基于决策树的语音基元语境特征权重训练算法

杨鸿武[1]，郭威彤[1]，蔡莲红[2]，吴志勇[2]
(1. 西北师范大学 物理与电子工程学院，甘肃 兰州 730070；
2. 清华大学 计算机科学与技术系 普适计算教育部重点实验室，北京 100084)

摘 要：提出了一种基于决策树的语音合成基元的语境特征权重训练算法. 对语音数据库中的每个带调音节，利用语境相关的问题集和候选基元的频谱距离建立决策树. 对每个要合成的音节，根据其语境特征，获得语音合成系统选择的基元的语境特征 F^* 和该语境特征下决策树叶子结点中基元的语境特征 F'. 统计 F 中每一个语境特征相对于 F^* 的变化，根据语境特征变化的概率对权重进行调整. 实验结果表明，这种方法能够训练出合理的语境特征权重，使得合成语音的自然度有一定提高. 同时，利用这种方法还可以对语音合成系统进行实时优化.

关键词：语音合成；文语转换；基元选取；权重训练
中图分类号：TP 391 **文献标识码**：A **文章编号**：1001-988 X(2007)04-0050-05

CARD based context specified weights training algorism for unit selection in speech synthesis

YANG Hong-wu[1], GUO Wei-tong[1], CAI Lian-hong[2], WU Zhi-yong[2]
(1. College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou, 730070, Gansu, China;
2. Key Laboratory of Pervasive Computing, Ministry of Education; Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: The paper introduces a context specified weights training algorism for contextual features of speech unit in speech synthesis based on Classification and Regression Tree (CART). A CART is created for each tonal syllable with the spectral distance of each candidate unit and the context dependent question set. The increments of contextual features are counted by comparing the units selected by TTS and given by leaf node of CART. The weights of contextual features are then adjusted in accordance with their probability of increment. The experiments demonstrate that a set of reasonable weights can be trained by the algorism so the naturalness of synthetic speech can also be improved. The algorism can also be used to optimize the speech synthesis system online.

Key words: speech synthesis; text-to-speech; unit selection; weight training

近年来，数据驱动的拼接语音合成技术获得了快速发展[1,2]. 在拼接语音合成系统中，合成语音是通过拼接从大规模语音数据库中选取的语音基元实现的，拼接过程中基本不进行语音信号处理，因此可以获得高自然度的合成语音. 在自然语流中，同一个音节在不同的语境(context)中具有不同的韵律特征和音质特征. 为了提高合成语音的自然度，目前高质量的汉语文语转换(Text-to-Speech，TTS)系统都保存同一音节在不同语境下的多个样本作为候选基元集. 在合成语句时，根据语句中的音节所处的语境，从候选基元集中选取最佳基元，使最终的最佳基元序列拼接后获得整体最优的自然

收稿日期：2007-01-02；**修改稿收到日期**：2007-05-28
基金项目：西北师范大学科研骨干培育项目(NWNU-KJCXGC-03-42)
作者简介：杨鸿武 (1969—)，男，甘肃合作人，副教授，在读博士研究生. 主要研究方向为高表现力语音合成.
E-mail: yang-hwo3@mails.tsinghua.edu.cn

度．从候选基元集中选择最佳基元序列是一个搜索问题，其目的是选出使整体代价最小的候选基元序列[1]．代价包括目标代价和拼接代价[1]．目标代价指选择一个与当前语境特征相匹配的最佳基元的代价，它通过目标代价函数来定义[1]

$$C(t,u_i) = \sum_{j}^{p} \omega_j c_j(t,u_i). \qquad (1)$$

(1)式中，t 为目标基元，u_i 为第 i 个候选基元，$C(t,u_i)$ 为第 i 个候选基元的目标代价，$c_j(t,u_i)$ 是第 i 个候选基元的第 j 个语境特征的目标代价，ω_j 是第 j 个语境特征的权重，它反映了第 j 个语境特征在基元选取中的重要性．

由(1)式可知，语境特征的权重设置对基元选取以及最终的合成结果有很大的影响．对于语境特征权重的设置，已有的研究采用机器学习、人工听辨等方法来训练权重．Hunt[1]和 MERON[3]利用权重空间的遍历和线性回归方法对权重进行训练．Alias[4]等基于遗传算法对随机生成的权重样本进行筛选．Park[5]等将基元选取看作模式识别中的分类问题，采取语音识别中区分训练的方法训练权重．以上的方法虽然可以快速地训练出语境特征的权重，但是没有考虑人的感知与语境特征之间的关系．语音合成的目标是要尽量满足人的听觉要求，因此吴志勇[6]在汉语语音合成中提出了基于听辨指导的权重训练方法．这种方法的好处是训练的权重与人的感知结果一致．然而，由于人工听辨是一件极其耗费时力的工作，而且很难保证不同听辨人或同一听辨人在不同时间听辨的一致性．此外，目前的语音数据库包含数万个候选基元，某些音节的候选基元多达几百个，而另外一些音节只有一个候选基元，这就导致人工听辨方法的准确性受到限制，而且很难处理数据稀疏的问题．

目前广泛应用于语音合成中的决策树(Classification and regression tree，CART)[7]可以将语境特征和声学特征相似的候选基元聚成一类，并且能够有效解决数据稀疏的问题．因此，针对汉语语音合成中人工听辨方法的不足，笔者利用决策树来训练基元选取中语境特征的权重．在权重训练过程中，利用决策树将自然语句中的音节按照其在语句中的语境聚类，从而使具有相同语境的音节处在决策树的同一个叶子结点中．由于决策树是利用自然语句建立的，决策树叶子结点中音节的语境特征最符合该音节在自然语句中的语境特征，因此可以用来代替相同语境下人工听辨的结果，从而能够自动进行语境特征权重的训练，以解决人工听辨方法的不足．

1　基于决策树的候选基元聚类

分类与回归树(CART)[7]被广泛应用于语音合成的语音基元选取[8]，它是一个二叉树，每个结点都绑定着一个“Yes/No”问题，所有允许进入根结点的候选基元都要回答结点上绑定的问题，根据回答结果选择进入左枝还是右枝．最终，每个进入根结点的候选基元均根据对一系列结点问题的回答进入到一个叶子结点中．进入同一个叶子结点的候选基元被认为具有相似的语境特征和声学特征．决策树结合了基于数据驱动的方法和基于知识的方法．与基于数据驱动的方法相比，它能够对训练数据稀少的基元给出适当的参数估计；与基于知识的方法相比，它能够弥补专家知识的不足．

1.1　决策树问题集的设计

问题集是供决策树构造使用的问题的集合．结点分裂时所选中的问题与此结点绑定，从而决定哪些基元进入同一个叶子结点．问题集的好坏会影响到语境相关的基元选取的性能．本文中采用基于语音学知识[6]的语境特征作为决策树的问题集．根据语音学知识，一个语音基元的语境特征被划分为若干类，每一类作为决策树的一个问题．在本文中，由于候选基元是音节，所以针对音节设计问题集．问题集采用音节的声调、声韵母、所处位置和边界 4 个方面的 12 个语境特征作为问题(12 个语境特征及其取值如表 1 所示)．

表 1　决策树问题集所用问题及其取值

Table 1　Question set used by CART

语境特征	含义	取值
声调特征	当前音节、前一音节和后一音节的声调类型	0 无；1 阴平；2 阳平；3 上声；4 去声；5 轻声
声母类型	当前音节和后一音节的声母类型	0 无；1 滑音；2 擦音；3 元音；4 鼻音
韵母类型	当前音节和前一音节的韵母类型	0 无；1 开口呼；2 启齿呼；3 撮口呼；4 合口呼．
边界类型	当前音节的前后边界类型	1 音节边界；2 韵律词边界；3 韵律短语边界；4 语句边界；
位置特征	当前音节在韵律词、韵律短语和语句中的位置	1 首；2 中；3 尾

1.2　声学距离的计算

为了对同一音节的不同候选样本进行聚类，需要确定两个候选样本之间的声学距离，以测量聚类

时的不纯度. 声学距离用频谱参数和基频参数来计算. 频谱参数采用加权 mel 倒谱系数[9], 对语音信号用 25 ms 的 Blackman 窗和 5 ms 的帧移, 计算出包括 0 阶在内的 25 阶加权 mel 倒谱系数及其一阶和二阶差分; 基频参数采用对数基频值及其一阶和二阶差分. 频谱参数和基频参数组成 81 维的特征向量, 利用(2)式计算 2 个基元的频谱距离[8]

$$\mathrm{Adist}(U,V) = \frac{WD * |U|}{|V|} * \sum_{i=1}^{|U|}\sum_{j=1}^{n} \cdot \frac{(\mathrm{abs}(F_{ij}(U) - F_{(i*|V|/|U|)j}(V)))}{SD_j * n * |U|}, \quad (2)$$

$|U|$ 是基元 U 中的帧数, $|V|$ 是基元 V 中的帧数, F_{ij} 是基元 U 中第 i 帧的第 j 个特征, SD_j 是特征 j 的标准差, WD 是 2 个基元的时长补偿因子.

1.3 决策树的构造

首先将一个音节的所有候选基元放入决策树的根结点中, 然后选择结点分裂后声学距离减小最大的问题作为本结点绑定的问题, 并对当前结点进行分裂. 当分裂后的结点中候选基元的数目小于阈值(一般为 10～20), 或者当本结点分裂后声学距离的减小小于阈值时, 停止分裂. CART 算法隐含了处理数据稀疏问题的方法, 保证了只分裂具有足够多基元的结点.

2 权重训练算法

2.1 基元选取方法

基元选取的目标是根据待合成文本的语境特征, 从语音数据库中的候选基元集中选出语境特征最为匹配的候选基元. 语音数据库音库中的每个音节都有 m 个候选基元, 每个候选基元都有一个 n 维的语境特征 F, 其分量分别对应表 1 中的 12 个语境特征. 第 i 候选基元的语境特征为 $F_i = \{f_1^i, f_2^i, \cdots, f_n^i\}$. 合成时, 对每一个要合成的音节, 通过文本分析可产生一个语境特征 $F_i = \{f_1, f_2, \cdots, f_n\}$, 需要根据 F 从候选基元集中选择一个最优的候选基元, 使得该候选基元的语境特征 F^* 最接近文本分析获得的语境特征 F, 即找出

$$F^* = \arg\min_{F_i} \|F - F_i\| = \arg\min_{F_i} \sum_{k=1}^{n} (f_k - f_k^i)^2. \quad (3)$$

由于 f_k、f_k^i 基本上只取 0、1 两种值, 表示该语境特征在基元选取中是否有用, 所以可将(3)式中的欧氏距离改成汉明距离

$$F^* = \arg\min_{F_i} \sum_{k=1}^{n} |f_k - f_k^i|. \quad (4)$$

实际操作时引入一个权重系数 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, 以控制不同的语境特征对基元选取的贡献

$$F^* = \arg\min_{F_i} \sum_{k=1}^{n} \omega_k |f_k - f_k^i|. \quad (5)$$

(5)式中, ω 满足归一化条件: $\sum_{k=1}^{n} \omega_k = 1$, $\omega_k \geq 0$.

2.2 权重训练过程

权重训练的目的是通过训练数据获得最优权重系数 ω, 使得 TTS 选择的基元的语境特征最接近训练数据的语境特征. 在权重训练过程中, 对于每一个训练语句中的音节, 根据文本分析获得其语境特征 F, 通过(5)式在所有的候选基元中找到一个语境特征最接近 F 的基元 S, 设其语境特征为 F^*; 同时, 根据 F 查找决策树, 看 S 是否出现在 F 所对应的决策树的叶子结点中. 由于决策树是根据训练数据的语境特征和声学特征建立的, 因此决策树叶子结点中的基元是最适合当前语境的基元. 如果 S 未出现在决策树的叶子结点中, 说明 S 的语境特征 F^* 与训练语句的语境特征不匹配, 需要调整语境特征的权重, 使得 TTS 根据新的权重选取的基元的 F^* 接近决策树叶子结点中某个基元(一般为类中心)的语境. 为此, 需要根据 F^* 和 F 调整(5)式中的权重 ω, 找出一个新的 ω', 使得对大部分决策树叶子结点中选择的基元 $d' > d^*$ 成立, 其中

$$d' = F'^{\mathrm{T}} \cdot \omega', \quad d^* = F^{*\mathrm{T}} \cdot \omega'.$$

2.3 权重调整算法

假设原权重系数为 $\omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$, 新的权重系数为 $\omega' = \{\omega_1', \omega_2', \cdots, \omega_n'\}$, 每个音节有 m 个 TTS 系统选择的结果 F^*, 相应地, 有 m 个决策树叶子结点中选择的结果 F'.

由于要使大部分 d 大于 d^*, 而权重系数 ω 和 ω 又要满足归一化条件, 因此, ω 的某些分量要增加, 某些分量要减小. 由于

$$d^* = F^{*\mathrm{T}} \cdot \omega' = \sum_{i=1}^{n} f_i^* \omega_i', \quad d' = F'^{\mathrm{T}} \cdot \omega' = \sum_{i=1}^{n} f_i \omega_i', \quad d' > d^*, \quad (6)$$

所以对于 F^* 和 F 中的每一个分量 f_i^* 和 f_i ($1 \leq i$

$\leq n)$，如果 $\sum_{k=1}^{m} f_i'' - f_i^{*k}$ 越大，则相应的 $\omega_i' - \omega_i$ 越大．说明从当前候选基元的统计特性上来看，第 i 个语境特征 f_i 具有较强的区分力，属于较关键的语境特征，对权重的修正是正向的，如果

$$\sum_{k=1}^{m} f_i^* - f_i''$$

越大，则相应的 $\omega_i' - \omega_i$ 越小，对权重的修正是负向的；而如果 $\sum_{k=1}^{m} f_i'' - f_i^{*k} = 0$，则 $\omega_i' - \omega_i = 0$，其对权重修正的贡献为 0.

设 $P = \{p_1, p_2, \cdots, p_n\}$，$p_i$ 为 f_i 变大的候选基元数．设 $Q = \{q_1, q_2, \cdots, q_n\}$，$q_i$ 为 f_i 变小的候选基元数．设 $C = P - Q = \{c_1, c_2, \cdots, c_n\}$，$c_i$ 为正，表示 f_i 整体上变大，$|c_i|$ 为变大的净候选基元数．c_j 为负，表示 f_j 整体上变小，$|c_j|$ 为变小的净候选基元数．

第 i 个语境特征的权重调整可以表示为

$$\omega_i' = \omega_i + \Delta\omega_i, \quad (7)$$

式中，$\Delta\omega_i$ 是第 i 个语境特征的权重变化量．

由归一化条件，可得

$$\sum_{i=1}^{n} \omega_i' = \sum_{i=1}^{n} (\omega_i + \Delta\omega_i) = 1, \quad (8)$$

所以

$$\sum_{i=1}^{n} \Delta\omega_i = 0. \quad (9)$$

即所有增加的权重与所有减小的权重相等．令其为 σ，则

$$\sigma = \sum_i \omega_i = \sum_j \omega_j.$$

式中，i 为 c_i 为正的语境特征的下标，j 为 c_j 为负的语境特征的下标．

因为增加的总权重要分配到每个需要增加权重的语境特征上，其中较关键的语境特征(即 c_i 的值大的语境特征)权重的增量应该大，所以对于需要增加权重的语境特征 i 来说，增加的权重为

$$\frac{c_i}{\sum_i c_i} * \sigma \quad (i \text{ 为 } c_i \text{ 为正的韵律特征下标}). \quad (10)$$

同样，对于需要减小权重的语境特征 j 来说，减小的权重为

$$\frac{|c_j|}{\left|\sum_j c_j\right|} * \sigma \quad (j \text{ 为 } c_j \text{ 为负的韵律特征下标}). \quad (11)$$

根据归一化条件，权重的最大增长是使得所有要减小的语境特征的权重全部为 0，所以 σ 的上界不超过 $\sum_j \omega_j$．由于 $0 \leq \omega_k \leq 1$，所以 σ 又满足

$$\begin{cases} \omega_i + \dfrac{c_i}{\sum_i c_i} * \sigma \leq 1, \\ \omega_j - \dfrac{|c_j|}{\left|\sum_j c_j\right|} * \sigma \geq 0. \end{cases} \quad (i, j \text{ 的定义同前}) \quad (12)$$

因此 σ 的上界和下界为

$$0 < \sigma < \min\left(\frac{(1 - \omega_i) * \sum_i c_i}{c_i}, \frac{\omega_j * \sum_j c_j}{c_j} \right). \quad (13)$$

如果 ω 的每个元素的初始值都设为 $\frac{1}{n}$，则

$$0 < \sigma \leq \min\left(\frac{(n - 1) * \sum_i c_i}{n * c_i}, \frac{\left|\sum_j c_j\right|}{n * |c_j|} \right). \quad (14)$$

具体计算时，首先根据训练结果统计出 P 和 Q，并计算出 C，根据 (13) 或 (14) 式计算出 σ 的上界，然后在 0～σ 的上界之间搜索出一个 σ 值，利用 (10) 和 (11) 式调整权重，其过程如图 1 所示．

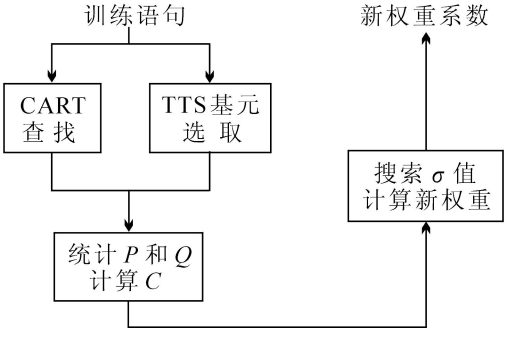

图 1　权重训练过程

Fig 1　Weights training procedure

搜索的约束条件是使得调整后的权重 ω 能够使满足 (6) 式的决策树选择的结果的个数最多．也可以在 0～σ 的上界之间根据实验估计一个 σ 值来直接调整权重．前一种方法适合于离线训练，后一种方法适合于在线训练．

3　实验

共收集了 5 000 多句自然语句语料，由说标准普通话的女性播音员在录音棚录制．语料包含近 85 000个音节，覆盖了汉语 1 268 个有调音节及多种语境特征的搭配关系．对语音语料自动进行了音

节切分和基频标注并进行手工校对. 对文本语料自动进行了韵律边界的预测[10]并手工校对. 利用这些自然语料, 抽取了一个大规模的语音数据库作为实验基础. 根据文本分析的结果和韵律边界的预测结果, 获得每个音节的语境特征, 建立 1 268 个带调单音节的决策树. 建立好决策树后, 让 TTS 系统合成所有的自然语料, 并记录每一个音节的基元选取结果. 同时, 根据句子中音节的韵律特征, 从决策树中找出该音节所在的叶子结点, 选择叶子结点的类中心作为决策树选择的结果. 然后根据第 3 节描述的权重调整算法调整语境特征的权重. 每一个语境特征的初始权重设为$\frac{1}{n}$.

为了评价权重训练算法的效果, 让 TTS 系统在权重训练前和权重训练后分别合成所有训练用的自然语句, 并且统计 TTS 系统从语音数据库中选择出自然语句中相应音节的个数. 结果显示, 在权重训练前有 51. 2 %的音节选自相应的自然语句, 在权重训练后, 这一比例提高到 82. 3 %. 可见, 权重训练后, TTS 系统正确选择出自然语句中音节的个数有显著的增加.

同时, 设计了 ABX 感知实验来进一步评价权重训练前后 TTS 系统合成语音的自然度. 选择了 100 个语句, 邀请 14 个被试参加感知实验. 实验中, 让被试听 3 个内容相同的语音文件: ①录制的自然语音; ②权重调整前 TTS 系统合成的语音; ③权重调整后 TTS 系统合成的语音. 每一个语句按照 ①—②—③或 ①—③—②的顺序播放. 在听的过程中, 要求被试仔细判断 ②或 ③哪一个的自然度与 ①更接近. 实验结果显示, 在权重训练前, 只有 26 %的合成语句被认为更自然, 而在权重训练后, 有 74 %的合成语句被认为更自然. 由此可以看出, 本文提出的韵律特征权重训练算法能够提高 TTS 系统合成语音的自然度.

4 结论

在拼接语音合成系统中, 语音基元的语境特征对合成语音自然度的影响各不相同. 为了提高合成语音的自然度, 需要确定选取语音基元的目标代价函数中各个语境特征的权重. 本文提出的权重训练算法, 根据语境特征利用决策树对音节聚类, 对语句中的每个音节, 通过语音合成系统选择的语音基元和决策树叶子结点中语音基元的语境特征的变化概率对语境特征的权重进行调整. 实验结果表明, 这种方法能够训练出合理的权重, 使得合成语音的自然度有一定提高. 同时, 利用这种方法还可以对语音合成系统进行实时优化.

参考文献:

[1] HUNT A, BLACK A. Unit selection in a concatenative speech synthesis system using a large speech database [C]// *ICASSP*96, Atlanta: IEEE Press, 1996: 373-376.

[2] DONOVAN R E. *Trainable Speech Synthesis* [D]. Cambridge: Cambridge University, 1996.

[3] MERON Y, HIROSE K. Efficient weight training for selection based synthesis [C]// *Euro Speech* 99. Budapest: ISCA Press, 1999: 2319-2322.

[4] FRANCESC A, XAVIER L. Evolutionary weight tuning based on diphone pairs for unit selection speech synthesis [C]// *Euro Speech* 2003. Geneva: ISCA Press, 2003: 1333-1336.

[5] PARK Seung-Seop, KIM Chong-Kyu, KIM Nam-Soo. Discriminative weight training for unit-selection based speech synthesis [C]// *Euro Speech* 2003. Geneva: ISCA Press, 2003: 281-284.

[6] 吴志勇, 蔡莲红, 蔡 锐. 语音合成中基于听辨指导的权重训练算法[J]. **清华大学学报: 自然科学版**, 2005, **45**(1): 52-56.

[7] BREIMAN L. *Classification and Regression Trees* [M]. Pacific Grove, CA: Wadsworth, 1984.

[8] BLACK A, TAYLOR P. Automatically clustering similar units for unit selection in speech synthesis [C]// *Euro Speech* 97. Rhodes: ISCA Press, 1997, **2**: 601-604.

[9] YANG Hong-wu, HUANG De-zhi, CAI Lian-hong. Perceptually weighted mel-cepstral analysis of speech based on psychoacoustic model [J]. *IEICE Transactions on Information and Systems*, 2006, E89-D(12): 2998-3001.

[10] ZHANG Xiao-nan, XU Jun, CAI Lian-hong. Prosodic boundary prediction based on maximum entropy model with error-driven modification [C]// Garbonell J G, Siekmann J. *Lecture Notes in Artificial Intelligence*. Berlin: Springer-Verlag, 2006: 149-160.

(责任编辑 孙晓玲)