文章编号：1003-0077（2011）00-0000-00

基于声调核参数及 DNN 建模的韵律边界检测研究*

林举[1]，解焱陆[1]，张微[1]，张劲松[1]

（1.北京语言大学 信息科学学院，北京市 邮编 100083）

摘要：韵律边界对言语表达的自然度和可理解度有着重要作用。韵律建模也是语音合成、语音理解中的重要方面。本文从相邻声调的相互作用角度出发，提出基于深度神经网络（DNN）及声调核声学特征的汉语韵律边界检测方法。该方法首先采用声调核部分的声学特征来计算边界检测相关参数。然后，利用深度神经网络进行建模。作为对比，实验中采用了以整个音节的声学特征为输入特征的基线系统。结果表明，只使用调核部分声学特征的系统优于使用整个音节的系统，韵律边界检测正确率相对提高了 4%，这表明本文提出的汉语韵律边界检测方法的有效性。

关键词：韵律边界建模；声调核；深度神经网络

中图分类号：H01，TN912.3　　**文献标识码：**A

Automatic Mandarin Prosody Boundary Detecting Based on Tone Nucleus and DNN Model

LIN Ju[1], XIE Yanlu[1], ZHANG Wei[1], ZHANG Jinsong[1]

(1. College of Information Sciences, Beijing Language and Culture University, Beijing 100083, China)

Abstract: Prosody boundary plays an important role in naturalness and intelligibility of speech of verbal expressions. Thus, prosody modeling is important aspect of speech synthesis and understanding. In this paper, from the angle of interaction of adjacent tones, we propose a method of prosody boundary detecting based on tone nucleus and DNN model. This method firstly calculated the boundary-related parameters by applying the tone nucleus features. Then, the parameters were modeled by deep neural network. For comparison, the baseline system was used the acoustic features of syllable. The experimental results showed that the proposed method used tone nucleus features outperformed the baseline system, with a relative 4% improvement. It demonstrated the efficiency of the proposed method.

Key words: Prosody Boundary Modeling; Tone Nucleus; Deep Neural Network

1 引言

人们在进行言语交际的时候，除了字面的文字信息之外，话语韵律变化也是相互传递的一个重要信息。合理有效地组织话语韵律结构不仅有助于说话者更清楚地表达，而且听话者也能够更清楚、准确地理解说话人的意图。从语音学角度来看，韵律间断或韵律边界表示相邻音节的疏远程度。韵律边界通常是将一串语流切分成大小不同的韵律单元，如韵律词、韵律短语等。它不仅减轻人脑理解加工的负担，也方便机器处理。韵律边界在人类言语表达的自然度和可理解度方面扮演着非常重要的角色。近年来，由于韵律边界信息在语音合成、语

* **收稿日期：**　　**定稿日期：**

基金项目：北京语言大学梧桐创新平台项目资助（中央高校基本科研业务费专项基金）（编号：16PT05）；北京语言大学研究生创新基金资助项目（中央高校基本科研业务费专项资金）（16YCX163）

作者简介：林举（1990——），男，硕士，主要研究方向为计算机辅助发音教学；解焱陆（1980——），男，博士，副教授，主要研究方向为计算机辅助语言习得、语音信号处理；张微（1993——），女，硕士，主要研究领域为计算机辅助发音教学；张劲松（1968——），通讯作者，男，博士、教授，主要研究方向为语音习得、韵律建模、语音识别、实验语音学、计算机辅助发音教学。

音理解等领域起到重要的作用，越来越多的人关注韵律边界的自动检测。

声学特征、词典和语法方面的特征在边界检测中被广泛用来建模。对于英语，Ostendorf 在用语音识别提供的音节或者音素强制对齐的基础上，利用时长、基频以及能量的特征构建决策树模型，以预测间断的类型，取得了 77.0%的正确率[1]。Hasegawa-Johnson 等人[2]利用多层感知机（MLP）对基频、时长等声学信息进行建模，同时利用 SVM 对词典和语义信息进行建模，最后在波士顿大学的广播新闻语料库 BURNC）上的间断检测率为 91.1%。Chen 等人利用上下文相关的隐马尔科夫模型（CD-HMMs）和 bigram 先验分布的方法在 BURNC 语料库上取得边界标注的 F-score 值为 79.6%[3]。对于汉语，胡伟湘等人[4]利用分类决策树（CART）在 ASCCD 的韵律标注语料库上，通过声学和文本特征对韵律边界进行建模，实验表明该方法也能够达到较好的预测正确率。倪崇嘉等人[5]利用声学特征和词典、语法特征，采用基于韵律间断层级的韵律间断分类算法，在综合测试集上取得了 78.25%正确率。杨辰雨等人[6]首先使用 CD-HMMs 模型对频谱、基频和音素时长进行建模，然后借助训练得到的模型采维特比解码完成韵律短语边界的自动标注，该方法标注时的 F-score 值达到 77.64%。

汉语是声调语言，基频负载声调和语调的信息，同时能量和时长对韵律也有一定影响。本文从相邻声调的相互作用角度出发，采用声学特征进行汉语韵律边界的自动检测。相邻的声调之间是否存在协同发音的影响与韵律边界密切相关[7]。基于声调核模型，可以准确提取相邻声调的基频重设范围以及调阶等相关参数，这些线索对于韵律间断检测提供重要帮助[8，9]。因此我们采用调核部分的声学特征以期望来提高边界检测率。调核的自动检测采用张劲松[8]提出的方法。近年来，深度神经网络（DNN）在语音识别，语义理解等领域表现出了优越的性能。本文也采用 DNN 来对调核部分声学特征进行建模。

2 调核模型

2．1 调核

有研究表明，在一个音节中，声调的负载并不是均匀分布的。张劲松[8]在声调识别任务中提出声调核模型，即一个音节的基频曲线可以分成潜在的目标部分和发音过渡部分，而潜在的目标部分就是调核部分（图 1）：

a) 潜在的目标部分代表要实现 F0 的目标值，并且在声调感知中提供主要声学线索。
b) 发音过渡部分主要出现在实现潜在的目标部分前需要经过一个过渡部分或在实现目标部分后的过渡部分。

图 1：声调核模型说明

2．2 调核与间断

在连续语流中，韵律短语中相邻的声调之间往往会存在协同发音的影响，比如连续的去声与阴平，阴平的“高”目标常常要比前面去声的“高”的目标要低很多，这是由于当两个声调在一个词内，阴平“高”的目标受到前面去声尾部“低”的目标 carryover 的影响[10]。但是当两个声调之间存在韵律边界时，这种 carryover 的影响就会消失，后面阴平的“高”

的目标就会达到相应的比较高的位置，如图 2 所示。这经常导致会有上升过渡段（CD）来达到阴平“高”的目标值。然而，基于声调核模型，上升的 CD 段属于发音过渡部分，声调信息主要负载在声调核 DE 段，并且调核段仍然符合潜在的基频目标值。如果两个声调的调核部分（AB 段和 DE 段）检测出来，阴平的音高重置范围（h）将为韵律间断检测提供重要线索。

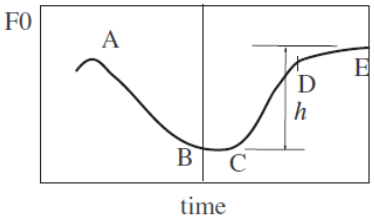

图 2：连续的去声与阴平 F0 曲线，并且中间存在韵律短语边界。

2．3 调核的自动检测

声调核的自动检测，主要分两步：提取声调核候选 F0 轨迹对应的韵律特征，然后从这些候选 F0 轨迹中选择调核。具体的做法如图 3 所示，使用分段 K-means 算法[11]聚类 F0 轨迹，依据是否符合 F0 斜率等均值假设检验来合并相邻的分段。对于最终 F0 曲线分割后只有两段的，利用线性判别分析方法（LDA）[12]设计一个区分函数来预测声调核的位置。对于最终分割后有三段的，中间的一段属于调核段，但是根据语音学规则，中间一段必须大于 50ms。否则，分割的段数将减少到 2 段，然后重复前面的操作。

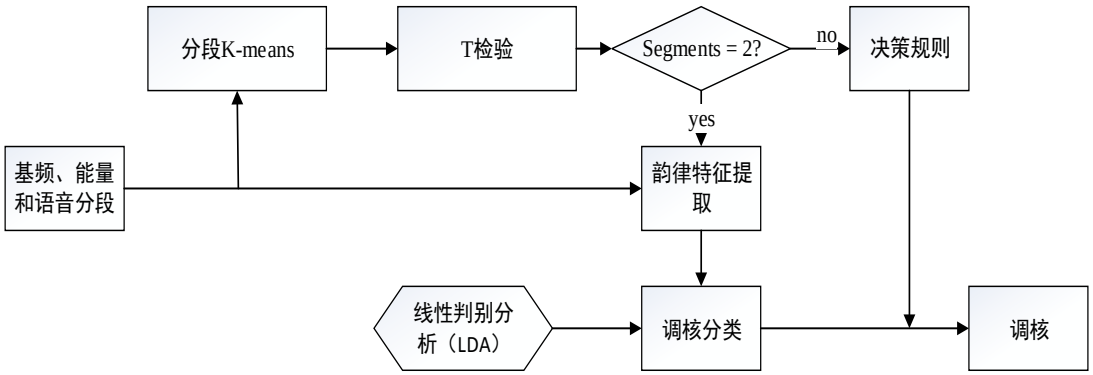

图 3：调核自动检测框架

3 深度神经网络（DNN）

DNN 的结构是一种多层感知机(multi-layer perceptron, MLP)。在给定观测向量 o 条件下，L-层的 MLP 用来对输出标签 l_{label} 的后验概率 $P(l_{label}|o)$进行建模。如图 4 所示，第一层是原始特征输入层，2...L-1 层为隐含层，每个隐含层是在给定上一层输入向量 v^l 对隐层节点 h^l 的后验概率进行建模，最顶层 L 用 softmax 来计算所有标签的后验概率：

$$P_{h_j|v}^{l} = \left(h_j^l \middle| v^l\right) = \frac{1}{1+e^{-z_j^l(v^l)}} = \sigma\left(z_j^l(v^l)\right), 1 \le l < L \tag{1}$$

$$P_{S|V}^{L}(l_{label}|v^L) = \frac{e^{z^L_{l_{label}}(v^L)}}{\Sigma_{l_{label}} e^{z^L_{l_{label}}(v^L)}} = softmax\left(z^L(V^L)\right) \tag{2}$$

$$Z^l(v^l) = (W^l)^T v^l + a^l \tag{3}$$

其中，W^l 和 a^l 表示对于隐层 l 的权重矩阵和偏置向量，h_j^l 和 $z_j^l(v^l)$ 分别表示第 l 层的第 j 个组件和它对应的激活函数值。

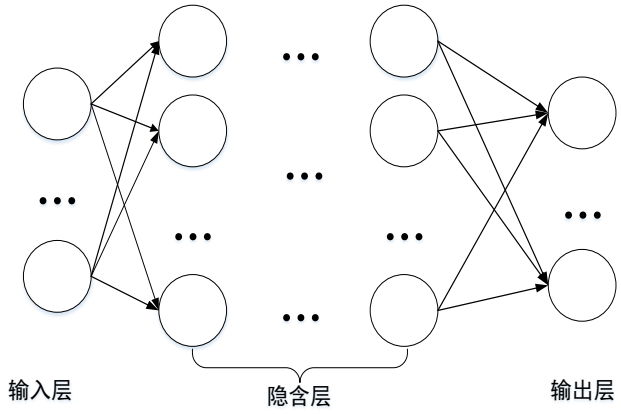

图 4：DNN 框架

4 声学特征

在对韵律边界建模时，除了时长相关特征，我们只采用调核部分对应的声学特征。同时为了减少不同说话者对声学特征的影响，用 z-score 算法对各特征进行规整。

4．1 时长相关特征

语言学上理论表明，时长相关特征对于韵律间断的建模有很大帮助。倪崇嘉[5]研究也表明：在韵律边界处，很多地方出现停顿；韵律边界处音节的时长比其他位置要长等。因此，我们把时长方面的特征用于韵律边界建模当中。

对于每一个音节，计算下列时长相关特征：

a) SilD_f：音节之后的静音段的时长；

b) SylDur：音节的时长；

c) SylDurRatio_foll：当前音节的时长与其后面一个音节时长的比值；

d) SylDurRatio_pre：当前音节的时长与其前面一个音节时长的比值。

4．2 基频相关特征

音高是非常重要的韵律特征，音高的变化反应了语调、声调、重音、信息焦点等非常复杂的韵律信息。音高重置的程度与韵律间断的层级密切相关，各级的韵律间断处的音高重置程度存在明显差异，韵律间断的层级越高，其间音高重置的程度就越大，韵律间断的层级越低，其间音高重置的程度就越小[9]。

F0 的计算使用 ESPS 中的 get_f0 命令（参数设置为：wind_dur=0.01, min_f0=60, max_f0=650）。对于每一个音节不仅计算其调型、调阶相关特征，还要考虑其与相邻音节的特征的比较。

对于每一个音节计算如下音高特征：

a） 用 $\mathrm{f}(x)=a+bx+cx^2$ 来拟合调核部分基频曲线，{a，b，c}来表示基频轮廓特征；

b） PMax：调核部分音高最大值；

c) PMin：调核部分音高最小值；

d) PRange：调核部分音高范围；

e) PMean：调核部分音高均值；

f) PMRatio：调核音高均值与其后音节调核音高均值比值；

g) PRatio：调核音高曲线的最后一个 F0 值与该音节之后音节调核部分第一个 F0 值得比值；

h) Delta_Max：调核音高最大值与其后音节调核音高最大值之间的差值；

i) Delta_Min:调核音高最小值与其后音节调核音高最小值之间的差值。

4．3 能量相关特征

与计算基频相关特征类似，计算能量相关的特征。能量是通过 praat 软件中的“To Intensity”提取，参数设置为 65, 0.01：

对于每一个音节提取如下能量相关的特征：

a) EgMax：调核部分能量最大值；

b) EgMin：调核部分能量最小值；

c) EgRange：调核部分能量的范围；

d) EgMean：调核部分能量的均值；

e) EgRatio:调核部分能量的均值与其后音节调核部分能量的比值。

5 实验与结果

5.1 实验语料

ASCCD 语料由语篇语料、语音数据和语音学标注信息组成，内容包括 18 篇文章。语音数据由 10 位北京地区标准普通话发音人录制而成。声音文件采用 16kHz 采样、16bit 数据。双声道 WAV 格式存储。语音学标注信息采用人工标注方法完成，内容包括拼音、声韵母、韵律间断、重音等。

语音学标注信息采用 C-TOBI 相应符号，利用 praat 软件完成标注[13, 14]。标注文件中标注了四层信息，我们主要关注第三层间断指数层（BI）。间断指数数值划分：0（缺省值，未标出），韵律词内的音节边界；1，韵律词间断；2，次要韵律短语（minor phrase）间断；3，主要韵律短语（major phrase）间断；4，语调组间断。具体分布如表 1 所示。

表 1：ASCCD 中各间断的分布情况。

总数	B0	B1	B2	B3	B4
87628	61518	16334	8442	7449	4072
100%	62.9%	16.7%	8.6%	7.6%	4.2%

本文实验选择每个说话人的前 58 个段子综合为训练集，其余的作为测试集。其中训练集的 10%用来作为验证集。

5.2 实验配置

我们的基线系统是采用音节层级的声学特征。在训练深度神经网络模型时，我们采用

Keras 工具包，为了得到最佳的性能，我们对比了不同的隐层数和节点数（1，2，3，4 层以及 384，512，1024 节点数）在验证集上的结果，调参结果如图 5 所示。最终我们的 DNN 网络结构为：

a） 20 个单元的输入层；

b） 3 个隐层，每个隐层包含 512 个 sigmod 单元；

c） 有 5 个 softmax 单元的输出层。

DNN 在训练时迭代 100 次，使用随机梯度下降（SGD）进行参数调整，批处理大小为 128，在输入层的 dropout[15]为 20%，隐层的 dropout 为 40%，目标函数是交叉熵。

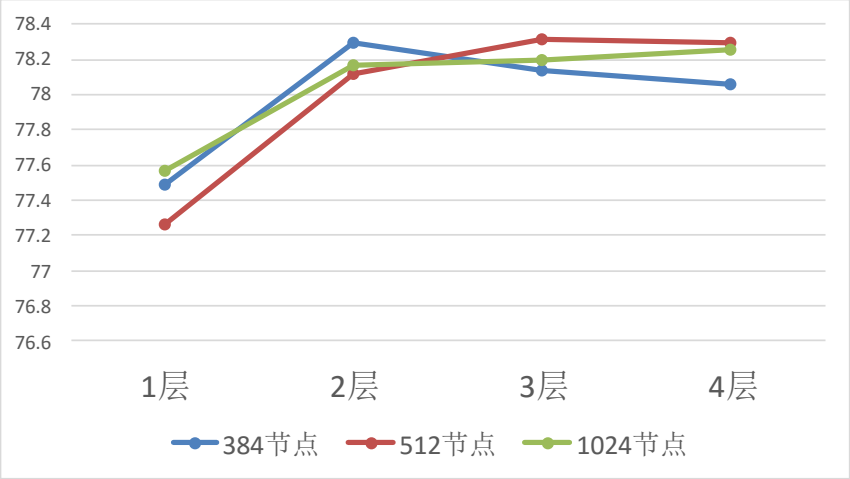

图 5：不同网络拓扑结构只使用调核部分声学特征的 DNN 系统检测性能。

5.3 实验结果

虽然我们的采用的训练集和测试集与倪崇嘉等人的研究有些许不同，但都是基于 ASCCD 语料库，可以看做是相同的实验数据，所以可以近似的比较。倪崇嘉[5]等人采用的特征包括声学特征、词典以及语法等方面的特征，采用决策树作为建模模型。我们的基线系统是使用整个音节部分的声学特征，采用 DNN 为建模模型。实验结果如表 2 所示。

表 2：不同系统的各间断检测性能比较（SY 表示音节，TN 表示调核，下同）(%)

系统	B0	B1	B2	B3	B4
SY-CART[5]	90.9	48.6	50.9	80.8	61.7
SY-DNN	96.1	21.2	44.1	75.3	74.6
TN-DNN	**96.1**	22.6	47.6	**83.6**	**75.1**

表 3：不同系统的间断检测整体性能比较(%)

系统	SY-CART[5]	SY-DNN	TN-DNN
正确率	78.3	76.5	77.34

6. 讨论

通过表 2 我们可以看到，使用调核部分声学特征的 DNN 系统在除 B0 外的各个间断层级上的检测正确率都要优于只使用整个音节声学特征的系统。B0 之所以没有提高的原因可能是因为 B0 间断一般出现在词内，容易受到协同发音的影响使得声调难以达到相应的标准模式，这也导致基于调核模型计算相应参数时缺少了针对性。而其他层级的间断，基于声调核模型在计算边界检测相应参数（基频重设和调阶等）更加精确。同时在表 3 中，只使用调核部分的声学特征相对于使用整个音节的声学特征的检测性能相对提升了 4%。这说明了我们使用的调核思想在韵律边界检测是有帮助的。

最后我们的实验结果与倪崇嘉[5]的结果相比，性能非常接近，而且我们只使用了声学特征，这一方面说明 DNN 在间断分类方面具有优势，另一方面也体现调核起到了一定作用。通过表 2 进一步可以发现，B1 的检测率相对于倪崇嘉等人的结果比较低，而 B1 对应的是韵律词边界，而韵律词边界与分词后的词边界有很大对应，因为我们只使用了声学特征，缺失词边界信息，这可能是导致 B1 检测率比较低的原因。

7. 结论

本文提出了基于深度神经网络(DNN)使用调核部分的声学特征进行韵律边界自动检测方法，并通过实验验证了该方法的有效性。在以后的工作中，将会把词典和语法特征也加入到韵律间断检测方法中，并且加大语料库来得到鲁棒性更强的模型。

参考文献

[1] C. W. Wightman, M. Ostendorf. Automatic labeling of prosodic patterns [J]. *Speech and Audio Processing*, 1994, 2(4): 469-481.

[2] M. Hasegawa-Johnson, K. Chen, J. Cole, et al, Simultaneous recognition of words and prosody in the boston university radio speech corpus [J]. *Speech Communication*, 2005, 46(3): 418-439.

[3] Q. Chen, Z. H. Ling, C. Y. Yang, et al, Automatic phrase boundary labeling of speech synthesis database using context-dependent HMMs and N-Gram Prior Distributions [C]// Sixteenth Annual Conference of the International Speech Communication Association, 2015: 1581-1585.

[4] W. X. Hu, T. Y. Huang, B. Xu, Study on prosodic boundary location in Chinese mandarin [C]// IEEE International Conference on Acoustics, 2002: 501-504.

[5] 倪崇嘉、张爱英、刘文举，等. 基于韵律间断层级的汉语韵律间断分类[J]. 计算机应用研究，2011，28(7): 2452-2454。

[6] 杨辰雨、朱立新、凌震华,等. 基于 Viterbi 解码的中文合成音库韵律短语边界自动标注[J]. 清华大学学报（自然科学版），2011, 51(9)：1276-1281。

[7] J. S. Zhang, and H. Kawanami, Modeling carryover and anticipation effects for Chinese tone recognition [C]// European Conference on Speech Communication and Technology, Eurospeech 1999.

[8] J. S. Zhang, and K. Hirose, Tone nucleus modeling for Chinese lexical tone recognition [J]. *Speech Communication*, 2004, 42(3): 447-466.

[9] 熊子瑜、林茂灿 语流间断出的韵律表现[C]// 第六届全国人机语音通讯会议论文集, 2006。

[10] Y, Xu, and Q. E. Wang, Pitch targets and their realization: Evidence from Mandarin Chinese [J]. *Speech communication*, 2001, 33(4): 319-337.

[11] L. Rabiner, B. H. Juang, Fundamentals of speech recognition [M]. 1993.

[12] R. O. Duda, P. E. Hart, and David G. Stork, Pattern classification [M]. Wiley, 2000.

[13] X. X. Chen, A. J. Li, S. G. Hua, An application of SAMPA-C for standard Chinese [C]// *Sixth International Conference on Spoken Language Processing*, 2000.

[14] A. J. Li, Chinese prosody and prosodic labeling of spontaneous speech [C]// *Speech Prosody*, 2002.

[15] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co- adaptation of feature detectors," arXiv preprint arXiv: 1207. 0580, 2012

林举（1990—），男，硕士，主要研究方向为计算机辅助发音教学。
Email: linjucs@163.com；

解焱陆（1980—），男，博士，副教授，主要研究方向为计算机辅助语言习得、语音信号处理。
Email :xieyanlu@blcu.edu.cn；

张微（1993—），女，硕士，主要研究领域为计算机辅助发音教学。
Email:wei_zhang_mail@126.com；

张劲松（1968—），通讯作者，男，博士、教授，主要研究方向为语音习得、韵律建模、语音识别、实验语音学、计算机辅助发音教学。
Email: Jinsong.zhang@blcu.edu.cn。