基于统计模型的韵律建模方法 *

陶建华 蔡莲红 吴志勇

清华大学计算机系人机交互与媒体集成研究所 100084

jhtao@tsinghua.edu.cn

摘要

本文论述了采用概率模型进行韵律建模的思路。同时，本文还进一步分析了韵律代价函数及其在基元选取中的体现或采用其它可训练韵律模型的方式，同时分析指出这些方法可以用概率模型得到很好的解释。在此基础上，论文还分析了韵律特征间相互作用对音节基元选取的影响。同时还对与基元选取息息相关的韵律节奏的预测，进行了一定的阐述。最后，本文分析了概率模型的进行韵律预测的误差分布情况，指出这种方法对于韵律预测可计算性研究的重要性。

1. 引言

语音合成技术自六十年代Klatt语音合成器诞生以来，进入了一个语音数字计算的时代，伴随着计算机技术的飞速发展和大量普及，语音的整体研究水平获得了不断的进步。尤其是进入20世纪九十年代以来，在人工智能、自然语言理解、信号处理、随机过程、模式识别等领域获得了很大的进展，这些技术在语音处理中也得到了非常成功的应用，导致了语音技术在多项关键性技术上的突破。同时，语音技术的研究，也变成了一个系统工程，所涉及的知识也远远超出了语音信号研究的本身，而越来越广。

在这种背景下，韵律建模的研究，也同样融入了许多新的概念。本文论述了采用概率模型进行韵律建模的思路（第二节）。目前，在许多基于数据驱动的语音合成系统中，选音算法变得越来越重要。本文对选音算法也进行一些分析（第三、四节），并分析了韵律代价函数的实现及权重确定和训练算法，同时，本文认为选音算法归根结底，是韵律建模的一种具体实现。论文还分析了韵律特征间相互作用对音节基元选取的影响，并通过实验进行了比较。本文在第三节中分析得到，在影响选音的因子中，音节位置信息和韵律短语信息占据了非常重要的影响因素。因而，本文还就韵律短语的自动预测机制，进行了一定的阐述（第五节）。最后，本文分析了基于统计的韵律模型的误差分布情况（第六节），指出这种方法对于韵律预测可计算性研究的重要性。

* 本课题受国家自然科学基金（69875008）资助

2. 韵律建模的概率描述

人在不同的语境下，会有不同的韵律特征，语境与韵律特征之间具有很强的相关性。谈到语音和句法以及语义之间有密切关系的时候，林焘先生强调“绝对不能把语言的这三方面割裂开来孤立地进行研究。”语境与韵律特征之间具有很强的相关性。“语调构造由语势重音配合而形成。它是一种语音形式，它通过信息聚焦来实施超语法的功能语义。”[5] Katherine Morton[2]在他的对话系统的语音合成模块中，根据几个基本的上下文模式，加入情感的变化，使合成出的语音变得有些生动。

韵律特征参数分布受着语境信息的影响，这种影响又满足一定的概率关联关系，而不是一个简单的函数映射。正如，在汉语轻声的研究工作中，沈炯先生在“从轻音现象看语音与语法研究的关系”[7]一文中阐述：“当我们说一种语音现象很明显的时候，主要是指它的离散成分很容易把握，一般并不指它的音理是否容易认知。离散成分是平常人都能把握的形式——在言语使用过程和语言教学中分别把握它。只要它能被把握，语音研究就可以在离散基础上进行，可以用文字或其他表音手段来列举语言样品，讨论问题就方便了。”

从概率的角度，对于一个已知的语句参数，与之相对应的韵律特征参数，为所有韵律特征参数中出现概率最大的一组，即为：

$$\hat{\vec{Y}} = \arg\max_{n} P(\vec{Y}_n \mid \vec{A}) \tag{1}$$

由 Bayesian 公式可以得到：

$$\hat{\vec{Y}} = \arg\max_{n} P(\vec{Y}_n \mid \vec{A}) = \arg\max_{n} \frac{P(\vec{A} \mid \vec{Y}_n) P(\vec{Y}_n)}{P(\vec{A})} \tag{2}$$

由于 $P(\vec{A})$ 表示语境信息的统计分布，可视其为常数，将其忽略，公式(2)将进一步转换为：

$$\hat{\vec{Y}} = \arg\max_{n} P(\vec{Y}_n \mid \vec{A}) = \arg\max_{n} P(\vec{A} \mid \vec{Y}_n) P(\vec{Y}_n) \tag{3}$$

公式(3)表明， 为求 $P(\vec{Y}_n \mid \vec{A})$ 这样的后验概率，被转换为求 $P(\vec{A} \mid \vec{Y}_n)$ 这个先验概率，而 $P(\vec{Y}_n)$ 则体现了韵律特征本

身的分布情况，而这种分布又通过韵律特征间出现概率和相互作用来体现。因而从公式(3)中体现了，韵律特征不仅受语境信息的影响，同时韵律特征本身也存在着相互影响的情况。

假设 1：对于 $\vec{A}$ 存在 $\vec{Y}_j$，使得 $\begin{cases} P(\vec{Y}_{i=j} \mid \vec{A}) = 1 \\ P(\vec{Y}_{i\neq j} \mid \vec{A}) = 0 \end{cases}$ 其中，$(j \in I, n \in N)$

则 $\vec{Y}$ 可以表示成 $\vec{A}$ 的函数形式，即：

$$\vec{Y} = \varphi(\vec{A}) \tag{5}$$

由此，韵律模型被简化成了简单的函数形式，这正是传统的基于公式映射的规则模型的设计基础。

假设 2：$P(\vec{Y}_n) = 常数C$

则公式(3)变为：

$$\hat{\vec{Y}} = \arg\max_n P(\vec{Y}_n \mid \vec{A}) = \arg\max_n P(\vec{A} \mid \vec{Y}_n) \tag{6}$$

公式(6)表明在基于 $P(\vec{Y}_n) = 常数C$ 假设的下，即假设韵律特征本身不产生相互作用的情况下，语境信息对韵律特征的影响和韵律特征受语境的制约是等价的。$P(\vec{A} \mid \vec{Y}_n)$ 作为先验概率的组成部分，从理论上讲，只要是具有数值训练并具有记忆能力的模型，均可以作为其实现的基础。

3. 韵律代价函数

3.1 音节基元选取中的韵律代价函数

汉语中，常用的语境信息通常包括：语音特征，其中有声调、重音、音长等信息；以及语言特征，其中有音节内部信息，如音节内的音联关系等，和超音节信息，如词性、音节位置、词位置、韵律短语、以及所含音节或词的信息等。根据Selkirk在1984[1]年提出的一种韵律分层模型，认为韵律结构从低到高的分层依次是音步、音节、韵律词、韵律短语和语调短语。

设定韵律代价函数为：

$$S = \sum_i \gamma_i V(a_i)，\ 其中，\ \gamma_i = f(\omega_i) \tag{7}$$

韵律代价函数的结果为一组语境信息的加权统计值。通过权值的记忆和训练达到适应不同语料库的目的。由公式(6)可以进一步得到：

$$\hat{\vec{Y}} = \arg\max_n P(\vec{A} \mid \vec{Y}_n) = \arg\max_n (S_n) = \arg\max_n \left[\sum_i \gamma_i V_n(a_i) \right]$$

通常情况下认为：$\gamma_i = f(\omega_i) \approx \omega_i$

其中 ω_i 为不同语境参数产生贡献的影响因子或称权值。函数权值 ω_i 的初始值确定，韵律处理的影响很大，虽然进一步的权值调整可以通过训练机制来实现。

3.2 韵律代价函数的权值初始值确定

本文介绍一种利用神经网络的权抑制的方法，较为有效而迅速的确定初始权的值。将语境信息作为神经网络模型的输入，韵律的声学参数作为输出参数。同时，在输入层（语境参数）和中间隐层之间加入权抑制层，如图1所示。则神经网络的构成函数由传统型变为：

$$\overline{F}(w) = F(w) + \lambda \sum_{\{k \mid \omega_k \in w_{sel}\}} \omega_k^2 \tag{8}$$

$$\overline{w}^{i+1} = \overline{w}^i - \eta \nabla \overline{F}(w) = \overline{w}^i - \nabla \left[\eta F(w) + \eta \lambda \sum_{\{k \mid \omega_k \in w_{sel}\}} \omega_k^2 \right] \tag{9}$$

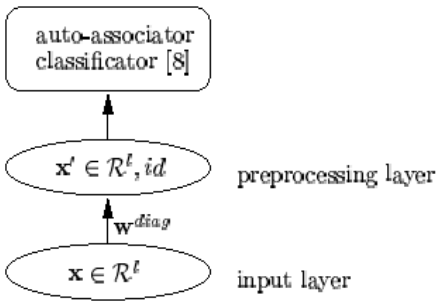

图1：具有权抑制层的神经网络模型

图2为经过一定训练后得到的与语境相关的权值分布图。通过这些数值，可以非常方便的确定韵律代价函数的权值初始值。针对汉语，实验表明，位置信息、韵律短语边界、声调信息、重音、词性等语境信息，占据着非常重音的作用。

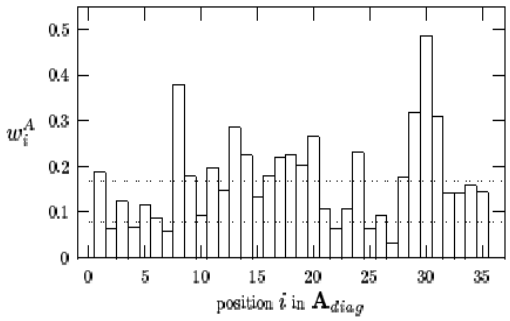

图2：权值抑制的训练结果

3.3 韵律代价函数的权值训练

虽然，韵律代价函数的权值的最终确定可以依靠设计者的经验，并经过大量测试来人工确定。但依靠自动训练的机制，却能给模型带来很大的弹性，并能产生自动适应语料库的能力。本文阐述的是通过输出误差的权值自动调节方法。

由初始权值的韵律代价函数确定的音节，其韵律参数的误差为：

$$E(\gamma)=E(\gamma,\vec{Y})=\frac{1}{M}\sum_{j=1}^{M}(\hat{\vec{Y}}_j-\vec{Y}_j)^2$$

在训练过程中，函数的权重调节则通过下式进行：

$$\gamma^{p+1}=\gamma^p+\eta^p\cdot d^p$$

其中 η 为权重调节的步长，d 则表示权值调节的方向。因而，训练的过程转换为根据学习规则从输出误差 $E(\gamma)$ 中得到权值调节方向 d 的过程。

对于 γ^p，曲面变换最陡的方向即为 $\nabla E(\gamma^p)$ 所指的方向。由此，权值调节方向 d 应为 $\nabla E(\gamma^p)$ 的函数。即：

$$d=f(\nabla E(\gamma^p))$$

为避免训练过程过快，导致权值较难稳定，本文引入了一个特殊的因子，定义为：

$$\beta_k=\frac{1}{\sqrt{\sum_{t=1}^{T}(\frac{\partial E^t}{\partial\gamma^p}-\hat{E})^2}}\qquad \text{其中}\quad \hat{E}=\frac{1}{T}\sum_{t=1}^{T}\frac{\partial E^t}{\partial\gamma^p}$$

权值调节方向为：

$$d=-\begin{bmatrix}\beta_1 & 0 & 0\\ 0 & \cdot & 0\\ 0 & 0 & \beta_k\end{bmatrix}\cdot\nabla E$$

权值变换公式则变为：$\gamma^p=\eta\cdot d^p$

4. 语音合成中选音模型的实现

在传统的韵律模型的设计中却往往忽略了韵律特征本身的相互作用，即 $P(\vec{Y}_n)$ 所起的作用。而很多情况下，$P(\vec{Y}_n)$ 所导致的韵律特征本身的相互作用又是非常明显。例如：当一个音节本身被重读时，通常会影响后续音节的发音等。这一结论，在吴宗济先生的“普通话三字组变调规律”[4]，以及林茂灿的“北京话轻声的声学性质”[10]、“普通话轻声与轻重音”[8]的论文中均提到了韵律特征间的互相影响。如：林茂灿提到的“轻声音节F_0曲线的形成，是由于它跟前面重读音节发生声调协同发音所致”，就是这一现象的典型反映。又如：汉语中出现的词中，其音节基频曲线的形状，往往受着前后发音的基频曲线或发音轻重的影响。而这些因素，却常常被目前语音合成系统的韵律模型所忽视。公式(3)则将其整合到一个统一的数学模型中。因而，一个好的韵律模型必须要能够反应韵律受语境的制约情况，而且还因能够体现韵律自身相互作用的现象。

在采用韵律代价函数进行音节基元选取的韵律模型中，当考虑相邻音节的韵律特征关系时，其整句评估因子可以表示为：$Q=\sum_j(\sum_i\gamma_iV_j(a_i))P_{j-1,j}$。其中，$P_{j-1,j}$ 表示韵律单元的转移出现概率，它体现了韵律特征之间的相互作用。然而，由于 $P_{j-1,j}$ 只表示了相邻韵律特征间的作用情况，从整句上看，应该下列公式来替代 $P_{j-1,j}$：

$$P'=P(\vec{Y}_j\mid\vec{Y}_{j-1}\vec{Y}_{j-2}\cdots\vec{Y}_1)P(\vec{Y}_{j-1}\mid\vec{Y}_{j-2}\vec{Y}_{j-3}\cdots\vec{Y}_1)\\ \bullet P(\vec{Y}_{j-2}\mid\vec{Y}_{j-3}\vec{Y}_{j-4}\cdots\vec{Y}_1)\cdots P(\vec{Y}_1)\qquad(10)$$

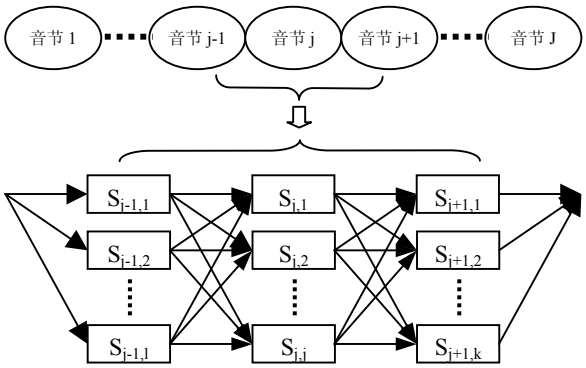

图 3：考虑韵律转移概率后的韵律代价函数在基元选取时的示意图

由于韵律参数预测，除了要有较好的语境信息到韵律参数之间的预测模型外，必须将整个模型放在整句甚至是篇章的整体环境中考虑。而在连续的语句中，韵律特征间的耦合效应则不可避免的出现，因而在韵律建模中，引入反应韵律特征转移的函数，对从整体上完善韵律模型，并提高韵律参数预测质量，将会起到较大的作用。具体实现起来，则通过在韵律模型中增加反馈的方式，或通过将整个韵律模型转换为，根据韵律参数出现的同现概率来实现。运用韵律参数出现的同现概率，实际上就是：在一组带转移概率的矩阵中，搜索最佳概率路径的问题。

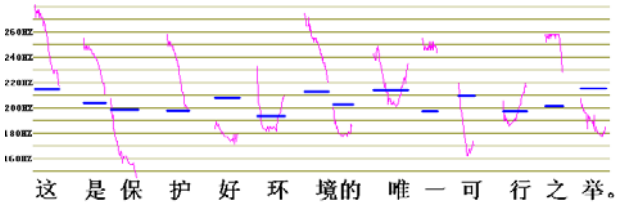

图 4：不考虑韵律特征相互作用选音的结果

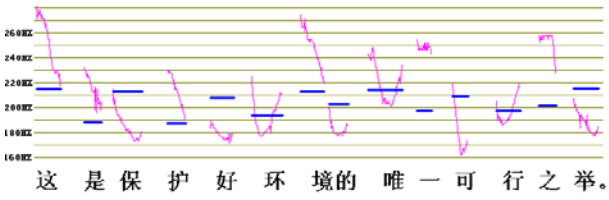

图 5：考虑韵律特征相互作用选音的结果

图 4 和图 5 显示了考虑韵律特征相互作用前后的选音结

果图。其中“是”、“保”、“护”、“环”等音节在考虑韵律特征相互作用后，其调域出现了下降或一些其它变化，从而使该句发音更为平滑，并体现了更好的节奏感。

5. 韵律节奏的预测

5.1 韵律节奏预测的一般性问题

韵律节奏作为一个重要的韵律特征，对语音合成的自然度和可懂度产生这重要影响，它通过人的自然呼吸将语句按不同的语气分成一些片断，既增强了语句的节奏感、流畅性，同时也便于消除一些歧义。在以选音算法为主要思路的语音合成系统中，韵律短语和重音在语境信息的组成中，占有非常重要的地位。建立韵律短语预测机制，较为完美的方法当然应该是从语义的分析入手，然而现代文本分析技术还很难从深层次中获取语句或篇章的完整语义信息，即便是这样，直接建立在语法结构的韵律建模思路，也是可行的。

5.2 基于分类决策树的汉语韵律短语预测

目前对韵律短语预测的主要方法有：采用统计模型，利用韵律边界相对语法结构的统计结果，来进行韵律短语边界的预测；另一种方法，采用分类决策树技术的规则学习算，该方法利用语料中手工标注的韵律边界信息建立语法信息到韵律短语边界的预测模型。

由于可以融入一定的人工经验，因此，采用决策树方法对韵律重音和短语界定是一种较好而可行的方法。这种方法基于大量的已标注了分词、拼音和词性的语料的统计，通过实现根据经验提出的大量问题集而自动训练决策树，并使用训练后的决策树对给定的文本进行节奏的预测，其结果相较传统的方法要了较大的改善。图 6 所示为利用分类决策树法实现韵律短语预测的示意图。

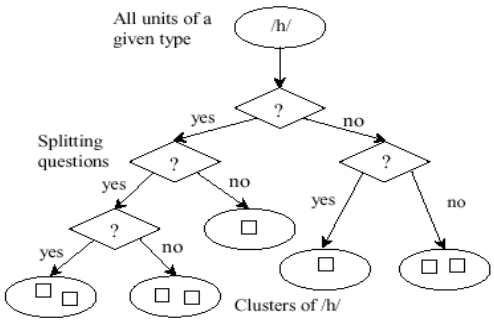

图 6：韵律重音和短语预测决策树示意图

已分词的文本：	|一 | 一 | 奉　陪 | 吃　喝 |
词性标注：	\ z \ z \ v \ v \
韵律边界标注：	/一　一 / 奉　陪　吃　喝 /
拼音信息：	|yi1　yi1|feng4　pei2| chi1 he1|

表 1：用于韵律短语边界预测的文本及相应标注实例

而利用的规则信息则根据不同的上下文信息来确定，如：词性、位置、音节声调等。决策树的叶子评估函数，可以采用输出目标的距离来衡量。该算法的特点要求设计的规则简练并多样化。

5.3 问题集的提出

在汉语韵律短语的预测中，决策树节点的特征向量包括了 13 个数值型或分类型的变量，从这些变量中提出了 95 个问题。这些问题包括了两种类型，即枚举型和数值型。无论何种问题，其答案都是二值化的。这些问题的合理选择能够使得构造的决策树生长平衡，即一定程度上防止决策树右偏现象的发生。

选择的问题分为三大类：

- ✧ 拼音相关类：如：当前字和前后字音调的判断，当前字的声母、韵母和前接的韵母和后接的声母的判断等。
- ✧ 词性相关类：预留了 50 个词性的判断，实际语料的词性标注只使用了 29 个词性。
- ✧ 位置信息类：如：词内字位置，句内词位置和字数、词数的判断等。

由于汉语语法结构的独特性，一些研究表明，汉语的虚词，如：在、和、的、地、于等，以及一些常用的关键词，如：如果、但是、虽然等，在汉语的韵律短语或语法从句起着重要作用。因而，进一步分析这些词在韵律节律中的体现，对最终改善规则学习的算法，将起到很大的作用。本文通过实验表明，辅以助词分析的决策树模型，获得了 86%（相对于人工标注，在本文的语料标注中，语料标注为经过训练的听音人，其人工标注统一性在 88%左右）的韵律短语预测的精确度。因而该模型的运算精确度，基本能够满足语音合成系统的韵律处理的要求。

6. 基于概率的韵律模型的误差分析

用类条件概率或联合分布表示韵律建模的方法实际上是内插和外插数据的一种方法，也是基于最大后验概率的分类过程。最佳的判决是基于最大后验概率的判断。但最佳的判决并不意味着没有误差，只能认为总的系统误差是最小的。

在模型中，问题不在于是否判断出错，而是要尽可能减少错误。为了评价一个判决规则的好坏，必须计算误识概率，即将一个韵律量化样本错分到其它类别中去的概率。

连续联合概率分布非常简便地显示出系统误差的基本性质。现考虑一个二分情况，如果是基于最大后验概率原则

进行判决，则对于语境参数 A，我们判决

$A \in Y_1$ ，当且仅当 $P(Y_1 \mid A) > P(Y_2 \mid A)$ (11)

否则， $A \in Y_2$ 。

即，在 $P(Y_1 \mid A) > P(Y_2 \mid A)$ 的情况下，可以由语境参数 A 得到韵律量化参数 Y_1 。

对于任意 A，误差概率密度由下式给出

误差概率密度＝ $P_l(A)$

$$= \min\{P(Y_1 \mid A), P(Y_2 \mid A)\} \quad (12)$$

特别地，对于一维情况，可以用两个联合概率分布 $P(Y, A)$ 的光滑曲线来描述，如图 7 所示。

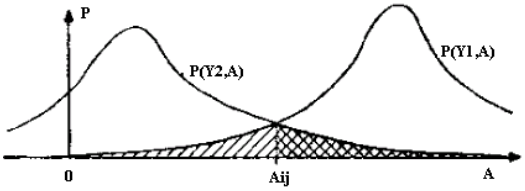

图 7：阈值判断图

令 A_{ik} 表示两个联合概率分布相交点地 A 值，则根据表达式(12)，当且仅当 $A > A_{ik}$ 时，判决 $A \in Y_1$ ，当且仅当 $A < A_{ik}$ 时，判决 $A \in Y_2$ ；对于 $A = A_{ik}$ 的特殊情况，可以判属任意一类。

误差概率为在所有 A 取值上的平均。

$$\text{系统误差} = \int_{-\infty}^{A_{ik}} P(Y_1 \mid A)P(A)dA + \int_{A_{ik}}^{+\infty} P(Y_2 \mid A)P(A)dA$$

$$= \int_{-\infty}^{A_{ik}} \frac{P(A \mid Y_1)P(Y_1)}{P(A)} P(A)dA + \int_{A_{ik}}^{+\infty} \frac{P(A \mid Y_2)P(Y_2)}{P(A)} P(A)dA$$

$$= \int_{-\infty}^{A_{ik}} P(A, Y_1)dA + \int_{A\infty}^{+\infty} P(A, Y_2)dA \quad (13)$$

即为图 7 的阴影。这种情况下音节样本的分类和选取最终不需要知道概率分布的详细情况，所要求的只是一个数 A_{ik} ，将其作为一个决策函数。对于更复杂的情况，相应的处理会困难一些。这是也许需要更多的决策值，如图 8 所示，当 $A_1 < A < A_2$ 或 $A > A_1$ 时， $A \in Y_1$ 。对于高维的情况，阈值可能是直线、平面或超平面。尽管分类是基于统计规则的，但它的表现形式是确定的，只有在分析系统误差时，才会想起其统计的原因。

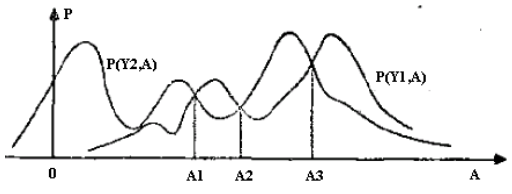

图 8：多维情况下阈值判断图

7. 结论

本文采用统计中概率的方法对韵律建模思路进行了阐述，在此基础上指出，本文分析了采用韵律代价函数进行直接基元选取的算法，并指出这一算法是韵律建模的一种具体实现。同时，本文利用该概率模型，指出了韵律特征间的相互作用在韵律建模中的重要作用，以及对提高整句合成语音自然度的重要影响。本文的工作，虽然在很多方面，还有待进一步细化，但希望通过本文的工作，将韵律建模的研究引入更多的数值化、可计算化和可训练化的方法。新一代的语音合成系统正向着概念到语音或意念到语音的方向发展，其语料库的组成也越来越庞大，信息含量也变得更为丰富。语音合成需要逐步摆脱过渡依赖设计人经验，变得规范化，在这种发展趋势中，有关整个语音的研究可计算化，尤其是语音合成各组成核心模组的数值模型化，将会变得越来越重要。

8. 参考文献

[1] Selkirk, E. (1984) Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT press.

[2] Katherine Morton, “Adding Emotion to Synthetic to Synthetic Speech Dialogue Systems”, ICSLP97, P675-678

[3] Achim Mueller, Jianhua Tao, Ruediger Hoffmann, “Data-driven importance analysis of linguistic and phonetic information”, ICSLP2000.

[4] 吴宗济，普通话三字组变调规律，中国语言学报，第二期，1985

[5] 沈炯，北京话上声连读的调型组合和节奏形式，中国语文，1994年第4期

[6] 沈炯，汉语语调模型议，语文研究，1992，VOL 4，P16-24

[7] 沈炯，从轻音现象看语音与语法研究的关系，吕叔湘等著，马庆株编《语法研究入门》，商务印书馆1999，158页

[8] 林茂灿、颜景助，普通话轻声与轻重音，语言教学与研究，1990年第3期。

[9] 陶建华、蔡莲红等，“汉语TTS系统中可训练韵律模型的研究”，声学学报,第26卷,P67-72

[10] 林茂灿、颜景助，北京话轻声的声学性质，方言，1980年第3期

[11] Andrew J. Hunt and Alan W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database”, ICASSP 96,