第19卷第1期
1994年1月

声 学 学 报
ACTA ACUSTICA

Vol. 19, No. 1
Jan., 1994

合成言语自然度的研究*

吕士楠 齐士钤 张家騄

(中国科学院声学研究所,北京 100080)

1992年2月26日收到

摘要 在过去的十年中,中国科学院声学研究所建立了一个文语转换系统,它包括语音库,声调模型和基本合成规则. 无限词汇的汉语合成问题初步解决,但合成言语的自然度必须进一步改进. 我们对语言的音段特征和超音段特征对合成言语自然度的影响做了研究,结果表明影响合成言语自然度的基本因素是语言的节奏和协同发音. 本系统所采用的声调模式适合于单句合成,对于大于单句的语言单元的合成,必须十分仔细地控制语调才能达成高自然度. 本文介绍利用主观评价对合成语言自然度进行研究的方法和结果.

Study on the naturalness of synthetic speech

LU Shinan QI Shiqian and ZHANG Jialu

(*Institute of Acoustics, Academia Sinica*)

Received February 26, 1992

Abstract In the past 10 years a text-to-speech system including a phonetic library, tone model and basic synthetic rules had been established in IAAS. At present the speech synthesis of unrestricted vocabulary has been attained, but the naturalness has to be improved. The effect of segmental and supra-segmental feature of speech upon naturalness of synthetic speech had been studied. This study shows that the rhythm and coarticulation occupy a dominant position for improving the naturalness of synthetic speech. The tone model for synthesis of short sentence is suitable. But if the synthesis of larger linguistic unit than simple sentence is considered, the intonation must be controled carefully. In this article the subjective assessment method and the results of the study on the naturalness of synthetic Chinese will be represented.

一、前 言

在言语工程领域中,言语识别和言语合成是近二十年来特别受到重视的领域,因为它是人

* 声学所所长基金支持项目

和计算机之间最理想的接口. 言语合成和言语识别不同，因为任何语言，音素数量有限，要做到无限词汇合成在原则上没有问题，对汉语即使以音节作为基本合成单元，实施起来也不困难. 可以通过规则合成出各种音色的语音，使其具有个人特征，国外已有成功的报道[1]. 但要使合成言语听起来和人说话一样自然生动，言语合成与言语识别一样都面临着对人的言语从发声到感知一系列的认识不足的问题. 合成言语的自然度成为目前言语合成中的主要难点. 对合成言语的自然度尚未有确切定义和定量的表示方法，因为它是一个主观量，容易受到个人因素影响. 自然度粗略地可理解为合成言语与自然言语在主观感受上的相似程度. 在本实验中将合成言语自然度规定为五级：1. 很自然、2. 比较自然、3. 一般、4. 不自然、5. 很不自然.

利用这五级评价标准，分别由17位听音人对五种合成条件下的合成言语做了自然度评价，同时也对其可懂度进行了测量. 实验结果表明在单句和短文两个语言学层次上影响自然度的因素不同，对于前者主要是节奏和协同发音；对于后者除这两个基本因素外，语调有重要作用.

二、试验方法

1. 试验材料

试验材料是以国际音标委员会推荐的伊索寓言“北风与太阳”的汉语普通话文本制作的合成言语. 为了鉴别声学参数对合成言语自然度的影响，使用由中国科学院声学研究所研制的KX型共振峰语音合成器，制作下述五种条件的合成言语：

(1)由无调音节库取所需音节的合成控制参数，字调由固定模式给出.

(2)由无调音节库取所需音节的合成控制参数，字调由固定模式给出，音节和间隙长度按朗读的“北风与太阳”的自然言语调整.

(3)由无调音节库取所需音节的合成控制参数，音节、间隙长度和基频曲线按自然言语调整. 基频曲线由音高仪(VISI-PITCH)提取.

(4)用ILS数字信号处理通用软件从自然言语提取共振峰参数，音节时长与自然言语相同，字调模式与(1)条件相同.

(5)全部合成参数由ILS软件从自然言语提取.

制作上述合成材料需对合成控制参数进行大量编辑工作，这是用一个以C语言编写的KX型合成器合成语音参数编辑软件KX-HANDL完成的[2].

五种合成条件的合成言语试验材料参数序列形成五个版本，分别存入磁盘文件中. 为了进行单句可懂度和单句自然度评价，将每个版本分成18个短句，共90句，按随机排列构成测试材料的主体. 另外从每个版本中任选两句，共10句，随机组合后放在测听材料开始部分，以克服测听时的学习效应，这10句测试结果不统计. 五种合成条件的短文——“北风与太阳”全文和100句无规排列的短句组成全部测听材料.

2. 测听试验

测听试验是在测听室进行的. 测听室混响时间0.5秒，环境噪声＜35dB(A)，信号声压级约70dB. 听音人共17人，均受过语音学训练，大部为青年人，说流利普通话，无听力障碍.

测听试验分句子和短文两部分. 测定单句可懂度与自然度时，句子按无规顺序放音，每句

连放两遍，中间间隔2秒．句与句之间停顿24秒，供听音人记录全句每个音节并给出自然度评价．每25句为一组，每组开始和结束都有提示，并留有较长时间停顿．短文试验时只要求听音人作出对全文自然度的评价．测听时先将第一至第五种合成条件短文依次放音，然后再按无规顺序放音一遍，在两短文之间停顿两分钟，供听音人对刚听到的合成言语的自然度作出评价．

三、试验结果

不同条件下合成言语自然度评价的统计平均结果分别按“短文”和“单句”两种情况列于图1．

在“短文”中，条件(5)的自然度比其他条件有显著差别，至少高出0．6级，有近半数的听音人判定为“很自然”．条件(5)的合成参数全部是从“北风与太阳”的朗读自然语音中提取的．这表明要使合成言语达到高音质，听之自然，必需使合成言语的音段特征和超音段特征都接近自然言语，任何因素的偏离都会造成自然度下降．如在条件(4)中的语调特征，和在条件(1)(2)(3)中的音强及音节间的过度特征等的偏离都会造成自然度下降．条件(1)的自然度最差，而合成条件与它接近的条件(2)的自然度比它要高半级．比较条件(2)和(1)，其唯一差别是在条件(2)中将音库中取来的音节的音长和句中的意群、呼吸群间的停顿调整到和自然言语相符合，就可使自然度有显著改善，这一事实表明语音的节奏是影响自然度的最基本因素之一．值得特别指出的是条件(3)比条件(2)的自然度低，说明在利用音节库合成的言语中简单地加上从自然言语中提取的基频曲线对改善自然度没有作用．

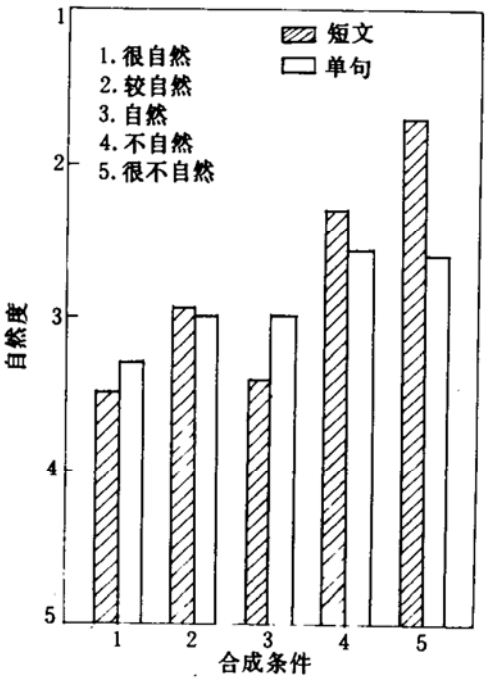

图1　不同条件下合成言语短文和单句的自然度

对于“单句”，其结果和上述情况不同．在五种合成条件下自然度的差异不如“短文”显著．“短文”自然度平均值上、下差1．8级，而单句仅差0．7级．另外在条件(2)和(3)之间，及(4)和(5)之间，几乎没有差别．这说明语调是语言学高层次上的特征，对单句的影响较小．所建立的声调模式和简单的语调模型对合成简单句是合适的，但不能满足合成由多个句子组成文本的要求．

可懂度试验结果表明单句可懂度都比较高，达到接近饱和程度，五种合成条件可懂度几乎相同，各句可懂度一般在97％左右，只有第10句为77．6％，第7句为88．4％和第15句为89．7％，这三句偏低．从单句测试中还可以看出，可懂度和自然度有一定关系，见图2．高可懂度是达到高自然度的必要条件．

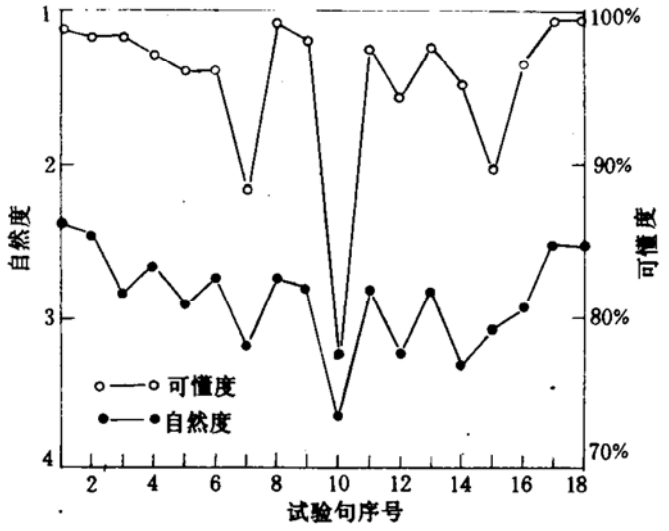

图2 单句的自然度和可懂度

四、讨 论

实验结果表明，将汉语音节机械地拼接起来所合成的文本的语音输出是可以听懂的，但自然度差. 如果使合成言语中音节的音长、音高、音强和共振峰值逐步与连续语言相接近，自然度也呈现逐步提高的趋势. 我们以“北风与太阳”这个文本为对象，具体分析合成条件(1)和(5)之间合成参数的差异.

在音长方面，音库中的音节是从同样语音环境中在该音节发阴平声时截取的. 音节长度的变化只与音节本身的结构有关[3]. 但在实际语流中的音长除音节结构外还会受到声调、语法功能和语意的影响. 从上述两个合成条件所对应音节的音长比的统计结果可以看出这种影响是很明显的，音节长度比可以从0. 3到1. 8的范围内变化. 在所统计的短文中缩短最显著的是“谁的本事大”中的“事”，因为“事”在这里读轻声，弱而短. 后续音节“大”是强调音节，按对比规则，将“事”压得更短. 被延长的音节是那些有强调意义的重读音节，如“他把袍子裹的越紧”中的“袍”，和“北风没了法子”中的“北”. 这种和语法、语义相关联，在音长上大幅度有规律的伸缩，形成不同节奏，增强了言语表达能力，也使人感到语言本身的活力. 用计算机产生出言语，要使其听之自然，首先必需在这个基本点上和自然言语取得一致.

关于音高及其变化在这两个合成条件中也有显著的差异. 由于从基频曲线上很难确定每个音节的调域. 图3是试验文本中第一句话从实际发音中测量的基频曲线和作者凭经验勾勒的调域.

由图可见即使在一个句子中，字调调域的变化也是经常出现的. 句子的主要成分调域宽，附加成分调域窄，重读音节调域要加宽，调域下限在句末有下降的趋势. 对全部短文的调域的统计结果得到调域的上限接近正态分布，平均值、最高值和最低值分别为180Hz、250Hz 和130Hz，变化范围接近一个倍频程. 下限平均值不计句末词时为114Hz，变化范围比上限窄，约为半个倍频程. 句末调域下限变化从80Hz 到96Hz 平均90Hz，变化不大，这也是由于文本中只

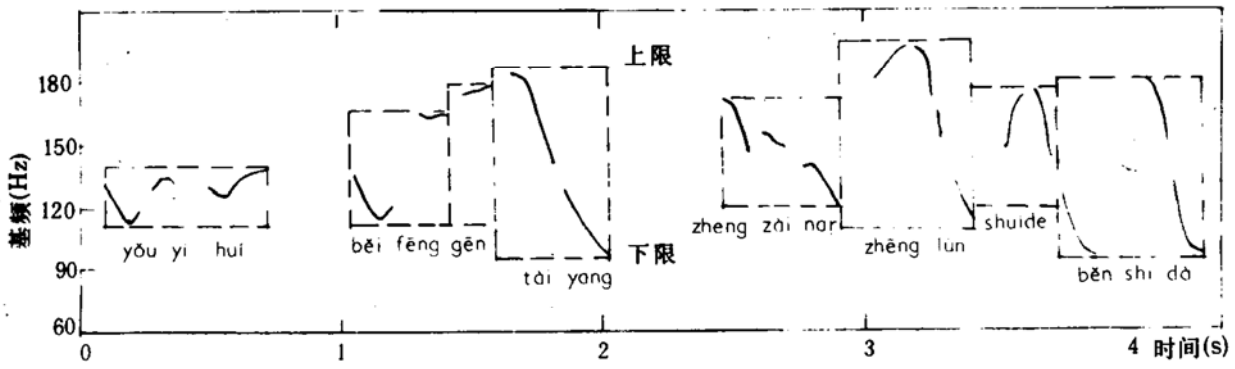

图3　“北风与太阳”文本中第一句话实际发音的基频曲线和调域估计

包含陈述句的缘故. 调域的宽度平均为73Hz，最小值为22Hz，最高值达160Hz，变化范围接近三个倍频程，调域变化十分显著. 与合成条件(1)比较，因为条件(1)采用了固定声调模式，没有这些变化，结果使合成言语缺乏条件(5)那样的抑扬顿挫，丧失了语言的音乐感.

这两种合成条件中的音强和共振峰频率方面的差异难以象音长和音高那样用统计数据做全面的阐述. 但我们发现在音节与音节的接合部，上述两个版本的共振峰控制参数有很大差别，这具体反映了协同发音的作用在合成过程中是被合成系统如实地模拟了，还是被忽略了. 实验证明这种协同发音的作用和音长调节一样，对汉语文语转换系统输出语音的自然度有重要的影响. 图4举出上述文本中词组“有一回”为例子，作简要说明.

在图4(a)表示合成条件一，没有考虑协同发音，音节参数机械地拼合后，共振峰频率曲线(F_1,F_2,F_3)在音节间有跳变. 特别是F_2，因为“有(you)”的F_2的终点很低，而“一(yi)”的F_2很高. 后续音节“回(hui)”中声母(h)没有自己固有的能量集中区，它的F_2在孤立音节中取决于后续元音. 韵母(ui)的F_2取由低向高伸展态势，起点也很低. 简单拼接的结果，在音节间必然有跳变. 共振峰振幅曲线(ALF,A_1,A_2,A_3,A_4)在音节间出现一个明显的谷. (b)表示合成条件五，共振峰频率曲线均以圆滑形式过度，这意味在发前一音节的末尾音位时，舌位已开始向发第二个音节开头的音位逐步移动，体现了发音器管协同工作的实际生理过程. 共振峰振幅曲线在前两音节之间虽有下降，但并不明显. 而后两音节之间因为有辅音/h/，振幅曲线出现低谷是正常的. 此外在基频曲线上也不相同：(a)中的基频曲线变化是生硬的；而实际发音中，在声调协同发音的作用下的基频曲线如(b)所示，呈圆滑过渡状. 因为这是一个三音节词组，次音节又轻读，这个去声调只看到有一点凸起，和典型的去声相差甚远，它只起到了承前启后的过渡作用. 这个词组在句子中是一个附加成份，如图3所示，调域狭窄，但这26Hz 的基频变化范围已完全达到合成音节的辨义效果，而且自然度比条件(1)高得多.

除了这种在音节接合部的显著的差别外，在音节内部我们也观察到两个版本之间的差异，如“厉害”这个词，在文本中“它刮得越厉害”，其中“害”是轻声，除了音长和音强的变化外，在合成条件(5)中，可以看到元音/A/被央化了，“害(hai)”实际已读成接近“嘿(hei)”. 这种元音减缩(Vowel reduction)现象也是汉语合成中不可忽视的.

比较两种合成条件下声学参数的差别，在某种程度上也反映了孤立音节和连续言语之间的差别，其结果对言语工程技术具有普遍的指导意义.

言语是人在脑的统一控制下通过发音器官的协调动作发出来的声音，因此在声学参数之

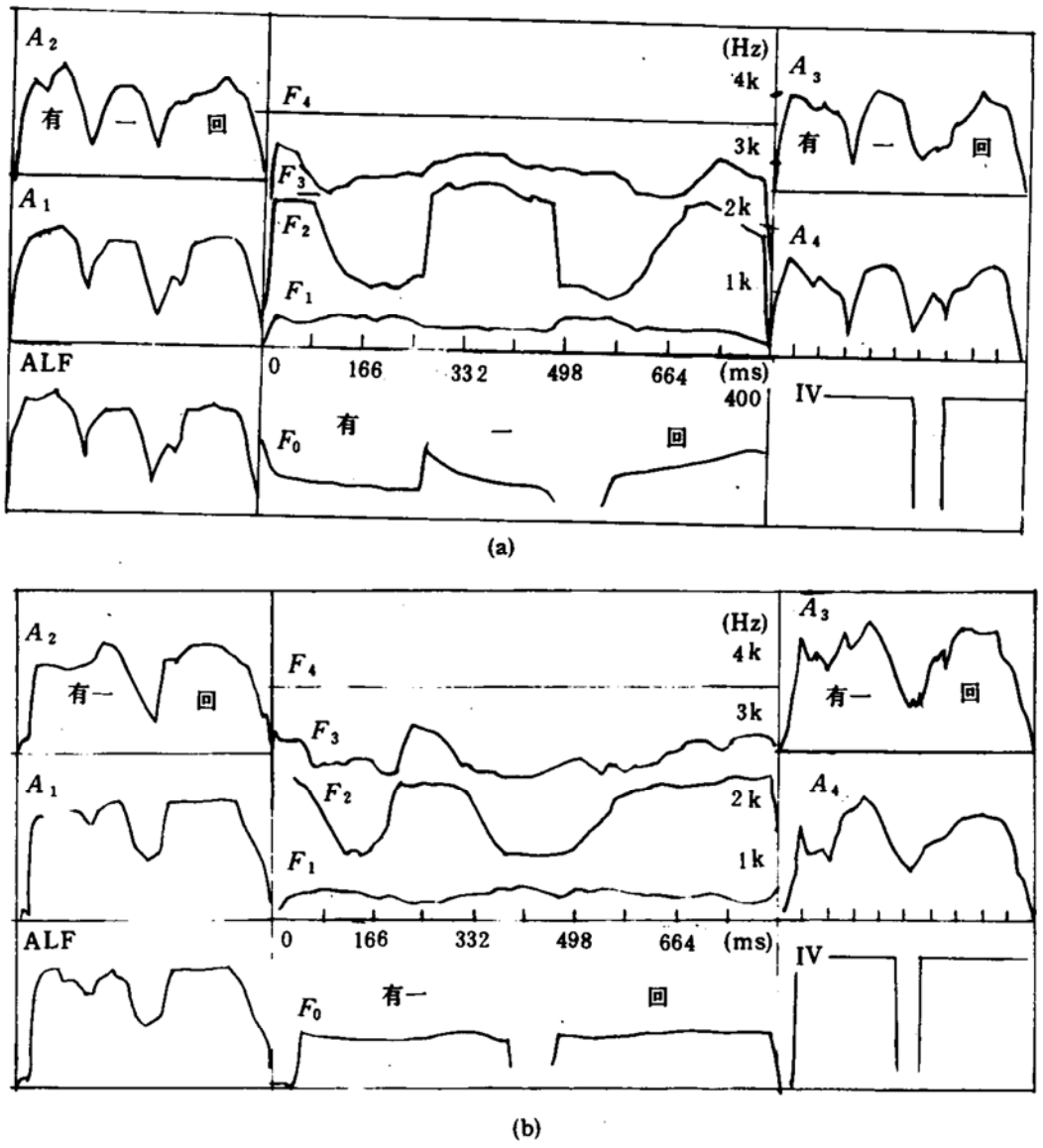

图4 词组“有一回”的合成控制曲线(变化范围:振幅0－80dB 基频 0－400Hz)
(a)合成条件一、(b)合成条件五

间必然包含着有机的联系.在目前大多数言语合成系统中由于失去了这种联系,使得合成言语的自然度受到损伤.这种联系包括历时的和共时的.这种影响使得单纯通过某一个声学参数的调整,不一定对提高自然度有作用.正如实验结果所表明,合成条件(3)的文本输出语音的自然度比条件(2)低.与此相对照的是合成条件(4)经过同样处理得到的合成条件(5)的自然度却有显著的提高.其声学参数方面的差别,对于条件(4),除基频曲线外,参数都是从自然发音中提取的,参数间历时的和共时的关系除基频外都保持下来了.将基频曲线恢复后,自然度就明显提高.而对于从音库中取得共振峰参数做了音长调整的条件(2)来说,在协同发音,音强和音变控制方面正如前面指出和自然言语还有许多不同,这时即使将基频曲线调整到和实际发音符合也不能达到预期的目的.

五、结 论

综上所述，试验证明影响合成言语自然度的基本因素是节奏和协同发音，通过对音节音长和共振峰参数的适当控制，可以达到改善合成言语自然度的目的. 语调对自然度影响比较复杂，从自然言语中观察到基频值的相对变化远比其他参数大. 声调域和中心位置都与文本内容密切相关，因此，只有在语言理解的基础上才能实现对合成言语的语调的完善控制. 试验结果表明，语调对文本语音输出的自然度的影响比单句大. 因为汉语有独立音节个数有限的特点，在汉语文语转换系统中[4]，采用以音节为基本合成单元的方案，可以提高清晰度和简化规则系统是可取的. 但如果象处理书面文字一样，以排版方式实现文语转换就必然造成自然度降低. 因此音节库中以音节为单位的合成参数必须是可控制的，在合成句子或文章时，按音节所在语境必须对其合成参数由协同发音规则作出灵活调整. 控制语音的节奏，做到音节间合成参数的合理过渡是最基本的，语调模式对文语转换系统具有特别重要的意义. 为满足上述要求，共振峰语音合成器具有明显的优势.

感谢北京师范大学周同春教授和该校同学协助进行评价试验.

参 考 文 献

[1] D. H. Klatt and L. C. Klatt, "Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers", *J. Acous. Soc. Am.*, **87**(2), Feb. 1990

[2] 吕士楠，齐士钤，周同春，"语音合成技术在声学语音学研究中的应用"，第五届语音图象通讯信号处理会议论文集，94－97.

[3] 齐士钤，张家骅，"汉语普通话辅音音长分析"，声学学报，7，(1982)，No. 1，8－13.

[4] 张家骅等，"汉语文语转换系统的研究"，信号处理，**5**(1989)，No. 1，1－17.