第 24 卷第 5 期　　声 学 学 报　　Vol. 24, No. 5
1999 年 9 月　　ACTA ACUSTICA　　Sept., 1999

多路实时、高音质数字串合成系统*

刘庆峰　滕永盛　王仁华
(中国科学技术大学电子工程与信息科学系　安徽合肥　230027)

1998 年 2 月 27 日收到
1998 年 12 月 23 日定稿

摘要　根据汉语普通话中数字串发音的韵律规则和特点，利用 LMA 语音合成器，实现了一种全新的数字串报号系统。本系统可以在不足 300 kbytes 的极小的音库容量下，通过采用预先计算、查表拼接快速处理方法，在各种特定应用场合下多路实时实现高自然度、高音质的任意多位的数字号码的合成语音。测听实验和用户反馈信息均表明，合成输出语音的听觉效果已经可以与播音员原始发音相媲美。

PACS 数：　43.70

A speech synthesis system of digit strings with high quality and multi–channel real–time way

LIU Qingfeng　TENG Yongsheng　WANG Renhua
(*Department of Electronic Engineering and Information Science, University of Science & Technology of China* Anhui Hefei　230027)

Received Feb. 27, 1998
Revised Dec. 23, 1998

Abstract　Based on the prosodic characteristics of Chinese digit strings and the application of the LMA synthesizer, a new speech synthesis system is introduced in this paper. Using pre–calculating and table–searching methods, the system can produce synthesized speech for digit strings with various lengths in a multi–channel and real time way. Besides the high speed, little speech library (less than 300 Kbytes), high naturalness and clarity are the three other outstanding features of this system. Listening tests and users' evaluation showed that the perceptual effect of the output from the new system was much close to the original speech.

引言

语音是人类最便利的信息交换手段，也是人与机器之间最理想的交流信息的途径之一。随着计算机领域和通信领域技术的日臻成熟，人机语音交流已经成为可能，其中一些关键技术正在逐步进入生活，涌现出一大批带有语音处理功能的智能型系统。汉语普通话数字串语音合成技术是

* 国家自然科学基金项目 (编号：　69575616) 和 863 项目 (306 主题)

汉语语音合成技术中关键一环和重要组成部分，在 114 自动电话号码查询、电话银行、机场、铁道信息查询等场合有着重要的应用前景。

适应市场的迫切需求，目前已经有很多汉语普通话自动数字报号系统问世。这些系统的实现方法有两种：规则合成方法和录音编辑合成方法。规则合成以韵律规则为指导，利用语音合成器对存储音库中少量基元的韵律参数进行调整，以获得在各种语言环境下的相应音变基元，然后将这些音变基元拼接得到连续语音。传统的规则合成方法因为语音合成器性能的局限，在实现语音高自然度 (由韵律规则决定) 的同时会伴随着音质的明显下降。同时，韵律规则的总结也是一长期艰巨的任务，规则总结不够完善，合成语音的自然度也不会令人满意。受到以上两方面因素的制约，传统的规则合成在实用系统中较少采用。大多数系统采用录音编辑合成方法，以数字短语或单个音节为合成单元，录音后直接进行数字化编码存放在存储器中。重放时，由语音库中取出所需合成单元，经解码还原后，编辑在一起作为输出语音。录音编辑合成具有清晰度高，单音音质好的优点，但自然度较差，存在“一字一顿”或“一词一顿”现象，不能很好满足人们对语音合成自然度的要求 – 希望能听到象播音员发音一样自然流畅的合成语音。例如，在电话号码升至 7 位和 8 位以后，国家有关部门已提出 114 报号必须具有 4 位连读功能。倘若对所有需要连读的号码都进行全面存储，则存储容量将会非常庞大，应用场合也将受到极大限制。

本文介绍一种新的方法，使得系统在存储容量、单音音质和语句自然度上能够同时具备传统规则合成和录音编辑合成的各自优点。本文系统从以下两方面改进了传统的按规则合成方法：(1) 针对数字串的韵律特点，总结提出了一套简化规范的韵律规则，大大提高了合成语音自然度。(2) 利用 LMA 语音合成器作为分析合成工具 [1]，得到所有合成必备的数字音变单元韵律特征。然后采用从原始语流中截取为主，合成器补充为辅的方法得到这些数字音变单元，从而大大提高了合成音质。基于这些改进，系统合成语音从自然度和音质上都已酷似原始发音。由于韵律规则的简单规范，存储所有数字音变单元只需极少空间 (280 Kbytes)。系统同时采用预先计算、查表拼接的快速处理方法，成功地支持了 114 系统中的 128 路实时实现。其功能模块目前已被深圳市华为技术公司的 114 电话查询系统所采用，在用户中获得了广泛好评。

1 LMA 语音合成器简介

LMA(对数振幅近似) 语音合成器，是以 LMA 滤波器为基础构造的一种源 / 滤波器型合成器。LMA 滤波器是利用输入语音信号的倒谱系数，按照给定公式构造的一组级联的指数函数形式的滤波器，其优点是可以在最小均方误差准则下无限逼近输入信号的对数幅度谱 [2]。在 LMA 语音合成器中，用 LMA 滤波器来模拟声道；声门波激励中的准周期部分 (浊音部分) 用类三角波来模拟，清音部分则直接取自原始语音的清音残差。其模拟框图如下图 1 所示。

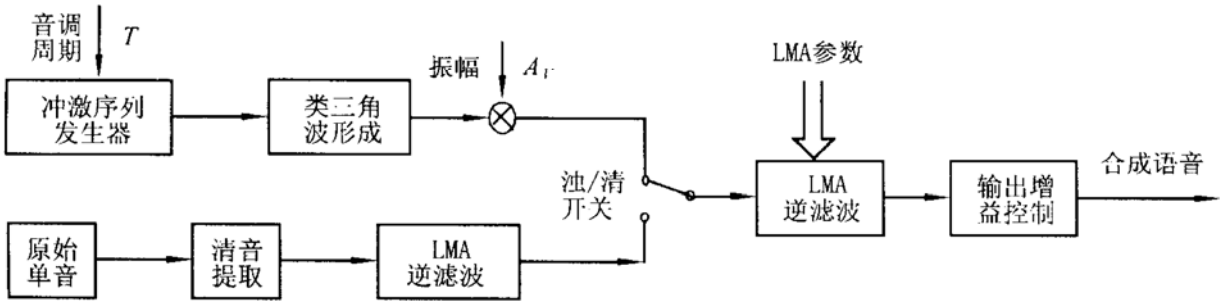

图 1　LMA 语音合成器框图

LMA 合成器的工作流程如下：首先根据待合成音节的声调特性构造出相应的声门波激励源；然后再根据协同发音、速度变换等音变信息在原始声道的基础上构造出新的 LMA 声道参数；最

后将声门波激励源送入新的声道模型中，其输出就是符合给定韵律特性的合成语音。由于 LMA 滤波器的高精度模拟和源 / 滤波器型结构的固有优点，使得这种新型合成器可以方便地实现语音学规则所需要的各种韵律参数调整。从实验语音学的基本原理出发，可以系统地给出对时长、声调等参数的调整以及对协同发音现象的模拟方法[3]。根据本实验室的测听结果表明，其对各种参数的调整范围和合成语音质量都优于目前占主导地位的 PSOLA 和共振峰等合成方法。LMA 合成器的上述优点，不仅保证了我们可以高音质地实现已有的韵律规则，也为我们总结和试验新的规则提供了技术手段。

2 新的按规则合成数字串方案

为了适应市场和用户的进一步要求，新的数字串报号系统应当既具有录音编辑合成中高清晰度的音质，又具有规则合成中低存储容量的特点，同时，还需要我们发展原有的韵律规则以获得更高的自然度。LMA 语音合成器的优良性能为系统提供了音质上的保证，而针对数字串发音特点总结的韵律规则，保证了合成输出的高自然度。

数字串比一般连续语音具有以下四个重要特点：(1) 短语中各成份间没有语意的轻重之分。(2) 短语的节奏划分可以用一个统一的模式，如 4 字短语可以都采用双双节奏，5 字短语都采用双三节奏。(3) 不存在轻声和轻读 [4] 的特殊处理需要。(4) 音节数目少 (包括 “yao1” 在内也仅仅 11 个)，协同发音的种类不多。这些特点，为下文韵律规则的简化提供了依据。

由于数字串服务场合下通常的要求是清晰、流畅、舒缓 (每秒钟约 2 个字)，所以我们在设计数字串连读模式时规定每个呼吸群最多包含两个词组，每个词组最多包含四个数字，词组间通过调域的对比差异来得到呼吸群的句调特征。实践表明，这种设定进一步简化了韵律规则的提取，同时也能很好地满足实用要求。

基于以上设定，对语调规则采用了如下分级处理原则：

1. 首先总结出所有的 4 字以内的短语变调规则。具体有以下几种：

(1) 二字、三字短语采用通用的词组变调规则，以前的系统对此已有较成熟方案[1]。

(2) 四字变调规则：由上述数字短语的第 2 个特点，可以对四字变调采用统一的双双节奏模式，由二字变调为基础产生。生成过程为：首先由二字变调规则得到前后两个二字词中各个字的调形和调域；再将 2 、 3 两个字组成一个新的二字词，由二字变调规则得到它们新的调形和调域；第 2 、 3 两字调域由新调域与原调域取平均得到最终值，调形以原调形为基础，将第 2 字的调尾根据新调形稍作调整，第 3 字的头部根据新调形稍作调整。如此得到的四字调联，不但具有调形上的自然变化，同时具备自然发音中语势渐降的特点，在实践中取得了很好的效果。

2. 在短语变调的基础上，再根据短语间的调联规则，获得整个呼吸群的调形。

根据设定，每个呼吸群最多包含两个四字以内的数字短语，所以在同一呼吸群内最多只需一级调联处理。调联处理的过程为：首先以短语联接处的两字为中心，向前先后各取一字以构成四字词组；再以四字变调规则得到中间两字的调型，分别作为第一词组尾字和第二词组首字的调型；获得第二词组首字调形后，根据其相对未作调联前的调域变化情况对该词组中其余各字作相同比例的移调处理。

3. 包含多个呼吸群的长句中，各呼吸群单独得到各自的调形曲线，互不影响，简单拼接后得到整句调形。

根据数字串短语发音的特点 1 和特点 2，时长规则可采用以下简化模式：

(1) 时长参数对自然度的影响，主要表现在各单音之间的时长对比上，只要各部分比例合适，就能达到自然的合成效果。对于二、三、四字词组其中各字的时长比例分别取为：1:1.1 、1.1:1:1.2 、1:1.1:1:1.2 。以上时长指的是各音节相对其正常单音时长的变化比例。

(2) 长句或呼吸群中各字的时长，完全由其所在词组决定，与词组所处位置无关。

考虑包括“幺”在内的所有 11 个数字音节的所有的声母和韵母种类，根据声、韵母的舌位和共振峰轨迹可给出所有协同发音搭配类型[4]。为了简化处理，各数字音节能量均采用正常单音能量，不考虑当前语流环境对能量的影响，合成实践表明如此简化基本不影响合成语音自然度。

有了上述各项规则，可以系统地给出各个音节的所有可能的音变情况。针对各个音变情况下的韵律参数，利用 LMA 合成器首先将相应的音节都合成出来，然后再用替代法进行检验，以测听结果为依据，将可以互换的音变音节都用一个代替。这样便得到了我们在上述数字串韵律规则下，各个音节在连续语流中的所有必备音变单元。实践表明，对每个音节最多只需要 10 个左右的音变单元，就可以拼接合成出极其自然的数字串号码。最后结果表明，包括“幺 (yao1)”在内的 11 个数字总共只需 95 个音变单元。LMA 合成器的优良性能，保证了各个音变单元具有很好的音质。

为了高速多路处理考虑，在我们将所有音变单元预先计算好存放在系统音库中，实际合成时，只要查找韵律规则表，根据查表得到相应的音变音节名称，然后将它们直接拼接就可以得到高自然度、高清晰度的语音输出。

为了进一步提高合成音质，我们又根据上述简化韵律规则设计出录音词表，尽量从原始发音中截取到相应的音变单元，部分难以从原始发音中得到的单元，则由 LMA 合成器生成。下图 2 给出了数字串“3601249”的合成音与原始发音的对比实例。

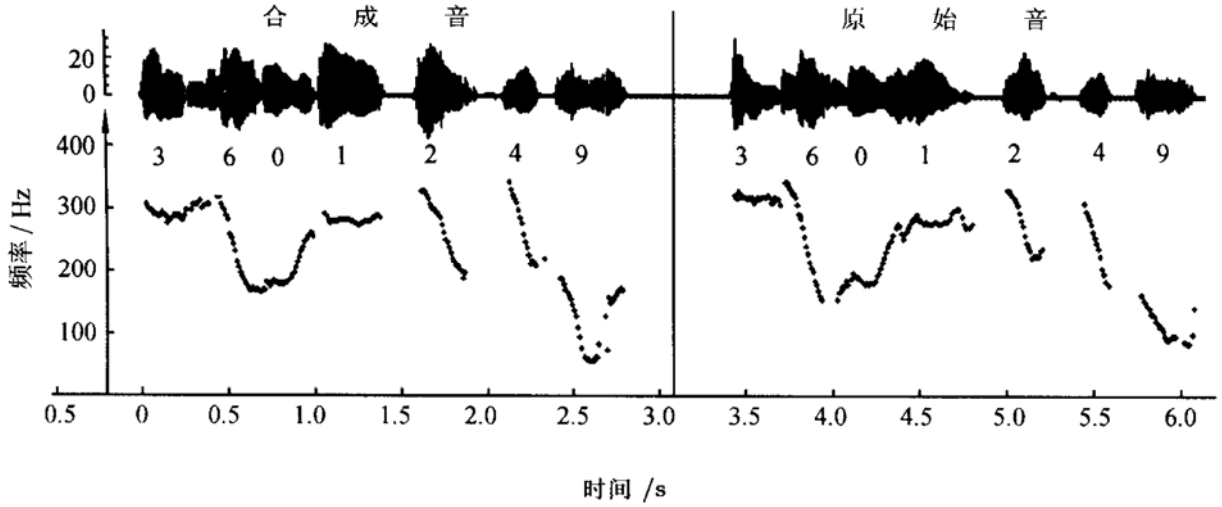

图 2　数字串“3601249”合成音与原始音的对比波形

根据简化韵律规则下的录音词表，图 2 中用于合成的各个基元分别来自以下数字词组：3572、7605、9407、7189、279、544、9039。合成过程中仅用 LMA 合成器对基元“2”的调域作了适当调整，其余基元则直接拼接。由图 2 可见，合成音和原始音在声调、时长和能量等韵律参数变化上都非常相似。听辨实验也表明：两者无论从自然度还是清晰度上都难以区分究竟谁更象原始发音。

3　系统性能评价

本系统具有以下显著特点：

(1) 音库小。音库中共存有 95 个经处理的合成单元 (包括从 0 ～ 9 以及“幺”在内的 11 个音节)，采用电话上传输的标准数据格式 (8 kHz 采样，8 bit 量化)，仅需要 280 Kbytes 的存储容量。

(2) 音质好。系统的合成输出具有可与原始发音相媲美的语音质量。

(3) 自然度高。本合成器采用了一整套完善的韵律规则，具有 2 字、 3 字、 4 字和句调的连读规则，可以高自然度地合成出任意位长的数字串，其效果与播音员原始发音非常接近。

(4) 平稳性好。在传统的录音编辑合成中，即使愿意付出巨大的存储空间，在音库中存放较长的数字串，整句的自然度也存在平稳性上的问题。因为在所需录音数目很大的情况下，播音员很难保证其所有录音词组的调域和音色都相同，当两个调域有差别的数字串短语连接在一起时，将给人一种很不平稳的感觉。在本系统中，因为韵律规则已考虑到句调问题，所以无论合成多长的数字串，都可以保持高自然度的整句输出。

(5) 合成速度快。本系统将语音合成器要做的大量计算预先算好后存放在音库之中，合成时采用查表的方式直接调用波形文件进行简单拼接，速度非常快；适于多路实时实现。这种高速度的合成方案，成功地支持了华为公司 114 报号的 128 路实时实现。

为了给出系统性能的定量评价，我们根据声调组合和协同发音类型设计了一组测听实验词表。因为不存在轻声，数字串四字词组共有 256 种调型组合方式，三字词组共有 64 种调型组合方式。这里首先以调型相同的数字 (如： 2 、 4 、 6 同属去声单音， 5 、 9 同属上声单音) 等概出现为原则设计出这 320 种三、四字词组。在此基础上，再根据协同发音类型补充 40 个四字词组。最后，将这些词组组成 120 条测听数字串，每条数字串包含 3 个词组，以电话报号的方式输出，第一个词组作为区号，后两个词组构成一个 7 或 8 位号码。

实验中邀请了 20 位青年学生作为测听人员，各人均对本合成系统没有先验知识，听力正常，能说较标准的普通话。本实验的主要目标，是为了定量描述合成音与原始音在听感上的相似程度。测听记录前，首先播放 10 句原始录音数字串作为参照，在 120 句合成音中，每 4 句混入一句由四字和八字录音拼接而成的原始音，位置随机，用以提供动态参考，合成音与原始音以同样的标准共同测试。测试指标和结果见下表 1 、表 2 所示。

表 1 中的自然度平均值，是以“极好”为 5 分、“很好”为 4 分，依次递减到“很差”为 0 分，将各项分数加权平均所得。

表 1 合成音与原始发音自然度对比表

自然度	极好	很好	较好	一般	较差	很差	平均值
原始音	13.3 %	68.9 %	17.8 %	0 %	0 %	0 %	3.96
合成音	5.0 %	79.5 %	15.5 %	0 %	0 %	0 %	3.90

表 2 合成音与原始音可懂度对比表

	判为原始音	判为合成音	可懂度
原始音	57.8 %	42.2 %	100 %
合成音	30.6 %	69.4 %	100 %

由以上测听结果可见，本系统的合成语音从自然度和音质上已经非常接近四字和八字原始录音的直接拼接，以原始音本身的测听结果为依据，合成音与原始音的听觉混淆程度达 53 % (30.6 % : 57.8 %) 。从自然度平均值上来看，两者几乎完全相同，倘若以“极好”为满分，则两者相差仅为 1.2 % (0.06: 5) 。从自然度指标的分布上来看，合成音比原始音分布更为集中，这说明合成音从听觉的整体感觉上比原始音更加平稳。值得指出的是，以上原始音仅仅是从播音员的少量录音中选取的，倘若请播音员朗读所有可能的可能出现的数字串，然后再从中随机挑选，则随着录音时间和播音员身体状况的不同，其原始发音的自然度会更加不平稳。

倘若采用录音编辑合成实现 4 位连读 (包括 4 位以内连读) 的数字报号功能，则音库中必需包含 4 位、 3 位、 2 位和 1 位词组的所有录音，对于长度为 n 的词组，每个位置上有 11 种音节

的可能性 (含 yao1 在内), 则共有 11^n 种可能的词组，考虑到其本身长度，则相当于有 $n \times 11^n$ 个单音音变单元。则以上所有录音相当于需要 $4 \times 11^4 + 3 \times 11^3 + 2 \times 11^2 + 11 = 62810$ 个单音音变单元。而本系统在自然度和音质几乎不变的情况下仅用 95 个音变单元就可实现这一功能，存储容量少了 600 倍以上！

4　总结

本文在 LMA 滤波器的基础上，首先利用简化分级的语调生成规则，从自然连续语流复杂多变的发音方式中找到一种简单规范的韵律构成模式；然后利用分析合成的方法提取出其中必需的音变单元，并用替代测听的方法将音变单元的冗余度控制在最少，最后在此基础上构成一种全新的实时的数字报号系统，达到了高清晰度、高自然度的听觉效果。

实验结果表明，采用本文所述方法，系统可以在数百倍地降低存储容量的同时，使合成语音高度逼近原始发音的音质和自然度。值得指出的是，LMA 滤波器的优良性能，为我们总结和验证新的韵律规则提供了技术保障，同时也使我们有能力补充和修正原始音库，系统地得到了自然录音中难以全面选取的典型音变单元。本文方法的提出，为文语转换系统在有限词汇合成领域内以真正接近播音员发音的效果进入实用领域，迈出了坚实有力的一步。

参　考　文　献

1　WANG Renhua, LIU Qingfeng, TANG Difei. A new Chinese text-to-speech system with high naturalness. In: Proc. of ICSLP, Philadelphia, USA, 1996, 1441—1444

2　Satoshi, Tadashi. Speech analysis synthesis system using the log magnitude approximation filter. *Electronic communication institute thesis*, 1978; *J61-A*(6): 527—534

3　刘庆峰，王仁华. 基于 LMA 声道模型的语声合成新方法. 声学学报，1998; **23**(3): 271—278

4　WANG Renhua, LIU Qingfeng, HE Runzhong. A new method of the phonetic control based on LMA synthesizer. In: Proc. of CJSLP, Huangshan, China, 1997, 141—146

5　WU Zongji, WANG Renhua, LIU Qingfeng. Towords a project of all-phonetic-labelling-text(APLT) for TTS synthesis of spoken Chinese. In: Proc. of CJSLP, Huangshan, China, 1997, 26—27

6　吴宗济，林茂灿. 实验语音学概要. 北京：高等教育出版社，1989