第21卷 第4期增刊 声 学 学 报 Vol. 21 No.4 Suppl.
1996 年 8 月 ACTA ACUSTICA August 1996

一种高清晰度、高自然度的汉语文语转换系统

初 敏 吕士楠

(中国科学院声学所，北京 100080)

1994 年 9 月 20 日收到

摘要 以基音同步叠加技术为基础，以汉语单音节为合成单元，有一包含词调模式、重音模式和句调模式的韵律规则库的汉语文语转换系统，可合成出高清晰度和高自然度的汉语语音。研究表明，影响汉语合成语音的自然度的主要因素是音高和音强随时间的变化、各音节的音长分布以及音节间的协同发音，其中以音高和音长的影响最为显著。时域基音同步叠加技术提供了一种在时域改变语音波形的音高和音长的方法，从而使在用波形拼接法合成汉语时，进行词一级和句一级的韵律调节成为可能。对新闻广播语言的声学特征的分析，为建立汉语合成的韵律调节规则提供了理论依据。本文介绍新的汉语文语转换系统的结构及流程、对广播语言韵律特征的初步研究结果、汉语合成规则及合成系统语音质量的评测结果。

A Chinese text-to-speech system with high intelligibility and high naturalness

CHU Min and LU Shinan

(*Institute of Acoustics, Academia Sinica*, Beijing 100080)

Received Sept. 20, 1994

Abstract A Chinese text-to-speech system, which is based on the time domain Pitch-Synchronous-Overlap-Add(PSOLA) method, a Chinese syllable dictionary and a prosodic-rule dictionary, can produce very clear and natural Chinese speech. Research work on naturalness of synthetic Chinese show that, when synthesizing Chinese, pitch, energy, syllable duration and coarticulation between syllables are main factors which affect the natrualness. Among them pitch and duration play the most important roles. The time domain PSOLA Scheme provide a method to modify the pitch and duration of a speech segment in time domain,and this makes it possible to adjust the prosody of speech in word level and sentence level,when synthesizing Chinese using waveform concatenation technique. Acoustics analysis of news broadcast speech provides theoretical basis for building up prosodic rules. this paper presents the flowchart of the new Chinese text-to-speech system, the research result of acoustics analysis of news broadcast speech, prosodic rules of the new system, and the evaluation results of speech quality of the new system.

一、引 言

合成一种语言时，只有使合成单元的音段特征和超音段特征都与自然语言相近[1]，合

成出的语音才能清晰、自然，二者缺一不可。就现有合成技术来讲，参数合成技术在语音合成中能灵活改变合成单元的音段特征和超音段特征，从理论上讲是最合理的。但是，由于参数合成技术过分依赖于参数提取技术的发展，并且至今言语生成模型的研究还不够完善，因此合成语音的清晰度往往达不到实用目标。与此相反，用波形拼接技术合成语音时，能很好地保持拼接单元的语音特征，因而在有限词汇合成中，得到广泛的应用，如语音表。但是，在简单的波形拼接技术中，合成单元一旦确定就无法对其做任何改变，当然也就无法根据上下文来调节其韵律特征。因此将这种方法用于合成任意文本的文语转换系统时，合成语音的自然度不高。八十年代末，由 F. Charpentier 等人提出的基音同步叠加技术 Pitch-Synchronous-Overlap-Add(PSOLA)[2]，既能保持原始发音的主要音段特征，又能在拼接时灵活调节其音高和音长等韵律特征，给波形拼接技术带来了新生。

汉语音节的独立性较强，但汉语音节的音高、音长和音强等韵律特征在连续语流中变化复杂，而这些韵律特征又是影响汉语合成语音自然度的主要因素。因此汉语很适合采用基于 PSOLA 技术的波形拼接法来合成。中央人民广播电台和电视台的新闻广播语言是汉语普通话的典范，应当是汉语合成系统的主要模拟目标。以对广播语言声学特征的分析为依据建立的汉语合成规则，使合成语音既规范又自然。本文介绍一个以 PSOLA 技术为基础，以汉语单音节为合成单元的汉语文语转换系统。该系统有一个具有词调模式、重音模式和句调模式的韵律规则库。它能对输入文本中的句子进行词一级和句一级处理，从而达到通盘规划一句话中的各音节的音高和音长等韵律特征的目的。该系统能将带有少量韵律符号的汉语文本转换成流畅的具有新闻广播风格的汉语口语输出。

二、基于 PSOLA 技术的波形拼接汉语文语转换系统

采用基于 PSOLA 技术的波形拼接法实现的汉语文语转换系统的核心部分是：①由汉语全音节采样值及其基音同步标记构成的音库；②能通盘规划一句话中各音节所应具有的音长和基音曲线的韵律规则库；③能将从音库中取出音节的音长和基音曲线调节到由韵律规则规划的目标值的 PSOLA 韵律调节模块。系统流程图如图 1 所示。系统启动后，首先进入文本扫视模块。在这里，系统对读入的文本进行分词处理、确定出多音字的发音、分离出外加的韵律控制符(如果有)，并记录各音节在词、呼吸群和句中的相对位置。然后，系统将准备好的数据送入韵律规则库，由韵律规则为每个音节规划其应有的音长和基音曲线。接着，系统根据音节名从音库中取出音节采样值及其基音同步标记，连同韵律规则库为之规划的目标音长和基音曲线一起送入 PSOLA 处理模块。最后，将处理好的音节拼接起来，并加入适当的停顿控制，送入 D／A 转换器，就可以得到清晰度和自然度相当高的连续合成语音。

图 1 所示的汉语文语转换系统中的韵律规则库，将在下两节中详细讨论，这里先介绍一下音库和 PSOLA 韵律调节模块。

(1)音库

利用汉语音节独立性较强和数目有限的特点，以音节为合成单元，建立由汉语带调音节采样数据及其基音同步标记构成的音库。音库的组成为：

全部汉语基本音节	1278 个
轻声音节	240 个
儿化音节	20 个

音库中只有部分常用的轻声和儿化音节。音库建成开放式的，可随时加入新成员。

(2) PSOLA 韵律调节模块

时域基音同步叠加技术提供了一种在时域改变语音波形基音曲线和音长的方法[3]，使在波形拼接汉语合成系统中进行韵律控制成为可能。其实施步骤是：

①将原始语音波形与一系列基音同步的 Hanning 窗相乘得到一系列基音同步的有重叠的短时分析信号。Hanning 窗的长度取基音周期的两倍。

②对这些短时信号做必要的修正，形成一系列短时合成信号。首先根据原始语音波形的基音曲线和音长及其目标基音曲线和目标音长，确定出合成波形与原始波形之间基间音周期的映射关系；其次由此映射关系确定合成所需要的短时合成信号系列。

③将合成短时信号系列，与目标基音周期同步排列并重叠相加得到合成波形。此时，合成语音波形就具有所期望的基音曲线和音长。

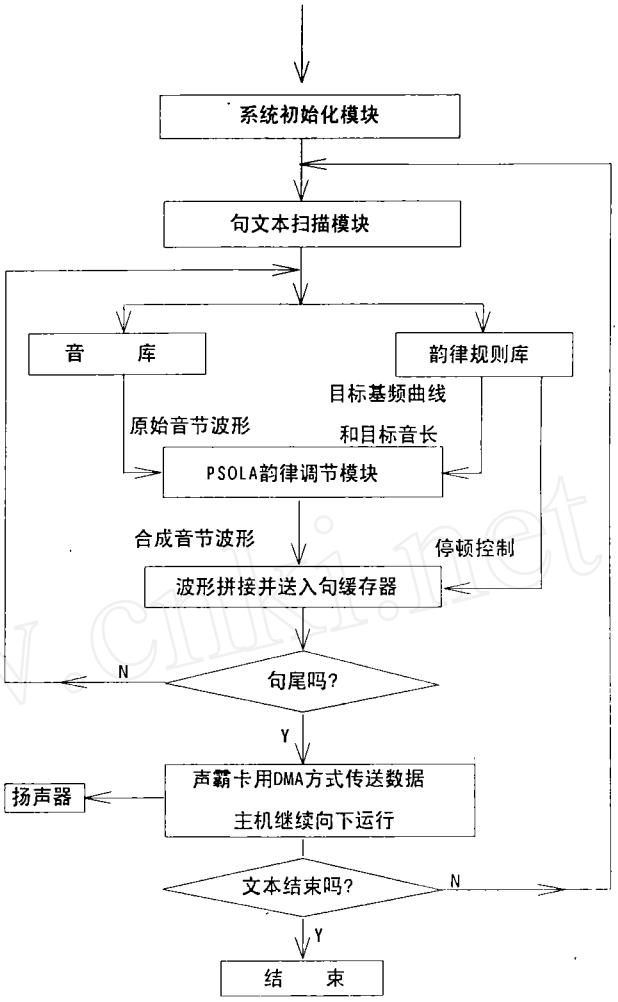

图 1　基于 PSOLA 技术的波形拼接汉语文语转换系统的流程图

图 2 所示为基音同步叠加技术的示意图。图中的(a)、(b)、(c)、(d)分别对应于基频的升高和降低，音长的缩短和延长。(a)和(b)中伴有音长变化，(c)和(d)所示为基频不变时的时长变化。在实际应用中，通常是基频和音长都要变，可将二者一起考虑，找出基音周期的对应关系，然后一次变成。

图 1 所示系统在一台 486PC 机上实现，所需的附加硬件是一块声霸卡和一个音箱。利用声霸卡的 DMA 传输方式，并采用多缓存器交替工作，可实时输出语音。这是一个

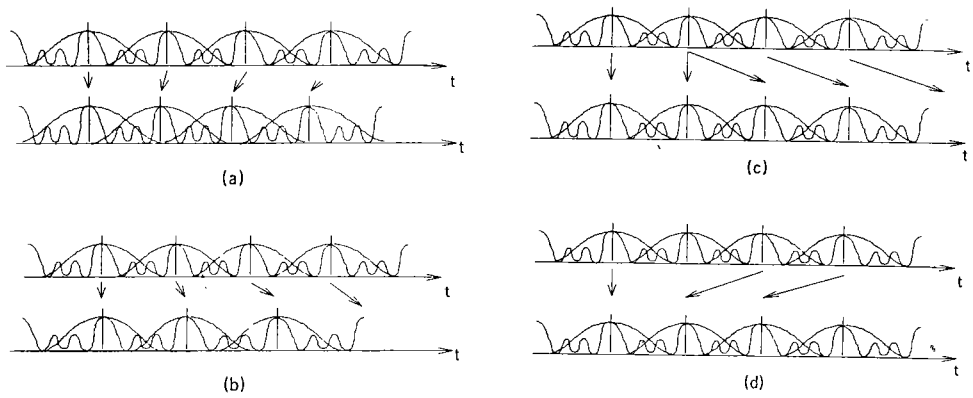

图2 用PSOLA技术改变原始波形的基频和音长的示意图
(a)基频提高；(b)基频降低；(c)音长延长；(d)音长缩短

高清晰度和高自然度，并且很实用的计算机汉语文语转换系统，它能将带有少量韵律的汉语文本转换成流畅的汉语口语输出。

三、广播语言声学特征的研究

汉语是一个语言大家族，做一个包括所有方言的汉语合成系统几乎是不可能的。中央人民广播电台和电视台的新闻广播语言是标准汉语的代表，也是我们的汉语合成系统的楷模。因此，我们以新闻广播语言为研究的对象，分析其声学特征，从中归纳出合成规则，力图使合成语言接近新闻广播语言风格。

我们录制整理了六个小时的电台新闻广播语音材料，并对其中近一小时的材料进行了声学分析。另外又请广播学院高年级学生男女各一名，做了半小时选定文本的录音，对这些材料也进行了声学分析，研究工作还在继续中。从前一阶段的分析数据中已得到一些初步结论，并根据这些初步结论建立起了一个具有词调模式、重音模式和句调模式，能对输入文本各句子进行词一级和句一级处理，从而达到通盘规划一句话中的各音节的音高和音长等韵律特征的韵律规则库。该韵律规则库用于我们的基于PSOLA技术的波形拼接汉语文语转换系统取得了良好的效果。初步的研究结果可归纳为以下几点结论:

1. 广播语言中的词调搭配关系比较稳定

汉语普通话有四个声调，分别是：阴平(55)、阳平(35)、上声(214)和去声(51)*。虽然各种声调独立时有相对稳定的基频曲线，它们在连续语流中的调型要受词内相邻音节调型的影响。这一点是早有共识的。社科院语言所吴宗济对二字词、三字词和四字词中各音节的基频曲线变化已有多次描述[4, 5]，这些描述对我们的研究工作起了指导作用。在前人基础上，我们对词中的声调组合关系做了进一步的定量研究，发现：上声和去声后的阴平调通常较底，用五度法记，只能到44或33。如图3中所示的“亚洲”的“洲”，“各家”的“家”，“组织”的“织”，“马家军”的“家”；若去声不在词尾，则去声不是51，而是52或53。如图3中所示的“亚洲”的“亚”，“四人”的“四”，“记者”的“记”，“运动”的“运”，“戴国红”的

* 这里采用了赵元任在《中国字调和语调》[6]一文中提出的五度值记调法。

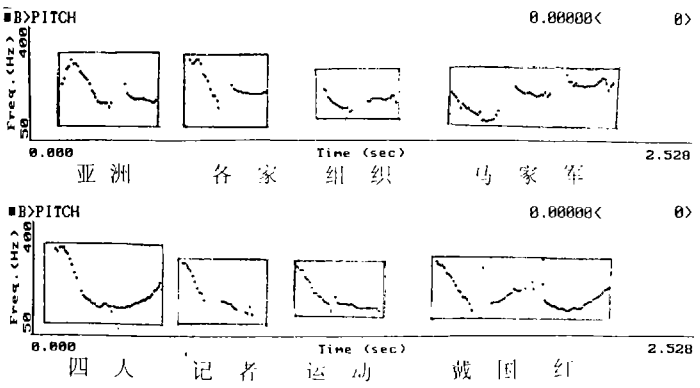

图 3　广播语言中声调间相互影响的词例

"戴"。另外，吴宗济指出，在口语中，三字词中间音节往往失去原有调型，而其调型变成首尾音节调型的过渡[5]。如把"西红柿 xi1hong2shi4"读成"xi1hong1shi4"。

在广播语言中，词中的不同声调组合对字调的影响较大，但对同一种声调组合的不同词，这种变化是比较稳定的。

2. 自然语言中调域主要以词为单位变化，其大小与词重音相关联

赵元任很早就用音域的概念来描述汉语的语调和字调[6]。沈炯则建议采用声调音域，他认为语调是以句子为单位的声调音域系列[7]。他还研究了语势重音对声调音域高音线和低音线的影响[8]。用声调音域的高、低音线的移动来描述象汉语这样的声调语言的语调具有很多优越性。因此本文采用双线语调模型。

对广播语言的分析表明，调域一般以词为变化单位，而其高音线移动与语句重音相关联。重音词的调域可能比非重音词的调域大一倍还要多。如图 4 所示，在"今年我国十大体育新闻"这句话中，重音在"今年"和"十大"上。

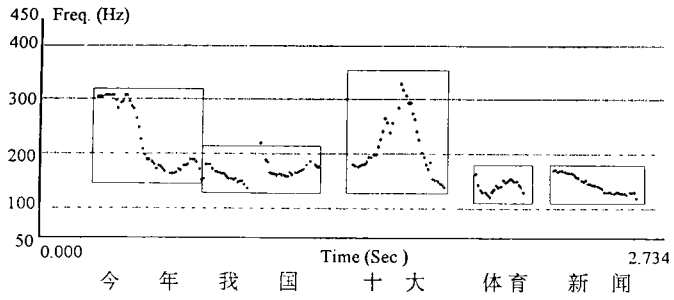

图 4　"今年我国十大体育新闻"的调域分析

我们将重音分成强重音、重音、中重音、轻音和弱轻音五个等级，统计了指定文本录音材料中的不同重音等级词的调域，并以统计结果为依据构成了合成系统的重音控制模式。

3. 自然语流中词内各音节的音长分布规律性不强

研究自然语流中词内音节的音长分布对提高合成语音的自然度很有益处。徐世荣认为汉语双字词大多数是后重[9]，后重就可能导致后音节较长。而王晶和王理嘉的结论是前长

规则[10]。我们统计了“北风与太阳”、“十大体育新闻”和“修复圆明园围墙”等三篇文稿的录音中双字词的音长分布，其结果如表 1:

表 1 双字词中音长分布统计结果

名 称	前长音节数目	后长音节数目
北风与太阳	41(59%)	28(41%)
十大体育新闻	72(46%)	84(54%)
修复圆明园围墙	24(44%)	30(56%)
合 计	137(49%)	142(51%)

由统计结果，我们不能得到简单的前音节长或后音节长结论。这是因为，双字词的音长分布不仅与音节位置有关，而且还与词重心位置以及词在语流中的位置有关。如呼吸群尾、分句尾或句尾的双字词通常后长。如图 5 所示，“围墙”和“修复”两词，在句中时是前长，而在句尾时后长。

在三字词的音长分析中，我们发现三字词的中间音节通常较短，末音节最长。在四字词及多字词中，有音节长短相间的趋势。

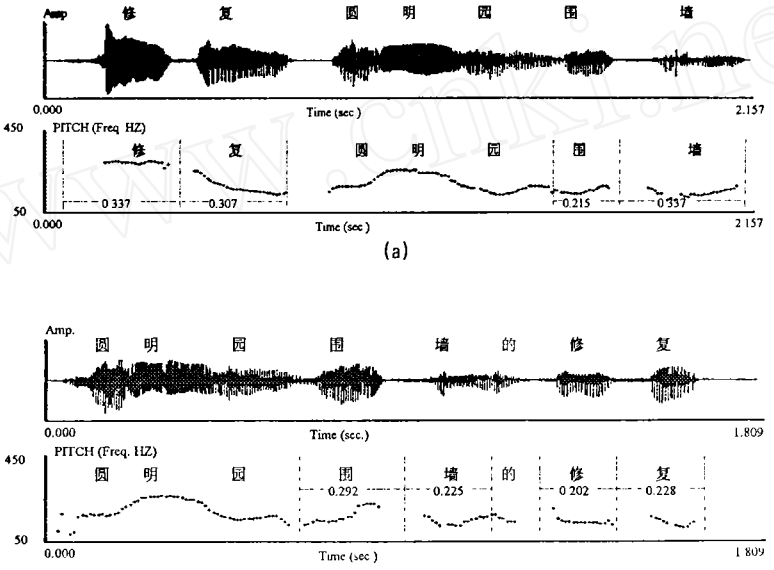

图 5 “修复圆明园围墙”和“圆明园围墙的修复”的音长分析

4. 在广播语言中词的调域低音线在呼吸群中呈下降趋势

通过对广播语言中词的调域分析，我们发现，在语流中不同重音级别的词通过调域高音线的变化来改变其音域，词的调域低音线有一共同的变化趋势，即：词的低音线在一个呼吸群中有明显的下降趋势，而每个呼吸群首词的低音线要回升。见图 6。

低音线的起伏与语流中的停顿和节奏是一致的。一个呼吸群通常表达一个相对独立的意思，是一个节奏单元，低音线在其内逐渐下降，其后通常有个小停顿。当一个呼吸群结束，低音线回升，表示一个新节奏单元的开始。

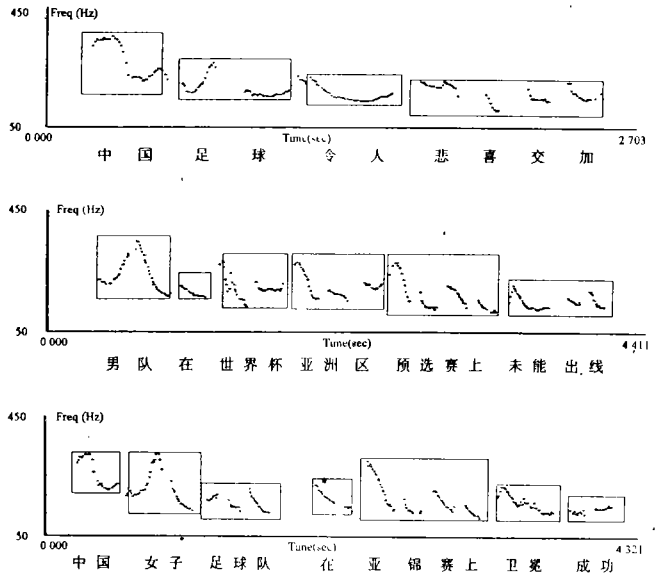

图 6　"中国足球令人悲喜交加，男队在世界杯亚洲区预选赛上未能出线，中国女子足球队|在亚锦赛上卫冕成功。"中调域低音线的走向。
(注:"|"是呼吸群边界符)

5. 句末低音线的升降和音域的伸缩是语气的主要表达手段

通过对陈述句、疑问句和感叹句的调域分析，我们发现句末低音线的升降和调域的伸缩是语气的主要表达手段。在陈述句句尾，低音线下降，调域缩小；在疑问句尾，低音线上升，调域也缩小；在感叹句中，低音线下降，而调域扩大。对上述变化，我们进行了初步统计，并以此为根据在韵律规则库中建立了语调模式。

四、汉语合成规则

PSOLA 技术为我们提供了对合成语音进行韵律控制的可能性，但如何进行韵律控制，才能提高合成语音的自然度，是汉语合成研究的核心问题。基于对广播语言的初步分析结果，我们建立了一个包含词调模式、重音模式和句调模式的韵律规则库。它能对一句话中各音节的音高和音长进行通盘规划，从而大大提高了合成语音的自然度。韵律规则库具体从以下五方面描述:

词的声调模式：设置汉语双字词、三字词及多字词的不同声调组合的声调模式。其中双字词和三字词在汉语中的出现概率最高，它们的各种声调组合模式都单独设定。多字词的声调组合方式很多，但出现概率不高，系统只为四声在首字、中字和末字分别设了几种不同的调式。所有这些音高模式的设定都是基于对新闻广播语言材料的分析结果。这些模式中的声调曲线都归一化到正常重音的调域。

词的音长模式：音长规则采用简单的后长规则，设置汉语双字词、三字词、四字词及多字词各音节的音长系数（音节音长／初始化时设定的标准音长）如表 2。

表 2 汉语双字词、三字词和四字词

字 序	1	2	3	4
双字词	0.95	1.00		
三字词	0.93	0.84	0.97	
四字词	0.90	0.77	0.86	0.93

在分析中发现随词中音节数的增加，音节音长普遍缩短，系统以四字词为基础，设定每增加一字，单个音节音长减 3%，即多字词音长缩减系数：$K=1-0.3\times(N-4)$，N 为词长。为保持汉语发音短、长间隔的基本节奏，设置首字音长为 $0.9\times K$，末字音长为 $0.93\times K$，中字音长按 $0.77\times K$ 或 $0.86\times K$ 交替变化。

词的重音模式：研究证明影响汉语重音的听觉感知因素有调域、音长和音强，其中音强的影响较小，系统只考虑词的调域宽度和音长的控制。系统设置五级重音模式，分别是：强重音、重音、中重音、轻音和弱轻音。正常重音调域由合成程序初始化时设定，用调域扩大一倍表达强重音；扩大 50%表达重音；减小 50%表达轻音；减小到 25%表达弱轻音。汉语有时也用音长变化作为重音的表达手段，调域和音长的变化对重音表达的作用是互补的，系统提供了音长调节符号，以便需要时改变词的整体音长。

系统根据统计数据在呼吸群中设置了缺省重音和音长分布形式，在没有重音和音长调节符时，自动设置重音和音长。如果在输入文本中加入适当的重音和音长控制符，可以进一步提高合成语音的自然度。

语调模式：系统设计将文本分成句，逐句进行文语转换。然后按树形结构将一个句子分成若干个分句；分句分成若干个呼吸群；呼吸群又分成若干个词。语调用句中各词的调域高、低音线的移动来描述。其中高音线的移动表示轻重音，低音线的移动表示节奏。系统设置了陈述句、疑问句和感叹句三种基本句型。

停顿模式：系统设置词与词之间的停顿为 10 ms；呼吸群间为 100 ms；分句间为 200 ms。陈述句后设 500 ms 的停顿，疑问句和感叹句后设 700 ms 的停顿。

五、合成系统评测结果

本系统参加了一九九四年国家高技术智能计算机系统专家组组织的汉语合成系统的评测。其中包括对语言清晰度和自然度的评测。清晰度测试采用国家标准测试方法（试行）。测试材料包括音节表、词表和句表各两张。清晰度测试包括音节清晰度、词清晰度和句可懂度。自然度测试只对句进行。假定播音员的发音的自然度为 10，分 10 级作主观评价。部分评测结果见图 7。其中系统 KX-PSOLA 是本文提出的基音同步叠加波形拼接汉语言语转换系统；KX-FSS 是一共振峰合成系统，它与 KX-PSOLA 系统采用同样的韵律规则库；TH-SPEECH 是另波形拼接合成系统；CELP 和 VQ-LPC 是两个基于线性预测的合成系统。

总共有 16 个测听人员。KX-PSOLA 的平均清晰度和自然度分别是 94.1%和 7.8。由图 7(a)看出，两个波形拼接系统 KX-PSOLA 和 TH-SPEECH 的平均语言清晰度比其他

合成系统高，而两个具有韵律控制的系统 KX-PSOLA 和 KX-FSS 的自然度比较高。KX-PSOLA 的清晰度和自然度都最高。合成系统的韵律控制不仅对自然度有重要影响，而且也影响清晰度。由图 7(b)可以看出，KX-FSS 的音节和单词的清晰度以及句可懂度逐渐提高。它清楚地表明了韵律控制对提高合成语音音质的贡献。

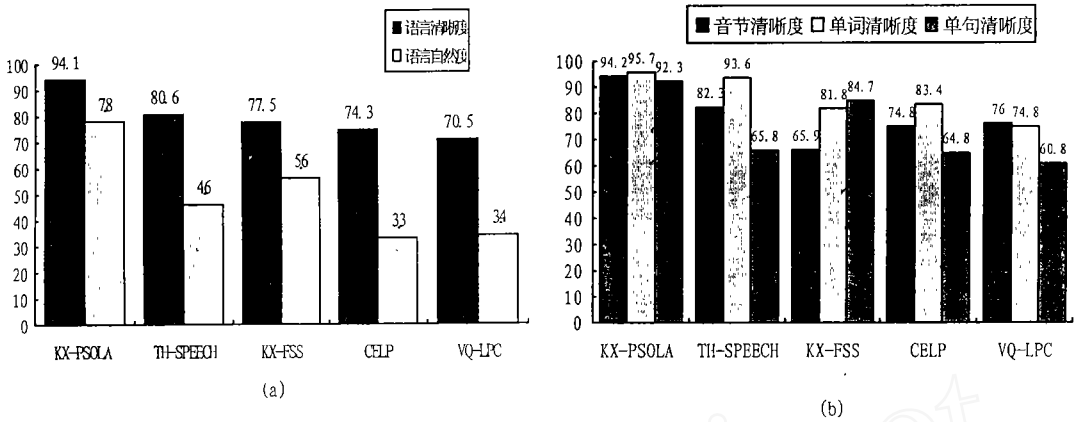

图 7　　汉语合成系统评测结果

六、总　　结

由上述分析看出，如果有一个适合的韵律规则库，波形拼接汉语合成系统能合成出高质量的汉语语音。基于 PSOLA 技术的波形拼接汉语文语转换系统具有两方面的优点：首先，系统的音库取自自然语音，保持了自然语音的音段特征，因此系统输出的清晰度高；其次，以对新闻广播语言的声学分析为基础建立的韵律规则库，较好的规划了一句话中各音节的音高和音长，而 PSOLA 技术又使系统能根据韵律规则去改变取自音库中的音节的基频曲线、音长等韵律特征，从而使系统输出具有良好的自然度。目前提高汉语合成语音质量的关键是韵律控制。继续深入研究广播语言的韵律特征，从而进一步完善合成系统的韵律规则，是我们下一阶段的目标。

参　考　文　献

[1] 吕士楠、齐士钤、张家骏. 汉语合成语音自然度的实验研究，声学学报，1994，19(1)。
[2] Charpentier F, Stella M. Diphone synthesis using an overlap-add technique for speech waveforms concatenation, Proc. Int. Conf. ASSP, 1986, 2015-2018.
[3] Hamon Ch, Moulines E, Eharpentier F. A diphone synthesis system based on time-domain prosodic modification of speech, Proc. Int. Conf. ASSP, 1989, 238-241.
[4] 吴宗济. 普通话语句中的声调变化，中国语文，1982 年，(6)：439-449.
[5] 吴宗济. 普通话三字组变调规律，中国语言学报，(4)：1984，70-92.
[6] 赵元任. 中国字调和语调，史语所集刊 4 本 2 分，1933，121-134.
[7] 沈炯. 北京话声调的音域和语调，北京语音实验录，1985，73-130.
[8] 沈炯. 汉语语势重音的音理，语文研究，1994 年，10-15.
[9] 徐世荣. 双音节词的音量分析，语言教学与研究，1982 年，(2).
[10] 王晶、王理嘉. 普通话多音节词时长分布模式，中国语文，1993 年，(2)：112-116.