基于汉语韵律参数的语音基元选取

吴志勇 蔡莲红 陶建华

清华大学计算机科学与技术系
wuzy99@mails.tsinghua.edu.cn

摘要

本文探讨了如何基于汉语韵律参数，在大规模语音数据库中优选语音基元。本文分析了基于韵律层级结构的汉语韵律模型参数集，在汉语语音基元选取的研究中引入匹配代价和拼接代价的概念，分析了利用韵律代价函数进行基元选取的算法，针对该算法计算复杂度大的缺点，对其进行了简化和优化考虑，提出了全匹配代价模型的韵律代价函数和基于语音词的基元选取算法。实验证明，这种方法能够在较低的计算复杂度的情况下，很大的提高合成语音的质量，达到比较令人满意的效果。

1. 引言

在基于语音数据库的文语转换（TTS）系统中，语音基元选自包含有大量语句的语音数据库。由于这些语音基元来自于自然语句，隐含了本单元所在语境下的语音特征，从而体现了该情况下的音段和韵律特性，因此基于这些单元拼接出来的语流，其自然度和表现力较为满意。但是在这样的系统中，如何选择合适的单元（即基元选取）则是一个需要研究的课题。

关于基元选取方面的研究，日本 ATR 实验室的学者在日文语音合成系统 CHATR 中，提出了基于匹配代价（Target Cost）和拼接代价（Concatenation Cost）的基元选取算法[1]。

在中文语音合成方面，台湾大学的李琳山（Lin-shan Lee）教授等人提出了基于决策树和韵律修正代价的基元选取算法。他们将汉语的韵律特征参数归结为：音节序号（SID）、词中位置（LIW）、声调特征（TID）等 13 个参数，并采用惰性决策树（Lazy Decision Tree）的方法来进行基元的选取[2]。

清华大学陶建华等针对汉语的韵律特征表现的多层次性，语境相关性，分析归纳了汉语韵律的语境参数，研究并提出了一种适用于汉语 TTS 系统的神经网络韵律模型，提出了适合汉语韵律处理的神经网络拓扑结构和输出优化方法[3]。

在中文语音合成的基元选取研究中，本文中针对汉语的特点，提出了基于韵律层级结构的汉语韵律模型参数集；然后说明了如何利用韵律代价函数来进行韵律特征参数的匹配度量；针对该算法计算复杂度大的缺点，对其进行了简化考虑，并针对汉语的特点进行了算法的优化；最后提出了全匹配代价模型的韵律代价函数和基于语音词的基元选取算法。

本课题得到国家自然科学基金项目（69875008）资助

2. 语音基元选取的问题

2.1. 基元选取算法概述

假设已知一串文本序列，如何选择相应的语音基元序列，可以使用下面的简单算法加以描述。

假设：

已知的文本序列：$x_1, x_2, ..., x_j, ..., x_n$

选定的相应语音基元序列：$y_1, y_2, ..., y_j, ..., y_n$

其中 j 为音节的序号，n 为文本序列中文本单元的数目。

定义：

语音基元选取是语境相关的，文本序列中的某个文本单元 x_j 可能有多个候选的语音数据在语音数据库中，比如：$y_{j1}, y_{j2}, ..., y_{jk}, ..., y_{jm}$，其中定义 y_{jk} 为第 j 个语音基元即 x_j 的第 k 个候选者，m 为本组候选单元的个数。

定义语音库中候选基元的音段和韵律代价距离函数：$F_{y_{jk}}$，用于度量候选基元 y_{jk} 与相应的文本单元 x_j 所在的语境信息的不匹配程度。该函数值越大，说明候选基元越不符合相应的语境要求。

选取算法：

在进行基元选取时，从 Bayes 的准则出发，为每个文本单元 x_j 计算其对应的每个候选基元的音段和韵律代价距离函数，并且选择距离函数值最小者：$F_{y_{jc}} = \min\{F_{y_{j1}}, F_{y_{j2}}, ..., F_{y_{jk}}, ..., F_{y_{jm}}\}$，则 y_{jc} 为最终的候选基元，即成为和 x_j 对应的 y_j。

2.2. 语音基元选取的问题

由上述算法可以看出，进行基元选取的研究，必须要解决两个非常关键的问题：（1）找出影响语音韵律信息的韵律特征参数，即判断度量候选基元 y_{jk} 与相应的文本单元 x_j 之间不匹配程度主要由那些语境信息来决定；（2）构筑韵律代价函数，即如何使用上述韵律特征参数，反映这些语境信息之间的相互关系和各参数对整体韵律特征的影响。

3. 汉语韵律模型参数集

语音基元的选取是语境相关的，语境信息受到语音学、语言学规则的约束，在本文中用音节韵律特征参数来加以描述。

汉语具有区别于其他语言的一些特征：汉语是声调语言；汉语的语调和声调相互区别又相互关联；汉语的音节间相互影响并存在严重的同化和逆同化现象；汉语音节的发音时长、停顿和发音轻重等都具有表情达意的作用等等。因此关于音节韵律特征参数的选取，可以从

音联过渡、声调、语调、前后音特征以及时长、停顿等方面加以考虑。这些信息直接反映了汉语和汉语音节的特点，同时也综合考虑到了语境上下文的信息。

本课题组在 TTS 系统的研究过程中，提出了影响汉语韵律的 17 个语境参数[3]。这 17 个参数较好的反映了音节韵律特征的上下文信息，但对韵律参数的考虑还较少。

初敏在研究的过程中，为了反映音节在不同的韵律和音联环境中的变化，用一个六维的环境特征矢量来描述每个音节所处的环境[4][5]。

3.1. 韵律层级结构

汉语是一个韵律层级结构十分明显的语言。目前，关于汉语的韵律层级结构，比较公认的层级有：音节、音步（Foot）、韵律词、韵律短语、语调短语和句子[6]等。

在本文中，从语音合成的应用角度出发，我们将汉语的韵律层级结构规定为：音节（Syllable）、语音词（Speech Word）、韵律短语（Chunk）、语句（Sentence）四个层级。表 1 给出了语法词、语音词和韵律短语之间关系的表述。

表 1： 语法词、语音词及韵律短语信息

语 法 词	\ 一九 \ 九二 \ 年 \ 体育 \ 建筑 \ 奖 \
语 音 词	| 一九 九二 | 年 | 体育 | 建筑 奖 |
韵律短语	/ 一九 九二 年 / 体育 建筑 奖 /

3.2. 语音词

这里提出的“语音词”（Speech Word），可以理解为：汉语语音发声时紧密连在一起发音的若干音节的组合。它由“语法词”[7]经处理后得到，又不能等同于“韵律词”。

语音词既可以是语法词，比如“体育”；也可以是非语法词，比如“也就”、“将之”等；还可以是较短的短语等，如“一九九二年”。

语音词与和韵律词有很大的关联，语音词更多是从声音的角度来加以界定的。语音词的提出，能够很好的反映人们实际说话过程中普遍存在的音节（语法词）连读或节奏现象，比如“一九九二年”、“也就”等；从语音学的角度来看，语音词内部往往具有某些共同的特征，是紧密联系而难以区分的，而且音节的协同发音等语音学现象也主要出现在语音词的内部。而大多数的情况，语音词可以按照韵律词进行处理。

为了验证提出语音词的必要性，我们进行了人工标注的语音词与传统的语法分词对比的相似度实验。实验材料为 1997 年人民日报语料库中任意抽取的 10 篇短文，包含约 450 个句子、5000 多个汉字，要求本课题组内的 3 位测试人员根据平时自己朗读的感觉，进行词语边界的标注。另外，基于传统的词典分词的办法，进行了程序的自动分词。最后统计的实验结果表明，3 个人工标注的结果之间有很大的相似性，相同的标注达到了 91.1%。而传统的语法分词结果和人工标注之间相似度则较低，只有 68.6%的结果是一致的。

3.3. 汉语韵律模型

人们通常将表征一个音节特征的特征集参数描述为一个向量，并且将之称为韵律模型。韵律模型的构筑，是进行基元选取的基础。在研究中，我们认为影响汉语韵律特征的参数可以从语言学、语音学以及韵律参数等方面加以考虑。

在本文中，将汉语的韵律模型描述为：

$$P = \{\vec{C}, \vec{P}, \vec{G}, \vec{A}\}$$

其中 $\vec{C}$ 表示音段参数集、$\vec{P}$ 表示位置参数集、$\vec{G}$ 表示关联参数集、$\vec{A}$ 表示韵律参数集，他们分别对应着不同层次韵律模型的参数向量。

3.4. 韵律模型参数集

在本文中，关于韵律模型的参数集，基于本课题组的研究成果[3]，增加了一些参数，引入了语音词的概念，使之更符合汉语的韵律层级结构模型。而且实验表明，这种扩展对于语音韵律特征的描述和模拟等都有较好的效果。

将文本上下文信息按照“音节—语音词—韵律短语—语句”的层级结构模型进行分析，可以得到一系列的韵律参数。这些参数根据其功能性，可以归结为如下参数向量：

* 音段参数 $\vec{C}$（当前音节的声母类型 c_1、韵母类型 c_2、声调类型 c_3，前音节的韵母类型 c_4、声调类型 c_5，后音节的声母类型 c_6、声调类型 c_7）；
* 位置参数 $\vec{P}$（当前音节在语音词中位置 p_1、在韵律短语中的位置 p_2、在语句中的位置 p_3、到前一个重音的位置 p_4、到后一个重音的位置 p_5、语音词在韵律短语中的位置 p_6、语音词在句中位置 p_7、韵律短语在句中位置 p_8）；
* 关联参数 $\vec{G}$（与前音节的耦合度 g_1、与后音节的耦合度 g_2、所在语音词的信息 g_3、所在韵律短语的信息 g_4、所在语句的信息 g_5）；
* 韵律参数 $\vec{S}$（时长 d、基频特征 p、幅度 a、重音属性 w、语调信息 t）。

稍微解释一下上面部分参数的意义：关联参数中当前音节和前后音节的耦合度即当前音节和前后音节的相关联程度，可以根据和语音词、韵律短语等的关系确定为三个等级（语音词内部、韵律短语内部、句子内部）；语句信息、短语信息和语音词信息反映了整个句子从各个层次描述的节奏紧凑程度；音段参数、位置参数和关联参数可以直接根据文本上下文分析得到，而韵律参数，在语音数据库中可以直接根据语音波形的数据计算得到，而在合成时，可以通过韵律预测模块根据文本分析的结果预测而得到。

4. 全匹配代价模型的韵律代价函数

对于语音库中的候选单元，可以通过程序计算和人工标注的方法得到本音节和语境上下文的标注信息。对于待合成的音节，也可以通过文本分析和韵律建模模块的分析得到相应的目标韵律参数。候选单元和待合成音节之间通过韵律代价函数进行韵律特征参数匹配的评估，选择代价最小的语音单元作为最终的合成用基元。

考虑到计算复杂度的问题，本文对韵律代价函数进行了修正和简化考虑，提出了全匹配代价模型的韵律代价函数。

4.1. 韵律代价函数

语音基元的选取，不仅仅需要进行候选单元和待合成音节韵律参数的匹配，即匹配代价（Target Cost），还需要考虑到待合成句子的整体效果，即相邻的语音候选单元之间在拼接时它们之间在拼接处的连接情况，即拼接代价（Concatenation Cost）。

对于具有 n 个待合成单元的输入序列，韵律代价函数可以表示为[1]：

$$C(t^n,u^n)=\sum_{i=1}^{n}C^t(t_i,u_i)+\sum_{i=2}^{n}C^c(u_{i-1},u_i)$$

其中匹配代价为：

$$C^t(t_i,u_i)=\sum_{j=1}^{p}w_j^tC_j^t(t_i,u_i)$$

拼接代价表示为：

$$C^c(u_{i-1},u_i)=\sum_{j=1}^{q}w_j^cC_j^c(u_{i-1},u_i)$$

其中 p 表示所有考察的韵律特征参数的个数，而每个代价的权重 w_j^t 为相应韵律特征参数的权重。q 表示所考虑的拼接处的特征参数的个数，w_j^c 表示相应特征参数的权重。

4.2. 基元选取

在定义了韵律代价函数以后，语音基元的选取就是从语音库中选择候选单元，使得韵律代价函数的总体代价达到最小。

这个过程，可以通过 Viterbi 最佳路径搜索的算法进行，关于基元选取的示意图，如图 1 所示。

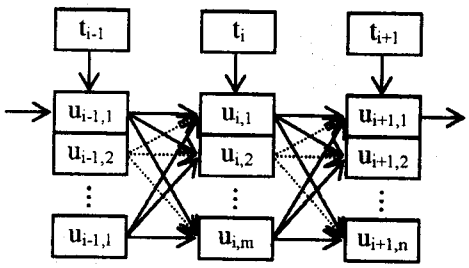

图 1 基元选取示意

4.3. 计算复杂度

在进行韵律代价函数的计算时，拼接代价的计算是很一个计算量很大的过程。

如果一个待合成的句子包括 n 个音节，而每个音节 p_i 在语音库中有 m_i 个语音基元可供选择。对于音节 p_i 的每个候选语音单元，都需要使用 Viterbi 算法来搜索其前一个音节的所有候选音节单元并且计算出拼接代价，则这个过程将会需要进行 $m_{i-1}\times m_i$ 次的计算。这样对于一个具有 n 个音节的句子来说，假设每个音节的平均候选基元的数目为 $\overline{m}$，则整个计算复杂度将是 $n\times\overline{m}^2$。因此，这样的计算复杂度差不多是该句子所有音节候选基元数目的平方级的，即随着句子包含的音节数目的增加，其计算复杂度也随着按照平方级而迅速扩大。

对于实时的合成系统而言，特别是速度要求比较高的场合，这种情况是不容允许的。因此需要对算法进行一定的简化和优化。

4.4. 全匹配代价模型

代价函数的计算复杂度主要在于拼接代价计算的 Viterbi 搜索过程，因此为了降低计算复杂度，提高系统的效率和速度，可以从简化拼接代价的计算入手，甚至考虑不使用拼接代价函数，而用其他方法来加以替代。

4.4.1. 简化考虑 1

在本文的研究过程中，对于音库的设计和录音工作进行了一些特殊的规定和处理，从而使得语音数据库中不同音节候选单元之间的相互影响降低到了最低的限度。这样在进行语音基元的选取时，相邻音节之间进行波形拼接时的拼接代价也降低到了最小。

只要上述处理得当，就可以在计算韵律代价函数的时候，忽略拼接代价的影响，而不进行拼接代价的计算，而且这样处理对整个研究的最终合成效果也不会造成很大的影响。

4.4.2. 简化考虑 2

拼接代价的引入，是为了度量相邻语音候选基元之间的拼接不平滑度，即平滑损失的，这种损失可以通过拼接处的倒谱、基频以及幅度能量等的过渡特性来加以表征。

在现有语音数据库中，这种平滑损失主要可以考虑两个候选基元在语音库中是否相邻的事实。如果属于相邻基元，则其平滑损失必然达到最小，甚至为 0。

为了描述候选基元在语音库中是否相邻的事实，在韵律模型参数集的音段参数中引入两个新的参数：当前音节的前后音节的语音库代码表示，前音节语音库代码 m_1、后音节语音库代码 m_2。

通过上述这种考虑，可以把拼接代价的平滑损失转换为韵律特征参数来进行计算。

4.4.3. 算法的简化：全匹配代价模型

通过上述简化考虑，韵律代价函数其实就是匹配代价的了，其中 p 为扩展了的韵律模型参数的个数。则韵律代价函数转化为全匹配代价的模型：

$$C(t^n,u^n)=\sum_{i=1}^{n}C^t(t_i,u_i)=\sum_{i=1}^{n}\sum_{j=1}^{p}w_j^tC_j^t(t_i,u_i)$$

基元选取的目标，如上所述，应该是的韵律代价函数的代价值达到最小。而上述公式中，待合成语句的总体代价是语句中各个音节自己匹配代价的代数和，因此要达到韵律代价函数最小的目标，只要使得每个音节的匹配代价达到最小，即只要使得每个音节的候选基元都达到最优的即可。这样大大简化了基元选取的模型，提高了整个合成系统的效率，而且对于系统的整体合成效果不会产生较大的影响。

5. 基于语音词的层次基元选取

考虑到汉语语音生成的层次模型，即语音的韵律层级结构，在研究的过程中，提出了和“音节—语音词—韵律短语—语句”的层级模型对应的层次基元选取策略。将每个因素的考虑和计算局限在相应的层级内部进行，使得基元选取的算法得到了进一步的优化。

研究发现，在连续语流中，语音词内部相邻音节的协同发音较为严重，而词间次之。另外，人们在听辨时，要求语音词内部相对比较紧凑和连续，语音的自然度高、轻重对比明显；而语音词之间，则因其间有一定的停顿，连续性要求相对比较低一些。

因此，层次基元选取算法中，语音词一级的选取是整个基元选取的基础，也是基元选取的难点所在。基于此，结合韵律代价函数的考虑，我们提出了基于语音词的基元选取算法。

基于语音词的基元选取算法，主要由两步完成：一是语音词内部的基元的预选，二是语音词的精选。首先根据韵律代价函数以及语音词相关的韵律特征参数，考虑语音词内部的基元选取，在确定了该语音词的所有候选基元后，再根据韵律短语、句子等上下文信息，考虑各个语音词候选之间的匹配性，进行语音词候选单元的进一步精选，最终确定各个语音词的语音基元，从而确定整个合成语句的基元。

6. 基元选取的权重模型

基元选取设计和研究的一个重要方面，就是如何评估韵律模型特征参数中的任何一个对于基元选取的作用，在数学特征上表现为韵律代价函数中各个参数权重设置的问题。

韵律参数权重的设定，可以采用回归训练、神经网络等基于数据驱动的机器学习的自动训练算法。

限于算法复杂度的考虑，以及实验数据的局限性，在本文中，对于基元选取的权重设定，进行了一定的简化，采用了一种人工指导下的机器学习的训练算法。

本算法主要包括两个步骤：（1）初始权重的设定，基于语言学和语音学规则的总结，人为设定不同参数之间的权重属性；（2）人工指导下的机器学习，采用了人工评测加机器自动学习的算法，通过人工听辨的方法，对现有的基元选取结果进行人为的评测打分，然后机器根据人工评测的结果进行训练，自动进行权重的调整修正。

在实验中，我们一共收集了 5000 句的自然语音数据：其中 3000 句长句（平均音节数 20 个）用于语音数据库的抽取，剩下的 2000 句短句（平均音节数 7 个）用来进行权重的训练。实验时，用现有权重的合成系统分别合成 2000 句的文本，由人工根据自然录制的相应语句进行基元选取好坏的评估，最后用评估的结果去进行训练，产生新的权重。

事实证明，这种方法有效地解决了权重的设定问题，很好的实现了一个能够比较客观正确的反映各个韵律特征参数作用的权重模型，对于基元选取的研究起了很大的促进作用。

7. 总结与讨论

本文基于汉语韵律层级结构模型，提出了表征韵律信息的汉语韵律模型参数集，扩展了现有的研究成果。基于此，文中还进行了韵律代价函数的分析，并从降低其计算复杂度的考虑入手，进行了韵律代价函数的简化考虑和优化，提出了一种基于全匹配代价模型的韵律代价函数，和基于韵律层级结构和语音词的层次基元选取策略。

实验表明，这种算法的简化考虑和优化，有效地降低了计算复杂度，并且合成效果也能够同时达到令人满意的效果。

上述研究，还是一个初步的结果，其中还有许多问题需要进行进一步的工作，比如说：文中提出了表征韵律特征的汉语韵律模型参数集，这只能说是一个充分的条件，其中的必要参量有哪些？也就是说能不能找到最少的决定韵律特征的韵律参数？另外，人类的感知是一个奇妙的过程，基元的选取能不能和人类感知结合起来，从感知的角度来研究韵律特征参数，这应该是以后进行进一步研究的指导思想。还有，关于权重模型的研究，还仅仅是一个初步的结果，还有很多工作需要进行。

8. 参考文献

[1] Andrew J. Hunt, Alan W. Black. Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech DataBase. ATR Interpreting Telecommunications Research Labs

[2] Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee. Selection of Waveform Units for Corpus-based Mandarin Speech Synthesis Based on Decision Trees and Prosodic Modification Cost

[3] 陶建华，蔡莲红，赵世霞，吴志勇. 汉语文语转换系统中可训练韵律模型的研究. 声学学报，V26 N1 2001. 67-72

[4] Min Chu and Hu Peng. An Objective Measure for Estimating MOS of Synthesized Speech. Eurospeech 2001

[5] 初敏. 韵律研究与合成语音的自然度. 第五届全国现代语音学学术会议. 新世纪的现代语音学. 295-301

[6] 冯胜利. 汉语的韵律、词法与句法. 北京：北京大学出版社，1997

[7] 胡明扬. 说“词语”. 语言文字应用，1999，3

[8] 吴志勇，蔡莲红. 汉语 TTS 中基于语音词的基元选取. 全国人机语音通讯技术会议 2000 论文集. 62-66

[9] Zhiyong Wu, Lianhong Cai, Tongchun Zhou. Research on Dynamic Characters of Chinese Pitch Contours. ICSLP2000. VIII 686-689

[10] 吴宗济等. 实验语音学. 高等教育出版社

[11] 吴宗济. 中国音韵学和语言学在现代言语处理中的应用