语音合成自然度的客观度量实验研究

姜涛 [1,2], 吴志勇 [1,2], 蔡莲红 [1,2]

(1. 清华大学深圳研究生院, 清华大学-香港中文大学媒体科学、技术与系统联合研究中心, 深圳 518055;
2. 清华信息科学与技术国家实验室(筹), 清华大学计算机科学与技术系, 北京 100084)

摘要：随着语音合成技术的发展，合成语音的音质和可懂度不断提高，而如何进一步提高其自然度成为语音合成方面的重要问题。本文总结了近年来主流的波形拼接式语音合成系统在自然度方面影响主观感受的四类问题，分析了它们对自然度主观感受的影响、产生的原因以及进行测试和改进的方法。考虑到不自然点的定位与检测对于自然度问题发现与改进的重要作用，本文针对其中与语音相关的两类问题，音调连续性和结尾自然度，分别提出了基频连续性和波形包络在停顿前的连续性两种不自然点的客观度量方法，并在此基础上进一步设计了语音自然度中不自然点的自动定位与检测算法。实验数据表明，人工听辨发现的音调不自然点都可以通过基频不连续点检测出来，并且在较为挑剔的自然度评判中，该算法有较高的准确率；而通过波形包络在长停顿前的不连续点检测，可以非常精确地发现结尾不自然的点。在语音合成系统的评测和改进工作中，本文提出的客观度量和不自然点自动检测方法提供了比人工听辨更可信的数据参考，具有较高的实用价值。

关键词：语音自然度；客观度量；基频连续性；波形包络

1 引言

语音合成是将文字符号转换成声音，让计算机像人类一样说话。而如何不断提高自然度、清晰度和可懂度是语音合成的一个核心问题。语音合成主要分为三种方法：基于声道模型的共振峰参数合成，基于波形拼接的拼接式语音合成，以及基于 HMM 模型的可训练的参数合成。不管采用何种语音合成方法，其自然度与自然语音相比都有着很大差距。随着语音合成技术的发展，合成语音的音质和可懂度不断提高，而如何能够进一步提高合成语音的自然度成为语音合成方面的重要问题。

基于波形拼接的拼接式语音合成是近年来语音合成方面的主流方法，也是清华大学 Crystal 中文语音合成系统所采用的合成方法。其基本原理是录制大规模语料的音库并切分为语音单元，通过文本分析和选音算法从音库中挑选合适的语音单元，最后拼接得到合成语音。由于这种合成语音中的每一个基元都来自自然语音，因此在清晰度、可懂度方面有比较好的表现，但是由于每一个基元还带有上下文的韵律特性，直接拼接后的合成效果不够稳定，在整体自然度上仍然需要提高。

要改进合成语音的自然度，首要问题是进行语音自然度的评测以及不自然问题的发现和分析。目前合成语音自然度的评测主要依赖于人的主观感受，普遍采用 MOS 主观印象打分[1][6]和多个合成系统间的对比测试[6]等，其优点是直接反应了人对语音自然度的主观感受，但是也有着费时费力和缺乏灵活性的缺点，而且由于评测结果受人的主观影响比较大，每次评测结果相对独立，可比性和重复性较差。为了给主观评测的同时提供更客观可信的评分参考，在客观评测方面已经一些研究成果。目前针对合成语音自然度的客观评测方法主要有：根据分段覆盖所导致的拼接代价评测[2]，根据语音参数距离评测[3]，根据人耳听觉特性进行评测[4]等方法。另外，初敏等人对韵律与自然度的关系进行了研究[5]。以上这些评测方法对于语音自然度的评分基本都与 MOS 主观印象打分做了对比，有比较高的相关性。

但是以上这些客观评测方法，或者需要相应的自然语音进行参数距离计算，这限制了测试语料是比较难搜集的；或者是对语音整体进行自然度的分析，没有从具体算法的自然度改进上提出建议。因此目前已有的客观评测方法没有很高的实用价值。而影响语音自然度的原因有很多，只有对各种不自然的问题进行具体的分析，才有助于算法整体的改进。

本文在清华大学 Crystal 中文语音合成系统的基础上，首先采用人工听辨的方法总结了影响自然度主观感受的四种主要因素，说明了语音自然度中不自然点的定位和检测对于改进合成算法的重要性，并且针对语音本身的参数比较了原始合成语音和人工调整选音后相对自然的语音，从基频和波形包络两种客观数据上提出了自然度的客观度量方法，以及语音不自然点的自动检测方法。

2 影响自然度主观感受的因素

本文采用人工听辨的方法，对现有语音合成系统经常出现的不自然问题做了归纳，总结了四类影响自然度主观感受的因素：读音、音调错误，韵律结构不好或停顿错误，音调连续性不好，长停顿前语音边界（即语音结尾）自然度不好。下面将对每一种因素对自然度主观感受的影响、产生的原因、以及可能的改进方法进行介绍。

2.1 读音、音调错误

受到听音人先期知识的影响，明显的读音错误（正确和错误读音不容易混淆）通常都能在人工听辨中发现，并且被归为可懂度的问题。在 MOS 自然度主观印象打分中，是不考虑读音错误的[6]。而对于音调错误和混淆程度较高的读音错误，由于听音时间和次数的限制，很难被听音人发现，但因为人在听音的时候同时在理解语音的含义，错误的读音和音调与听音人潜意识里预期的读音和音调不同，这就对语音的自然度产生非常大的影响。

造成读音、音调错误的原因有两点：制作音库时切词和标注的错误，导致音库中包含了错误的样本；或者由于合成算法前端文本分析算法（包括中文分词、多音

字处理等）出错，导致字音转换错误，而出现的问题也可以在字音转换的结果中直接反映出来。要解决这一类不自然的问题，就必须优化音库和前端文本分析算法，保证字音转换结果的正确性。

2.2 韵律结构不好、停顿错误

在汉语中很多情况下不同的韵律结构意味着不同的语义，而好的韵律结构可以在很大程度上增加语言的表现力，也是自然语音的重要特征。如果合成算法中韵律预测的结果出现了偏差，就会导致合成语音中出现错误或者过多、过少的停顿，直接影响自然度。

造成这类问题的原因可能是由于韵律预测算法直接造成的，也有可能是由于韵律预测依赖的分词算法造成的，而出现的问题也可以在韵律预测的结果中直接反映出来。要解决这类问题，就必须优化中文分词和韵律预测等相关算法，保证韵律预测结果的正确性。

2.3 音调连续性不好

语音是否流畅自然主要体现在音调连续性上，而音调连续性是影响语音整体自然度的最主要因素，而且在拼接式语音合成中经常出现这类问题。因为汉语是声调语言，而且前后声调组合时还会产生音调的不同变化，音库录制的不同样本之间很难做到音调一致，拼接后往往会出现音调连续性不好的问题，使合成语音听起来有很多跳变的点，严重影响自然度。

造成这类问题的可能原因也复杂，中文分词、韵律预测和选音算法都会影响音调连续性。这类问题也只能在语音合成的最终结果中才能体现出来，而选音算法是生成合成语音的直接模块，要解决这类问题，除了要综合考虑前端的文本处理算法以外，主要是对选音算法进行优化，提高音调连续性。

2.4 长停顿前语音边界自然度不好

语音结束时实际上是声带逐渐停止震动的过程，如果长停顿前语音边界没有做专门的处理，就会给听音人一种戛然而止的感觉，直接影响自然度。造成这类问题的原因在于选音算法，必须保证选中的音节样本同样来自于韵律结构结尾处，才能保证该处的自然度。

综合以上四类问题，通过对合成语音中的不自然点进行具体分析，可以从中文分词、韵律预测、字音转换、选音等算法上，有针对性地改进自然度。但是这也意味着，MOS 自然度主观印象打分这样的自然度评测方法只是从整体上对合成语音进行打分，对于提高自然度的合成算法改进工作作用不大。而为了从语音自然度上改进合成算法，首先需要对合成语音中的不自然点进行定位与检测，这是分析问题和解决问题的前提条件。

3 不自然点的定位与检测

在语音合成的评测与算法改进工作中，要发现合成算法存在的自然度问题就需要首先进行不自然点的定位和检测。而不自然点的定位与检测的现有方法主要是人工听辨。虽然人工听辨的方法可以精确定位不自然点，并且很容易进行后续的问题分析，但是往往需要花费大量的时间进行听音，并且在长时间听音后很可能精神疲劳或者适应合成语音，而不能正确定位不自然点。因此，在不自然点的定位与检测时，基于客观参数的度量自动进行不自然点的定位与检测的方法可以在很大程度上简化人工的工作，并且使在较大规模测试集上进行实验成为可能，并且如果需要高的精确度，则可以在自动检测的结果上再进行人工听辨筛查。

前面提到的影响自然度主观感受的四类因素中，读音、音调错误和韵律结构不好或停顿错误这两类问题都出在前端文本分析的算法上，在形成最终语音前的文本分析结果中就已经体现出来，实际上不需要对语音结果的分析。并且这类问题的检测必须依靠语言学规则才能检测出来，在当前语境下错误的读音或者韵律在其他语境下反而可能是正确的，比如发音“dei3dao4”在“火势得到控制”中是错误的，但是在“我得到学校去了”中却是正确的。因此本文提出的不自然点的客观度量方法只针对音调连续性和长停顿前语音边界自然度这两类问题。

3.1 音调连续性的客观度量

音调的连续性主要是前后音节的音调变化体现出来的，而语音音调的变化最直接体现就是语音基频的变化。虽然人在听音时，音调的感知是一个复杂的过程，相同的基频变化在不同的语境中会有不同的音调感受，但是音调的连续性感受与基频曲线的连续性具有很高的相关性。

人在发音过程中，声带的震动变化是一个渐变的过程，而且前后音节的基频一般也是连续的，基频产生突变的点一般会给人带来不流畅的感觉。因此可以通过基频曲线的连续性来度量音调连续性。

在语音信号处理中，尚没有算法能够做到基频的精确提取。本文利用 praat 软件对语音的基频信号进行提取，并进行个别点的人工校准。以“惠州大亚湾石化区中海油炼油基地”与“现场冒出黑烟”两段文本为例，本文对比了原始合成语音和人工调整选音语音的基频曲线，如图 1 和图 2 所示.其中红色曲线表示合成语音，蓝色曲线表示对比语音。

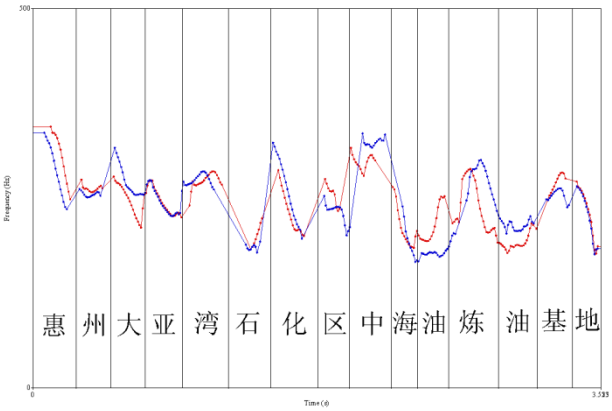


图 1“惠州……”合成与人工调整后语音的基频曲线

在“惠州大亚湾石化区中海油炼油基地”一句的合成语音的人工听辨标注结果中，“中”字、“海油”之间、“油炼”之间都有比较明显的音调不自然，而且以“炼”字最为明显。在图 1 中也可以观察到红色曲线的相应位置在短时间内出现了明显的波动或者不连续，而“炼”字前半部分出现了基频的一段突然降低，与前后基频曲线都不连续。而在人工调整后的语音中，对于这些标注的不自然点进行了选音调整，使其听起来流畅自然。在图 1 中也可以观察到蓝色曲线相应位置的连续性比调整前有了提高，而“炼”字原本出现的基频波动也消除了。

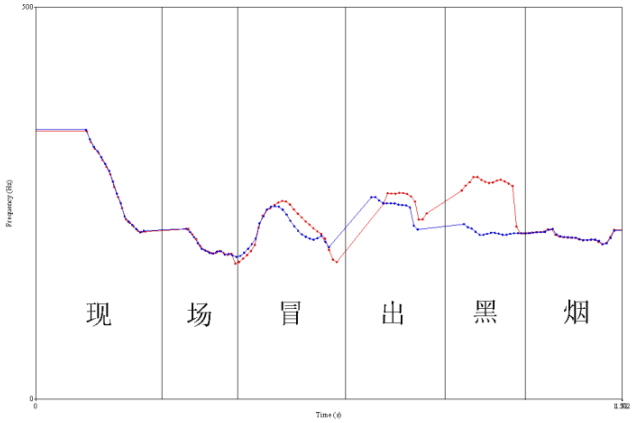


图 2 “现场……”合成与人工调整后语音的基频曲线

在“现场冒出黑烟”一句的合成语音的人工听辨标注结果中，“冒出”之间、“出黑”之间、“黑烟”之间有非常明显的不自然，而人工调整时将“出”字和“黑”字的选音进行了修改，调整后“冒出”和“黑烟”的语音样本来自于音库中已经存在的“冒出”和“黑烟”两词。对调整后语音的人工听辨发现“冒出”和“黑烟”之间的不自然被消除了，但“出黑”之间仍然有较高的音调跳变。在图 2 中也可以观察到相应位置基频曲线连续性上的变化。

对于离散信号，一般采用一阶差分与二阶差分来分析信号的变化。由于基频信号在时间上并不是均匀分布的，而且部分语音段是没有基频，所以求得的基频点中除了包含一个基频值以外，还包含一个时间值。因此本文在求基频曲线一阶差分和二阶差分的时候，都将基频差分数值除以时差的数值。定义基频曲线上两个相邻时间点(t_1, p_1)和(t_2, p_2)，其中 $t_1<t_2$，那么在 t_2 时刻的时差：

$$\triangle t = t_2 - t_1 \tag{1}$$

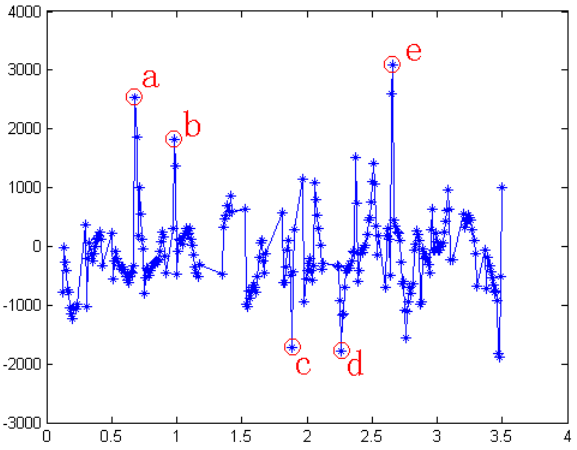
在 t_2 时刻的一阶差分：

$$diff(t_2) = (p_2 - p_1) / \triangle t \tag{2}$$

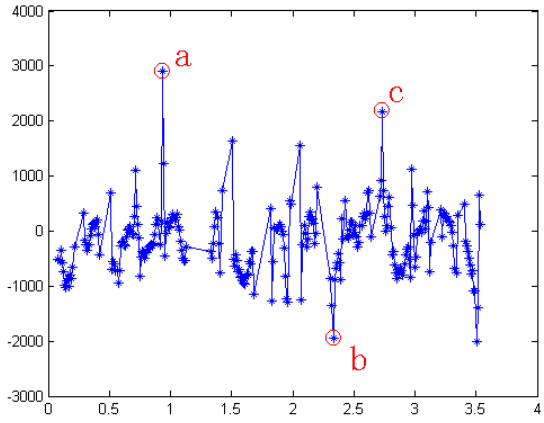
在 t_2 时刻的二阶差分

$$diff2(t_2) = (diff_{t2} - diff_{t1}) / \triangle t \tag{3}$$

根据以上定义可以求得基频曲线的一阶差分，“惠州……”一句合成语音和人工调整后语音的基频一阶差分曲线如图 3 所示。



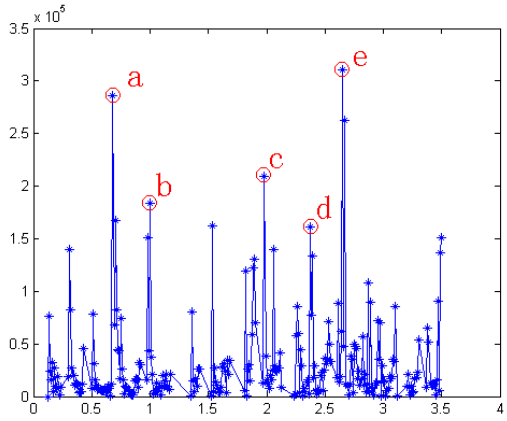
a. 合成语音



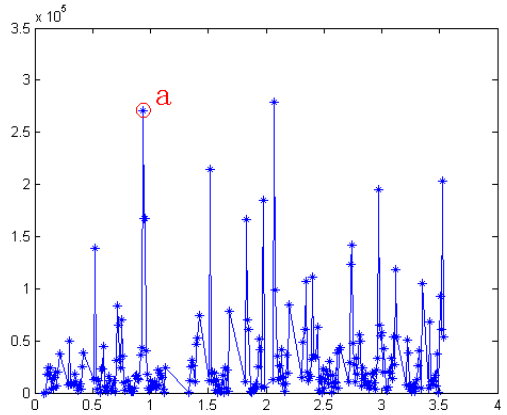
b. 合成语音

图 3 “惠州……”基频一阶差分

从一阶差分曲线可以看出，人工调整后的语音中较大的一阶差分值个数比合成语音更少，而且总体来说一阶差分的绝对值也更小。一阶差分表现出在该时刻基频变化的大小，但是在正常的声调变化中也会出现一阶差分较大的情况，对于一阶差分很大的情况，如果相邻点的一阶差分同样很大，那么在这个时刻基频有一个较大的变化趋势，是正常的声调变化，而不是不连续的跳变。二阶差分的绝对值表现了基频变化率与相邻基频变化率的差，可以排除正常声调变化的点。“惠州……”一句合成语音和人工调整后语音的基频二阶差分绝对值曲线如图 4 所示。



a. 合成语音



b. 合成语音

图 4 “惠州……”基频二阶差分绝对值

由图 3 和图 4 可以看出，综合考虑基频一阶差分和二阶差分可以比较好的定位图 1 中基频不连续的点，只有当两个值的绝对值都比较大时，基频也就不连续。例如对合成语音来说，图 3-a 中的 a、b、c、d、e 五点的绝对值较大，同时在图 4-a 中的值也较大，因此可以被检测为不自然点，对应了“惠州……”句中“亚湾”之间、“湾石”之间、“中”字、“海油”之间、“油炼”之间这五个位置，将人工听辨发现的不自然点都检测出来。而对于人工调整后的语音来说，图 3-b 中的 c 和 d 两点，在图 4-b 中就因为比较小而被排除掉了，而且在人工调整后的语音中也确实没有发现不自然点。

另外考虑到人在说话时，在韵律短语边界，会有比较明显的基频重设的现象；也即停顿后呼吸的调整使停顿后基频一般比停顿前高。这种停顿前后基频的不连续并不影响自然度的主观感受，因此为了让自动检测的结果更符合自然度的主观感受，需要设定时差阈值，排除长停顿前后的基频不连续点。例如图 3-a 中的 e 点，即“湾石”之间实际上人工听辨的结果是自然的，而此处是韵律短语边界，合成语音中设置的停顿，利用时差阈值可以排除该点。

同时我们也注意到，“亚湾”之间的音调变化是自然的，但是合成语音和人工调整后的语音通过上述检测方法都被错误地判定为不自然。这是因为这两个音节的音调之间原本就有一个跳变，即使在自然语音（人工调整后“大亚湾”一词已经取自同一词组样本）中基频曲线也不是连续的。受到先期知识的影响，听音人对这类语义正确的音调突变容忍度较高。对于测试句语法语义的分析超出了本文的研究范围，因此在后面的实验中不再具体讨论。

为了使算法适用于不同合成系统和语音，本文采用差分门限比率 α 和时差门限系数 τ 作为算法的基本参数，通过这两个参数以及语音自身基频的统计数据来求得一阶差分、二阶差分和时差的门限值。令 $P_{\max}$ 为该语音段的最高基频，$P_{\min}$ 为该语音段的最低基频，那么如果某一时刻的一阶差分满足

$$|diff| > (P_{\max} - P_{\min}) \bullet \alpha \bullet P_{\min} \quad (4)$$

二阶差分满足

$$|diff2| > (P_{\max} - P_{\min}) \bullet \alpha \bullet P_{\min}^2 \quad (5)$$

并且时差满足：

$$\triangle t < \tau / P_{\min} \quad (6)$$

则该时刻的基频曲线不连续，可以认为该点是一个音调不连续点。

3.2 长停顿前语音边界自然度的客观度量

对于出现这类问题的合成语音，在语音结尾都缺少一个声音逐渐降低的过程，振幅瞬间从一个较高的值降为 0。以“目前，消防车仍在作业”一句为例，合成语音经过人工辨识发现了“前”和“业”两处不自然的结尾，其波形如图 5 所示。

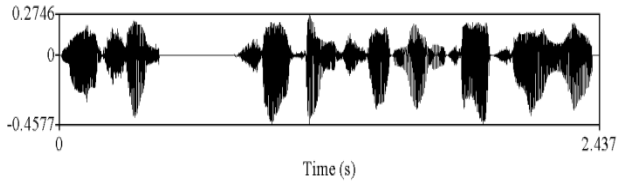


图 5 “目前……”合成语音波形图

从图 5 可以看到在中间的长停顿和语音最后，波形有明显截断的痕迹。经过人工调整选音后，将“前”和“业”的选音改为音库中原本就处于结尾处的样本，调整后的语音波形如图 6 所示。

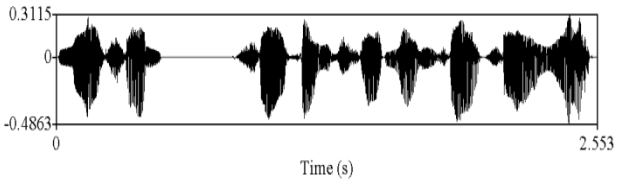


图 6 “目前……”人工调整后语音波形图

从图 6 中可以发现，调整后的语音在中间的长停顿和语音最后，振幅和声强逐渐降低，与结尾自然度的主观感受一致。

振幅和声强的变化有多种分析手段，可以通过振幅曲线或声强曲线，也可以利用波形包络进行分析。但是一般的振幅曲线因为是从声强曲线得到的，而声强曲线的计算过程中在时域上进行了加窗运算，因此振幅曲线在语音边界处仍然是连续的，并不能正确表现结尾处的振幅和声强突变。使用 praat 软件计算得到“目前……”一句的振幅曲线如图 7 所示

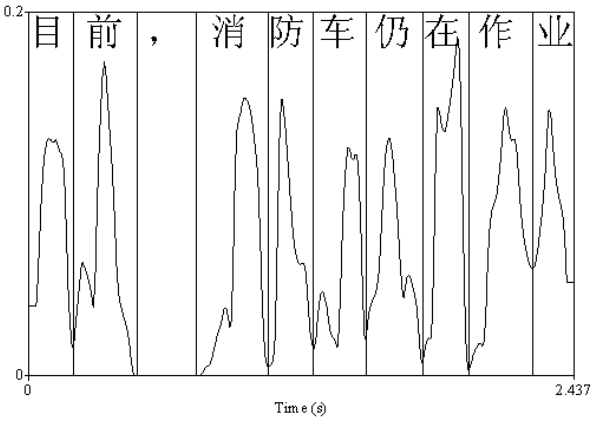


图 7 “目前……”振幅曲线

从图 7 中我们可以看到，中间的间隔前振幅曲线是平缓下降的，与波形图的明显截断有很大区别，因此并不适合用来做结尾自然度的检测。相比而言，因为波形包络曲线是由波形直接计算得到，并没有在时域上进行加窗运算，因此能够更好地表现结尾处振幅和声强的突变。

但是对于语音这样包含复杂频率成分的离散波形数据，尚没有一种好的算法准确求出波形包络。在以往的研究中，波形包络一般就是只取波形数据中的极值点（即波峰和波谷）。但是这样一来就会保留了过多的共振峰极值点和清辅音的噪声极值点，这样获得的包络线仍然具有很大的波动性，每一点差分绝对值都相对较高，难以满足结尾不自然点检测的要求。本文设计了一种迭代算法，通过设定了曲线分辨率 w 和二阶差分上限 a 两个参数，不断降低包络曲线的波动性，从而求得语音片段的相对平滑的波形包络。

令波形曲线离散点集为 S，包络线离散点集为 S_e。沿用基频曲线中的一阶差分与二阶差分定义，则包络线上任意 t 时刻的点 $(t, s) \in S_e$ 都满足下列条件：

$$\triangle t > w \quad (7)$$

$$(t, s) \in S \quad (8)$$

$$diff_S(t - \triangle t) > 0 \text{ 且 } diff_S(t) < 0 \quad (9)$$

若 $diff_{Se}(t-\triangle t)<0$ 且 $diff_{Se}(t)>0$，

则 $\left|diff2_{Se}(t)\right|<a$ (10)

具体的算法步骤如下：

1) 遍历波形数据，将采样点数据改为其绝对值（若不做此修改，则可以用类似的方法求波形的上包络线和下包络线）；

2) 再次遍历音频数据，保留所有的极值点数据（不小于前后的数值），其他置为空；

3) 遍历所有非空数据，对于所有极小值点，计算前后一阶差分相减的绝对值，若该值大于二阶差分上限 a，或者前后时差之一小于分辨率 w，则将该点置为空。在遍历中若有置空操作，则重复该步骤。

经过上述算法，求得“目前……”一句的波形包络如图 8 所示：

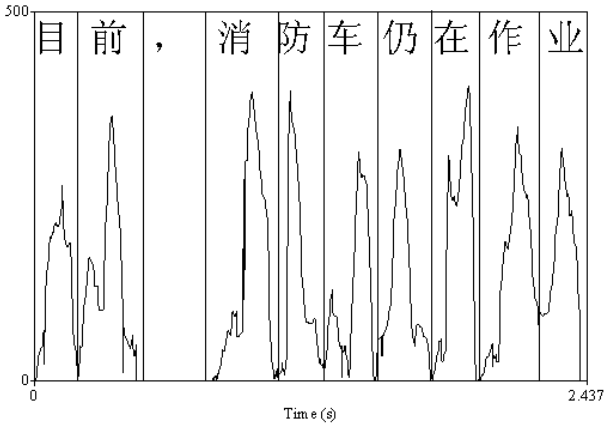


图 8 “目前……”波形包络图

在波形包络的基础上，类似于基频不连续点的检测方法，本文采用一阶差分门限率 β 与时差门限系数 τ 来检测不自然的语音结尾。令 $E_{\max}$ 为波形包络最大值，$P_{\min}$ 为该语音最低基频值，那么当某一结尾时刻的一阶差分满足：

$$diff > E_{\max} \bullet \beta \bullet P_{\min} \quad (11)$$

并且时差满足：

$$\triangle t < \tau / P_{\min} \quad (12)$$

则该结尾是一个不自然的结尾点。

4 实验结果及分析

4.1 实验过程

本文从 1998 年人民日报、2011 年 7 月互联网新闻以及旅游景点介绍中，随机抽取的长度从 10 到 150 不等的完整句子，并且筛选只包含常用汉字和阿拉伯数字（无生僻字以及外文字符等需要文本正则化的内容），最后保留 20 句作为最终的测试语料。采用的语音合成算法基于现有比较成熟的清华大学 Crystal 中文语音合成系统，其中包括女播音员录制的标准发音的普通话音库。

实验数据除了采用直接的系统合成结果，还采用了人工调整选音结果后感觉比较自然的语音。在人工调整选音结果时主要针对如下两种情况进行修改：

1、词组在音库中已有的，调整选音到已有词组；

2、读音或音调错误，调整正确的发音；

3、韵律结构内部连续性和韵律边界自然度不好的，调整使其听起来自然；

需要说明的是，这里手工调整的结果在一定程度上提高了合成语音的自然度，主要是为了与直接合成语音进行对比，在修正一部分不自然问题的同时可能会在一些细节上引入其他问题，但是在总体的自然度上要明显高于人工调整前的合成语音。

本文首先对实验数据进行人工听辨，标注其中的不自然点，然后利用前面说明的客观度量方法进行不自然点的自动定位和检测。

4.2 实验结果分析

本文采用基频差分门限率 α =7%，波形包络差分门限率 β =30%，以及时差门限系数 τ =5 作为算法参数对合成语音与人工调整后的语音进行不自然点的自动检测，实验后统计的算法准确率与召回率如下表所示：

表 1 不自然点客观检测算法的准确率与召回率

人工标注不自然点数：简写为 T

准确率：简写为 P

召回率：简写为 R

	不自然点	音调不连续			结尾不自然		
	T	*T*	*P*	*R*	*T*	*P*	*R*
合成语音	41	26	76.47%	100%	6	100%	100%
对比语音	14	11	91.67%	100%	0	--	--

从上面的实验结果可以看出，由于结尾不自然这种情况在波形包络上表现非常明显，针对这类问题的客观检测几乎可以做到完全正确。

而在音调不连续这种情况中，由于本文采用的差分门限值比较低，虽然数据集中的不自然点都被算法检测出来，但是也有一些可以容忍或者正常的音调变化也为误判为不自然点。另外注意到音调不连续检测算法在合成语音和对比语音这两个数据集上的准确率差别很大，这是因为对比语音的整体自然度比较高，有一些在较不自然语音中可容忍的不自然点变得很明显，这就导致两个数据集上的主观评测的标准不一样。

另外从标注的不自然点数目来看，音调不连续和结尾不自然占所有不自然点中的大部分，而只有少部分不自然问题是读音、音调错误、韵律结构和停顿的错误。因此，针对这两种不自然的问题做客观检测是有意义的，可以比较直接地帮助合成算法的自然度改进。

总的来说，客观检测方法和主观感受所标注的不自然点还是有非常高的相关性的，并且在实际应用中，针对不同的语音合成系统，可以先用少量测试语料调整门限参数，这样在使用客观检测方法就可以很好的定位和检测不自然点。

5 结论与展望

本文通过人工听辨的方法总结了波形拼接式语音合成中容易出现的四类问题：读音、音调错误，韵律结构不好或停顿错误，音调连续性不好，长停顿前语音边界自然度不好。

读音、音调错误和韵律结构、停顿方面的问题主要是由于合成算法前端文本分析算法造成的，而且这些问题依赖于上下文，在不同的语境中和不同的听音人的主观感受也不尽相同，必须依靠语言学规则才能对这部分的算法进行改进。除了可以通过算法排查音库标注和切词的部分错误外，一般只能靠主观测试来发现这方面的自然度问题，并且可以在合成系统外单独对前端文本分

析的算法进行测试和改进，与语音本身关系并不大。

对于音调连续性和边界自然度的问题，虽然原因很复杂，前端文本分析算法和后端的基元选取算法都会产生影响，但是都会在合成的语音中产生明显变化，可以通过基频连续性和波形包络边界连续性来进行判断。从实验结果来看，本文提出的客观检测的结果与主观感受一致，而且所用的基频差分门限率α、波形包络差分门限率β以及时差门限系数τ与具体语音段自身的参数进行计算，得到具体的差分门限值和时差门限值，使自动检测算法具有更高的适用范围，并且可以根据需求方便调整门限值的高低。。

本文提到的客观检测算法不仅可以用来协助发现合成语音中的不自然点，还可以作为合成系统中选音算法的一部分，将不自然点检测的结果作为选音代价函数的参数之一，在音调连续性和结尾自然度上提高语音的自然度。在接下来的工作中，将会继续研究门限值的设定与算法准确率、召回率的关系，并在语音合成系统中的选音算法中进行实际的应用，评测其对合成语音自然度的改进程度。

另外在本文研究的基础上，可以设计综合主观和客观测试方法的语音评测方案，为语音的自然度提供更客观可信的评测标准，从而可以更科学地衡量语音合成系统的水平。

6 致谢

本文研究得到了国家自然科学基金（60805008，60928005，61003094，60931160443）以及教育部博士点新教师基金(200800031015)的经费支持。

参 考 文 献

[1] V. Karaiskos, S。 King, R. Clark and C. Mayo,“The Blizzard Challenge 2008,” in Proc. of the Blizzard Challenge, 2008.

[2] R. Batusek. An objective measure for assessment of the concatenative tts segment inventories. In Proceedings of Eurospeech 2001 — Scandinavia, Aalborg, Denmark, Sept. 2001.

[3] Cuntai, J. X., Li, G. H. 2002. An Objective Measure for Assessment Of A Corpus-Based Text-To-Speech System. In Proceedings of 2002 IEEE Workshop on Speech Synthesis, pages 179- 182

[4] 陈静，周毅刚，周建林. 符合人耳听觉特性的语音音质的客观评价方法[J]. 哈尔滨工业大学学报，1998-06

[5] 初敏. 韵律研究与合成语音的自然度[C]. 见：第五届全国现代语音学学术会议-新世纪的现代语音学，北京：清华大学出版社，2001: 295-301

[6] 赵博;中文语音合成系统的评测方法研究[D];清华大学;2005