第 26 卷第 1 期　　声 学 学 报　　Vol. 26, No. 1
2001 年 1 月　　ACTA ACUSTICA　　Jan., 2001

汉语文语转换系统中可训练韵律模型的研究*

陶建华　蔡莲红　赵世霞　吴志勇
(清华大学计算机系　北京　100084)

1999 年 9 月 7 日收到
1999 年 12 月 22 日定稿

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人工神经网络，并用其来构筑汉语 TTS 系统的韵律模型。大量测试表明，该人工神经网络的拓扑结构相较传统的人工神经网络模型更能反映出汉语的韵律特点。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，在基频调节中的作用和方式。SPiS 参数能够反映出汉语的声调特点，且方便了网络模型的建立和汉语韵律的控制。

PACS 数：　43.70

The study of the trainable prosodic model for Chinese text to speech system

TAO Jianhua　CAI Lianhong　ZHAO Shixia　WU Zhiyong
(*Department of Computer Science and Technology Tsinghua University* Beijing 100084)

Received Sept. 7, 1999
Revised Dec. 22, 1999

Abstract　Mandarin prosody is characterized by its hierarchical structures when it is influenced by the context. An artificial on this, a neural network, with specially weighted factors and optimizing outputs, is described and applied to construct the Mandarin prosodic model in a TTS system for Chinese. Extensive tests show that the structure of the artificial neural network characterizes the Mandarin prosody more exactly than traditional models. Learning rate is speeded up and computational precision is improved, which makes the whole prosodic model more efficient. Furthermore, the paper also stylizes the Mandarin syllable pitch contours with SPiS parameters (Syllable Pitch Stylized Parameters), and analyzes them in adjusting the syllable pitch. It shows that the SPiS parameters effectively characterize the Mandarin syllable pitch contours, and facilitate the establishment of the network model and the prosodic controlling.

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，并已成功地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。鉴于人工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，将人工神经网络用于语音韵律模型的构造，愈来愈受到重视。目前，国外有关基于人工神经网络的 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。然而，将人工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，与其它西方语系相比，无论是韵律描述方面，考虑韵律的基本单位方面，还是语境

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

信息的归纳方面，均有非常大的区别，因而，其韵律的处理方法也有着很大的不同。

台湾交通大学的陈信宏等人[3]采用了一种 RNN 网络来建立汉语的韵律模型，他们把 RNN 网络分成两层，其中一层用以生成汉语语句和短语的韵律信息，而另一层则生成音节韵律信息，这一结构基本能反映汉语的韵律特点，并取得了一定的成果。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，也分别从音长、二字词基频曲线以及汉语重音分类等不同的侧面应用了人工神经网络，虽然他们的工作主要集中在汉语韵律的一些特定的参数上，但对人工神经网络在汉语韵律中的应用进行了许多有意义的探索。由于，传统的人工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。这里，本文则详细研究并较全面的阐述了一种带特殊加权因子的人工神经网络在汉语韵律模型中的应用和实现方法。与传统的模型相比，该网络结构能更好的反映汉语的韵律特征，它提高了模型本身的收敛速度和运算精度，较大的改善了整个韵律模型的质量。同时本文还采用了高斯参数分解方法对人工神经网络的输出进行优化，一定程度上增强了网络的容错性。整个文章包括 5 个主要部分：(1) 分析汉语的韵律特征受语境信息影响的情况，并获得对汉语韵律特征产生重要影响的语境参数。(2) 提出适合网络学习和应用的 SPiS(Syllable Pitch Stylized Parameters) 参数。该参数较好的反映了音节基频的特征，且提取简洁，非常适合于用大语料进行训练。同时本文还较详细地阐述了用 SPiS 参数控制音节基频曲线的算法。(3) 提出了一种带特殊加权因子的人工神经网络，同时还较详细的描述了该网络的训练算法。(4) 通过高斯分解方法对网络的输出进行优化，进一步提高了网络的容错性。(5) 利用该人工神经网络模型，描述了一个完整的汉语韵律模型的结构，并对其输出结果进行了仔细的分析。

1 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：

$$\boldsymbol{P} = F(\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3), \tag{1}$$

其中，$\boldsymbol{P}$ 表示韵律参数矢量，$\boldsymbol{A}_1$、$\boldsymbol{A}_2$ 和 $\boldsymbol{A}_3$ 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 3 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。

找出对韵律特征产生重要影响的语境参数是生成良好的人工神经网络韵律模型的基础。语境参数选取的好坏，将直接影响网络的收敛性。本文根据的文本的上下文信息，按照其对汉语韵律特征不同层次的影响，将其沿着语句 (Sentence)- 韵律短语 (Phrase)- 音节 (Syllable) 的思路划分开。分为 5 组，共得到了 17 个语境参数：当前音节信息 $\boldsymbol{C}$(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 和与后音节耦合度 c_6)；相邻前音节信息 $\boldsymbol{L}$(韵母类型 l_1 和声调类型 l_2)；相邻后音节信息 $\boldsymbol{N}$(声母类型 n_1 和声调类型 n_2)；音节所在韵律短语信息 $\boldsymbol{W}$(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 $\boldsymbol{S}$(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。

语句信息和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。

2 汉语音节声调规格化模型

杨顺安[4]、Fujisaki[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，在这些方法中，它们的参数往往较难从训练语料的基频曲线中直接提取，尤其不方便通过大语料对人工神经网络进行训练。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 SPiS 参数。该参数对基频的描述比较直观，且提取方法简洁，因而非常方便人工神经网络模型的建立和训练。

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 3 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 3 个部分：弯头段 (头部)、调型段 (中部) 和降尾段 (终尾)。

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：$\boldsymbol{P}=(B,H,N_1,N_2,F,E)$ 。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。

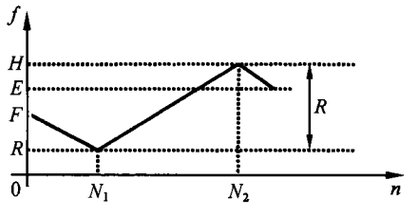

图 1　调型曲线

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表述如下。

算法 1:

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

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) 计算基频的变换率：$\lambda_1=N_1/\widehat{N}_1$ ，$\lambda_2=|N_2-N_1|/|\widehat{N}_2-\widehat{N}_1|$ 以及 $\lambda_3=|T-N_2|/|T-\widehat{N}_2|$ 。其中，T 为调型结束点位置。

(3) 计算基频音域变换率：$\eta_{BH}=|\widehat{H}-\widehat{B}|/|H-B|$ ，$\eta_F=|\widehat{F}-\widehat{B}|/|\eta_{BH}(F-B)|$ 和 $\eta_E=|\widehat{E}-\widehat{B}|/|\eta_{BH}(E-B)|$ 。

(4) 进行合成单元基频的变换：

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

(5) 进行首尾基频变换，进而可以得到：

$$\widehat{p}(t)=\begin{cases}\dfrac{\eta_F(N_1-t)[p'(t)-\widehat{B}]}{N_1}+\widehat{B}, & t\in(0,N_1),\\ p'(t), & t\in(N_1,N_2),\\ \dfrac{\eta_E(t-N_2)[p'(t)-\widehat{B}]}{T-N_2}+\widehat{B}, & t\in(N_2,T),\end{cases}\tag{2}$$

其中，$\widehat{p}(t)$ 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。

3　人工神经网络拓扑结构及训练算法

人工神经网络的输入和输出分别为语境参数 $\boldsymbol{X}=(\boldsymbol{C},\boldsymbol{L},\boldsymbol{N},\boldsymbol{W},\boldsymbol{S})$ 和韵律控制参数。其中，韵律控制参数又包括了 SPiS 参数 $\boldsymbol{P}$ 和音长参数 L ，因而整个输出为：$\boldsymbol{Y}=(\boldsymbol{P},L)$ 。网络的拓扑结构如图 2 所示，网络基本可以分为 3 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：

$$\boldsymbol{Y}=F(\boldsymbol{X}).\tag{3}$$

在实际工作中，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，通过对人工神经网络中输入参数对输出参数的灵敏度分析，可以发现，整个语境参数则可以分为两组：

$$\boldsymbol{X}_1=(c_3,c_4,l_2,n_2,w_2,w_3,s_1),$$
$$\boldsymbol{X}_2=(c_1,c_2,c_5,c_6,l_1,n_1,w_1,w_4,w_5,s_2),$$

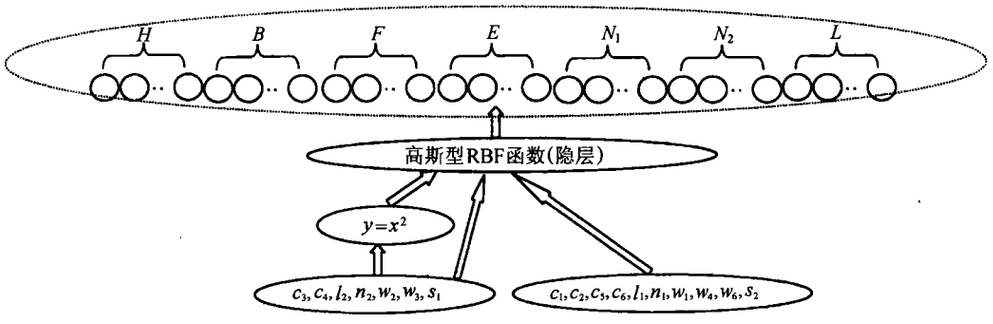

图 2　韵律人工神经网络模型

其中，$\boldsymbol{X}_1$ 基本上决定着音节的轻重特性，它对网络输出的 SPiS 参数有相对较大的影响，而 $\boldsymbol{X}_2$ 则对音节基频的平滑过渡起着重要作用，其影响相对较弱。同时，音系学的研究表明，汉语的韵律特征较其它语言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 $\boldsymbol{X}_1$ 和中间隐层之间，加入一个特殊的加权隐层以突出 $\boldsymbol{X}_1$ 的权重，该隐层的神经元函数为：

$$y = x^2$$

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 (3) 进而变为：

$$\boldsymbol{Y} = F(\boldsymbol{X}_1^2, \boldsymbol{X}_1, \boldsymbol{X}_2). \tag{4}$$

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径向基函数 (RBF)。其函数可表述为：

$$x_i^{(2)} = f[I_i^{(2)}] = \left[\sqrt{2\pi}\sigma_i^{(2)}\right]^{-1} e^{-I_i^{(2)}}, \text{ 其中：} \sigma_i^2 > 0, \tag{5}$$

$x_i^{(2)}$ 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：

$$I_i^{(2)} = \varphi\left[\boldsymbol{X}^{(1)}, \boldsymbol{W}_i^{(2)}\right] = \frac{1}{2}\frac{\left\|\boldsymbol{X}^{(1)} - \boldsymbol{W}_i^{(2)}\right\|^2}{\left(\sigma_i^{(2)}\right)^2}, \tag{6}$$

其中，$\boldsymbol{X}^{(1)}$ 为输入矢量，$\boldsymbol{W}_i^{(2)}$ 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：

$$y_i = x_i^{(3)} = f\left[I_i^{(3)}\right] = I_i^{(3)}, \tag{7}$$

$$I_i^{(3)} = \varphi\left[\boldsymbol{X}^{(2)}, \boldsymbol{W}_i^{(3)}, \theta_i^{(3)}\right] = \boldsymbol{X}^{(2)} \cdot \boldsymbol{X}_i^{(3)} + \theta_i^{(3)}, \tag{8}$$

其中 $x_i^{(3)}$ 为输出层第 i 个单元的输出，$I_i^{(3)}$ 为其输入。$\boldsymbol{X}^{(2)}$ 为隐层的输出矢量，$\boldsymbol{W}_i^{(3)}$ 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 $\left(\widehat{\boldsymbol{X}}_{(1)}^{(3)} = \widehat{\boldsymbol{Y}}_{(1)}, \widehat{\boldsymbol{X}}_{(1)}^{(1)} = \boldsymbol{X}_{(1)}\right), \cdots,$ $\left(\widehat{\boldsymbol{X}}_{(M)}^{(3)} = \widehat{\boldsymbol{Y}}_{(M)}, \widehat{\boldsymbol{X}}_{(M)}^{(1)} = \boldsymbol{X}_{(M)}\right)$，其中 M 为样本的个数，则网络的离线参数学习算法分下列 3 步进行：

(1) 运用公式 (4)，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 $\boldsymbol{W}_i^{(2)}$ 用 LBG 算法学习 (无监督学习)。

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

4 人工神经网络输出参数优化

在韵律模型中应用人工神经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，本文利用概率分布的原理，采用输出离散化并取其质心的方法，对人工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。

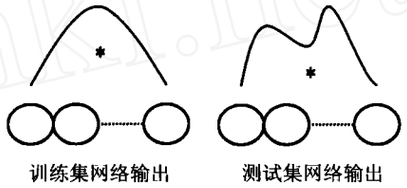

图 3 网络模型每一个输出参数对应 10 个量化间距相等的神经元

算法 3:

(1) 将人工神经网络输出层的每一个神经元，均通过高斯函数 $f(x) = E/(\sqrt{2\pi}\sigma)e^{-x^2/2\sigma^2}$(取 $\sigma = 1$)，分解成 10 个间距相等的神经元。

因而，对每一个训练目标值 $\widehat{Y}$，将分解成：

$$\widehat{Y}_1 = \frac{E}{\sqrt{2\pi}}e^{-x_1^2/2}, \widehat{Y}_2 = \frac{E}{\sqrt{2\pi}}e^{-x_2^2/2}, \cdots, \widehat{Y}_{10} = \frac{E}{\sqrt{2\pi}}e^{-x_{10}^2/2}, \tag{9}$$

其中，$\widehat{Y}$ 为这 10 个值的质心，即：

$$\widehat{Y} = \sum_{i=1}^{10} \widehat{Y}_i.$$

(2) 用 $\widehat{Y}_1, \cdots, \widehat{Y}_{10}$ 对人工神经网络进行训练。

(3) 而在人工神经网络工作阶段。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：

$$Y = \sum_{i=1}^{10} Y_i. \tag{10}$$

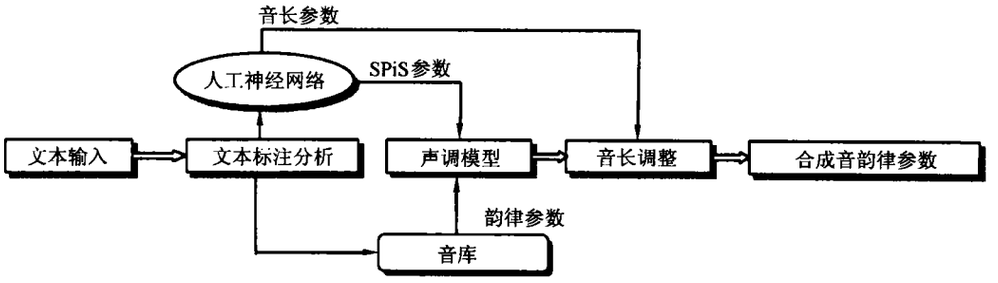

图 4　用于语音合成时的韵律模型

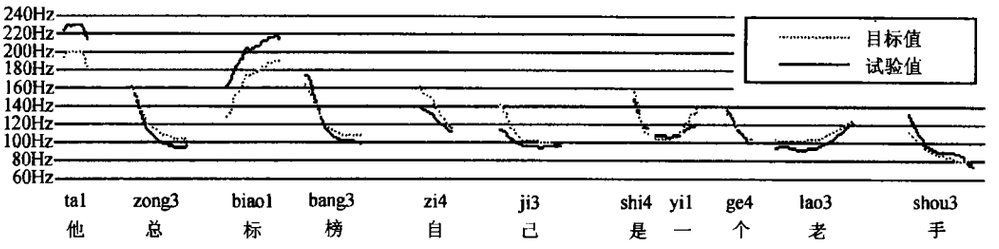

图 5　陈述句 SPiS 参数的测试结果

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，通过人工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。

5　韵律模型及结果分析

运用人工神经网络而构成的完整韵律模型如图 4 所示。

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 16 kHz 。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。

在实际训练过程中，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得多的效果。图 5 和图 6 反映了一个陈述句基频和音长的测试结果。

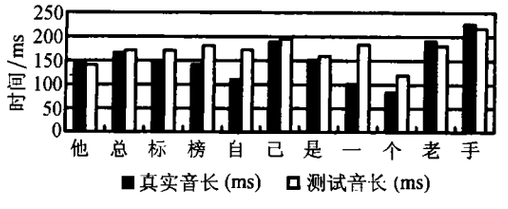

图 6　陈述句音节音长参数的测试结果

5.1　基频控制参数 (即 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，人工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。

5.2　连续语句中音长参数的测试结果

人工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。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 ～ 50 ms，约 14% 的音节输出误差在 50 ～ 120 ms，而只有约 5% 的音节输出误差会超过 120 ms 。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。

6　小结

将人工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, 本文提出的汉语人工神经网络韵律模型, 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 系统中, 输出了较为满意的合成语音。对二、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。我们将进一步研究不同长度和韵律结构的语句。

参 考 文 献

1 TAO Jianhua, CAI Lianhong, ZHONG Yuzuo. The context-based method of creating Chinese prosodic model. ISSPR'98, 1998: 271—276

2 Karaali O *et al.* Text-to-speech conversion with neural networks: a recurrent TDNN approach. Proc. Eurospeech, 1997: 561—564

3 CHEN Sinhorng *et al.* An RNN-based prosodic information synthesizer for mandarin text-to-speech. *IEEE Transcations on Speech and Audio Processing*, 1998; **6**(3): 226—239

4 YANG Shunan. A tonal model for synthesizing polysyllabic words and phrases in standard Chinese. Essays on Linguistics, 1990: 65—79

5 Fujisaki H, Hirose K. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *J. Acoust. Soc. Jpn.(E)*, 1984; **5**(4): 233—242

6 XU Chingx, XU Yi, LUO Lishi. A pitch target approximation model for F0 contours in mandarin. ICPHS99, San Francisco, 1999: 2359—2362

7 HUANG Yan, HUANG Taiyi. A neural learning approach for duration parameter generation in mandarin speech synthesis. ISCSLP'98, 1998: 118—121