如此，电脑说话和听话的用途是相当广泛的，人—机—人的语音通信的理想是很有诱惑力的，所以，人们依然孜孜不倦地探索着。随着人们对言语机制的认识的逐步深化，随着科学技术日新月异的进步，电脑的言语能力必定会逐步提高，总有一天，电脑将会同我们对答如流地滔滔不绝地谈天说地。

1.2 语音合成技术的类别及基本原理

语音合成技术可细分为四类：波形编码合成、参数式分析合成、规则合成和文—语转换。下面，简要地介绍每一类的基本原

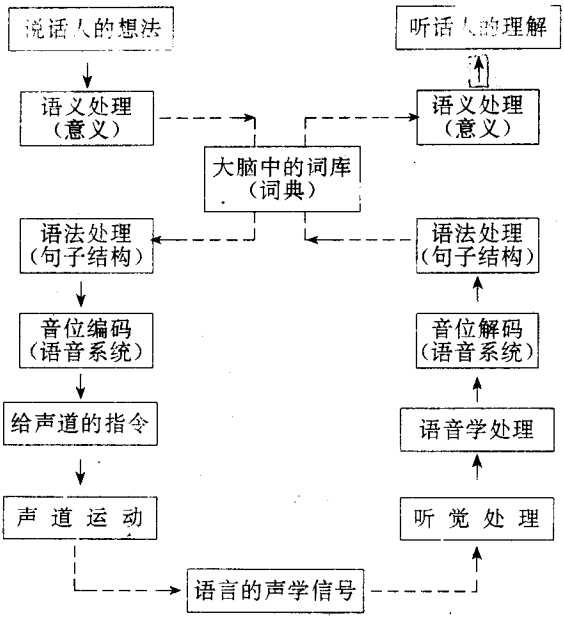

图1. 1 人类的语音通信过程的框图（引自：Bailey，1984，P. 194）。

理：

1.21 **波形编码合成**（waveform coding synthesis）

在这种语音合成方式中，以语句、短语、词或音节为合成单元，录音后直接进行数字化编码，经适当的数据压缩，这些单元的语音数据就驻留在存储器中，组成一个合成语音库；重放时，根据待输出的信息，由语音库中取出一个一个的单元的波形数据，串接（concatenation）或编辑在一起，经解码还原出语音。这种合成方式也叫录音编辑合成。

例如，在一种自动报时系统中，设计者事先将下列词语“录入”电脑存储器中：“现在的时间是”、“零点”、“一点”、“两点”，……“十”、“二十”、“三十”，……，“一分”、“二分”…“九分”，总共30多个单元。当需要报告的时刻为11：37时，就从存

储器中顺序调出下列四个单元："现在的时间是"、"十一点"、"三十"和"七分"，经解码还原成话音输出出去。可以想见到，对这种方式来说，一个个单元的输出音质是会相当清晰和自然的，但多个单元拼接出来的语句就会多多少少有些断续的和不自然的感觉。因为，在自然发音时，象上述十几音节组成的短句，是非常流利地一气呵成的。当然，为了提高连贯性，我们可以把"X 点 X 分"作为一个单元，但这样一来，就需要增加 $24 \cdot 60 = 1440$ 个单元，这是很不经济的。好在，对自动报时一类用途来说，只要话音清晰，自然度稍差一点也是可以接受的。

这种方式的合成音质较好，系统结构简单，价格低廉，但合成语音的数码率较大，合成的语汇量很有限。在自动报时、报号、报站或报警等装置中，多采用波形编码合成技术，现已开发出多种合成芯片以资选用。

1.2.2 **参数式分析合成** (parametric analysis—synthesis)

在这种语音合成方式中，多以音节、半音节或音素为合成单元，首先，按照语音产生模型，对所有合成单元的语音进行分析，一帧一帧地提取出有关语音参数，这些参数经编码后组成一个合成语音库；输出时，根据待合成的语音的信息，一个个语音单元的相应参数自合成语音库取出，经编辑和连接，顺序送入语音合成器；在合成器中，在合成参数的控制下，再一帧一帧地重新生成语音波形。主要的合成参数有：控制音强的幅度、控制音高的基频和控制音色的声学参数。由于在合成语音库中存的是表征语音特性的较少参数，而不是语音波形，所以这种方式的数码率比波形编码式的小得多，但系统结构也复杂一些，合成音质也差一些。目前，此种方式已开发出专用的芯片和插件板。汉语语音的音节性较强，以音节为合成单元，采用此种合成方式，可以合成出无限的语句。

1.2.3 规则合成（synthesis—by—rule）

在这种语音合成方式中，合成语音库中所存的是较小的语音单位（如音素、双音素、半音节和或音节）的声学上的合成参数。合成时，输入一串代码来指定每一语音单元的音色、音高、音强和音长，合成系统中有一套合成规则，对其合成参数进行必要的修改和调节，而后，由语音合成器合成出连续的语句来。合成规则是在分析每一语音单元出现在不同环境中的协同发音效应后，归纳其规律而制定出来的。与分析合成方式相较，规则合成方式的合成语音库的存储量更小，而所能合成的词语是无限的。这种方式涉及到许多语言学和语音学模型，系统结构较复杂。汉语是一种声调语言，合成规则中的韵律规则尤为重要。目前，由于合成规则还不完善，其合成音质一般较差，基本上处于研制阶段。

1.2.4 文—语转换系统（text—to—speech conversion system）

这是一种以文字串为输入的规则合成系统。在这种语音合成方式中，输入的文字串就是通常文本中的字串，不带任何特殊标记。这种系统中有一文本分析器，首先根据系统中的发音词典，将输入的文字串分解为带有属性标记和读音符号的词，再根据语义规则和语法规则，为每一个词和每一音节确定声调和重音等级以及断续度等等，这样，文字串就转变为代码串，再用规则合成系统就可以合成出带有抑扬顿挫和不同语气的语句。文 — 语转换系统集语言学、语音学和语音合成技术的研究成果于一身，是功能最全面，应用前景最广的语音合成技术。

对于英、德、法、俄等语种来说，将一个个音素或音节拼连起来，是成不了话语的。但对汉语这种音节性很强的语言来说，将一个个音节拼连起来，倒是听得懂的，只是听起来断断续续，不太连贯，不太自然。在一些对合成音质要求不高的应用场合，这样的合成系统也能将就着用了。在较为完备的文 — 语转换系统

中，有一整套繁杂的语义学、语法学、词汇学、音系学和语音学的规则，输入的文字串经过这套规则的加工后，得到一连串合适的合成参数，再由合成器一句一句地连贯地合成出来。显然，这样的系统的合成音质应该是相当好的。但是，正如前面所说，人的言语能力是非常高明的，言语机制是非常复杂的，再繁杂的规则系统也不能囊括人们话语中种种变化规律。所以，迄今为止，即使是对研究历史较长的英语来说，也未能开发出一套相当满意的文 — 语转换系统。人们正在从各个领域（特别是语言学）去探索。

此外，在语言学中，语音合成又是研究语音特性的一种重要手段。人们可以利用合成技术人为地产生出各种语音，通过对这些语音的听辨，从而进一步探讨语音产生和语音感知的机制。“如今，没有经过合成的验证，没有谁敢于发表语音产生方面的重要理论”（Coker，1972，p. 319）。

1.3 汉语语音合成技术的进展

随着电脑科学技术的发展，从80年代初，一些在国外作访问的中国学者开始了汉语普通话的语音合成技术的研究（如：Huang，et al.，1982；Lee，et al.，1982），以此同时，国内一些部门，也开展了这方面的探索。（如：李昌立等，1981；华一满，1984）

下面、大体上按编辑合成（包括波形编码合成和参数式分析合成）和规则合成（包括规则合成和文—语转换）两大类，分别介绍一下汉语普通话方面的语音合成技术的发展概况。

1.3.1 普通话的编辑合成技术

从技术角度来说，编辑合成方式比较地容易实现，人们可以使用现成的和专用的芯片，以较短的时间开发出实用的语音合成

装置。目前，在普通话编辑合成系统中，使用较多的有 TM5200 语音合成片和 TI 语音板等，这类专用硬件及其开发系统原是为合成英语而设计的，但也有一定的通用性。从 80 年代中期，国内一些部门相继开发若干语音合成装置，用于电话授时台的自动报时、查号台自动报号、公共电汽车的自动报站等场所。采用合成技术的报时、报站等装置，没有一般录音技术中不可回避的机械结构和磁头、磁带的磨损等麻烦问题，可靠性好，使用寿命长。

中国科技大学王仁华（1985）研制出一种语音合成自动报时装置。该装置共存了 48 个有关报时方面的音节，全部 LPC 合成数据仅占 4582 字节，能按 24 小时时制不停地报出当前的准确时间。报不同的点、分值时，数字音调各不相同，使得报时语调比较自然。该装置已在邮电系统的 117 电话报时台推广使用。

在一种电脑自动查询和报号系统中（王仁华，1986），话务员将用户所要讯问的部门名称，按其拼音的第一字母顺序从 IBM—PC 的键盘打入，经 PC 机自动检索，查出电话号码后，启动一个由 Z—80 CPU 和 TM5220 语音合成片组成的语声响应系统，把号码报给用户。系统存有报号所须的 0 到 9 这十个数字、“幺”、“拐”、“您好”、“没有电话”等，其 LPC 数据总共占 1772 字节。

在北京理工大学 1988 年开发的一种 HB—1 型语音合成报站器中，用一片 5220 语音合成片和一片 27256 RAM，它能较清晰和自然地报出一百多词语。

东北财经大学郭全等（1990）采用 T6668 单片机对普通话的 407 个基本音节的波形进行了观测，从波形的相似性出发，将一个音节分为前后两半，选取了 55 个前半音和 76 个后半音。对这些单元进行自适应增量调制（ADM）编码，采样频率为 16k 赫，前半音节取 25 毫秒，数据量为 $55 * 16k * 0.025 = 22k$ 比特；后半音节取 200 毫秒，数据量为 $76 * 16k * 0.2 = 343k$ 比特，总计 365k 比特（约 46k 字节），用 3 片 41256 存储这些数据。在 Z80 CPU 和 IBM—PC/XT 的控制下，根据给定的音节选出相应的前后半音节

进行拼合，合成出一个个音节来。

同济大学计算机系统（1990）开发出一种“CSSS—1 通用汉语语音合成系统”，以声母、过渡音和韵母为合成单元，分别采用矢量量化和自适应差值量化进行语音波形的压缩编码，用单片机和存储芯片实现了普通话的全音节合成，该合成系统可用作 IBM—PC 机的语音输出部件。

在国外，陆续开发出一些具有语音识别和语音合成功能的专用板，将其插在 IBM—PC 机中，经过较简单的再开发，就可以使用。例如，美国 TI 公司开发的 TI—SPEECH 语音板就是使用较广的一种，该板采用 TMS32010 数字信号处理器，有较强的实时处理能力，用该板对语音作 LPC 分析，逐帧取得基频、能量和 10 个反射系数，或建立用于语音识别的模板，或建立用于语音合成的语音库。国内一些部门也相继进行过此项技术的研究，开发出一些可供实用的系统。下面，就语音合成方面的内容举例介绍一下。

利用 TI 公司的语音板，赵珀章等在“CSIPS—2000 型汉语语音信息处理系统”中，建立了一个 9600 比特/秒的女声汉语音调节语音库，能高效地用于电脑文本的校对等。

1988 年 5 月，北京信息工程学院中文信息处理研究中心研制出一种“中英文语音合成系统”，该系统的中文合成是通过一套软件，借助于“TI—SPEECH”的支持而实现的。选用了 1200 多个普通话音节，能覆盖一、二级国标汉字，所建立的语音库有两种：2400 比特/秒（占 300 千字）和 16000 比特/秒（占 1.6 兆字）。

在林伟等（1989）研制一种“SDS 中文语音合成开发系统”中，利用“TI—SPEECH”板来建立语音库，以 8 千赫的采样频率，25 毫秒一帧，对待合成语音进行 LPC 分析，取得语音参数，一帧 12 个参数，每一参数占 2 字节，数据存入语音参数文件。语音合成系统使用 TM 5220C 合成器构成。

随着价格不断下降的通用性的数字信号处理器芯片的大量问世，近年来，用这类数字信号处理器芯片陆续开发出一些语音合

成器。在开发中，研制者可以灵活地利用处理器的各种功能，所以开发出来的合成器的合成音质有可能比用现成的合成芯片或合成板的好一些。

中国科学院声学研究所的莫福源等（1989）采用 TM32010 处理器开发出一种有限词语的实时语音合成器。在这种合成器中，将待输出的话音先进行 LPC 分析，逐帧求出基频和反射系数，用 256 个矢量的码本，作矢量量化，码率为 1200 比特/秒，其输出音质比不用矢量量化的 2400 比特/秒的还好。

1.3.2 普通话的规则合成技术

1980 年，中国科学院声学所李子殷在西德对普通话的规则合成技术作了初步的探索（李子殷，1981），他采用一种并联式的共振峰合成器，以“双音素”（diphone）作合成单元，每一单元存有 30 个参数。在 LSI 电脑上，该作者合成了约 400 个汉字的一篇短文，经试听，平均句子可懂度为 90%。

1982 年，中国科学院声学所张家騄在瑞典作访问学者期间，利用能合成瑞典语和英语等语种的 OVE—3 型级联式共振峰语音合成系统，针对汉语的特点，对参数和规则作了一些增删，合成出一些普通话词语和短文。

1982 年，中国科学院自动化所黄泰翼等在美国作访问学者期间，研究过以 LPC 系数为语音参数的文—语转换系统。该系统存储了 21 个声母和 37 个韵母的 LPC 参数，四种直线式的声调模式，从而合成出普通话的音节，再将单音节拼接成语句。（Huang，T. et al.，1982）

VOTRAX 语音合成器原来是一种以共振峰频率为参数，以音位为合成单元来合成英语的装置。合肥工业大学徐士林等在美国工作期间，对于 VOTRAX 无法合成的 j、q、x、zh、ch、sh、r、z、c 和 u，研制了专用的硬件，这样，构成了以 8080 微机为基础的一种普通话语音合成实验系统。（Lee，S. et al.，1982）