[19] 中华人民共和国国家知识产权局

[51] Int. Cl.
G10L 13/02 (2006.01)
G10L 13/08 (2006.01)

[12] 发明专利说明书

专利号 ZL 200510086340.1

[45] 授权公告日 2007 年 11 月 7 日

[11] 授权公告号 CN 100347741C

[22] 申请日 2005.9.2
[21] 申请号 200510086340.1
[73] 专利权人 清华大学
地址 100084 北京市北京 100084 - 82 信箱
[72] 发明人 蔡莲红 叶振兴 倪 昕 黄德智
[56] 参考文献
CN1420486A 2003.5.28
CN1099165A 1995.2.22
CN1118493A 1996.3.13
审查员 吴 芸

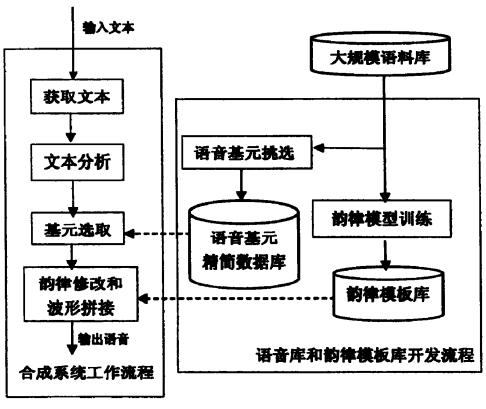
权利要求书 6 页 说明书 16 页 附图 4 页

[54] 发明名称

移动语音合成方法

[57] 摘要

移动语音合成方法属于通信中文字与语音之间信息转换和处理技术领域。 其特征在于，它涉及一种在移动终端设备上进行文语转换的技术。 包括移动语音合成系统的语音库构建、韵律模型的训练、合成方法等。 它包括使用决策树 CART——Classification and RegressionTrees 方法从大规模语音语料库中挑选基元样本，快速地建立适用于移动终端的语音基元精简数据库；一个基于大规模语音语调库的韵律模型训练方法，可以从自然语音中提取基频曲线以生成韵律模板。 从而在移动终端上实现了从文本到语音的转换的方法和系统。 对于待转换的文本，基于 CART 方法选取适当的基元，比对韵律模块修改语音基元，拼接成自然流畅的合成语音。



1. 移动语音合成方法，其特征在于：该方法是一种在移动通信终端设备上进行文字与语音之间相互转换的方法，所述方法是在内存有本方法软件的存储卡、中央处理器和语音输出设备依次串接的系统中实现的，所述方法分为以下两个阶段，即开发阶段和语音合成阶段：

开发阶段包括离线建立的语音基元精简数据库，简称语音库，和训练韵律模板库的开发阶段，该阶段依次含有以下步骤：

步骤 A：从大规模语音语料库内挑选合适的基元样本组成该语音库，该基元样本是指在该语音语料库中的汉字句子的实例，本发明使用汉语有调音节作为语音合成系统的合成基元，基元样本是基于公知的*CART*决策树模型来挑选的，*CART*是 *Cassification and* Re*gression Tree*的缩略语，中文名是分类与回归树，*CART*决策树即利用该语音语料为每个基元训练出来的，通过对该树相应基元的所有样本进行聚集，把韵律和频谱特征相似的样本聚到一起，再选取聚类中心而建成该语音库，步骤*A*依次包含以下子步骤：

步骤 A1：选择能反映一个有调音节的韵律变化、前后音联的九个特征组成该*CART*方法所需的问题集，用*Q*表示在该具有二叉树结构的*CART*树中，每一个非叶子节点包括一个问题，根据对所给问题的回答为“是”与“否”分别指向左、右两个子节点，叶子节点则包括可属于同一基元的所有样本数据；所述九个特征描述如下：

1. *PosInWord*：当前音节在所在韵律词中的位置，共有词首、词中、词尾三种位置；所述 “韵律词”，从韵律学的角度定义为：“最小的能够自由运用的语言单位”，其声学表现是发音连续且紧密，一般包括2-3个音节，音节间没有感知到的停顿；

2. *PosInPhrase*：当前音节在所在韵律短语中的位置，共有短语首、短语中、短语尾三种位置，所述韵律短语由一组韵律词组成，边界处有停顿感或音高的变化；

3. *PosInSentence*：当前音节在所在句子中的位置，共有句首、句中、句尾三种位置；

4. Pr*eTone*：前面音节的音调类型，共有高、低、轻、无四种类型，前面音节的声调为1声或2声时，音调类型取高；前面音节的声调为3声或4声时，音调类型取低；前面音节的声调为轻声时，音调类型取轻；前面没有音节时，音调类型取无；

5. *PostTone*：后面音节的声调类型，共有高、低、轻、无四种类型，后面音节的声调为1声或4声时，音调类型取高；后面音节的声调为2声或3声时，音调类型取低；后面音节的声调为轻声时，音调类型取轻；后面没有音节时，音调类型取无；

6. *LeftFinal*：左邻音节的韵母，所述韵母包含所有韵母；

7. *RightInitial*：右邻音节的声母，包含所有的声母和零声母，所述零声母用“0”表示；

8. *RightInitialClass*：右邻音节声母类别，共分为10类，取值为1,2,3,4,5,6,7,8,9,10，依次表示爆破音、爆破音不送气、爆破音送气、塞擦音、塞擦音不送气、塞擦音送气、擦音清音、擦音浊音、鼻音、边音；

9. *LeftFinalClass*：左邻音节韵母类别，共分为5类，取值为1,2,3,4,5，依次表示韵尾开口呼类、韵尾齐齿呼类、韵尾合口呼类、韵尾撮口呼类、韵尾鼻音；

上述9个特征是从所述语音语料库中文本部分的韵律标注中得出的；

步骤A2：从所述语音语料库的标注文件中提取基元的声学特征参数，用以在后面计算基元样本之间的距离，来度量样本间的相似度，所述特征参数包括：

时长用D表示，音节的时长以采样点个数计；

能量用U表示，音节的均方根能量：$U=\sqrt{\frac{1}{D}\sum_{i=1}^{D}|s(i)|^2}$，

$s(i)$为该样本第i个采样点的幅值；

基频向量用P表示，基频向量P包括三个分量：p_1,p_2,p_3，它们分别是该音节长度的0.15、0.5、0.85处的基频值，该基频值是根据该语音语料中对基音周期所作的标注得到的；

步骤A3：选择节点分裂标准，使得一个叶子节点分裂为两个子节点后，该两个子节点中的样本尽可能地集中，即相互之间距离尽可能地靠近；

本申请采用最大化方差减小量$\Delta E_q(t)$为分裂标准，分裂标准的值越大，则分裂效果越好；$\Delta E_q(t)$定义为：

$$\Delta E_q(t)=E(t)z(t)-\left[E(l)z(l)+E(r)z(r)\right]$$

其中，t为被分裂节点，l、r分别为分裂后的左、右子节点；

$z(t)$、$z(l)$、$z(r)$分别为节点t、l、r中的样本数占所有样本数的比例；

$E(t),E(l),E(r)$分别表示节点t,l,r的能量U、时长D、基频向量P的方差的加权和，以$E(t)$为例描述之，$E(t)$用下式表示：

$$E(t)=w_dE_d(t)+w_uE_u(t)+w_pE_p(t)$$

其中，w_d、w_u、w_p分别为时长、能量、基频向量的权值，是设定的；$E_d(t),E_u(t),E_p(t)$

分别为一个节点中所有样本的时长、能量、基频向量的方差；

步骤A4：构建语音基元精简数据库

对所述语音语料库中的每一个基元训练一棵*CART*树，该树的每一个叶子节点包含了具有相同韵律上下文和音联环境的，听感比较接近的若干基元样本，该步骤A4依次含有以下步骤：

步骤A41：把一个有调音节的所有基元样本作为一个初始类；

步骤A42：采用A1所述的特征，提取步骤A41所述所有基元样本的时长、能量和基频向量；

步骤A43：构建有一个根节点的决策树，该根节点把步骤*A*41所述所有基元样本$x_1, x_2, ...x_i...x_N$作为它的样本，样本x_i的特征向量X_i包括三个分量：D_i、U_i、P_i，它们分别为基元样本x_i的时长、能量和基频向量；

步骤A44：按下式计算步骤*A*43所述每一个节点的样本集中任意两个样本x_j、x_i之间的*Mahalanobis*距离，生成一个$N \times N$的*Mahalanobis*距离矩阵

$$Dist(i,j) = \sqrt{(X_i - X_j)^t S^{-1} (X_i - X_j)}$$

$Dist(i,j)$即*Mahalanobis*距离，S为样本空间协方差矩阵；

设置停止分裂的阀值η；

步骤A45：根据节点分裂标准，从所述根节点开始进行节点分裂；

对于任何一个允许分裂的叶子节点T，用问题集Q中的每一个问题q试分裂该叶子节点T，并计算所得方差减少量ΔE_q，取得所述ΔE_q值最大的问题把该节点T分裂为两个子节点；

步骤A46：继续执行步骤*A*45，一直到分裂后叶子节点中的样本数$\leq \eta$为止；

步骤A47：根据下面公式计算距离类中心最远的样本，并且将其删除，重复此步直到删除10%的样本；

$$k^* = \arg\max_{k=1..N} \sum_{i=1..N, i \neq k} Dist(i,k)$$

步骤A48：假设删除10%的最远的样本后，节点中剩余的样本集为$x_1, x_2, ...x_i..., x_M$；根据公式下面公式计算类的中心样本，取类中心样本作为该叶子节点中所有样本的代表样本；

$$k^* = \arg\min_{k=1..M} \sum_{i=1..M, i \neq k} Dist(i,k)$$

步骤A49：把每个叶子节点的类中的样本用*G*.723.1标准编码压缩，采用*CART*树作为各

叶子节点的索引，建立对各基元的*CART*树的总索引，把最终得到的精简音库复制到移动终端设备的存储卡中供合成使用；

步骤 B：韵律模板的开发，即韵律模型训练，所述步骤*B*依次含有以下各子步骤：

步骤 B1：采用与所述步骤 A1 中出相同的若干上下文环境特征组成*CART*树训练所需的问题集，包括与步骤 A1 所述的相同的九个特征；

步骤 B2：采用与所述步骤 A2 所述的方法提取声学特征参数：提取基元样本的时长作为时长模型的训练参数；提取基元样本在 0.1、0.3、0.5、0.7、0.9 时长处的基频值作为基频曲线模型的训练参数；

步骤 B3：采用步骤 A3 所述的最大化方差减小量$\Delta E_q(t)$作为节点t的分裂标准；

对于时长模型的训练而言：$\Delta E_{dq}(t)=E_d(t)z(t)-\left[E_d(l)z(l)+E_d(r)z(r)\right]$

对于基频曲线模型的训练而言：$\Delta E_{pq}(t)=E_P(t)z(t)-\left[E_P(l)z(l)+E_p(r)z(r)\right]$

其中，各物理量的定义与所述步骤*A*3 中的相同；

步骤 B4：时长模型的训练依次含有以下步骤：

步骤 B41：根据步骤 B1 至 B3 所定义的问题集、时长参数和分裂标准按以下步骤对每一个基元训练一棵*CART*树作为时长预测树；

步骤 B42：统计时长预测树每一个叶子节点中所有样本的时长，按正态分布做参数分布估计，剔除两倍方差之外的样本；

步骤 B43：取剩余样本的时长的平均值作为该叶子节点的时长模板；

步骤 B44：把各叶子节点的时长模板存入韵律模板库中，采用时长预测树作为其基元的索引；

步骤 B5：基频曲线模型的训练；

步骤 B51：根据步骤 B1 至 B3 所定义的问题集、时长参数和分裂标准按以下步骤对每一个基元训练一棵*CART*树作为基频预测树；

步骤 B52：假设*CART*树的一个叶子节点中的样本集为$x_1,x_2,...x_i...,x_N$，样本x_i的特征向量采用如步骤 B2 所述特征，该样本空间的协方差矩阵为S，根据下面公式计算任意两个样本x_j,x_i之间的*Mahalanobis*距离，生成一个$N\times N$的*Mahalanobis*距离矩阵；

$$Dist(i,j)=\sqrt{(X_i-X_j)^t S^{-1}(X_i-X_j)}$$

步骤 B53：根据下述公式计算每个叶子节点中距离类中心最远的样本k^{**}，并且把其删除，重复此步直到删除 10%的样本为止；

$$k^{**} = \arg\max_{i=1...N} \sum_{i=1...N, i=k} Dist(i,k);$$

步骤 B54：假设删除 10%的最远的样本后，节点中剩余的样本集为$\{x_1, x_2, ..., x_M\}$；根据下式计算每个叶子节点的类中心样本k^*，取所述类中的样本为该叶子节点中所有样本的代表样本；

$$k^* = \arg\min_{k=1...M} \sum_{i=1...M, i\neq k} Dist(i,k);$$

步骤 B55：把各叶子节点的类中的样本组成基频曲线模板，经过平滑处理以消除跳变点后，在存入韵律模板库，采用由上述各类中的样本构成的基频曲线预测树作为其基元的索引；

步骤 B57：建立所有基元的时长预测树和基频曲线预测树的总索引，把最终得到的韵律模板复制到移动终端的存储卡中供合成使用；

步骤 C：在语言合成阶段，依次含有以下步骤：包括文本获取、文本分析、基元选取、韵律修改和波形拼接合成：

步骤 C1：文本获取：从键盘输入，或从文件获得，或截取短信息；

步骤 C2：文本分析：首先，识别数字、简写符号或者特殊语义块，并给出对应的在词表上的规范写法；其次，进行分词和词性标注；最后，分析韵律结构，得到韵律词、韵律短语、语调短语三个韵律层级的信息；得到一个目标单元序列，其中，每个目标单元由拼音码和上下文特征组成，所述拼音码由拼音和声调构成；上下文特征包括：该音节在所处句子中的位置、该音节在所处韵律短语中的位置、该音节在所处韵律词中的位置、前音节的拼音码以及后音节的拼音码；

步骤 C3：基元选取，依次含有以下步骤：

步骤 C31：从步骤 C2 得到所含每个目标单元的上下文特征的目标单元序列；

步骤 C32：根据目标单元的拼音码在语音库中检索该基元的*CART* 树；

步骤 C33：根据目标单元的上下文特征对*CART* 树按所述步骤 *A* 1 进行迭代检索，从根节点开始，回答“是”或“否”一直到合适基元样本，再用*G*.723.1标准算法解码，得到原始的波形数据；

步骤 C4：韵律修改和波形拼接

步骤 C41：根据目标单元的上下文特征检索韵律模板库中的时长预测树和基频曲线预测树；

步骤 C42：根据从步骤 C33 所述预测树选取出的基频曲线和时长使用*TD – PSOLA* 算法对解码样本数据的时长和基频进行修改，对于基频而言，当修改量大于 10%停止修改；

步骤 C43：根据步骤 C42 得到的修改过的样本进行拼接；

步骤 C44：根据需要把合成语音输出到文件或声音输出设备。

2.根据权利要求 1 所述的移动语音合成方法，其特征在于：对所述步骤 A43 得到的 $N \times N$ 的 *Mahalanobis* 距离矩阵按下式进行坏节点消除操作：

按下式计算距离类中心最远的样本，并且将其删除；

$$k^{**} = \arg \max_{k=1...N} \sum_{i=1...N, i \neq k} Dist(i,k)$$

重复此步，直到删除 10%的样本。

移动语音合成方法

技术领域

移动语音合成方法属于通信中文字与语音之间信息转换和处理技术领域。本发明涉及一种在移动终端设备特别是智能手机上进行文语转换的技术。

背景技术

文语转换(*Test－To－Speech,TTS*)，是将文字转换成声音的技术，也经常被称为语音合成。目前主流的文语转换系统多采用基于大规模语音语料库的波形拼接式合成方式。为了获得高质量的合成语音，这类合成系统往往需要一个大规模的语音数据库，一个音库动辄几百*MB*，达到*GB*级的音库也已经十分常见；与此同时，随着信息技术的发展，各种各样的移动终端设备如手机、个人数字助理（*PDA*）、车载设备等逐渐得到普及；移动终端上的各种文字信息也呈现出爆炸性增长的趋势，一个典型的例子就是手机短信；但是由于包括手机在内的各种移动终端的屏幕都比较小，所以对文字信息的阅读造成了一定的障碍。如果能够将文字信息转换为语音信息，必将有效提高移动终端的人机交互性能。

基于语料库的拼接式合成技术可以查阅：1．中国发明专利：ZL94103372，发明人：清华大学蔡莲红等，发明名称："基于波形编辑的汉语文字——语音转换方法及系统"；2．中国重大科技成果数据库：编号941008，清华大学蔡莲红等，"汉语文语转换系统*TH－Speech*"；3．中国发明专利：ZL01123418，发明人：北京捷通华声语音技术有限公司吕士楠等，发明名称："一种动态汉语语音合成方法"。

由于移动终端存储空间和计算能力有限，而通用的基于大规模语音语料库的语音合成方法需要一个大规模的音库，基元搜索算法也比较复杂，不能完全满足移动终端的需要。为此我们设计了一种适用于移动终端设备的文语转换技术，称之为移动语音合成技术。

本发明的目的在于针对移动终端设备存储空间和计算能力有限的特点，及其与拼接式合成系统所需要的大规模语音语料库及复杂的基元选取算法之间的矛盾，提出一套语音合成方法和系统；该方法能够大幅度降低合成系统所需的存储空间，并且充分利用拼接式合成方法的优点，合成出具有较高的可懂度和自然度的语音。

发明内容

为了实现上述目的，本发明提出一套技术方案，主要包括三部分。首先，提出一种基于*CART*树的语音库构建方法，可以从大规模语音语料库中进行裁减和优化，得到一个精简的小型语音库；其次，本发明还提出了一个基于大规模语音语料库的韵律模型训练方法，该韵

律模型用来在合成过程中指导基元选取和韵律修改；最后，基于以上步骤所得到的语音库和韵律模型，本发明设计了一套相应的语音合成方法，可以在移动终端设备上实时合成高质量的语音。

1. 基于*CART*树的音库构建方法

移动语音合成系统的语音库构建是从原始语音语料库的大量基元样本中选取少量的最具有代表性的样本，从而达到缩小音库规模的目的。为此，对每个音节构建一棵决策树，决策树的问题集为若干影响音节的韵律特征的上下文特征值，节点分裂标准基于基元样本之间距离的声学度量。每个音节的样本根据决策树的问题集生成若干叶子节点，每个叶子节点代表韵律上下文一致、听感比较接近的若干样本，最后选出每个叶子节点的聚类中心样本代表该叶子节点中的所有样本。

2. 数据驱动的韵律模型训练方法

本发明的韵律模型包括了基频、时长两个模型。在各个韵律特征中，基频曲线有较强的表现力，对于语音的自然度有很大的影响。本发明设计了一个基于*CART*树的基频曲线预测模型，利用一个大规模语音语料库对其进行训练，得到一个基频曲线的韵律模板库。在实时合成的时候，能够根据目标单元的韵律符号描述对韵律模板库进行检索，得到与目标韵律最接近的基频曲线。对时长的预测与基频曲线的预测一样，也是基于*CART*决策树的预测模型。

3. 语音合成方法

包括了基元选取和韵律修改两个模块。考虑到移动语音合成系统在构建音库时裁减掉了大量的基元样本，同时考虑移动终端设备对算法复杂度的要求，设计了一种高效的基元选取方法，能够快速地从音库中选出与目标单元的韵律特征最为接近的基元样本。设计一种高效的韵律修改算法，能够将从文本分析得到韵律上下文环境映射为时长、基频曲线等韵律特征参数，并根据这些参数使用*TD－PSOLA*算法对目标单元进行修改。最后将修改后的目标单元序列进行拼接即得到最终的合成语音。

本发明的特征在于：

该方法是一种在移动通信终端设备上进行文字与语音之间相互转换的方法，所述方法是在内存有本方法软件的存储卡、中央处理器和语音输出设备依次串接的系统中实现的，所述方法分为以下两个阶段，即开发阶段和语音合成阶段：

开发阶段包括离线建立的语音基元精简数据库，简称语音库，和训练韵律模板库的开发阶段，该阶段依次含有以下步骤：

步骤A：从大规模语音语料库内挑选合适的基元样本组成该语音库，该基元样本是指在该语音语料库中的汉字句子的实例，本发明使用汉语有调音节作为语音合成系统的合成基元，基元样本是基于公知的*CART*决策树模型来挑选的，*CART*是 *Cassification and* Re*gression*

Tree 的缩略语，中文名是分类与回归树，*CART* 决策树即利用该语音语料为每个基元训练出来的，通过对该树相应基元的所有样本进行聚集，把韵律和频谱特征相似的样本聚到一起，再选取聚类中心而建成该语音库，步骤 *A* 依次包含以下子步骤：

步骤 A1：选择能反映一个有调音节的韵律变化、前后音联的九个特征组成该 *CART* 方法所需的问题集，用 *Q* 表示在该具有二叉树结构的 *CART* 树中，每一个非叶子节点包括一个问题，根据对所给问题的回答为“是”与“否”分别指向左、右两个子节点，叶子节点则包括可属于同一基元的所有样本数据；所述九个特征描述如下：

1. *PosInWord*：当前音节在所在韵律词中的位置，共有词首、词中、词尾三种位置；所述 “韵律词”，从韵律学的角度定义为：“最小的能够自由运用的语言单位”，其声学表现是发音连续且紧密，一般包括 2-3 个音节，音节间没有感知到的停顿；

2. *PosInPhrase*：当前音节在所在韵律短语中的位置，共有短语首、短语中、短语尾三种位置，所述韵律短语由一组韵律词组成，边界处有停顿感或音高的变化；

3. *PosInSentence*：当前音节在所在句子中的位置，共有句首、句中、句尾三种位置；

4. Pr*eTone*：前面音节的音调类型，共有高、低、轻、无四种类型，前面音节的声调为 1 声或 2 声时，音调类型取高；前面音节的声调为 3 声或 4 声时，音调类型取低；前面音节的声调为轻声时，音调类型取轻；前面没有音节时，音调类型取无；

5. *PostTone*：后面音节的声调类型，共有高、低、轻、无四种类型，后面音节的声调为 1 声或 4 声时，音调类型取高；后面音节的声调为 2 声或 3 声时，音调类型取低；后面音节的声调为轻声时，音调类型取轻；后面没有音节时，音调类型取无；

6. *LeftFinal*：左邻音节的韵母，所述韵母包含所有韵母；

7. *RightInitial*：右邻音节的声母，包含所有的声母和零声母，所述零声母用“0”表示；

8. *RightInitialClass*：右邻音节声母类别，共分为 10 类，取值为 1, 2, 3, 4, 5, 6, 7, 8, 9, 10，依次表示爆破音、爆破音不送气、爆破音送气、塞擦音、塞擦音不送气、塞擦音送气、擦音清音、擦音浊音、鼻音、边音；

9. *LeftFinalClass*：左邻音节韵母类别，共分为 5 类，取值为 1, 2, 3, 4, 5，依次表示韵尾开口呼类、韵尾齐齿呼类、韵尾合口呼类、韵尾撮口呼类、韵尾鼻音；

上述 9 个特征是从所述语音语料库中文本部分的韵律标注中得出的；

步骤 A2：从所述语音语料库的标注文件中提取基元的声学特征参数，用以在后面计算基元样本之间的距离，来度量样本间的相似度，所述特征参数包括：

时长用D表示，音节的时长以采样点个数计；

能量用U表示，音节的均方根能量：$U=\sqrt{\frac{1}{D}\sum_{i=1}^{D}|s(i)|^2}$，

$s(i)$为该样本第i个采样点的幅值；

基频向量用P表示，基频向量P包括三个分量：p_1,p_2,p_3，它们分别是该音节长度的0.15、0.5、0.85处的基频值，该基频值是根据该语音语料中对基音周期所作的标注得到的；

步骤A3：选择节点分裂标准，使得一个叶子节点分裂为两个子节点后，该两个子节点中的样本尽可能地集中，即相互之间距离尽可能地靠近；

本申请采用最大化方差减小量$\Delta E_q(t)$为分裂标准，分裂标准的值越大，则分裂效果越好；$\Delta E_q(t)$定义为：

$$\Delta E_q(t)=E(t)z(t)-\left[E(l)z(l)+E(r)z(r)\right]$$

其中，t为被分裂节点，l、r分别为分裂后的左、右子节点；

$z(t)$、$z(l)$、$z(r)$分别为节点t、l、r中的样本数占所有样本数的比例；

$E(t),E(l),E(r)$分别表示节点t,l,r的能量U、时长D、基频向量P的方差的加权和，以$E(t)$为例描述之，$E(t)$用下式表示：

$$E(t)=w_dE_d(t)+w_uE_u(t)+w_pE_p(t)$$

其中，w_d、w_u、w_p分别为时长、能量、基频向量的权值，是设定的；$E_d(t),E_u(t),E_p(t)$分别为一个节点中所有样本的时长、能量、基频向量的方差。

步骤A4：构建语音基元精简数据库

对所述语音语料库中的每一个基元训练一棵*CART*树，该树的每一个叶子节点包含了具有相同韵律上下文和音联环境的，听感比较接近的若干基元样本，该步骤A4依次含有以下步骤：

步骤A41：把一个有调音节的所有基元样本作为一个初始类；

步骤A42：采用A1所述的特征，提取步骤A41所述所有基元样本的时长、能量和基频向量；

步骤A43：构建有一个根节点的决策树，该根节点把步骤*A*41所述所有基元样本$x_1,x_2,...x_i...x_N$作为它的样本，样本x_i的特征向量X_i包括三个分量：D_i、U_i、P_i，它们分别为

基元样本x_i的时长、能量和基频向量；

步骤 A44：按下式计算步骤*A* 43 所述每一个节点的样本集中任意两个样本x_j、x_i之间的*Mahalanobis* 距离，生成一个$N\times N$的*Mahalanobis* 距离矩阵

$$Dist(i,j)=\sqrt{(X_i-X_j)'S^{-1}(X_i-X_j)}$$

$Dist(i,j)$即*Mahalanobis* 距离，S为样本空间协方差矩阵。

设置停止分裂的阀值η；

步骤 A45：根据节点分裂标准，从所述根节点开始进行节点分裂；

对于任何一个允许分裂的叶子节点T，用问题集Q中的每一个问题q试分裂该叶子节点T，并计算所得方差减少量ΔE_q，取得所述ΔE_q值最大的问题把该节点T分裂为两个子节点；

步骤 A46：继续执行步骤*A* 45，一直到分裂后叶子节点中的样本数$\leq\eta$为止；

步骤 A47：根据下面公式计算距离类中心最远的样本，并且将其删除，重复此步直到删除 10%的样本；

$$k^*=\arg\max_{k=1..N}\sum_{i=1..N,i\neq k}Dist(i,k)$$

步骤 A48：假设删除 10%的最远的样本后，节点中剩余的样本集为$x_1,x_2,...x_i...,x_M$。根据公式下面公式计算类的中心样本，取类中心样本作为该叶子节点中所有样本的代表样本。

$$k^*=\arg\min_{k=1..M}\sum_{i=1..M,i\neq k}Dist(i,k)$$

步骤 A49：把每个叶子节点的类中的样本用*G*.723.1标准编码压缩，采用*CART* 树作为各叶子节点的索引，建立对各基元的*CART* 树的总索引，把最终得到的精简音库复制到移动终端设备的存储卡中供合成使用；

步骤 B：韵律模板的开发，即韵律模型训练，所述步骤*B* 依次含有以下各子步骤：

步骤 B1：采用与所述步骤 A1 中由相同的若干上下文环境特征组成*CART* 树训练所需的问题集，包括与步骤 A1 所述的相同的九个特征；

步骤 B2：采用与所述步骤 A2 所述的方法提取声学特征参数：提取基元样本的时长作为时长模型的训练参数；提取基元样本在 0.1、0.3、0.5、0.7、0.9 时长处的基频值作为基频曲线模型的训练参数；

步骤 B3：采用步骤 A3 所述的最大化方差减小量$\Delta E_q(t)$作为节点t的分裂标准；

对于时长模型的训练而言：$\Delta E_{dq}(t)=E_d(t)z(t)-\left[E_d(l)z(l)+E_d(r)z(r)\right]$

对于基频曲线模型的训练而言：$\Delta E_{pq}(t)=E_P(t)z(t)-\left[E_P(l)z(l)+E_p(r)z(r)\right]$

其中，各物理量的定义与所述步骤 *A* 3 中的相同；

步骤 B4：时长模型的训练依次含有以下步骤：

步骤 B41：根据步骤 B1 至 B3 所定义的问题集、时长参数和分裂标准按以下步骤对每一个基元训练一棵 *CART* 树作为时长预测树。

步骤 B42：统计时长预测树每一个叶子节点中所有样本的时长，按正态分布做参数分布估计，剔除两倍方差之外的样本；

步骤 B43：取剩余样本的时长的平均值作为该叶子节点的时长模板；

步骤 B44：把各叶子节点的时长模板存入韵律模板库中，采用时长预测树作为其基元的索引；

步骤 B5：基频曲线模型的训练。

步骤 B51：根据步骤 B1 至 B3 所定义的问题集、时长参数和分裂标准按以下步骤对每一个基元训练一棵 *CART* 树作为基频预测树。

步骤 B52：假设 *CART* 树的一个叶子节点中的样本集为 $x_1,x_2,...x_i...,x_N$，样本 x_i 的特征向量采用如步骤 B2 所述特征，该样本空间的协方差矩阵为 S，根据下面公式计算任意两个样本 x_j,x_i 之间的 *Mahalanobis* 距离，生成一个 $N\times N$ 的 *Mahalanobis* 距离矩阵：

$$Dist(i,j)=\sqrt{(X_i-X_j)'S^{-1}(X_i-X_j)}$$

步骤 B53：根据下述公式计算每个叶子节点中距离类中心最远的样本 k^{**}，并且把其删除，重复此步直到删除 10%的样本为止；

$$k^{**}=\arg\max_{i=1...N}\sum_{i=1K\,N,i=k}Dist(i,k);$$

步骤 B54：假设删除 10%的最远的样本后，节点中剩余的样本集为 $\{x_1,x_2,...,x_M\}$。根据下式计算每个叶子节点的类中心样本 k^*，取所述类中的样本为该叶子节点中所有样本的代表样本；

$$k^*=\arg\min_{k=1K\,M}\sum_{i=1K\,M,i\neq k}Dist(i,k);$$

步骤 B55：把各叶子节点的类中的样本组成基频曲线模板，经过平滑处理以消除跳变点后，在存入韵律模板库，采用由上述各类中的样本构成的基频曲线预测树作为其基元的索引；

步骤 B57：建立所有基元的时长预测树和基频曲线预测树的总索引，把最终得到的韵律模板复制到移动终端的存储卡中供合成使用；

步骤 C：语言合成阶段。该阶段依次含有以下步骤：包括文本获取、文本分析、基元选取、韵律修改和波形拼接合成：

步骤 C1：文本获取：从键盘输入，或从文件获得，或截取短信息；

步骤 C2：文本分析：首先，识别数字、简写符号或者特殊语义块，并给出对应的在词表上的规范写法；其次，进行分词和词性标注；最后，分析韵律结构，得到韵律词、韵律短语、语调短语三个韵律层级的信息；得到一个目标单元序列，其中，每个目标单元由拼音码和上下文特征组成，所述拼音码由拼音和声调构成；上下文特征包括：该音节在所处句子中的位置、该音节在所处韵律短语中的位置、该音节在所处韵律词中的位置、前音节的拼音码以及后音节的拼音码；

步骤 C3：基元选取，依次含有以下步骤：

步骤 C31：从步骤 C2 得到所含每个目标单元的上下文特征的目标单元序列；

步骤 C32：根据目标单元的拼音码在语音库中检索该基元的 *CART* 树；

步骤 C33：根据目标单元的上下文特征对 *CART* 树按所述步骤 *A* 1 进行迭代检索，从根节点开始，回答“是”或“否”一直到合适基元样本，再用 *G*.723.1 标准算法解码，得到原始的波形数据；

步骤 C4：韵律修改和波形拼接

步骤 C41：根据目标单元的上下文特征检索韵律模板库中的时长预测树和基频曲线预测树；

步骤 C42：根据从步骤 C33 所述预测树选取出的基频曲线和时长使用 *TD－PSOLA* 算法对解码样本数据的时长和基频进行修改，对于基频而言，当修改量大于 10%停止修改；

步骤 C43：根据步骤 C42 得到的修改过的样本进行拼接；

步骤 C44：根据需要把合成语音输出到文件或声音输出设备。

本发明所述的移动语音合成方法，其特征在于：对所述步骤 A43 得到的 $N\times N$ 的 *Mahalanobis* 距离矩阵按下式进行坏节点消除操作：

按下式计算距离类中心最远的样本，并且将其删除；重复此步，直到删除 10%的样本；

$$k^{**}=\arg\max_{k=1\mathrm{K}\,N}\sum_{i=1\mathrm{K}\,N,i\neq k}Dist(i,k)$$

本发明提出一种针对移动终端设备的，基于拼接式合成方式的语音合成方法。

本发明提出一个移动语音合成系统的语音库构建方法，其特点是从现有的大型语音基元数据库出发，首先进行韵律特征的声学参数分析，然后对语音库中的每一个基元训练一课 *CART* 树，根据 *CART* 树的聚类结果挑选基元样本，最后采用参数编码算法对合成基元进行压缩，得到一个小型的嵌入式语音基元样板数据库--精简语音库。

本发明设计了一个适用于移动语音合成系统的韵律模型，包括时长、基频曲线模板库的训练方法及其在合成中的使用。

本发明设计了移动语音合成的核心合成方法，包括基元选取方法和韵律修改方法，基于这些方法，利用前面得到的精简语音库和韵律模板库能够生成高质量的合成语音。

为了检验合成系统的效果，我们对移动语音合成系统与*PC*机上的大规模通用*TTS*系统作了主观听辨实验并进行比较，比较结果如表1所示。

表1

	音库中的样本数目	音库大小	采样率	编码算法	可懂度	自然度
大规模通用TTS系统	93962	750MB	16k	PCM	90.2%	3.90
移动语音合成系统	6755	1.5MB	8k	G.723.1	86.5%	3.71

实验结果表明，在大幅度降低音库规模的前提下，本发明的移动语音合成系统的自然度和可懂度依然能够达到和大规模通用*TTS*系统比较接近的水平。

附图说明：

本发明的一个实例通过下列图表来说明：

图1是一个适用于本发明实施例的移动终端设备系统方框图。

图2是移动语音合成系统流程图，包括离线的语音库构建和韵律模型训练，以及在线的实时合成流程。

图3是音库构建流程图；

图4是基频曲线模型训练流程图；

图5是基元选取流程图；

图6是韵律修改和波形拼接的流程图。

具体实施方式：

图1描述了一个适用于本发明实施例的移动终端系统。该系统包括处理器、存储卡、通信电路和一些输入输出设备，如键盘、液晶显示器、声音输出设备等。实施本发明的软件存储在存储卡中，处理器可以对内存进行存取，并运行相应的指令，以获得本发明的实施结果。

图2是移动语音合成系统的流程图。从图2可以看出，本发明的技术方案包括两大部分：图中右半部分是离线准备工作流程，包括语音库构建和韵律模板库训练等过程；左半部分是在线实时合成的流程图，包括文本分析、基元选取、韵律修改和波形拼接等过程。下面将逐一解释之。

一、语音库构建过程

为了构建适用于移动终端的语音合成系统，首先需要构建一个小型的语音库。本发明采用基于统计学习的方法是从一个大规模语音语料库中，为每个基元挑选合适的基元样本，并

用此构建所需要的精简语音库。具体方法是，首先利用大规模语音语料库为每个基元训练一棵*CART*树，通过该树对该基元的所有样本进行聚类，可以把具有接近的韵律和频谱特征的样本聚到一起，然后选取各个聚类中心组成一个精简语音库。

基元是指拼接式语音合成系统中的最基本的合成单元，在本发明中使用有调音节作为基元，例如“shi4”就是一个基元。基元样本是指一个基元在音库中的实例，如“他是士兵”这句话就包括了基元“shi4” 的两个样本。

移动语音合成系统语音库的构建基于一个大规模语音语料库，该语料库包括约几千或更多的汉语句子，每个句子由文本和录音组成。文本部分包括汉字、拼音、韵律层级标注。其中韵律层级包括句子、语调短语、韵律短语、韵律词。录音部分包括普通话录音数据，音节边界标注，基频曲线标注。整个语料库覆盖了比较全面的韵律环境和前后音联环境。

*CART*是一种常见的决策树模型，该模型是一种统计学习工具，可以用来分类或者聚类。本发明使用的是它的聚类功能，通过*CART*把训练数据中具有相同的韵律上下文环境和前后音联环境、声学特征比较接近的基元样本分别聚到相同的类中。

*CART*采用了一种二叉树结构，树的每一个非叶子节点包含一个问题，根据对该问题的回答是“是”还是“否”分别指向左子节点和右子节点，每一个叶子节点则包括了属于同一类别的样本数据。*CART*的训练过程就是将一个包含了所有训练样本的节点反复分裂，最后生成一棵二叉树的过程。为了训练*CART*树，首先要解决两个方面的问题：（1）根据什么来分裂一个节点，为此需要选择一个问题集Q，问题集中的每一个问题代表了训练样本的一个特征；（2）选取哪个问题来分裂一个节点可以达到最好的效果，为此需要选择若干声学特征参数以计算不同样本之间的相似性，并设计一个分裂标准以衡量分裂效果的好坏。

下面分别介绍问题集Q的选择、声学参数的提取、分裂标准的设计、*CART*树的训练方法和音库的构建过程。

1．问题集Q的选择

问题集Q由合成基元的若干特征组成，选择问题集的特征需要满足两点：（1）该特征能够影响一个基元样本的听觉特性，（2）在合成系统中，该特征能够从文本分析模块获得。我们共选取九个影响一个音节的韵律变化和前后音联的特征值来组成问题集Q：

1. *PosInWord*：当前音节在所在韵律词中的位置，取词首--head、词中--body、词尾--tail三个值；

2. *PosInPhrase*：当前音节在所在韵律短语中的位置，取短语首--head、短语中--body、短语尾--tail三个值；

3. *PosInSentence*：当前音节在所在句子中的位置，取句首--head、句中--body、句尾--tail 三个值；

4. Pr*eTone*：前音节的声调类型，取高--high、低--low、轻--neutral、无--null 四个值，当前面音节的声调为 1 声或 2 声时取为 high，前面音节的声调为 3 声或 4 声时为 low，前面音节的声调为轻声时取 neutral，前面没有音节时取 null；

5. *PostTone*：后面音节的声调类型，取高--high、低--low、轻--neutral、无--null 四个值，当后面音节的声调为 1 声或 4 声时取为 high，后面音节的声调为 2 声或 3 声时为 low，后面音节的声调为轻声时取 neutral，后面没有音节时取 null；

6. *LeftFinal*：左邻音节的韵母，包含所有韵母：a, ai, ao, an, ang, o, ou, e, E, ei, en, eng, er, -i, i, ia, iao, ian, iang, ie, iu, in, ing, iong, iou, u, ua, uo, uai, uei, ui, uan, uen, uang, ueng, ong, v, ue, van, un, ve, ive, iuan, iue, vn, iun ；

7. *RightInitial*：右音节的声母，包含所有声母和零声母:b, ch, c, d, f, g, h, j, k, l, m, n, p, q, r, sh, s, t, x, zh, z, 0；

8. *RightInitialClass*：右邻音节声母类别，共分为 10 类，取值：1, 2, 3, 4, 5, 6, 7, 8, 9, 10，具体分类标准参考表 2；

9. *LeftFinalClass*：左邻音节韵母类别，共分为 5 类，取值：1, 2, 3, 4, 5，具体分类标准参考表 3；

表2 声母分类表

类别	内容
1 爆破音	"b","d","g","p","t","k"
2 爆破音不送气	"b","d","g"
3 爆破音送气	"p","t","k"
4 塞擦音	"z","zh","j","c","ch","q"
5 塞擦音不送气	"z","zh","j"
6 塞擦音送气	"c","ch","q"
7 擦音清音	"f","s","sh","x","h"
8 擦音浊音	"r"
9 鼻音	"m","n"
10 边音	"l"

表3 韵母分类表

类别	内容
1 韵尾开口呼类	"a","ao","o","ou","e","er","ia","ie",

	"iao","iou","ua","uo","ve"
2 韵尾齐齿呼类	"i","ai","uai","uei","ui","-i","ei"
3 韵尾合口呼类	"u"
4 韵尾撮口呼类	"v"
5 韵尾鼻音	"ian","in","iang","ing","iong","uan","uen","uang", "ueng","ong","un","an","ang","en","eng","van"

对于训练数据来说，以上特征之都可以从语料库的文本标注中得到，不同的特征值对不同的样本的分类所起的作用的重要性不同，*CART* 训练算法会自动选择效果最好的问题来分裂节点。对于合成系统来说，以上特征都能在文本分析模块中根据输入的文本对其进行预测。

2. 声学参数的提取

为了计算不同基元样本之间距离，以度量样本之间的相似度，我们选择了四种声学参数作为特征值：

时长 D，合成基元样本的时长，以采样点为单位；

能量 U，合成基单元的均方根能量，$U=\sqrt{\frac{1}{D}\sum_{i=1}^{D}|s(i)|^2}$，其中 $s(i)$ 为该单元的第 i 个采样点的幅值。

基频向量 P，基频是说话时噪音的频率表示，是声带振动的频率，感知为音高，反映在语音信号上是信号的准周期性；该周期即为基音周期，其倒数即基频；基频向量 P 包括三个分量：p_1, p_2, p_3，它们分别是该音节长度的 0.15、0.5、0.85 处的基频值，为了减小标注误差的影响，该基频值由前后若干基频标注点的值平滑后得到；对于零声母的音节，如果前面缺少基频标注则采用样条插值算法来计算缺少的基频值。

对于训练数据，以上特征值可以直接从语料库的标注文件中提取。

3. 节点分裂标准

在 *CART* 训练过程中，一个叶子节点中的样本可以根据其对问题集中某个问题的回答是“是”还是“否”而分裂为两个子节点。如果分裂后两个子节点中的样本比较集中，即相互之间的距离比较近，该问题的分裂效果比较好。为了衡量选取哪个问题可以更好的分裂一个节点，需要定义一个分裂标准，分裂标准的值越大，分裂效果越好。这里采用最大化方差减小量 $\Delta E_q(t)$ 为分裂标准，$\Delta E_q(t)$ 定义为：

$$\Delta E_q(t)=E(t)z(t)-[E(l)z(l)+E(r)z(r)]$$

其中 t 代表被分裂节点，l 和 r 分别代表分裂后的左右子节点，$z(t)$、$z(l)$、$z(r)$ 分别为节点 t、l、r 中的样本数占所有样本数得比例。

对于一个节点t，$E(t)$定义为能量U、时长D、基频向量P的方差的加权和：

$$E(t)=w_dE_d(t)+w_uE_u(t)+w_pE_p(t)$$

其中w_d,w_u,w_p分别为时长、能量、基频向量的权值；$E_d(t),E_u(t),E_p(t)$分别为一个节点中所有样本的时长、能量、基频向量的方差。

4．语音库构建过程

有了问题集Q、声学特征参数和分裂标准$\Delta E_q(t)$，可以根据图 3 对语料库中的每一个基元训练一棵*CART*树，树的每一个叶子节点包含了具有相同的韵律上下文和音联环境的、听感比较接近的若干样本。

因为移动语音合成系统音库的构建过程就是从一个大规模语音语料库中选取少量最有代表性的样本来组成一个小型的音库，因此利用这些*CART*树就可以构建移动语音合成系统所需要的精简音库。构建音库的工作就是从每个叶子节点中选出最有代表性的一个样本，而*CART*树则作为这些样本的索引。由于可能存在一些坏的样本，比如发音不完全或者标注有错误的样本，因此 我们把这个工作分两步进行，第一步是去除坏的节点，第二步是选取有代表性的样本。

（1）假设*CART*树的一个叶子节点中的样本集为$x_1,x_2,...x_i,...x_N$，样本x_i的特征向量为$X_i=\{D_i,U_i,P_i\}$，其中D_i,U_i,P_i分别为样本x_i的时长、能量、基频向量，该样本空间的协方差矩阵为 S，根据下面公式计算任意两个样本x_j,x_i之间的*Mahalanobis*距离，生成一个$N\times N$的*Mahalanobis*矩阵。

$$\mathrm{Dist}(i,j)=\sqrt{(X_i-X_j)'S^{-1}(X_i-X_j)}$$

（2）根据下面公式计算距离类中心最远的样本，并且将其删除。重复此步直到删除 10%的样本；

$$k^*=\arg\max_{k=1..N}\sum_{i=1..N,i\neq k}Dist(i,k) \qquad (0.1)$$

（3）假设删除 10%的最远的样本后，节点中剩余的样本集为$x_1,x_2,...,x_i,...x_M$。根据下面公式计算类的中心样本，取类中心样本作为该叶子节点中所有样本的代表样本。

$$k^*=\arg\min_{k=1..M}\sum_{i=1..M,i\neq k}Dist(i,k) \qquad (0.2)$$

最后，为了进一步降低存储空间，将所有样本以 8k 采样率做重采样，采用 G.723.1 标准算法做编码压缩，并且将二叉树树存储为二进制格式作为音库的索引。经过以上步骤，

得到一个适用于移动语音合成系统的精简音库，将其复制到移动终端的存储卡中供合成系统使用。

二、训练韵律模板库

韵律对合成语音的自然度有较大的影响，由于移动语音合成系统音库中的基元样本数量比较小，为了避免合成语音的韵律过于单调，需要一个韵律模型。韵律模型能够根据从文本分析模块得到的韵律上下文来预测韵律声学参数，如时长、能量、基频曲线等。由合成系统的基元选取模块选出基元的某个样本，然后根据预测出的韵律参数对选出的基元样本作韵律修改，以提高合成语音自然度。常见的韵律模型有基于规则的韵律模型和基于语料库的数据驱动的韵律模型。本发明的韵律模型属于后者，包括一个时长模型和一个基频曲线模型，均在大规模语音语料库的基础上采用*CART*算法训练生成。同音库构建过程一样，韵律模型的训练也包括问题集的选择、声学参数的选择、分裂标准设计、*CART*训练、生成韵律模板库几个方面，下面分别叙述各个步骤。

1 特征集的选择

（1）问题集的选择：

与构建音库的*CART*一样，共选择九个韵律环境和音联环境特征组成问题集Q：

PosInWord，*PosInPhrase*，*PosInSentence*，*PreTone*，*PostTone*，*LeftPhone*，*RightPhone*，*RightPhoneType*，*LeftPhoneType*。

（2）声学特征参数的提取：

对于时长模型来说，声学参数即取基元的时长D;

对于基频曲线模型，声学参数取基频曲线上的五个样值，分别是该音节长度的0.1、0.3、0.5、0.7、0.9处的基频值，为了减小标注误差的影响，该基频值由前后若干基频标注点的值平滑后得到。对于零声母的音节，如果前面缺少基频标注则采用样条插值算法来计算缺少的基频值。

（3）分裂标准

分裂标准亦采用最大化方差减小量。

时长模型：$\Delta E_{dq}(t)=E_d(t)z(t)-[E_d(l)z(l)+E_d(r)z(r)]$;其中$E_d(t)$为一个节点中所有样本的时长的方差。

基频模型：$\Delta E_{pq}(t)=E_p(t)z(t)-[E_p(l)z(l)+E_p(r)z(r)]$;其中$E_p(t)$为一个节点中所有样本的基频特征向量P的方差。

2 训练韵律模板

有了上述准备工作，可以根据图 3 的流程图对语料库中的每一个基元训练一棵基频曲线预测树。时长预测树的训练方法与基频曲线预测树相同。下面分别叙述如何从 *CART* 树生成所需的时长模型和基频曲线模型。

时长模型

首先统计时长预测树每一个叶子中的所有样本的时长，按照正态分布做参数估计，剔除两倍方差之外的样本；

取剩余样本的时长的平均值作为该叶子节点的时长模板；

将各叶子节点的时长模板存入韵律模板库中，采用时长预测树作为其索引。

基频曲线模型

假设 *CART* 树的一个叶子节点中的样本集为 $\{x_1, x_2, ..., x_N\}$，样本 x_i 的特征向量为 $X_i = \{F0_{i1}, F0_{i2}, ..., F0_{i5}\}$，该样本空间的协方差矩阵为 S。根据下面公式计算任意两个样本 x_j, x_i 之间的 *Mahalanobis* 距离，生成一个 $N \times N$ 的 *Mahalanobis* 距离矩阵。

$$\mathrm{Dist}(i,j) = \sqrt{(X_i - X_j)' S^{-1} (X_i - X_j)}$$

根据下面公式计算距离类中心最远的样本，并且将其删除。重复此步直到删除 10%的样本；

$$k^* = \arg\max_{k=1..N} \sum_{i=1..N, i \neq k} Dist(i,k)$$

假设删除 10%的最远的样本后，节点中剩余的样本集为 $\{x_1, x_2, ..., x_M\}$。根据下面公式计算类的中心样本，取类中心样本作为该叶子节点中所有样本的代表样本。

$$k^* = \arg\min_{k=1..M} \sum_{i=1..M, i \neq k} Dist(i,k)$$

对基频曲线模板作平滑处理，消除跳变点，存入韵律模板库。采用基频曲线预测树作为其索引。

最后，将生成的韵律模板库复制到移动终端的存储卡中，供合成系统使用。

合成方法

语音的合成包括文本获取、文本分析、基元选取、韵律修改和波形拼接等步骤。

1．文本获取

根据应用的不同，文本获取可能有不同的途径，比如键盘输入、从文件获取、截取短信息等。以手机短信语音合成为例，既可以实时截取刚刚从通信线路上接收到的短信息，也可以从存储卡中的收件箱里提取已保存的短信息。

2．文本分析

文本分析模块首先对获取的文本作规范化，把其中的数字，简写符号或者特殊语义块识别出来，并给出它们对应的在词表上的规范写法。然后进行分词和词性标注。最后作韵律结构分析，得到韵律词、韵律短语、语调短语三个韵律层级的信息。

文本分析模块最终得到一个目标单元序列，其中每个目标单元由拼音码（拼音+声调）和上下文特征组成，其中上下文特征包括：该音节在所处句子中的位置、该音节在所处韵律短语中的位置、该音节在所处韵律词中的位置、前音节的拼音码、后音节的拼音码等。

3．基元选取

基元选取的流程图如图 5 所示。基元选取模块从文本分析模块得到包含每个单元的上下文特征的目标单元序列，根据目标单元的拼音码在语音库中检索该基元的*CART* 树，根据目标单元的上下文特征对*CART* 树做迭代检索，将检索到的叶子节点中的样本数据加入到目标单元序列中。最后，输出更新后的目标单元序列给下一模块。

4．韵律修改和波形拼接

韵律修改和波形拼接的流程图如图 6 所示。韵律修改模块根据目标单元序列的上下文特征检索韵律模板库中的时长预测树和基频曲线预测树，然后根据选出的基频曲线和时长使用*TD－PSOLA* 算法对解码后的样本数据的时长和基频进行修改。由于使用*TD－PSOLA* 修改基频的幅度较大时会带来较大失真，所以设定一个修改门限，当修改量大于 1%的时候不再修改基频。*TD－PSOLA* 对时长的修改能力比较强，所以，所有的目标单元的时长都根据韵律模型的预测值进行修改。最后对修改过的样本进行拼接。

5．输出合成语音

根据实际需要，将合成语音输出到文件或者声音输出设备。

下面根据一个在智能手机上实现的移动语音合成系统为实例来说明本发明的实施方法：

第一步：准备一个大规模语料库，语料库包括 5000 多句取自人民日报的汉语语句，每个语句包括文本、拼音、韵律层级标注、16K 采样率 16bit 精度的普通话录音数据、音节切分标注、基频标注。

第二步：提取特征值，包括每个音节的，PosInPhrase，*PosInWord*，*PosInPhrase*，*PosInSentence*，Pr*eTone*，*PostTone*，*LeftPhone*，*RightPhone*，*RightPhoneType*，*LeftPhoneType*，共九个特征值，以及时长、能量、基频曲线、波形数据。

下面两步以基元“shi4”为例说明音库构建和韵律模型训练的过程。

第三步：“shi4”在语料库中共有 1166 个样本。根据每个样本的时长 D、能量 U、基频向量 *P* 所组成的特征向量，按照下面公式计算每两个样本之间的*Mahalanobis* 距离，生成 1166

×1166 的距离矩阵 M1。根据时长 D 计算每两个样本之间的欧式距离，生成一个 1166×1166 的距离矩阵 M2。根据基频向量 P，按照下面公式计算每两个样本之间的 *Mahalanobis* 距离，生成 1166×1166 的距离矩阵 M3；

$$\mathrm{Dist}(i,j)=\sqrt{(X_i-X_j)'S^{-1}(X_i-X_j)}$$

第四步：将用于音库构建的 *CART* 树的聚类比例设为 10：1，将韵律模型训练的聚类比例设为 5:1，分别根据距离矩阵 M1、M2、M3 训练三棵 *CART* 树——T1、T2、T3。T1 共包含 108 个叶子节点，T2 共包含 276 个叶子节点，T3 共包含 258 个叶子节点。

提取 T1 的每个叶子节点的中心样本，做 8k 采样率重采样，用 G.723.1 算法压缩，以 T1 为索引，将压缩后的样本数据存入音库中。

计算 T2 每个叶子节点的均值、以 T2 为索引存入韵律模板库；提取 T3 的每隔叶子节点的中心样本并作平滑处理，以 T3 位索引，存入韵律模板库。

第五步：建立音库和韵律模板库的总索引，分别指向每个基元的 *CART* 树索引。将音库和韵律模板库保存到手机的存储卡上。将合成系统的可执行程序安装到手机上。

下面各步以“我是中国人。”这句话来说明在移动语音合成系统中的合成过程：

第六步：文本分析模块首先进行文本分析，生成一个目标序列“wo3 shi4 zhong1 guo2 ren2”，其中每个目标单元包含了他的上下文信息。以“shi4”为例： $PosInWord=tail$，$PosInPhrase=tail$，$PosInSentence=body$，等

第七步：基元选取模块根据每个目标单元的上下文特征从音库中选取合时的基元样本。以“shi4”为例：首先根据音库的总索引检索到“shi4”的 *CART* 树，从树的根节点开始回答节点上的问题，如根节点的问题为 $PosInPhrase=head$，因为回答为“否”，所以取其右子节点，以此类推，直到找到一个叶子节点。用 G.723.1 算法解码，得到原始的波形数据。

第八步：采用和第七步同样的方法从韵律模板库中取得时长和基频曲线的预测值。使用 *TD-PSOLA* 算法根据预测的时长和基频曲线对波形数据作韵律修改。

第九步：将经过韵律修改的“wo3 shi4 zhong1 guo2 ren2”五个目标单元的波形数据拼接起来，输出到手机的声音输出设备中。

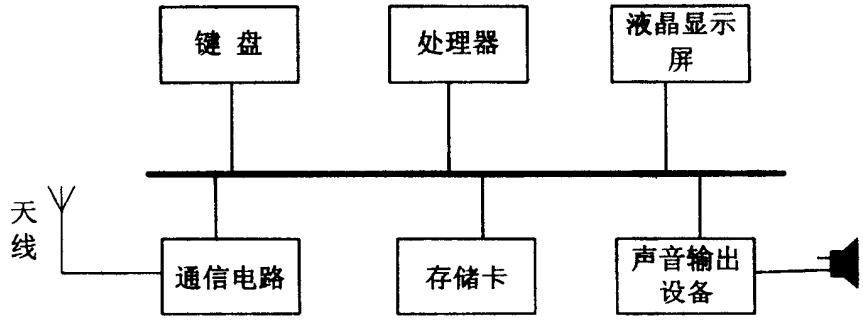

键 盘
处理器
液晶显示屏
天线
通信电路
存储卡
声音输出设备

图 1

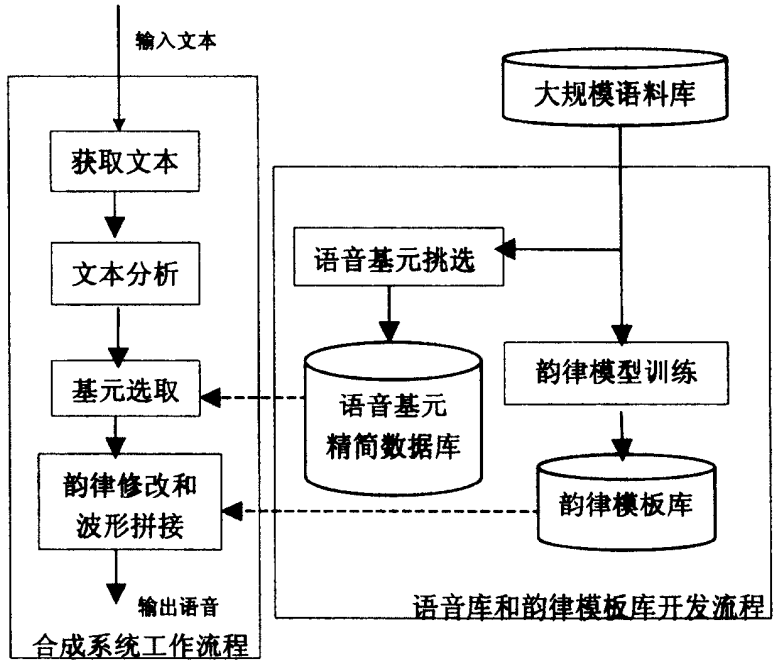

输入文本
获取文本
文本分析
基元选取
韵律修改和波形拼接
输出语音
合成系统工作流程
大规模语料库
语音基元挑选
语音基元精简数据库
韵律模型训练
韵律模板库
语音库和韵律模板库开发流程

图 2

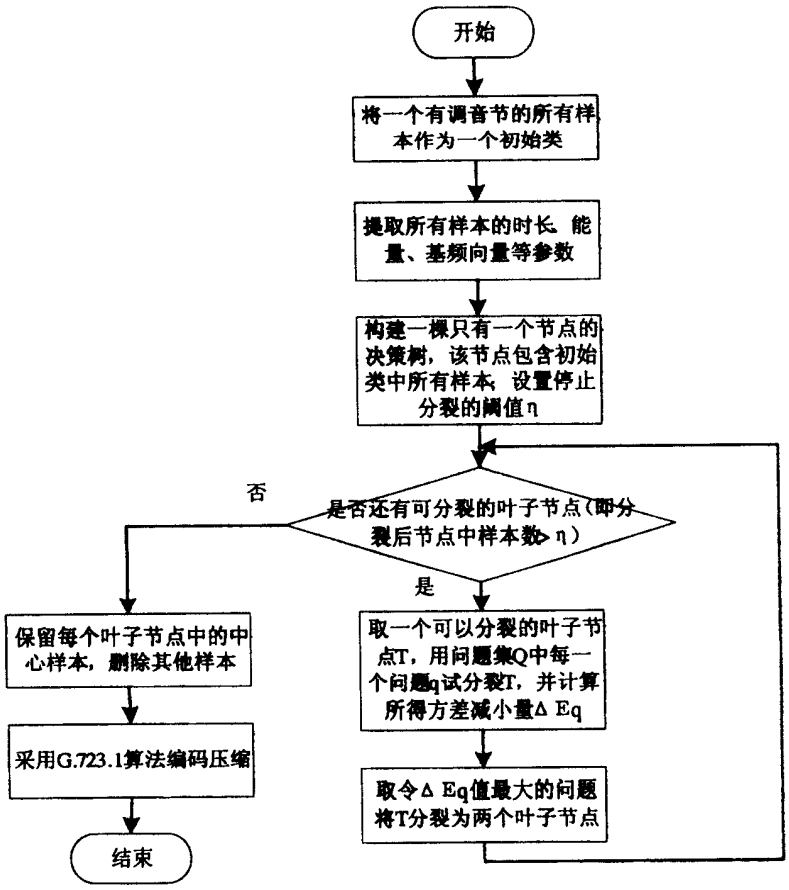

开始
将一个有调音节的所有样本作为一个初始类
提取所有样本的时长、能量、基频向量等参数
构建一棵只有一个节点的决策树，该节点包含初始类中所有样本，设置停止分裂的阈值η
是否还有可分裂的叶子节点（即分裂后节点中样本数>η）
否
是
保留每个叶子节点中的中心样本，删除其他样本
采用G.723.1算法编码压缩
结束
取一个可以分裂的叶子节点T，用问题集Q中每一个问题q试分裂T，并计算所得方差减小量Δ Eq
取令Δ Eq值最大的问题将T分裂为两个叶子节点

图 3

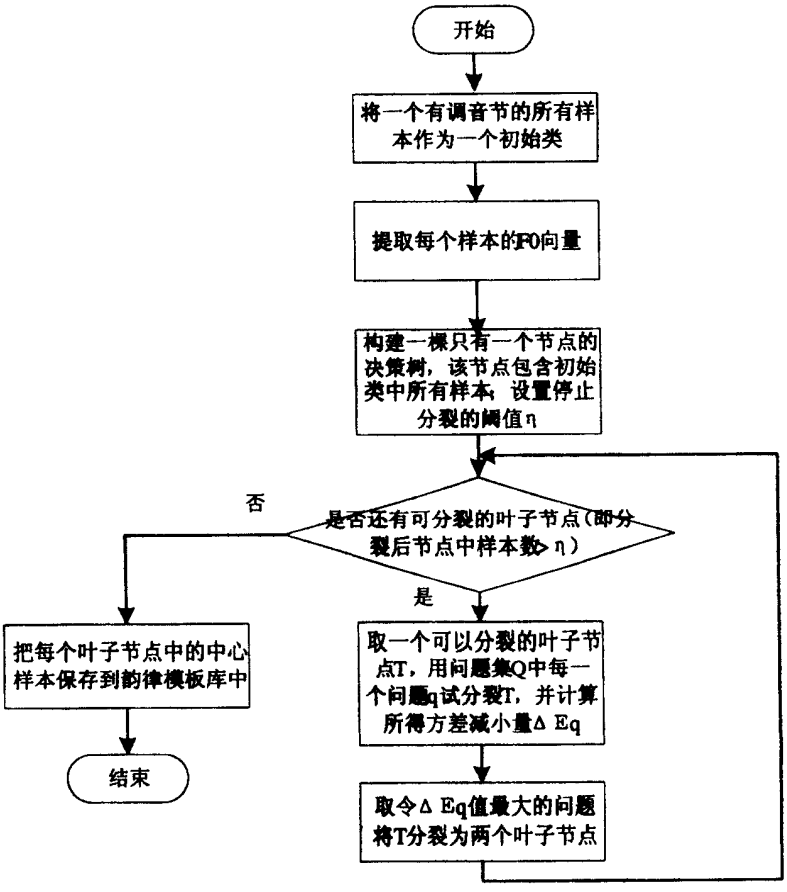

开始
将一个有调音节的所有样本作为一个初始类
提取每个样本的F0向量
构建一棵只有一个节点的决策树，该节点包含初始类中所有样本，设置停止分裂的阈值η
是否还有可分裂的叶子节点（即分裂后节点中样本数>η）
否
是
把每个叶子节点中的中心样本保存到韵律模板库中
结束
取一个可以分裂的叶子节点T，用问题集Q中每一个问题q试分裂T，并计算所得方差减小量Δ Eq
取令Δ Eq值最大的问题将T分裂为两个叶子节点

图 4

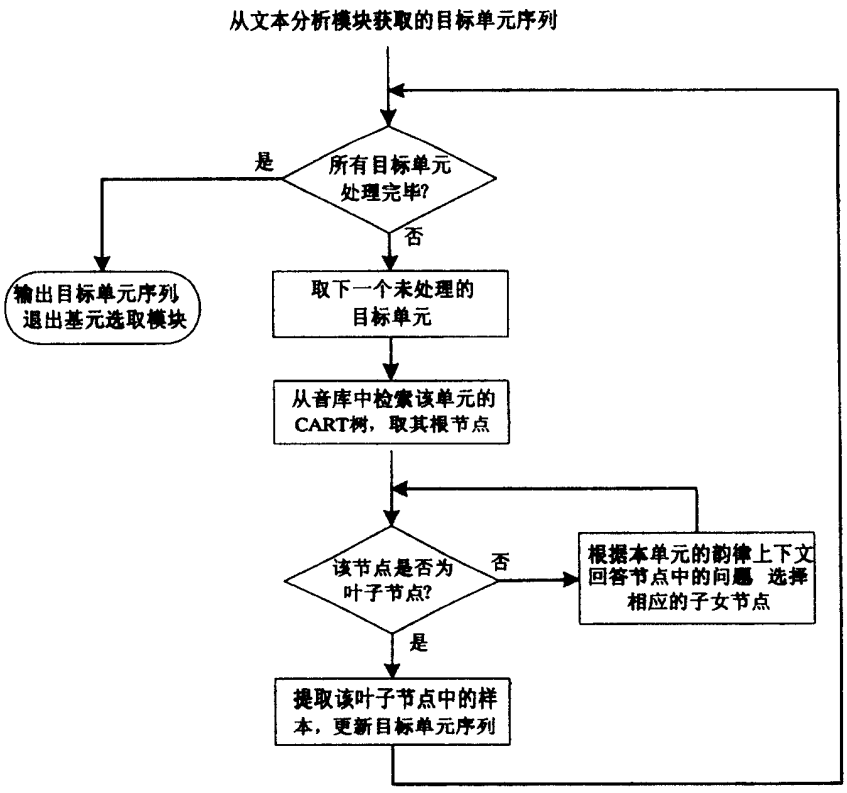

从文本分析模块获取的目标单元序列
所有目标单元处理完毕?
是
否
输出目标单元序列
退出基元选取模块
取下一个未处理的目标单元
从音库中检索该单元的CART树，取其根节点
该节点是否为叶子节点?
否
是
根据本单元的韵律上下文回答节点中的问题 选择相应的子女节点
提取该叶子节点中的样本，更新目标单元序列

图 5

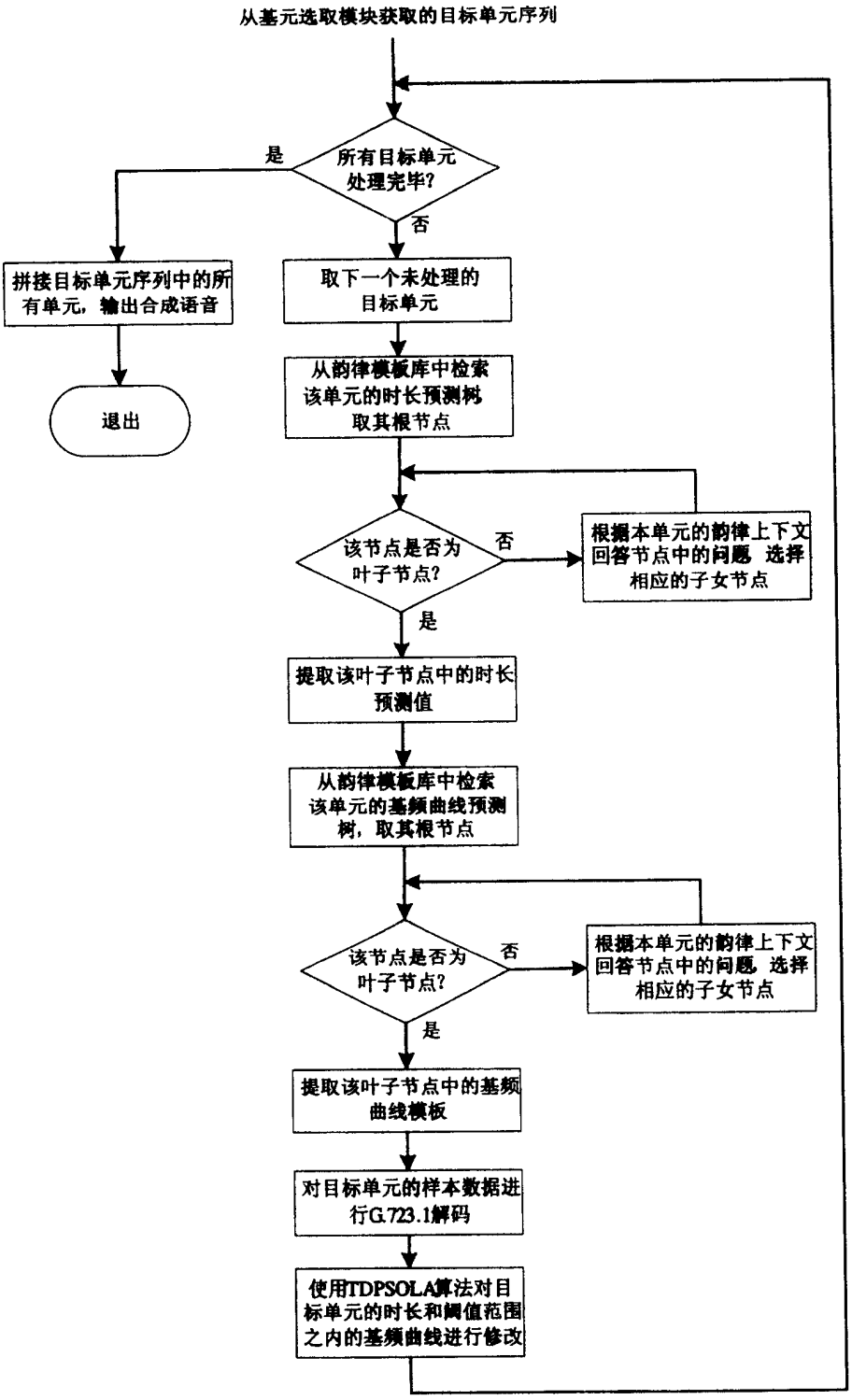

从基元选取模块获取的目标单元序列
所有目标单元处理完毕?
是
否
拼接目标单元序列中的所有单元，输出合成语音
退出
取下一个未处理的目标单元
从韵律模板库中检索该单元的时长预测树，取其根节点
该节点是否为叶子节点?
否
根据本单元的韵律上下文回答节点中的问题，选择相应的子女节点
是
提取该叶子节点中的时长预测值
从韵律模板库中检索该单元的基频曲线预测树，取其根节点
该节点是否为叶子节点?
否
根据本单元的韵律上下文回答节点中的问题，选择相应的子女节点
是
提取该叶子节点中的基频曲线模板
对目标单元的样本数据进行G.723.1解码
使用TDPSOLA算法对目标单元的时长和阈值范围之内的基频曲线进行修改

图 6