汉语 TTS 系统中可训练韵律模型的研究 *

陶建华# 蔡莲红 赵世霞

（清华大学计算机系　北京　100084）

摘要　针对汉语的韵律特征受语境参数影响时，表现出层次性的特点，本文描述了一种带特殊加权因子和输出优化功能的神经网络，并用其来构筑汉语 TTS 系统的韵律模型。大量测试表明，该神经网络的拓扑结构相较传统的神经网络模型更能反应出汉语的韵律特点。它提高了模型本身的收敛速度和运算精度，从而改善了整个韵律模型的质量。同时，本文还对汉语音节的基频曲线进行了规格化处理，较详细的分析了音节基频规格化参数—SPiS Parameters (Syllable Pitch Stylized Parameters)在基频调节中的作用和方式。SPiS 参数能够反应出汉语的声调特点，且方便了网络模型的建立和汉语韵律的控制。

关键字　韵律模型、神经网络、音节基频规格化参数（SPiS Parameters）、TTS 系统

一、引　言

随着语音学和计算机技术的发展，语音合成系统（TTS 系统）的研究已获得了重大的进展，并已成功的应用在许多不同的场合。但是，合成语音的结果依然与人自然流畅的发音相去甚远，其中的关键就在于语音韵律模型还不很完善。近几年来，随着计算机处理的进一步深入，从大量语料中提取连续语句的韵律特征，已逐渐成为可能。鉴于神经网络具有良好的自动学习和参数映射的特点，可以使系统具有不断地自我学习和输出优化功能，因此，将神经网络用于语音韵律模型的构造，愈来愈受到重视。目前，国外有关基于神经网络的 TTS 系统的研究要稍早一些，相较传统的规则合成方法，它们研究的结果都表明其合成语音的自然度均得到了一定程度的提高。其中较为典型的有，Motorola 的 O.Karaali[2]等人的工作。而，将神经网络用于汉语语音韵律的研究还刚刚处于起步阶段。由于汉语是一种有调语言，相较其它西方语系，无论是韵律描述方面，考虑韵律的基本单位方面，还是语境信息的归纳方面，均与西方语系有着非常大的区别，因而，其韵律的处理方法也有着很大的不同。

本文将详细研究神经网络在汉语韵律模型中的应用和实现方法。目前，研究汉语神经网络韵律模型较多的主要有台湾交通大学的陈信宏等人[3]，另外，中国科学院自动化所黄燕[7]、计算所朱廷劭和声学所许洁萍等人，也分别从不同的侧面对神经网络在汉语韵律中的应用进行了一定的研究。与这些人的工作相比，本文的工作则侧重于如下三个方面：1、针对汉语韵律表现出层次性的特点，提出了更适合汉语韵律处理的神经网络拓扑结构，并提出对其输出进行优化的方法。传统的神经网络在用于汉语韵律的学习时，其网络拓扑结构往往不能很好反应汉语的韵律特性，因而，用于建立汉语韵律模型时，其网络的收敛性和映射能力受到较大的限制。而，本文在神经网络内部引入了特殊的加权因子，从而使神经网络在汉语韵律的训练中，无论其收敛速度，还是效率都得到了较大的提高。另外，本文还利用了高斯参数分解方法，对神经网络的输出参数进行优化，一定程度上增强了网络的容错性。2、对汉语韵律特征受语境信息的影响，进行了一定的归纳和总结。3、提出了一种对汉语的音节基频曲线进行规格化处理的方法，该方法较为简洁，不仅适合于用大语料对神经网络进行训练，也非常适合汉语语音的基频控制。

*本课题受国家 863 高技术项目和国家自然科学基金（69875008）资助。

二、影响韵律特征的语言环境参数的选取

通常，汉语的韵律模型可以表示为：

$$\vec{P} = F(\vec{A}_1, \vec{A}_2, \vec{A}_3)$$

其中，$\vec{P}$ 表示韵律参数矢量，$\vec{A}_1$、$\vec{A}_2$ 和 $\vec{A}_3$ 则表示对应于不同层次韵律模型的语境参数。因此，由公式(1)可以看出汉语韵律模型的构造必须要解决三个非常关键的问题：1、必须找出影响语音韵律特征的语境参数；2、确定描述韵律特征的方法；3、构筑韵律模型。

其中，找出对韵律特征产生重要影响的语境参数是生成良好的神经网络韵律模型的基础，这将直接影响网络的收敛性。目前，有关这方面的报导比较少，本文在查阅了大量的音系学文献和进行部分实验的基础上，归纳了 17 个能对汉语韵律产生重要影响的语境参数。这些参数将帮助神经网络韵律模型的建立，同时也能对研究汉语韵律特征受语境的影响提供便利。

由于，汉语韵律特征有其特殊性，一方面，它的最小韵律单位为带声调的音节；另一方面，当多个音节组成词或词组而连续发音时，它们之间将会相互影响，形成较独立、完整的韵律块，这些韵律块的韵律特征对语音的自然度起着非常重要的作用，而不同音域的韵律块组合在一起，往往可以形成不同的语调。根据这些分析，我们将汉语的语境信息沿着语句（Sentence）--韵律短语（Phrase）--音节（Syllable）的思路划分开。共分为五组：当前音节信息 $\vec{C}$（声母类型 c_1、韵母类型 c_2、声调类型 c_3、在词中位置 c_4、与前音节耦合度 c_5 和与前音节耦合度 c_6）；相邻前音节信息 $\vec{L}$（韵母类型 l_1 和声调类型 l_2）；相邻后音节信息 $\vec{N}$（声母类型 n_1 和声调类型 n_2）；音节所在韵律短语信息 $\vec{W}$（音节数 w_1、在句中位置 w_2、重音类型 w_3、距前一个重音距离 w_4 和距后一个重音距离 w_6）以及语句信息 $\vec{S}$（语句类型 S_1 和韵律短语个数 S_2）。

其中，当前音节与前、后音节的耦合度。即为，当前音节与前、后音节的相关联程度。根据韵律短语的内部格局，以及短语间的情况，共分为 4 个等级。

语句信息、短语信息反映了整个句子的语气变化和重音的情况。所有这些标注信息共同决定着音节基频、音长等韵律参数的基本特性。

三、汉语音节声调规格化模型

在 TTS 系统中如何合理的表述汉语语音的基频曲线，一直是从事汉语韵律模型研究的较为棘手的问题。杨顺安[4]、Fujisakii[5]、许毅[6]和初敏等人均从不同的侧面，提出了对汉语基频描述或表达公式。但是，我们发现，它们的参数往往较难从训练语料的基频曲线中直接提取，因而不便于通过大语料对神经网络进行训练。这里，本文直接对汉语连续语句中音节基频曲线进行了规格化处理，得到了 SPiS 参数。该参数对基频的描述比较直观，且提取方法简洁，既方便了神经网络模型的建立和训练，也非常适合汉语语音的基频控制。

语音学的大量研究表明，普通话孤立音节的声调调型可分为三大类：平调，拱调和平拱结合调，但在连续语流中，其调型还会受到协同发音的影响，形成多种变体。通常将完整的汉语音节声调基频曲线分为三个部分：弯头段（头部）、调型段（中部）和降尾段（终尾）。

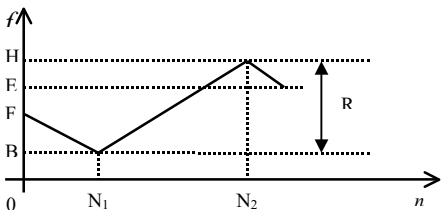

图 1： 调型曲线

根据汉语音节调型的这些特性，可以通过 SPiS 参数对其基频变化进行了描述（如图 1 所示）：B（基频曲线的最小值）、H（基频曲线的最大值）、N_1（最小值位置）、N_2（最大值位置）、F（基频曲线起始值）和 E（基频曲线

终止值)。它们共同构成矢量：$\vec{P}=(B,H,N_1,N_2,F,E)$。通过这些参数的调节，结合音长、能量参数，则较好的反映了汉语语气的轻、重、缓、急等特性。

特别需注意的是，在 TTS 系统中，合成语音的基频曲线并不是通过韵律模型生成的 SPiS 参数直接生成，韵律模型中得到的 SPiS 参数必须和合成音库中相应合成单元的基频曲线结合起来。通过 SPiS 参数对合成单元的基频曲线进行音域、斜率、首尾值等方面的综合调节，才能最终形成合成音的基频包络。这样处理的优点是，使最终生成的基频曲线保留了音库中合成单元基频的细微特征。

SPiS 参数对音库中合成单元的基频调节算法算法表述如下。

算法 1：

1、设韵律模型中输出的 SPiS 参数为：$\hat{B}$、$\hat{H}$、$\hat{N}_1$、$\hat{N}_2$、$\hat{F}$ 和 $\hat{E}$。

同时，设 $p(t)$ 为音库中合成单元的基频曲线，其 SPiS 参数为：B、H、N_1、N_2、F 和 E。先考虑 $N_1 < N_2$ 的情况。

2、计算基频的斜率变化率：$\lambda_1=\frac{N_1}{\hat{N}_1}$，$\lambda_2=\frac{|N_2-N_1|}{|\hat{N}_2-\hat{N}_1|}$ 以及 $\lambda_3=\frac{|T-N_2|}{|T-\hat{N}_2|}$。

3、计算基频幅度的变化率：$\eta_{BH}=\frac{|\hat{H}-\hat{B}|}{|H-B|}$，$\eta_F=\frac{|\hat{F}-\hat{B}|}{|\eta_{BH}(F-B)|}$ 和 $\eta_E=\frac{|\hat{E}-\hat{B}|}{|\eta_{BH}(E-B)|}$。

4、进行合成单元基频的音域变换：

$$p'(t)=\begin{cases}\eta_{BH}[p(\lambda_1 t)-B]+\hat{B} & t\in(0,N_1)\\ \eta_{BH}[p(\lambda_2 t)-B]+\hat{B} & t\in(N_1,N_2)\\ \eta_{BH}[p(\lambda_3 t)-B]+\hat{B} & t\in(N_2,T)\end{cases}$$

5、进行首尾基频变幻，进而可以得到：

$$\hat{p}(t)=\begin{cases}\frac{\eta_F(N_1-t)[p'(t)-\hat{B}]}{N_1}+\hat{B} & t\in(0,N_1)\\ p'(t) & t\in(N_1,N_2)\\ \frac{\eta_E(t-N_2)[p'(t)-\hat{B}]}{T-N_2}+\hat{B} & t\in(N_2,T)\end{cases} \quad [4]$$

其中，$\hat{p}(t)$ 为最终用于合成音的基频曲线。当 $N_1 > N_2$ 时，合成音的基频曲线同样可以算出。这里需指出的是当音节声调为阴平时，计算步骤 4 需作适当的调整，应以平移取代音域变换。

四、神经网络拓扑结构及训练算法

神经网络的输入和输出分别为语境参数 $\vec{X}=(\vec{C},\vec{L},\vec{N},\vec{W},\vec{S})$ 和韵律控制参数。其中，韵律控制参数又包括了 SPiS 参数 $\vec{P}$ 和音长参数 L，因而整个输出为：$\vec{Y}=(\vec{P},L)$。网络的拓扑结构如图 2 所示，网络基本可以分为三层，即，输入层（语境标注矢量层）、输出层（韵律控制矢量层）和中间隐层。整个结构可表示为：

$$\vec{Y}=F(\vec{X}) \quad [5]$$

在实际工作中，我们发现若将输入参数不加区别对待，常常导致网络在训练时较难收敛，或收敛很慢。由于汉语的韵律特征具有层次性，经过实验分析，这里我们进一步语境参数分

为两组：

$\vec{X}_1=(c_3,c_4,l_2,n_2,w_2,w_3,s_1)$，$\vec{X}_2=(c_1,c_2,c_5,c_6,l_1,n_1,w_1,w_4,w_5,s_5)$。

其中，$\vec{X}_1$基本上决定着音节的轻重特性，而$\vec{X}_2$则对音节基频的平滑过渡起着重要作用。

音系学的研究表明，汉语的韵律特征相较其它语言它更强调音节间的轻重搭配和语气的走势特性。一般认为，人对音节间音长和基频的相对高低反应比较敏锐。考虑到这些特性，本文进而在输入矢量$\vec{X}_1$和中间隐层之间，加入一个特殊的加权隐层以突出$\vec{X}_1$的权重，该隐层的神经元函数为：

$$y=x^2$$

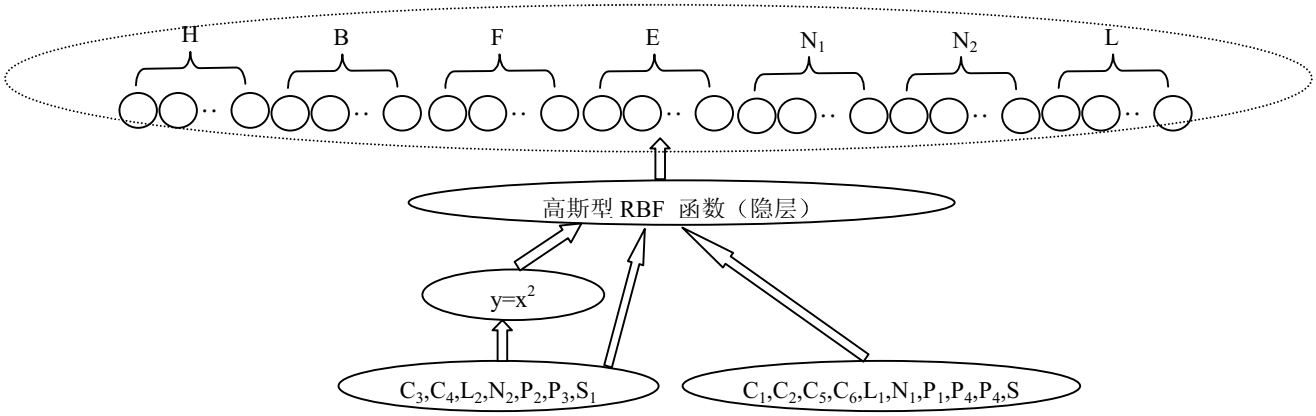

图 2：韵律神经网络模型

测试结果证明，加权隐层的引入进一步使网络结构体现了汉语的独特的韵律特点，使网络在收敛速度在原有的基础上约提高了 18%，从而较大的改善了网络的收敛性。

若将这个隐层折合到输入层，则公式[5]进而变为：

$$\vec{Y}=F(\vec{X}_1^2,\vec{X}_1,\vec{X}_2) \qquad [6]$$

不同于通常的 BP 网络，为改善网络的输出精度，本文中，网络中间层的神经元函数采用了高斯颈向基函数(RBF)。其函数可表述为：

$$x_i^{(2)}=f\left[I_i^{(2)}\right]=\left[\sqrt{2\pi}\sigma_i^{(2)}\right]^{-1}e^{-I_i^{(2)}} \qquad \text{其中：}\ \sigma_i^2>0 \qquad [7]$$

$x_i^{(2)}$为隐层第 i 个单元的输出，$I_i^{(2)}$为隐层第 i 个单元的输入，$\sigma_i^{(2)}$则为每一个单元的阈值权。

$I_i^{(2)}$又表示为：$I_i^{(2)}=\varphi\left[\vec{X}^{(1)},\vec{W}_i^{(2)}\right]=\dfrac{1}{2}\dfrac{\left\|\vec{X}^{(1)}-\vec{W}_i^{(2)}\right\|^2}{\left(\sigma_i^{(2)}\right)^2}$ [8]

其中，$\vec{X}^{(1)}$为输入矢量，$\vec{W}_i^{(2)}$为隐层第 i 个单元与输入矢量连接的权值。

网络的输出层函数则表述为：

$$y_i=x_i^{(3)}=f\left[I_i^{(3)}\right]=I_i^{(3)} \qquad [9]$$

$$I_i^{(3)}=\varphi\left[\vec{X}^{(2)},\vec{W}_i^{(3)},\theta_i^{(3)}\right]=\vec{X}^{(2)}\bullet\vec{W}_i^{(3)}+\theta_i^{(3)} \qquad [10]$$

其中$x_i^{(3)}$为输出层第 i 个单元的输出，$I_i^{(3)}$为其输入。$\vec{X}^{(2)}$为隐层的输出矢量，$\vec{W}_i^{(3)}$为输出层第 i 个单元与隐层的输出矢量之间的连接权，$\theta_i^{(3)}$为输出层第 i 个单元的阈值。

算法 2：

若训练集为$\left(\hat{\vec{X}}_{(1)}^{(3)}=\hat{\vec{Y}}_{(1)},\vec{X}_{(1)}^{(1)}=\vec{X}_{(1)}\right)\cdots,\left(\hat{\vec{X}}_{(M)}^{(3)}=\hat{\vec{Y}}_{(M)},\vec{X}_{(M)}^{(1)}=\vec{X}_{(M)}\right)$，其中 M 为样本的个数，则本文中神经网络的离线参数学习算法分如下三步进行：

1、运用公式[6]，将中间加权隐层的作用，变换到在输入层增加一个输入矢量。

2、隐层各神经元参数的$\vec{W}_i^{(2)}$用 LBG 算法学习（无监督学习）

3、在固定$\vec{W}_i^{(2)}$及 $\theta_i^{(2)}$的条件下，对于隐层至输出层的各个权值$\vec{W}_i^{(3)}$，$i=1\sim M$，用 BP 算法进行训练（有监督学习）。

五、神经网络输出参数优化

在韵律模型中应用神经网络网络的一个潜在问题就是，由于发音人在发音时易受一些人为因素的影响，将使连续语句的韵律特性具有一定的离散性，这些因素将不利于网络的训练。为保证网络输出的稳定，本文利用了概率分布的原理，采用输出离散化并取其质心的方法，对神经网络的输出进行优化。具体的方法为：每一个输出参数用十个量化间距相等的神经元取代，取输出目标值为这十个神经元的质心。如图 3 所示。具体的算法如下。

图 3：网络模型每一个输出参数对应十个量化间距相等的神经元

算法 3：

1、 将神经网络输出层的每一个神经元，均通过高斯函数$f(x)=\frac{E}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}$（取$\sigma=1$），分解成十个间距相等的神经元。

因而，对每一个训练目标值$\hat{Y}$，将分解成：

$\hat{Y}_1=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_1^2}{2}}$，$\hat{Y}_2=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_2^2}{2}}$，…，$\hat{Y}_{10}=\frac{E}{\sqrt{2\pi}}e^{-\frac{x_{10}^2}{2}}$ ［11］

其中，$\hat{Y}$为这十个值的质心，即：$\hat{Y}=\sum_{i=1}^{10}\hat{Y}_i$

2、 用$\hat{Y}_1,\cdots,\hat{Y}_{10}$对神经网络进行训练。

3、 而在神经网络工作阶段。通过网络运算得到另外一组输出值$Y_1,\cdots,Y_{10}$，则最终网络输出结果为：$Y=\sum_{i=1}^{10}Y_i$ ［12］

图 2 中虚框所示的部分，即为经过输出离散化改进的网络拓扑结构。实验表明，通过神经网络输出离散化方法，使网络的输出精度提高了约 7%，从而增强了网络输出值的稳定性，最大限度的减少因输入和输出参数的随机特性而导致的输出误差。

六、 韵律模型及结果分析

图 4 给出了完整的神经网络韵律模型的结构。

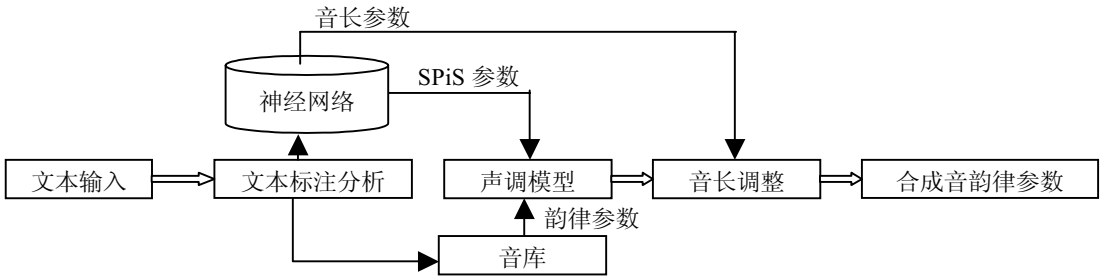

图 4：用于语音合成时的韵律模型

本文使用了 1000 个句子分别对模型进行了训练和测试。语句内容涵盖了语境标注中大部分所有可能出现的情况。其中包括：汉语中常见的句型、汉语中所有的音素、音节上下文的音联特性、音节声调组合情况、重音等信息。共有音节 10157 个。语音的采样率为 16KHz。录制的语音均经过了基频规格化和音长量化等处理。其中，75%的语料用来进行训练，而 25%的语料则用来测试。

在实际过程中，我们发现由于相邻音节间存在较强的联系，为能使网络能适应连续语句中相邻音节韵律特征的变化，在训练时以连续语句为一组单位连续进行训练可以获得比单音训练时好的多效果。图 5 和图 6 反应了一个陈述据基频和音长的测试结果。

1、基频控制参数(即 SPiS 参数)的测试结果

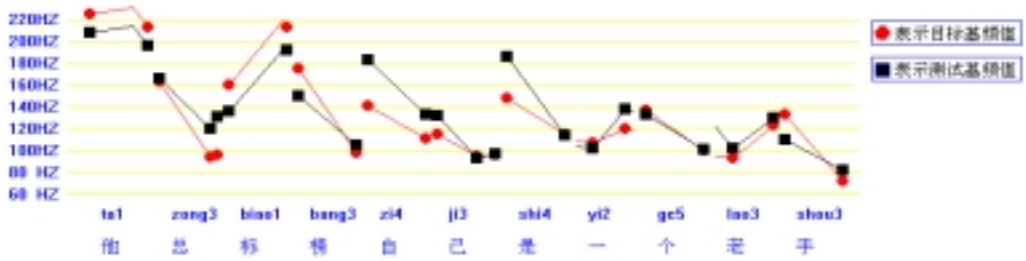

图 5：陈述句 SPiS 参数的测试结果

韵律模型的基频输出基本反应了汉语语句的韵律特征。由图 5 可以看出，其基频参数的测试的结果与真实的基频参数比较接近。其基频变化过程基本保持了陈述语气的下倾趋势，同时它还反映出了发音过程的韵律块特性。如，受发音停顿的影响，“是”作为一个韵律短语的开头，其基频和音域变得相对较高。另外，神经网络韵律模型还能很好的反映上声变调的现象。如“老手”中的“老”字，受后音的影响，由上声变为了阳平。

2、连续语句中音长参数的测试结果

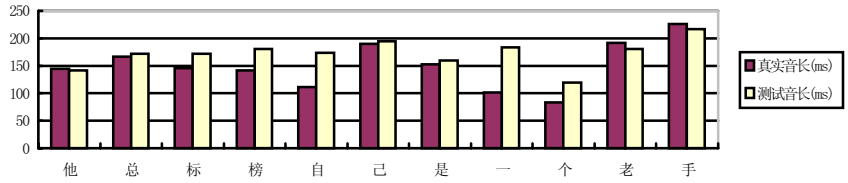

图 6：陈述句音节音长参数的测试结果

神经网络韵律模型同样输出了较好的音节音长参数，图 6 很好的反映出了语句的音长的变换趋势。由于音节音长参数在自然语句中，对控制音节发音的节奏和轻重，起着非常重要的作用。本文中，对所有的测试结果进行统计表明，81%的音节输出误差在 0～50ms，约 14%的音节输出误差在 50～120ms，而只有约 5%的音节输出误差会超过 120ms。从音长改变百分比上看：89.8%的音节，其音长输出误差占目标音长的百分比在 0～20%之间；另外，9%的音节输出误差百分比在 20%～50%之间，而只有 1.2%的音节输出误差百分比会超过 50%。因此，该模型的音长参数输出结果基本上满足了较高质量韵律控制参数的要求。

七、 小 结

将神经网络韵律模型与已有的 TTS 系统相结合，改变了传统的 TTS 系统的构筑方式。新系统合成语音的自然度得到了提高，同时也使语音合成系统中的韵律模型具有了更强的适应性和可训练性。新系统经过学习和训练，合成的语音便能体现不同的韵律特征，增加了系统的灵活性和风格的多样性。大量的测试表明，本文提出的汉语神经网络韵律模型，及其输出参数的优化方法，能适于汉语的韵律特征的处理。目前，这一模型已初步结合在我们已有的 TTS 中，输出了较为满意的合成语音。特别是，二、三、四音节组和部分较短的语句，其输出的语音自然度几乎可以和自然语音相比。

参考文献

1 Tao Jianhua, Cai Lianhong, Zhong Yuzuo, “The Context-based Method of Creating Chinese Prosodic Model”, ISSPR’98, pp271-276

2 O.Karaali etc., “Text-to-Speech Conversion with Neural Networks: A Recurrent TDNN Approach”. Proc. Eurospeech, 1997

3 Sin-Horng Chen et al., “An RNN-Based Prosodic information Synthesizer for Mandarin Text-to-Speech”, IEEE Transcations on Speech and Audio Processing, VOL.6, NO.3, 1998,5.

4 Shun-an. Yang, “ A Tonal Model For Synthesizing Polysyllabic Words and Phrases in Standard Chinese”, Essays on Linguistics, pp 65-79 (1990)

5 H. Fujisaki and K. Hirose, “ Analysis of Voice Fundamental Frequency Contours for Declarative Sentences of Japanese”, J. Acoust. Soc. Jpn.(E), Vol.5, No.4, pp 233-242, 1984

6 Ching X. Xu, Yi Xu, and Li-Shi Luo, “A Pitch Target Approximation Model for F0 Contours in Mandarin”, ICPHS99, San Francisco, pp2359-2362

7 Huang Yan, Huang Taiyi, “A Neural Learning Approach for Duration Parameter Generation in Mandarin Speech Synthesis”, ISCSLP’98, pp118-121

The Study of the Trainable Prosodic Model for Chinese TTS System

Tao Jianhua Cai Lianhong Zhao Shixia
(Department of Computer Science, Tsinghua University, Beijing 100084)

ABSTRACT Chinese prosody is characterized by its hierarchical structure when influenced by linguistic environments. Based on this, a neural network with specially weighted factors and optimizing outputs, is described and applied to construct the Chinese prosodic model in TTS system. Extensive tests show that the structure of the neural network characterizes the Chinese prosody more exactly than traditional models. Convergence is speeded up and computational precision is improved, which makes the whole prosodic model more efficient. Futhermore, the paper also stylizes the Chinese syllable pitch, and anlyzes the SPiS parameters (Syllable Pitch Stylized Parameters) in adjusting the syllable pitch. It shows that the SPiS parameters effectively characterize the Chinese tone, and facilitate the establishment of the network model and the prosodic controlling.

KEYWORDS Prosodic Model, Neural Network, SPiS Parameters, TTS System