

汉语重音的凸显度分析与合成*

孟凡博^{1,2} 吴志勇^{1,2,3} 贾 珈^{1,2†} 蔡莲红^{1,2}

(1 清华大学 计算机科学与技术系, 普通计算教育部重点实验室, 清华信息科学与技术国家实验室 北京 100084)

(2 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)

(3 香港中文大学 系统工程与工程管理学系 香港)

2013 年 7 月 21 日收到

2014 年 3 月 20 日定稿

摘要 重音是重要的语调特征, 重音合成技术可以提高语音的自然度和表现力。针对重音的局部凸显性, 该文提出了声学特征凸显度的表示方法, 分析了不同韵律位置(韵律词首、中、尾, 韵律短语首、中、尾等)重音音节的声学特征凸显度, 发现在韵律单元末(韵律词末音节和韵律短语末韵律词)的重音其基频最大值凸显度要低于非韵律单元末重音。提出了基于声学特征凸显度的非线性的重音声学参数生成算法, 解决了传统重音声学参数线性修改算法的修改幅度不足或过大的问题。采用该算法建立了基于隐 Markov 模型的支持重音合成的语音合成系统。实验表明, 该系统可以有效合成带有重音的语音, 提高了合成语音的自然度和表现力。

PACS 数: 43.72

The prominence analysis and synthesis of emphasis in Putonghua

MENG Fanbo^{1,2} WU Zhiyong^{1,2,3} JIA Jia^{1,2} CAI Lianhong^{1,2}

(1 Tsinghua National Laboratory for Information Science and Technology,
Department of Computer Science and Technology, Tsinghua University Beijing 100084)

(2 Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems,
Graduate School at Shenzhen, Tsinghua University Shenzhen 518055)

(3 Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong Hong Kong)

Received Jul. 21, 2013

Revised Mar. 20, 2014

Abstract Emphasis is an important feature of intonation. The technology of emphatic speech synthesis can improve the naturalness and the expressiveness of the synthesized speech. This paper defined the prominences of acoustic features, analyzed the prominences of the acoustic features of the emphasized syllables in different prosody locations, e.g. the head, the body, and found that the prominences the maximum F0 of the emphasized syllables at the end of prosody units(prosody words or prosody phrases) are lower than other emphasized syllables. A parameter generation algorithm of emphasized syllables based on the prominences of the acoustic features was proposed based on the analysis. It avoided the problem of oversize modification of the traditional linear modification algorithm. An emphatic speech synthesis system based on hidden Markov model (HMM) was built with the proposed algorithm. Experiments demonstrated that the system could synthesize emphasized speeches and improve the naturalness and the expressive of the synthesized speeches.

引言

随着汉语语音合成技术的发展, 汉语合成语音的清晰度和可懂度不断提高, 但自然度和表现力仍

不尽人意。合成语音的自然度和表现力主要与语调有关, 而重音则是最主要的语调特征之一。重音一般是指在词、短语乃至句子中某些成分的一种凸显, 它使得语音中基频包络起伏更为明显, 听起来更具表现力。有研究表明在合成语音中加入重音可以提高

* 国家 973 项目 (2013CB329304)、国家自然科学基金 (61375027, 61370023)、香港政府研究资助局项目 (N-CUHK414/09) 和国家社会科学基金 (13&ZD189) 资助

† 通讯作者: 贾珈, Email: jjia@mail.tsinghua.edu.cn

合成语音的自然度^[1-2]。因此重音分析与合成相关研究得到了广泛的关注。

重音的声学特征具有凸显性, 声学特征高于临近音节的音节更容易被感知为重音^[3]。一般的, 与汉语重音感知相关的声学特征主要有基频和时长, 并且重音的声学表现与多种因素有关。汉语重音的时长增长; 阴平、阳平和去声重音的基频增大, 并且重音后音节的基频下降; 而上声的基频下降, 其后的音节基频提高; 当上声位于句末时, 其基频调型更加完整^[4-6]。语句中重音的声学表现与其在句中的位置是相关的。文献 2 和文献 7 对处于不同位置的焦点重音进行了声学分析, 其中文献 7 的分析表明随着焦点重音位置的不同音高变化程度不一致, 重音在句首和句中的表现, 比在句末尾显著。此外, 文献 8 发现, 在重音的声学表现上, 基频和时长具有互补的作用, 呈负相关性。

在重音合成方面, 一些早期研究是基于波形拼接语音合成系统进行的。文献 9 分析了重音的时长模式并建立了基于规则的重音合成系统, 文献 10 分析了句子重音的韵律参数, 建立了基于规则的重音 TTS 系统, 文献 11 和文献 2 采用统计方法建立了重音声学特征预测模型进而指导合成系统选音。由于拼接系统中语音参数受到原始音库制约, 变化范围较小, 因此拼接系统难以满足重音合成的需要。而基于隐马尔可夫模型 (hidden Markov model, HMM) 的参数化语音合成方法^[13]由于灵活性高等优点, 成为近年来重音合成的热门研究方向。一些初期研究尝试直接在经典 HMM 语音合成框架中添加重音相关问题来合成带有重音的语音^[14-15], 然而由于一句话中仅有少数几个音节为重音, 导致在决策树聚类过程中重音相关问题的区分性不强, 最终训练的模型并不能有效合成重音。文献 16 提出了两种特殊结构的决策树 (two-pass decision tree 和 factorized decision tree) 来提高重音相关问题的区分性, 不过由于重音数据量的限制, 合成语音的重音强度和自然度仍不能满足要求。另一些研究采用参数转换的思想, 通过对非重音到重音的声学参数变化进行建模, 然后根据模型修改合成的非重音语音的声学参数, 从而达到生成重音的目的。文献 2 和文献 17 将重音等级分为轻、中和重 3 级并分别统计不同上下文情况下 (如声调) 非重音与重音音节基频和时长的比例, 文献 18 采用决策树和支持向量机对非重音和重音音素的基频、时长和能量的比值进行建模。该方法由于进行声学参数修改过程中没有考虑原始语音声学参数的大小, 容易导致修改幅度过小或过大的问题, 前者导致

重音感知不明显, 后者会降低重音的自然度。

局部凸显性是重音最重要的特征之一, 本文根据该特征, 提出了声学特征凸显度的表示方法, 尽量去除其他因素对声学特征的影响, 还原连续语流中重音本身的声学表现, 并基于凸显度对语料进行重音自动判定, 判定结果与人工听辨有 80% 的一致性。分析了不同韵律位置 (韵律词前、中、后, 韵律短语前、中、后等) 重音的基频和时长凸显度, 提出了基于声学特征凸显度的重音声学参数生成算法, 解决了传统线性修改算法修改幅度过大或过小的问题, 并实现了支持重音的汉语语音合成系统, 该系统采用分类回归树实现基于文本特征的重音预测, 通过 HMM 语音合成模型预测不含重音的语音声学参数, 然后采用基于声学特征凸显度的非线性的重音声学参数生成算法生成重音声学参数, 最后合成含有重音的语音, 实验表明, 该系统能有效生成带有重音的语音, 提高了合成语音的自然度和表现力。

1 重音声学特征的凸显度

在连续语流中, 语音的声学表现是多种因素相互影响的结果, 如语调下倾会使位置靠后音节的基频降低, 韵律短语边界前音节的时长延长, 这些因素影响重音凸显性的评估和定量表示。因此, 本文提出了一种重音凸显度的表示方法, 去除额外因素对声学特征的影响, 还原出重音本身的凸显特性。根据汉语重音的声学表现, 本文采用的声学特征包括: 基频最大值 (pitch maximum, 记为 P_{Max})、最小值 (pitch minimum, 记为 P_{Min}) 和时长 (duration, 记为 D), 考虑了语调、韵律结构和汉语声调对重音声学参数的影响后, 计算其相应的凸显度 ($\hat{P}_{\text{Max}}$, $\hat{P}_{\text{Min}}$ 和 $\hat{D}$)。然后计算得到音节的凸显度。

1.1 基频最大值和最小值的凸显度

本文采用当前音节声学特征与所在韵律短语的所有音节声学特征平均值的比值来表示当前音节该声学特征在所处上下文中的凸显度, 设韵律短语中有 n 个音节, 则第 i 个音节的基频最大值凸显度 $\hat{P}_{\text{Max}}^i$ 表示为:

$$\hat{P}_{\text{Max}}^i = \frac{P_{\text{Max}}^i}{\frac{1}{n} \sum_{j=1}^n P_{\text{Max}}^j}, \quad (1)$$

其中 P_{Max}^j 为第 j 个音节基频最大值。考虑到上声为重音时, 其声学表现为基频最小值下降, 即基频最小值越低, 越可能为重音, 因此, 音节 i 基频最小值凸显度 $\hat{P}_{\text{Min}}^i$ 表示为:

$$\hat{P}_{\text{Min}}^i = \frac{\frac{1}{n} \sum_{j=1}^n P_{\text{Min}}^j}{P_{\text{Min}}^i}. \quad (2)$$

考虑到汉字音节调型的声学特性, $\hat{P}_{\text{Max}}^i$ 仅计算阴平、阳平和去声音节, 并且阳平的基频最大值仅计算其基频低点之后的基频最大值, 避免计算由前一音节以高调结束导致的基频最大值, 而 $\hat{P}_{\text{Min}}^i$ 仅计算阳平、上声和去声音节, 由于连上而变调的上声按阳平处理。

1.1.1 语调下倾对基频最大值和最小值的凸显度计算的影响

陈述语调存在语调下倾现象, 该现象将导致在同一个韵律单元(韵律词、韵律短语或句子)内从前向后音节的基频逐渐下降, 从而影响到基频凸显度的计算。受声调本身基频的高低、轻重音、韵律边界基频重置等因素的影响, 直接根据一句话中各个音节的基频计算语调下倾会产生较大误差, 因此本文首先根据句中韵律短语个数及韵律短语在句中位置将韵律短语分为以下 6 类, 然后分别计算每类韵律短语内的语调下倾:

- (1) 单韵律短语句子的韵律短语;
- (2) 2 韵律短语句子的第 1 个韵律短语;
- (3) 2 韵律短语句子的第 2 个韵律短语;
- (4) 3 韵律短语句子的第 1 个韵律短语;
- (5) 3 韵律短语句子的第 2 个韵律短语;
- (6) 3 韵律短语句子的第 3 个韵律短语。

文献 19 和文献 20 论述了汉语语调的双线调节功能, 音高上线(对应于基频最大值)主要负载语调重音, 下线(对应于基频最小值)主要负载语调节奏。文献 21 和文献 22 的研究也支持双线调节功能的观点, 发现汉语中音步、韵律短语中普遍存在音高下倾现象, 并且韵律短语的音高下倾程度也与其在句中的位置有关, 此外, 如果区分声调考察音高下线, 可以发现每种声调都呈现均匀的音高下倾趋势。相似的, 文献 23 基于大规模语料分别对不同声调的音高下线进行了研究, 发现不同声调所承载的调高不同, 然而其总体下降趋势保持一致。因此可以采用公式 $Y = P_{I0} + \beta X$ 表示不同声调的音高下倾, 其中 P_{I0} 表示不同声调的调高, X 为音节在韵律短语中的归一化位置, β 为声调无关的降阶指数, 即不同的声调拥有相同的降阶指数。本文采用该方法计算语调下倾, 分别统计了语料库中阴平基频均值, 阳平、上声和去声的基频最小值, 设属于某类韵律短语且声调为 t (t 为阴平、阳平、上声或去声) 的音节共有 m 个, 其中音节 i 为所在韵律短语的第 j 个音节, 该

韵律短语共有 n 个音节, 则音节 i 在韵律短语内的从 0 开始的归一化时刻为 $(j-1)/(n-1)$, 例如由 7 个音节组成的韵律短语第 1, 3, 7 个音节的归一化时刻分别为 0, 0.33 和 1。本文将该类韵律短语的音节的归一化时刻 X_t^i 和音节的基频平均值 Y_t^i 组成点集 (X_t^i, Y_t^i) , 然后采用最小二乘法计算该点集的基频斜率 $P_{S,t}^{\text{PP}}$ 及基频平均值 $P_{\text{Mean},t}^{\text{PP}}$ 。计算方法如下:

$$\bar{X}_t = \frac{1}{m} \sum_{i=1}^m X_t^i, \quad \bar{Y}_t = \frac{1}{m} \sum_{i=1}^m Y_t^i, \quad (3)$$

$$P_{\text{Mean},t}^{\text{PP}} = \bar{Y}_t, \quad P_{S,t}^{\text{PP}} = \frac{\sum_{i=1}^m (X_t^i - \bar{X}_t)(Y_t^i - \bar{Y}_t)}{\sum_{i=1}^m (X_t^i - \bar{X}_t)^2}. \quad (4)$$

图 1 为单韵律短语句子四个声调的点集以及拟合直线示意图, 其中四个声调的基频斜率分别为 $-44.2, -39.1, -40.0$ 和 -41.3 , 可见不同声调音节具有相似的下倾幅度, 与文献 23 的结论一致。最终该类韵律短语的语调下倾的基频斜率和基频平均值分别为四个声调点集的基频斜率 $P_{S,t}^{\text{PP}}$ 和基频平均值 $P_{\text{Mean},t}^{\text{PP}}$ 的平均值:

$$P_S^{\text{PP}} = \frac{1}{4} \sum_t P_{S,t}^{\text{PP}}, \quad P_{\text{Mean}}^{\text{PP}} = \frac{1}{4} \sum_t P_{\text{Mean},t}^{\text{PP}} \quad (5)$$

根据斜率和均值该类韵律短语音节的语调下倾可以表示为如下形式:

$$Y = (X - 0.5) P_S^{\text{PP}} + P_{\text{Mean}}^{\text{PP}}, \quad X \in [0, 1], \quad (6)$$

其中, 设 $Y_1 = P_{\text{Mean}}^{\text{PP}}$, $Y_2 = P_S^{\text{PP}}(X - 0.5)$, 那么有:

$$\frac{\int_{X=0}^1 Y_2 dX}{1-0} = 0. \quad (7)$$

可见, 该类韵律短语音节的基频平均值由斜率为 0 的横线 $Y_1 = P_{\text{Mean}}^{\text{PP}}$ 和平均值为 0 的斜线 $Y_2 = P_{\text{Slope}}^{\text{PP}}(X - 0.5)$ 组成。那么要去除语调下倾只需要根据音节在韵律短语内的归一化位置减掉斜线 Y_2 上对应的幅值即可。

由于语调下倾句中位置靠前的韵律短语的基频均值要大于靠后的韵律短语, 本文以韵律短语数为 3 的句子的第 3 个韵律短语的基频均值为韵律短语基准基频, 那么设 p_i 为音节 i 基频曲线上一点的基频值, 去除句子级和韵律短语级语调下倾后该基频点的值 p'_i 为:

$$p'_i = p_i - \left(\frac{j-1}{n-1} - 0.5 \right) P_S^{\text{PP}} - \Delta P_{\text{Mean}}^{\text{PP}}, \quad (8)$$

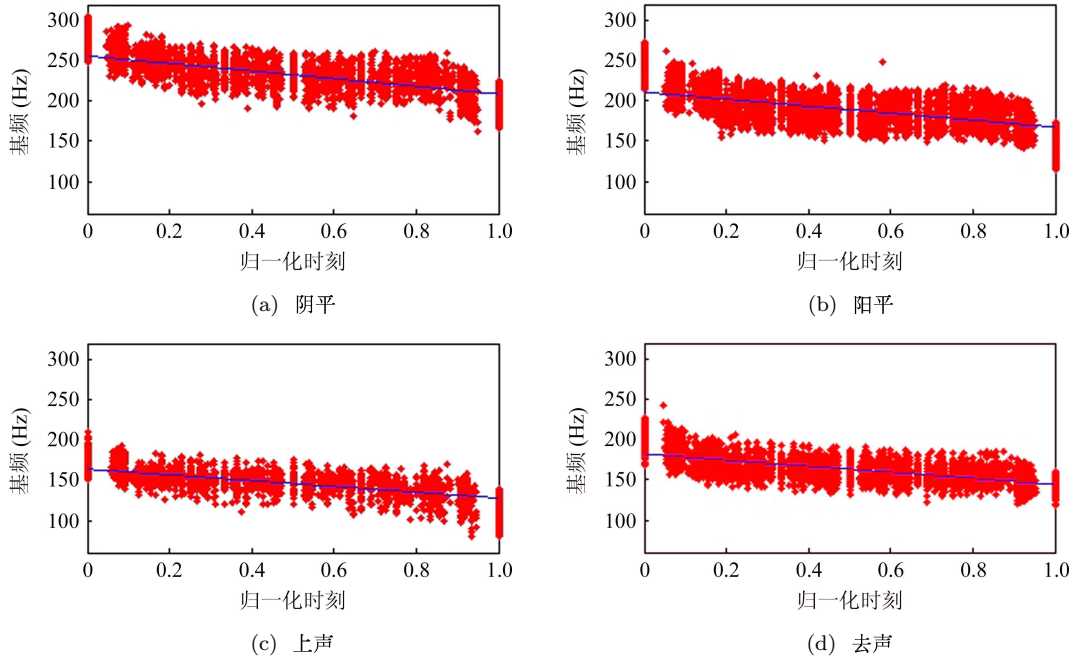


图 1 单韵律短语句子各声调的点集及拟合直线示意图

其中, $(j-1)/(n-1)$ 为该音节的归一化时刻, 括号项在不改变基频曲线平均值的条件下去除基频曲线斜率, $\Delta P_{\text{Mean}}^{\text{pw}}$ 为该类韵律短语的基频均值与韵律短语基准基频的差值, 该项将各类韵律短语的基频曲线平均值调整到相同频率。

同理, 本文根据韵律词中的音节个数将韵律词分为单音节词、双音节词和多音节词, 并以单音节词的平均基频为基准基频, 设该音节为韵律词中的第 j_{pw} 个音节, 且该韵律词长度为 n_{pw} ($n_{\text{pw}} > 1$), 采用式 9 去除非单音节韵律词内的语调下倾:

$$p_i'' = p_i' - \left(0.5 - \frac{j_{\text{pw}} - 1}{n_{\text{pw}} - 1}\right) P_S^{\text{pw}} - \Delta P_{\text{Mean}}^{\text{pw}}, \quad (9)$$

其中, P_S^{pw} 为该类韵律词的基频斜率 (计算方法与韵律短语的基频斜率相同), $\Delta P_{\text{Mean}}^{\text{pw}}$ 为该类韵律词基频平均值与韵律词基准基频的差值。

1.1.2 声调对基频最大值和最小值的凸显度计算的影响

汉语阴平、阳平、上声和去声的单字调可简述为高平调、中升调、降升调和高降调。然而受发音人个性化以及前后音节的影响, 实际连续语流中不同声调的调域往往有很大差异, 这种差异也将影响到基频凸显度的计算。因此, 本文采用式 (10) 和式 (11) 对音节基频最大值和最小值的凸显度进行修正。

$$\hat{P}_{\text{Max}}^i = \frac{P_{\text{Max}}''^i / T_{\text{Max}}^i}{\frac{1}{n} \sum_{j=1}^n P_{\text{Max}}''^j}, \quad (10)$$

$$\hat{P}_{\text{Min}}^i = \frac{\frac{1}{n} \sum_{j=1}^n P_{\text{Min}}''^j}{P_{\text{Min}}''^i / T_{\text{Min}}^i}, \quad (11)$$

其中, $P_{\text{Max}}''^i$ 和 $P_{\text{Min}}''^i$ 分别为去除语调下倾影响后音节 i 的基频最大值和最小值, T_{Max}^i 和 T_{Min}^i 分别为音节 i 基频最大值和基频最小值的修正系数, 前者为语料库中与音节 i 同样声调音节的基频最大值平均值与所有阴平、阳平和去声音节基频最大值平均值的比值, 通过该系数将三种声调音节的平均基频最大值调整到相同频率, 避免声调差异影响凸显度计算, 后者为语料库中与音节 i 同样声调音节的基频最小值平均值与所有阳平、上声和去声音节基频最小值平均值的比值, 通过该系数将三种声调音节的基频最小值平均值调整到相同频率, 避免声调差异影响凸显度计算。

1.2 时长的凸显度

与基频的凸显度计算方法相似, 时长的凸显度计算方法如下:

$$\hat{D}^i = \frac{D^i}{\frac{1}{n} \sum_{j=1}^n D^j}. \quad (12)$$

式 (12) 中并没有考虑音节本身的时长差异, 这些差异将影响时长凸显度的计算精度。考虑到韵母时长占总时长的比重较大, 并且重音相关的音节延长中声韵母延长是相对平衡的^[24], 本文根据韵母对音节进行分类, 韵母相同的音节归为一类, 然后计算各类音节的时长平均值与所有音节时长平均值的比

值,称为时长修正系数 α ,通过该系数将韵母不同的音节的平均时长调整到相同长度,避免音节本身的时长差异影响凸显度计算,采用时长修正后的时长凸显度计算如下:

$$\hat{D}^i = \frac{D^i / \alpha^i}{\frac{1}{n} \sum_{j=1}^n D^j}. \quad (13)$$

此外,汉语中存在韵律域末尾延长(边界前延长)现象,该现象在韵律短语末最为显著^[25],而对于句末音节是否存在延长仍存在一定争论^[24]。本文统计了不同韵律短语末音节平均时长与非末音节平均时长的比值 β ,通过该系数去除韵律短语边界前延长,修正后的时长凸显度计算如下:

$$\hat{D}^i = \frac{D^i / \alpha^i / \beta^i}{\frac{1}{n} \sum_{j=1}^n D^j}, \quad (14)$$

其中,对于非韵律短语末音节 β^i 等于 1。

1.3 音节的凸显度

在得到音节声学特征的凸显度之后,考虑到基频和时长对重音的互补作用^[8],采用音节各声学特征凸显度的加权之和表示该音节的凸显度。音节 i 的凸显度 $S\hat{w}^i$ 为:

$$S\hat{w}^i = w_1 \hat{P}^i + w_2 \hat{D}^i, \quad (15)$$

其中, $\hat{P}^i$ 为音节 i 的基频凸显度,因为阴平、阳平和去声重音的基频表现为高点升高,而上声重音的基频表现为低点降低,因此对于阴平、阳平和去声,该项为基频最大值的凸显度,对于上声,该项为基频最小值的凸显度, w_1 和 w_2 分别为基频凸显度和时长凸显度的权重。

2 基于凸显度的音节重音判定

在进行重音相关研究之前,首先要对语料进行重音标注。因为音节的凸显度去除了语调下倾等重音之外因素的影响,描述了音节的局部凸显特性,所以本文对大量语料进行了音节凸显度的计算,并根据音节凸显度的分布,确定重音判定的凸显度阈值,凸显度高于阈值的音节将被判定为重音。

2.1 语料库

本文研究采用的语料库包含一名女性专业播音员录制的 5726 句新闻风格的语音,每句语音包含 1~3 个韵律短语,平均长度 16 个音节,总音节数约 94000 个,时长超过 6 个小时。不考虑声调,语料库

包括了汉语音节的声母和韵母的所有组合。全部录音在专业录音室中完成,以 Microsoft Windows Wav 格式保存(16 位,单声道,16 kHz 采样率)。语料库标注了基频和音节的拼音、声调(含轻声)以及韵律边界(音节、韵律词和韵律短语)。

本文的目的是基于该语料库建立支持重音的汉语语音合成系统,因而,声学参数统计、重音自动判定、声学特征分析以及重音声学参数生成算法设计都是基于该语料库进行的。虽然文本分析的结论以及算法参数是数据(发音人)相关的,但本文所提出的基于凸显度的分析和参数生成方法是数据无关的,当得到其他发音人的大规模语料后,也可以采用该方法对其他发音人的重音进行分析并建立相应的支持重音的语音合成系统。

2.2 声学参数统计

表 1 显示了根据式(3)—式(5)计算得到的不同韵律短语基频曲线的基频平均值和时长归一化后的基频斜率,其中, pp_h , pp_b 和 pp_t 分别表示句首、句中和句末韵律短语。可见,句子越长,其首韵律短语的基频平均值越高,而其末韵律短语的基频平均值越低;句子首韵律短语的基频曲线斜率绝对值最大,末韵律短语次之,中部韵律短语最小。本文以韵律短语数为 3 的句子的第 3 个韵律短语的基频平均值为基准基频去除韵律短语间基频平均值差异。例如,当音节 i 为韵律短语数为 2 的句子的第 1 个韵律短语时,由表 1 得 $\Delta P_{\text{Mean}}^{\text{PP}} = 187.9 - 171.0 = 16.9$ 。

表 1 韵律短语基频曲线平均值及斜率

句子长度 (韵律短语数)	$P_{\text{Mean}}^{\text{PP}}$ (Hz)			$P_{\text{Slope}}^{\text{PP}}$ (Hz/s)		
	pp_h	pp_b	pp_t	pp_h	pp_b	pp_t
1	182.9	—	—	-41.2	—	—
2	187.9	175.0	—	-35.5	-20.9	—
3	192.5	179.9	171.0	-41.6	-16.8	-22.3

去除韵律短语内及韵律短语间语调下倾后,与韵律短语基频曲线的基频平均值和基频斜率计算方法相同,本文计算了单音节词、双音节词和多音节词的基频平均值以及双音节词和多音节词时长归一化后的基频斜率,如表 2 所示。不同长度的韵律词基频平均值相似,但多音节词的基频斜率绝对值要大于双音节词。本文以单音节词的平均基频为韵律词的基准基频。

表 2 韵律词基频曲线平均值及斜率

$P_{\text{Mean}}^{\text{PW}}$ (Hz)			$P_{\text{Slope}}^{\text{PW}}$ (Hz/s)	
单音节词	双音节词	多音节词	双音节词	多音节词
174.4	173.7	174.5	-6.5	-13.7

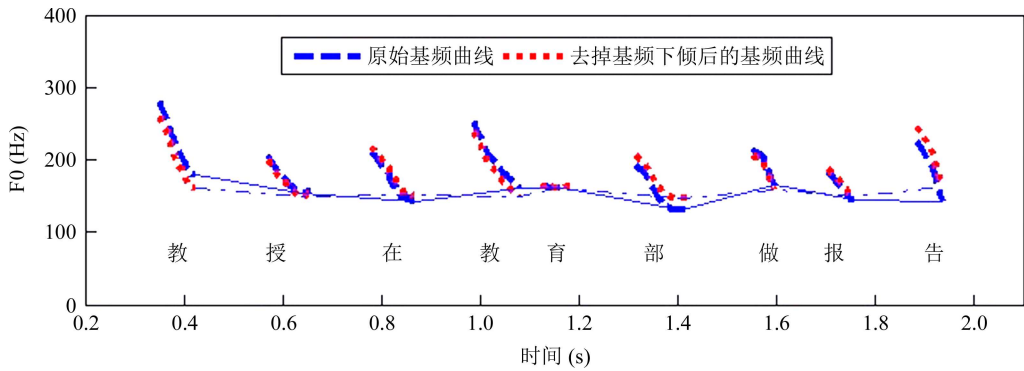


图 2 “教授在教育部做报告”去掉语调下倾前后基频曲线对比

如图 2 所示，粗线为句子“教授在教育部作报告”去掉语调下倾前后的基频曲线，两条细线分别为去掉语调下倾前后各音节基频曲线低点的连线。“||”为韵律短语边界，“|”为韵律词边界。可见该句由 1 个包含 3 个韵律词的韵律短语组成。由于去声重音的基频表现为高点提高，低点变化不大，因此可以通过低点观察语调下倾，通过高点观察重音。本句中原始基频曲线的韵律词内和韵律短语内可以观察到明显的语调下倾，而去掉语调下倾后，各音节基频低点的基频值相近。此外，去除语调下倾后，各音节高点根据基频高低可以分为 3 类，低：“授”、“育”和“报”；中：“在”、“部”和“作”；高：2 个“教”和“告”。可见 2 个“教”和“告”更可能为重音。对于句末音节“告”，其基频高点在去掉语调下倾前并未下降，反而高于同一个韵律词内前面 2 个音节的基频高点，说明说话人倾向于将该音节读作重音，但该音节的基频高点在整个韵律短语中凸显并不明显。去掉语调下倾前后该音节的基频最大值凸显度分别为 1.05 和 1.17，可见若不去除语调下倾，该音节可能被漏检。说明去除语调下倾对于计算基频凸显度具有重要作用。

表 3 显示了语料库中不同声调音节的基频最大值和最小值的平均值。可见，该语料库的发音人的阳平基频较低，阳平和去声的基频最小值也要明显高于上声。

表 4 显示了部分韵母音节的时长修正系数。可见不同韵母音节的时长有较大差异，例如韵母为 e 的音节的时长仅为所有音节平均时长的 0.81 倍。

语料库中不同韵律短语末音节平均时长与非末音节平均时长的比值 β 如表 5 所示。句末音节时长增长最大，句中最小。

表 3 不同声调音节的基频最大值和最小值

	阴平	阳平	上声	去声	平均值
P_{Max}	233.6	191.2	—	228.8	217.9
P_{Min}	—	176.6	123.7	153.0	159.7

表 4 部分韵母音节的时长修正系数

韵母	a	ai	an	ang	ao
修正系数	0.97	1.01	1.08	1.16	1.02
韵母	e	ei	en	beng	i
修正系数	0.81	0.91	1.00	1.01	0.90

表 5 韵律短语末音节平均时长与其他音节平均时长的比值 β

句子长度 (韵律短语数)	β		
	pp_h	pp_b	pp_t
1	1.52	—	—
2	1.25	1.47	—
3	1.28	1.16	1.44

2.3 基于凸显度的音节重音判定

汉语重音的主要韵律感知差异体现在基频和时长的变化^[2]。因此，汉语语音中重音的判定主要使用基频(如最大值、最小值、范围等)和时长作为特征^[26]，有的研究额外引入了词典、语法特征(如当前音节在词中的位置等)^[27]，其准确率一般在 80% 左右。文献 28 和文献 17 分别统计了大规模中性朗读风格语料中的重音分布，后者所统计的语料库风格为新闻风格，与本文相同。将重音等级分为轻、中、重后，重音在文献 28 和文献 17 语料库中所在的比例分别为 38% 和 36%。本文根据式 (15) 计算各音节凸显度，然后将凸显度从大到小排序，为了提高重音判定的准确率进而进行重音相关的声学特征分析和建模生成，取第 30% 的凸显度为重音判定阈值，最后将凸显度高于阈值的音节判定为重音。

本文对重音自动判定结果进行了验证性实验，从语料库中随机抽取了 3 句单韵律短语语音、4 句双韵律短语语音和 3 句 3 韵律短语语音，为避免短时记忆能力对听音人重音的判断产生干扰，将双韵律短语语音和 3 韵律短语语音按照韵律短语切分，最后共得到 20 句语音。听音人在听音之后，按照以下原

则标记重音：(1) 每句语音可以反复听辨；(2) 1 句语音中的重音数量没有上限，如果感知不到重音也可以为 0；(3) 重音落在具体的音节上；(4) 标记重音的同时尽量给出该音节感知为重音的声学特征（基频、时长）。

一共邀请 4 位长期从事语音方面研究的老师或研究生参加了实验，听音人之间的重音标注结果有 75% 的一致性。在根据听音人感知音节为重音的声学特征的比例对式 (15) 的权重系数进行调整之后，80% 听音人一致感知为重音的音节被算法判定为重音，85% 被算法判定为重音的音节被至少一位听音人感知为重音。因此，重音自动判定算法的结果与听音人感知具有较好的一致性，本文在此基础上进行重音的凸显度分析与合成建模。

3 不同韵律位置的重音的凸显度分析

本文根据韵律词在韵律短语中的位置以及音节在韵律词中的位置进行分类： pw_h 、 pw_b 和 pw_t 分别表示位于韵律短语首、中和尾部的韵律词， sy_h 、 sy_b 和 sy_t 分别表示位于韵律词首、中和尾部的音节。

处于不同韵律位置的重音的声学特征凸显度及其标准差如表 6 所示。其中，基频最大值凸显度仅计算阴平、阳平和去声重音，基频最小值凸显度仅计算上声重音。该表说明在不同韵律位置音节被感知为重音时其声学特征凸显度需要满足的条件。

在基频最大值凸显度方面，韵律词末音节的基频最大值凸显度低于其他音节，而韵律短语末音节的基频最大值凸显度最低，本文对韵律词末音节重音和非韵律词末重音以及韵律短语末韵律词重音和非韵律短语末韵律词重音的基频最大值凸显度进行了显著性分析，结果分别为 $p < 0.01$ 和 $p < 0.05$ 。说明在韵律单元末（韵律词末音节和韵律短语末韵律词）的重音其基频最大值凸显度要低于非韵律单元末重音。

需要说明的是，本文所得结论虽然与文献 2 和文献 7 的结论相似，但文献 2 和文献 7 在进行声学特征分析之前，并未去除语调下倾对基频的影响，因此其观察到的现象是语调下倾与重音综合作用的结果，而本文观察到的现象是去除语调下倾后的结果。其原因可能是由于靠近句末的音节其基频较低，发音人可以较小幅度的提高基频即可达到凸显重音的目的。

在基频最小值凸显度方面，非韵律短语末的上声重音的基频最小值凸显度大于 1，说明其基频最小值低于临近音节，其中位于韵律词中部的上声重音基频最小值凸显度最大（与非韵律词中部上声重音基频最小值凸显度的显著性分析结果为 $p < 0.01$ ），而韵律短语末的上声重音的基频最小值凸显度接近 1，说明其基频最小值凸显度与非重音相比无明显变化，这是由于韵律短语末的上声重音的基频表现为更完整的调型，而其他位置上声重音的基频表现为本身基频下降同时提高其后音节的基频。

在时长凸显度方面，所有重音的时长都有明显增长（时长凸显度大于 1），但与基频凸显度不同的是，韵律词中部的重音的时长增长幅度低于其他位置的重音（与非韵律词中部重音的时长凸显度的显著性分析结果为 $p < 0.05$ ），并且在去除韵律短语边界前延长后，韵律短语末重音（ pw_t 中的 sy_t ）的时长与韵律词末重音（ pw_h 和 pw_b 中的 sy_t ）时长没有明显区别。这体现了重音的声学表现中基频与时长的互补作用。

4 基于 HMM 的汉语重音合成

实验证明，在传统 HMM 模型训练过程中，直接在问题集中添加重音相关问题进行训练的方法不能有效生成重音。由于一句话中仅有少数音节为重音，重音数据量的限制导致一些基于改进决策树训练过程的重音合成模型仍存在着合成重音的强度和自自然度不足的问题。因此，本文采用不含有重音信息的语

表 6 不同韵律位置重音的时长和基频凸显度

声学特征		pw_h			pw_b			pw_t		
		sy_h	sy_b	sy_t	sy_h	sy_b	sy_t	sy_h	sy_b	sy_t
$\hat{P}_{\text{Max}}$	均值	1.09	1.11	1.08	1.09	1.10	1.08	1.08	1.09	1.06
	标准差	0.08	0.07	0.07	0.07	0.06	0.08	0.07	0.07	0.08
$\hat{P}_{\text{Min}}$	均值	1.05	1.12	1.04	1.12	1.16	1.12	1.13	1.19	1.03
	标准差	0.09	0.07	0.08	0.08	0.09	0.10	0.08	0.09	0.08
$\hat{D}$	均值	1.10	1.10	1.13	1.12	1.09	1.13	1.14	1.12	1.13
	标准差	0.07	0.06	0.07	0.06	0.05	0.07	0.07	0.06	0.07

$\hat{P}_{\text{Max}}$ 、 $\hat{P}_{\text{Min}}$ 和 $\hat{D}$ 分别表示基频最大值凸显度、基频最小值凸显度和时长凸显度。

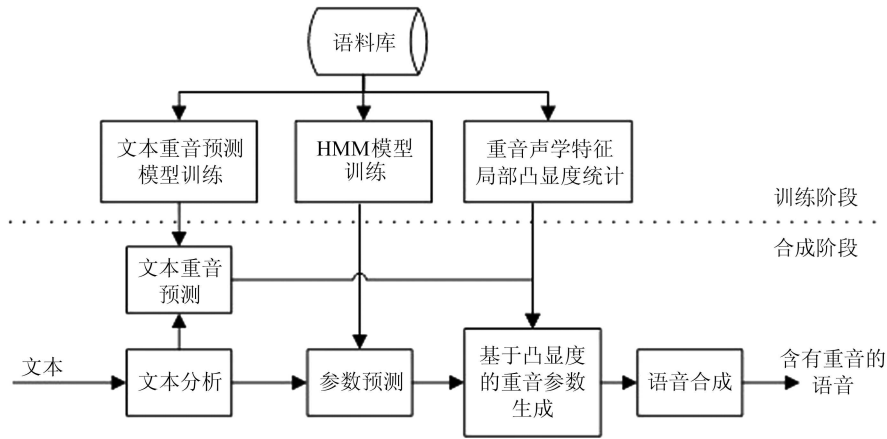


图 3 基于 HMM 的汉语重音合成系统流程图

音合成模型结合重音声学参数生成模型来生成含有重音的语音，系统流程图如图 3 所示。基于语料库，分别训练文本重音预测模型和 HMM 模型并统计重音声学特征凸显度。HMM 模型训练所使用的问题集不包括重音相关问题，即采用不含重音信息的 HMM 模型。在合成阶段，首先对输入文本进行文本分析，然后根据文本分析结果预测重音位置以及不含有重音的语音参数，并进一步根据预测参数、重音位置以及重音声学特征凸显度生成重音声学参数，最后合成含有重音的语音。其中文本重音预测模型的训练方法以及基于凸显度的重音声学参数生成算法将在 4.2 节详细介绍。

4.1 基于分类回归树的文本重音预测模型

文本重音预测是支持重音的语音合成系统的重要内容。研究表明，语句中常规重音的分布与句法结构有很大关系（如主谓结构中重音落在谓语部分）^[29]，然而目前正确预测语法结构还有一定困难，因此，已有研究大部分基于易于从文本分析得到的特征（如韵律结构）进行文本重音预测^[15,22]，准确率一般在 70%~80%。本文采用分类回归树（classification and regression tree, CART）进行文本重音预测，所用特征全部可由文本分析得到，包括当前音节以及前后音节的拼音、声调，音节在韵律结构中的位置（韵律词、韵律短语和句子）。

设当前节点 N 中重音音节和非重音音节个数分别为 n_s 和 n_{us} ，则当前节点重音音节的比例 $R_s(N)$ 为：

$$R_s(N) = \frac{n_s}{n_s + n_{us}}. \quad (16)$$

设待选特征集为 F ， N 被特征 f 分裂为 N_{fl} 和 N_{fr} ，其中一个节点重音音节比例提高、而另一个节点降低，提高的重音音节比例可以由 $\max(R_s(N_{fl}), R_s(N_{fr}))$ 表示，本文选择使节点重音音节比例提高

最多的特征 f_0 进行分裂：

$$f_0 = \arg \max_{f \in F} (\max(R_s(N_{fl}), R_s(N_{fr}))). \quad (17)$$

分类回归树分裂停止条件有两个：(1) F 里没有能够增大节点重音音节比例的特征；(2) 当前节点中音节个数少于一定阈值。

在预测阶段，模型根据音节的特征选择相应的节点，当该节点的重音音节比例高于 50% 时，音节被判定为重音。

本文采用 90% 的语料训练文本重音预测模型，其余 10% 的语料作为集外测试集。最终模型在集内测试集上的准确率为 89%，集外测试集上的准确率为 77%。

4.2 基于声学特征凸显度的重音声学参数生成算法

传统的根据固定比例对重音声学参数进行线性修改的模型^[15-16]，由于没有考虑被修改音节与邻近音节声学特征的大小关系，容易出现修改幅度不够或过于夸张的情况。针对这个问题，本文提出了基于声学特征凸显度的重音声学参数生成算法，既保证了修改后音节的凸显程度又解决了修改幅度过大的问题。算法具体步骤如下：

(1) 提取声学特征：设拟生成为重音的音节所在韵律短语共有 n 个音节，重音音节为第 i 个音节。 P^j 和 D^j 为 HMM 预测的第 j 个音节的基频和时长序列。从 P^j 中提取基频最大值 P_{Max}^j 和基频最小值 P_{Min}^j 。

(2) 去除语调下倾：由于语调下倾会影响声学特征凸显度的计算，因此根据式 (8) 和式 (9) 去除语调下倾，得到各音节去除语调下倾后的基频最大值 $P_{\text{Max}}^{\prime\prime j}$ 和最小值 $P_{\text{Min}}^{\prime\prime j}$ 。

(3) 根据声学特征凸显度计算目标的声学特征：设根据文本分析结果重音的声调和位置从表 6 数据

中得到重音音节 i 的目标基频最大值、最小值和时长的凸显度分别为 $\hat{P}_{\text{Max}}^i$, $\hat{P}_{\text{Min}}^i$ 和 $\hat{D}^i$, 根据表 3—表 5 基频最大值、最小值、时长以及末音节时长的修正系数分别为 T_{Max}^i , T_{Min}^i , α^i 和 β^i 。设 $P_{\text{Max}}^{\prime\prime i}$ 为根据基频最大值凸显度计算得到的目标基频最大值, 那么根据式 (10) 有:

$$\hat{P}_{\text{Max}}^i = \frac{P_{\text{Max}}^{\prime\prime i} / T_{\text{Max}}^i}{\frac{1}{n} \left(\sum_{j=1}^n P_{\text{Max}}^{\prime\prime j} - P_{\text{Max}}^{\prime\prime i} + P_{\text{Max}}^{\prime\prime i} \right)}, \quad (18)$$

整理得:

$$P_{\text{Max}}^{\prime\prime i} = \frac{\frac{1}{n} \hat{P}_{\text{Max}}^i \left(\sum_{j=1}^n P_{\text{Max}}^{\prime\prime j} - P_{\text{Max}}^{\prime\prime i} \right)}{\frac{1}{T_{\text{Max}}^i} - \frac{1}{n} \hat{P}_{\text{Max}}^i}, \quad (19)$$

其中, 括号项为该韵律短语内其他音节基频最大值之和, 一般情况下 (n 不小于 2 时) 分母大于 0, 则 $P_{\text{Max}}^{\prime\prime i}$ 与括号项呈正相关性, 即其他音节的基频最大值较高时, $P_{\text{Max}}^{\prime\prime i}$ 较大, 反之, $P_{\text{Max}}^{\prime\prime i}$ 较小, 避免了修改幅度不足或过于夸张的情况, 同时也满足了局部凸显的要求。此外, 为了增加算法适用性, 本文添加了 $P_{\text{Max}}^{\prime\prime i}$ 不能低于 $P_{\text{Max}}^{\prime\prime i}$ 的规则限制, 并且对于上声, 考虑到前后音节基频曲线的连续性, 令 $P_{\text{Max}}^{\prime\prime i}$ 等于 $P_{\text{Max}}^{\prime\prime i}$ 。

同理可得目标基频最小值和时长:

$$P_{\text{Min}}^{\prime\prime i} = \frac{\frac{1}{n} \left(\sum_{j=1}^n P_{\text{Min}}^{\prime\prime j} - P_{\text{Min}}^{\prime\prime i} \right)}{\hat{P}_{\text{Min}}^i / T_{\text{Min}}^i - \frac{1}{n}}, \quad (20)$$

$$D^{\prime i} = \frac{\frac{1}{n} \hat{D}^i \left(\sum_{j=1}^n D^j - D^i \right)}{\left(1/\alpha^i / \beta^i - \frac{1}{n} \hat{D}^i \right)}, \quad (21)$$

其中, 对于阴平令 $P_{\text{Min}}^{\prime\prime i}$ 等于 $P_{\text{Max}}^{\prime\prime i}$ 。

(4) 恢复语调下倾: 根据式 (8)、式 (9) 恢复语调下倾, 得到各音节带有语调下倾的基频最大值 $P_{\text{Max}}^{\prime\prime\prime i}$ 和最小值 $P_{\text{Min}}^{\prime\prime\prime i}$:

$$P_{\text{Max}}^{\prime\prime\prime i} = P_{\text{Max}}^{\prime\prime i} + \left(0.5 - \frac{i-1}{n-1} \right) P_{\text{S}}^{\text{PP}} + \Delta P_{\text{Mean}}^{\text{PP}}, \quad (22)$$

$$P_{\text{Min}}^{\prime\prime\prime i} = P_{\text{Min}}^{\prime\prime i} + \left(0.5 - \frac{i_{\text{pw}}-1}{n_{\text{pw}}-1} \right) P_{\text{S}}^{\text{PW}} + \Delta P_{\text{Mean}}^{\text{PW}}. \quad (23)$$

(5) 修改基频参数: 根据目标基频最大值和基频最小值修改重音音节的基频参数, 设 p_i 为 HMM 预测的音节 i 基频曲线上一点的基频值, 该点从音节起

始开始的时刻为 d_i , 则修改后的基频值 p_i' 和时刻 d_i' 为:

$$p_i' = (p_i - P_{\text{Min}}^i) \frac{P_{\text{Max}}^{\prime\prime\prime i} - P_{\text{Min}}^{\prime\prime\prime i}}{P_{\text{Max}}^i - P_{\text{Min}}^i} + P_{\text{Min}}^{\prime\prime\prime i}. \quad (24)$$

$$d_i' = d_i \times \frac{D^{\prime i}}{D^i}. \quad (25)$$

此外, 为了提高句末上声重音的感知, 本文采用二次曲线为句末上声重音构造了调型。

5 实验及结果分析

本文对语音合成系统的合成语音进行了偏向性实验。实验包括 3 个语音合成系统的合成语音: (1) 不含有重音信息的 HMM 合成系统; (2) 基于线性修改算法生成重音声学参数的 HMM 合成系统, 修改的比例为不同位置重音与非重音音节声学特征的比值; (3) 基于声学特征凸显度生成重音声学参数的 HMM 合成系统。其中, 系统 2 和系统 3 都使用 4.1 节介绍的文本重音预测模型进行重音预测。随机选取集内文本和集外文本 (皆为新闻风格的文本) 各 10 句, 分别采用 3 个系统合成语音, 共 60 句合成语音。将合成语音以随机顺序分别提供给听音人, 听音人在听音之后采用 5 分制的 MOS 分对合成语音的自然度进行评价。

共有 10 位听音人参加了实验, 实验结果如图 4 所示。可见支持重音系统 (系统 2 和系统 3) 的合成语音自然度要高于无重音系统 (系统 1), 而基于声学特征凸显度生成重音声学参数的系统 (系统 3) 的合成语音自然度要高于基于线性算法生成重音声学参数的系统 (系统 2) 的合成语音。这是由于当中性语音中待修改为重音的音节局部凸显度较高时, 系统 2 仍按照固定比例修改其声学参数, 导致修改幅度过大而降低了合成语音的自然度, 而系统 3 则根据局部凸显度动态调整修改幅度, 因而较少出现这种问题, 避免了自然度的下降。

图 5 是“然后搬到他从来没有住过的新选的别墅”的 3 个系统合成语音的基频曲线, 由于该句话强

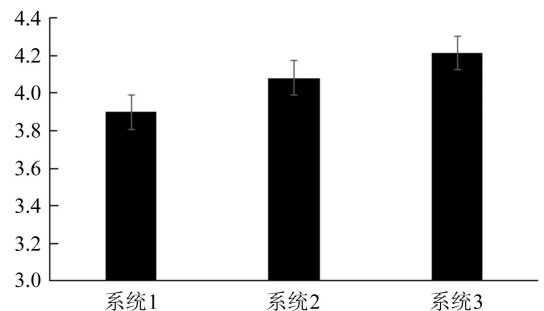


图 4 基于 HMM 的汉语重音合成系统流程图

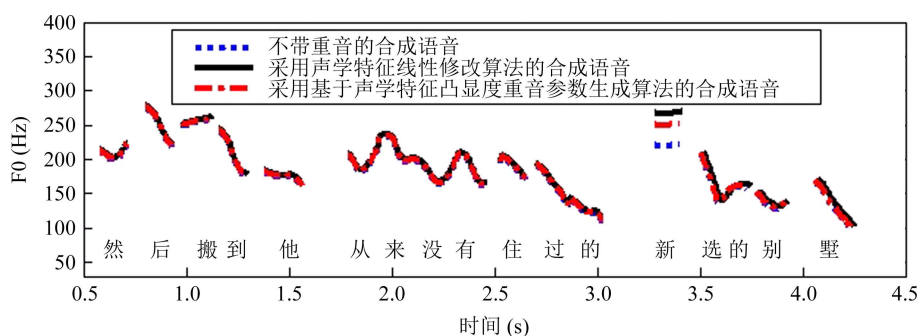


图5 “然后搬到他从来没有住过的 新 选的 别墅”不同系统合成语音的基频曲线

调别墅是新选的,因此“新”为重音。可以看到,与不带重音的合成语音相比,带有重音的合成语音基频曲线起伏更明显,具有更好的表现力。对于重音“新”,由于该音节为韵律短语首音节,且为阴平,其不带有重音时的基频较高,因此,若直接根据重音与非重音基频的统计平均比值进行修改,会有过修改的问题。而基于声学特征凸显度的重音参数生成算法由于考虑到了音节的原始基频凸显度,既达到了凸显重音的目的,又在一定程度上避免了过修改的问题。

6 结论

连续语流中重音的声学表现是多种因素相互影响的结果,如语调下倾。本文根据汉语重音的声学特点,提出了一种重音凸显度的表示方法,去除了语调下倾、声调和韵律短语边界前延长等因素,分析了不同韵律位置重音的基频和时长凸显度的变化规律,发现在韵律单元末(韵律词末音节和韵律短语末韵律词)的重音其基频最大值凸显度要低于非韵律单元末重音,而韵律短语末与韵律词末重音的时长凸显度无明显区别。提出了基于声学特征凸显度的重音声学参数生成算法,算法生成的重音声学参数与临近音节的声学特征呈正相关性,解决了传统线性修改算法的修改幅度过小或过大的问题。实现了支持重音的汉语语音合成系统。实验证明,本文提出的重音声学参数生成算法可以提高合成语音的自然度和表现力。

今后工作主要集中如下2个方面:(1)在文本重音预测模型中引入专家知识,例如“连/就/才”句式后的音节更可能为重音,双音节词中阴平音节更容易获得重音等;(2)对轻读音节的声学特征凸显度进行分析并建模,进一步提高合成语音的自然度。

参 考 文 献

- 1 Strom V, Nenkova A, Clark R *et al.* Modelling prominence and emphasis improves unit-selection synthesis. 2007, 2: 1169—1172
- 2 李雅, 卢颖超, 许小颖等. 连续语流中韵律层级和调型组合对重音感知的影响. 清华大学学报(自然科学版), 2011: 1239—1243+1248
- 3 Li K, Zhang S, Li M *et al.* Prominence model for prosodic features in automatic lexical stress and pitch accent detection. International Speech and Communication Association, 2011: 2009—2012
- 4 曹文. 汉语焦点重音的韵律实现. 北京: 北京语言大学出版社, 2010
- 5 贾媛, 熊子瑜, 李爱军. 普通话焦点重音对语句音高的作用. 2006: 61—68
- 6 陈玉东. 普通话上声重音的音高组合方式. 语言科学, 2005: 54—65
- 7 Xu Y. Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics*, 1999; **27**(1): 55—105
- 8 许洁萍, 初敏, 贺琳等人. 汉语语句重音对音高和音长的影响. 声学学报, 2000; **25**(4): 335—339
- 9 Aijun L. Duration characteristics of stress and its synthesis rules on standard Chinese. Beijing: 1994
- 10 周俏峰, 蔡莲红. 汉语的重音及在 TTS 系统中的模拟. 微型计算机, 1996; **16**(4): 16—19
- 11 Raux A, Black A W. A unit selection approach to F0 modeling and its application to emphasis. 30: 700—705
- 12 朱维彬. 支持重音合成的汉语语音合成系统. 中文信息学报, 2007: 122—128
- 13 Tokuda K, Yoshimura T, Masuko T *et al.* Speech parameter generation algorithms for HMM-based speech synthesis. 2000(3): 1315—1318
- 14 Morizane K, Nakamura K, Toda T *et al.* Emphasized speech synthesis based on hidden Markov models. 10: 76—81
- 15 Maeno Y, Nose T, Kobayashi T *et al.* HMM-based emphatic speech synthesis using unsupervised context labeling. International Speech and Communication Association, 2011: 1849—1852
- 16 Yu K, Mairesse F, Young S. Word-level emphasis modelling in HMM-based speech synthesis. Institute of Electrical and Electronics Engineers Inc., 2010: 4238—4241
- 17 李雅, 潘诗锋, 陶建华. 采用重音调整模型的 HMM 语音合成系统. 清华大学学报(自然科学版), 2011: 1171—1175
- 18 Meng F, Wu Z, Meng H *et al.* Generating emphasis from neutral speech using hierarchical perturbation model by decision tree and support vector machine. IEEE Computer Society, 2012: 442—448
- 19 沈炯. 汉语语调构造和语调类型. 方言, 1994(3): 221—228

-
- 20 沈炯. 汉语语调分类和标记方法试说. 语言文字应用, 1998(1): 104—106
- 21 黄贤军, 高路, 杨玉芳等. 汉语语调音高下倾的实验研究. 声学学报 (中文版), 2009; **34**(2): 158—166
- 22 王安红, 陈明. 基于言语数据库的汉语音高下倾现象研究. 声学学报, 2004; **29**(4): 353—358
- 23 王永鑫, 贾珈, 张雨辰等. 基于 HMM 语音合成的语调控制. 清华大学学报 (自然科学版), 2013(6): 781—786
- 24 曹剑芬. 音段延长的不同类型及其韵律价值. 南京师范大学文学院学报, 2005: 160—167
- 25 倪崇嘉, 刘文举, 徐波. 汉语韵律短语的时长与音高研究. 中文信息学报, 2009: 82—87
- 26 胡伟湘, 董宏辉, 陶建华等. 汉语朗读话语重音自动分类研究. 中文信息学报, 2005: 80—85
- 27 倪崇嘉, 刘文举, 徐波. 基于互补模型的汉语重音检测. 计算机工程, 2011: 20—23
- 28 倪崇嘉, 张爱英, 刘文举. 基于声学相关特征与词典语法相关特征的汉语重音检测. 计算机学报, 2011: 1638—1649
- 29 曹剑芬, 陈方. 基于文本信息的韵律结构预测及其在合成系统中的应用. 2003: 176—181
- 30 邵艳秋, 韩纪庆, 刘挺等. 自然风格言语的汉语句重音自动判别研究. 声学学报, 2006; **31**(3): 203—210