基于上下文的汉语韵律模型建模方法的研究

陶建华　　蔡莲红　　钟玉琢
清华大学计算机系信息组　100084
email: tjh@tts.cs.tsinghua.edu.cn

摘要

由于汉语的韵律特征与上下文的语境密切相关。本文针对汉语语言的这些特点，研究并提出了汉语声调的规格化模型。同时分析了词调的特点，用模糊相似矩阵的方法对词调进行了分类统计。另外、根据汉语语调与语言环境密切相关的特性，运用神经网络模型建立了语法以及发音环境与韵律特征之间的联系。并在此基础上建立了一种适用于汉语语音合成的语调模型。

~~基于上下文~~语境相关的汉语韵律模型建模方法的研究

陶建华　　蔡莲红　　钟玉琢
清华大学计算机系信息组　100084
email: tjh@tts.cs.tsinghua.edu.cn

摘要

~~由于~~汉语的韵律特征~~与上下文的~~表现为多层次的语境密切相关。~~本文~~作者针对汉语语~~言~~音的~~这些~~字调特点，研究并提出了汉语声调的规格化模型。~~同时~~作者还分析了汉语词调的特点，用模糊相似矩阵的方法对词调特征进行了分类统计。~~另外、~~同时根据汉语语调与语言环境密切相关的特性，运用神经网络模型建立了~~语~~法~~、以及发音环~~语境与韵律特征之间的~~联系~~映射关系。并在此基础上建立了一种~~适~~用于汉语语音合成的语调模型。

关键词：韵律模型、语音合成、神经网络、语境相关（context-dependent）

一、前言

众所周知，汉语是一种~~有~~声调语言，其单音节大致可分为阴平、阳平、上声、去声、轻声等五种声调，每一种调型都有自己相对稳定的音高变化模式。语调的形成好比就是大波浪上叠加小波浪（Zhao Yuan-Ren, 1932,1933,1968），大波浪就是声调变换的走势，而小波浪就是指字调或词调。而语调模式又可以看作由单字词、二字词、三字词和四字词等声调的模式加上语气的变化而形成（Wu Zong-Ji, 1982,1988,1992,1996）。任何韵律模型都必须先解决声调的数学模型问题。本文的第二单元将首先介绍我们针对汉语的韵律特点而设计的声调规格化模型(2)，由声调模型可以将语调分解为：词调、语气、基频偏移三个部分。从而语调模型也将建立在对这三个部分基础上进行研究和总结。

在本文中，作者采用了一种模糊相似矩阵聚类的方法，对大量的词汇发音进行自动分类统计，从统计结果可以看出人在词汇发音时的一些规律性，并直接将其结果运用到语调模型的建立中。

从现代音系学的研究表明，汉语的韵律特征与上下文，即与音节位置、重音类型、句法环境、音节音系结构以及音段的语音性质密切相关（Lin Mao-Can, 1998）。而汉语语言又是变化多端的，从这些语境环境中直接分析韵律特征将会非常困难。本文的重点就是试图在不同的语境环境中与韵律参数之间建立一种自动分析和聚类的方法。这里我们对大量的语句先进行语法上和韵律上的标注。通过一种带时延的神经网络模型，对已经分析出的标注信息进

行训练，从而在此基础上建立上下文语法与韵律上的联系，为提高语调模型的质量、和增加模型的适应性起着非常重要的作用。

由于汉语是韵律更强调音节间的连接特性、语气的走势特性。人对音节间韵律的断续、高低反映比较敏锐。本文采用一种统计加神经网络混合模型的方法，网络的训练单元为“词调”，对于合成系统，首先根据音节所在词的属性选取合适的词调模型，再在此词调的基础上，进一步运用神经网络模型的训练和生成结果，叠加而产生语调。

最后，我们在上面的基础上，建立了一个适于汉语语音合成的句调韵律模型，该合成系统采用 PSOLA 算法。

二、声调模型

$$F_n(t) = 10EXP(V_n + D_n(t) \times S_n(L \times t)) \quad (1)$$

公式 $S_n(t)$ 代表字或词单独发音时的字调或词调模型，它是基频曲线的对数值，而 $D_n(t)$ 不是一个固定值，而是随 $S_n(t)$ 中时间变量 t 同步变换的变换量，$D_n(t)$ 反映了说话人的语气变化，其中 $\max[D_n(t) \times S_n(t)]$、$\min[D_n(t) \times S_n(t)]$ 指出了该词调的调域范围，V_n 则是调域的中值曲线。由公式 1.1 可以看出一个完整的声调模型将由说话人的词调曲线、语气变化曲线及调域中值曲线三个部分组成。对于语调模型的建立也分解为分别求 $D_n(t)$、$S_n(t)$及 V_n 这三个参量的数学模型。根据人发音时的频变化特性，所有涉及基频的参数均采用对数坐标值。

三、语句的标注处理

由于汉语的韵律特征与语言环境密切相关。找出人在不同语言环境中的发音特征对语音合成系统具有非常重要的作用。在所有语言环境中，语句的语法属性是与韵律特征相联系的最直接且最容易分析的参数。

为了寻找这种联系，本文选用 265 个长句子做分析，分别对其进行标注分析和韵律参数提取。这些句子基本概括了汉语语言的不同句型。标注处理又分为两个部分：语法标注和韵律标注。

1、语法标注

语法标注的目的在于划分语句中的韵律及语法成份，并为每一个成份标上一定的语法信息，以及适当的韵律特征，为寻求韵律参数与汉语的语法特征的相互联系提供帮助。便于进行韵律分析时的聚类分析处理。

把给出的不同句型的文档，沿着篇章（SECTION）---语句（SENTENCE）---从句（CLAUSE）---词组（PHRASE）---词（VERBA）---音节（SYLLABLE）的思路划分开。其中词组和词的韵律特征对语句的整体自然度起着最直接的作用。对每一项又按照表 1 所示进行细分。

	篇章(T)	语句(S)	从句(C)	词组(P)	词(W)	音节(Y)

标记内容	1、篇章类型	1 语句类型 2、说话风格 3、语气 4、速度 5、音高 6、语句重音	1、语法成分 2、重音	1、词组词性 2、结构 3、音节数 4、重音 5、强调词位置	1、词性 2、音节数 3、格局 4、在语气中所起的作用 5、强调方式 6、强调音节位置 7、重音类型	1、强调方式 2、拼音码

每一个音节分别对应着一个语法标注矢量：

$$\vec{X}=(T_1,S_1,S_2,S_3,S_4,S_5,S_6,C_1,C_2,P_1,P_2,P_3,P_4,P_5,W_1,W_2,W_3,W_4,W_5,W_6,W_7,Y_1,Y_2)$$

共 23 个分量，每个分量的值分别代表着一个属性的状态。

3、韵律标注

韵律标注是为了反映文字与韵律曲线之间的最直接的关系，用户可以通过韵律标注符号，直接控制韵律的生成，产生满足用户特殊需求的语调特性。目前国际上已有一种较流行的韵律标注符号 TOBI 标注系统。在需要发音的文本中，嵌入 TOBI 符号，可以达到修饰韵律的目的。（由于汉语是一种有调语言，其单音节全部只有 1200 多个，且其词调、语调、音长、停顿等的特性较其它语种更具特色，TOBI 标注符并不能完全反映汉语的韵律特征，必须在参考别的标注符号的基础上，针对汉语的韵律特点，设计一种符合汉语韵律特点的韵律标注符。）

韵律标注矢量表示为：$\vec{A}$ 。

四、基于语法规则的韵律聚类分析

通过自行设计的一种能对语音波形进行自动韵律标注的工具。通过该工具对语句波形进行韵律参数的自动提取，不仅可以得出语句的韵律曲线，还可以得到不同音节、词的韵律曲线、重音及疑问语气等韵律参数。

1、词调曲线的聚类分析：

词调分析是进行语调变换的基础，而它的分析重点又在音联的分析上，即分析字在词中发生的变调规律以及字与字之间基频的过渡上。词调的分析又以二字词的分析为基础，三字词及四字词除有它的特殊性外，其分析思路可以从二字词上扩展得来，具体的分析方法将在下面介绍。

在已经划分出的每一个音节的基音曲线上均匀的选取 10 个点，t_0,t_1 ,～t_9。这 10 个点位置的基频值 $f(t_i)$ 基本代表了该音节在特定环境中的基频变换情况。

二字词分析：将所有待分析的二字词的前后音节按上述方法，各取 10 个点，得到前音节 $fr_j(t_i)$与后音节 $fn_j(t_i)$，其中 t_i 表示第 i 个点的值，fr_j 表示第 j 个词的前音节基频值，fn_j 则表示第 j 个词的后音节。对于同一种调型搭配的情况下，其基频平均值为：

$$\overline{f_j}=\frac{1}{20}\left[\sum_{i=0}^{9}\left(fr_j(t_i)+fn_j(t_i)\right)\right] \qquad (2)$$

令：　$fr_{ji}' = fr_j(t_i) - \overline{f_j}$　　(3)

则任意两个基频曲线之间的相关系数可用下式得到：

$$R_{jk} = \frac{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ki}' + fn_{ji}' \times fn_{ki}'\right]}{\sum_{i=0}^{9}\left[fr_{ji}' \times fr_{ji}' + fn_{ji}' \times fn_{ji}'\right] + \sum_{i=0}^{9}\left[fr_{ki}' \times fr_{ki}' + fn_{ki}' \times fn_{ki}'\right]} \quad (4)$$

由此可以得到一个 $M \times M$ 相似矩阵：

$$A = \begin{bmatrix} R_{00} & R_{01} & \cdots & R_{0M} \\ R_{10} & R_{11} & \cdots & R_{1M} \\ \cdots & \cdots & \cdots & \cdots \\ R_{M0} & R_{M1} & \cdots & R_{MM} \end{bmatrix} \quad (5)$$

其中 M 为分析二字词调的个数。

运用模糊聚类中的最大树法，可以对矩阵进行非常准确的分类。

通过聚类，相同调型搭配的情况下的词调模型可以分为几种不同的基频模型来表示 $S_{ij}(t)$，每一种曲线分别为该类的所有基频点取平均并经平滑得到，其中 i 表示不同的调型组合，j 表示同一种组合中的基频模型。由于 B 样条曲线具有良好的曲线平滑性能，以及端口二级平滑功能，非常适合声调曲线的平滑，因此平滑算法采用 B 样条曲线平滑算法。

由于在词一级分析词调问题，其核心就是分析音节之间的音联问题，因此对其他三字词、四字词的聚类方法采用与二字词相同的方法。

2、实际词调与词调统计曲线的发散性研究（待做）

3、韵律神经网络模型：

目前已有的语音合成系统中采用的神经网络模型，多是通过模型选取合适的合成基元。模型的工作方式一般是通过训练大量的合成样本，建立规模庞大的语音合成音库，语音合成是通过环境参数在神经网络模型中选取合适的合成基元，从而达到提高语音合成自然度的目的。一般训练的最小单位为音节(Kerao)，甚至小于一个音节而只有几十毫秒（Motorola），这样做往往也能产生较好的效果，但基于这种模型的合成系统一般伴随着大规模的合成音库，不便存储和推广，最关键的是这种模型很不方便根据韵律标注进行语句的韵律修改。而直接在韵律上建立模型，则能较好的解决这些问题，韵律模型不需要大的音库，所有的合成语句均是通过基本合成音节通过韵律修改而得到，且能产生高的自然度，另外，韵律模型的建立非常方便使用者根据韵律标注符号修改韵律特征，产生符合特殊韵律特征的连续语句。

由于汉语是韵律特征的特殊性，相较其它语种它更强调音节间的连接特性、语气的走势特性。人对音节间韵律的断续、高低反映比较敏锐，若以音节或比音节更小的单元作训练单元，很可能由于训练样本的限制以及网络模型的运算误差，导致计算出的音节韵律曲线之间的过渡不平滑或不合音联的要求，从而导致语音自然度下降。

这里采用一种统计加神经网络混合模型的方法可以较好的解决这一问题。网络模型如图（4.1）所示。网络的训练单元为“词调”，由于前面已经用聚类方法得出汉语的词调模型，对于合成系统，首先根据分词模型，以及音节所在词的属性选取合适的词调模型，再在此词调的基础上，进一步运用语气模型的训练和生成结果，叠加而产生语调。直接选取聚类出的词调模型的好处是，不会因网络的训练量基运算误差而将低音联特性的质量。具体的设计思

路如下：

设词在句子中的基频曲线为：$f_i(t)$，而词单独发音时的基频曲线由上面的聚类算法得出：$S_{ij}(t)$。由此可以得到语气变换曲线为：

$$D_i(t)=\frac{f_i(t)}{S_i(t)} \tag{6}$$

模型实际上就是求取人的发音语气，即语气变换矢量$\vec{D}$与语法标注矢量$\vec{X}$、韵律标注矢量$\vec{A}$之间的关系：

$$\vec{D}=F(\vec{X},\vec{A}) \tag{7}$$

考虑到韵律特征不仅跟当前音节、词的属性有关，还跟前后音节的特征有关，因此（7）式应该写为：

$$\vec{D}=F(\vec{X},\vec{A},\overrightarrow{XR},\overrightarrow{AR},\overrightarrow{XN},\overrightarrow{AN}) \tag{8}$$

其中$\overrightarrow{XR}$、$\overrightarrow{AR}$为前音节标注矢量，而$\overrightarrow{XN}$、$\overrightarrow{AR}$为后音节标注矢量。

根据上面的思路，输出矢量可以表示为：

$$\vec{D}=(D_1,D_2,D_3,D_4,D_5,D_6,D_7,D_8,D_9,D_{10},L) \tag{9}$$

其中 D_1~D_{10} 反映了语气变换特性，L 则表示音的长度。

在合成系统中，由 D_1~D_{10} 得出的点，还应根据前后语气变化用 B 样条曲线作平化处理。

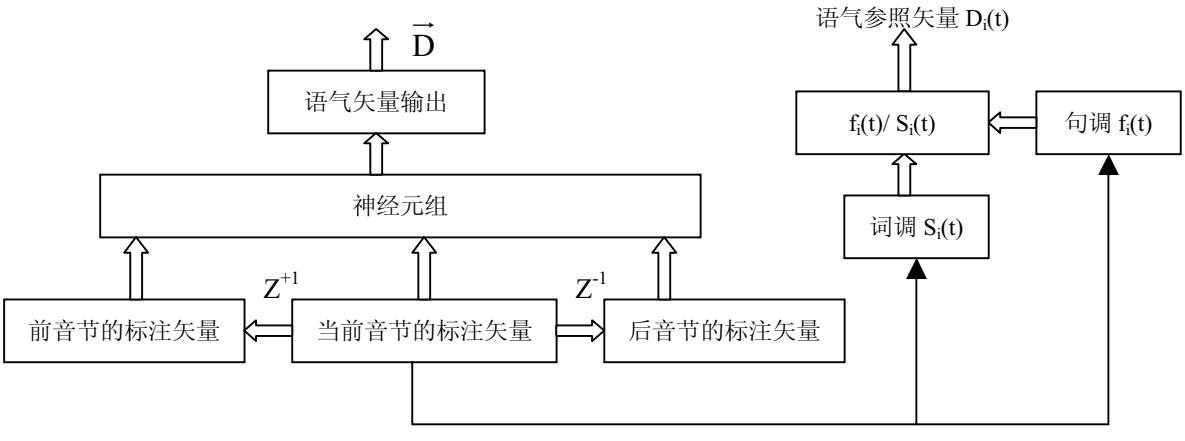

图 4.1、韵律神经网络模型

五、建立韵律模型

通过上面的分析，对于合成系统来说，韵律模型的建立分为词调模型、语气神经网络模

型、声调合成模型、时长调整等几个部分。其输入为经过文本分析处理的带语境信息（包括语法标注矢量、韵律标注矢量）的文本。

韵律模型首先根据分析后的文本中的音节所在的词，由词调模型选出合适的词调曲线，并由语境矢量输入神经网络模型得出该词的语气特性曲线和时长参数。通过公式（1）将这几部分的参数合成成语调曲线并输出。

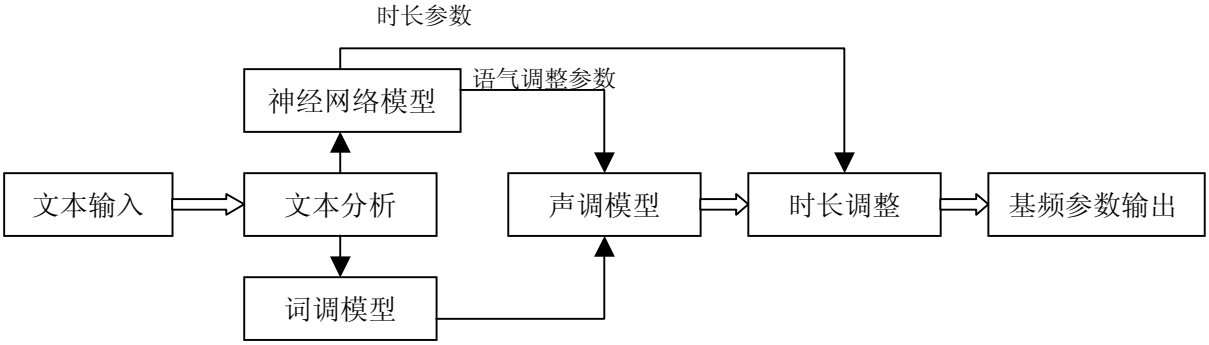

图 5.1、用于语音合成时的韵律模型

六、结果

参考文献：

1、