Computer Science and Application 计算机科学与应用, 2021, 11(4), 1081-1088
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114111

Hans汉斯

基于韵母结构的LSTM汉语韵律边界识别

魏新享，吴怡之，高文明

东华大学，信息科学与技术学院，上海
Email: wxx.email@qq.com, yz_wu@dhu.edu.cn, gaowm219@163.com

收稿日期：2021年3月26日；录用日期：2021年4月21日；发布日期：2021年4月28日

摘　要

随着语音合成的普及，人们对合成语音的自然度以及准确度的要求日益提高，而韵律边界就对这两个指标起着重要作用。在人们交流中，语句间停顿的部分即为韵律边界。如何提升韵律边界的识别率，仍是当前学术界的重要研究内容。本文在当今已有研究理论的基础上，提出了基于韵母结构的LSTM汉语韵律边界识别方法。该方法首先对语料库进行特征提取，然后利用韵母结构特征对韵母时长进行归一化，最后利用所得特征数据集对LSTM模型进行训练以得到具有较高识别率的韵律边界识别模型。结果表明，将韵母时长更换为归一化时长的模型其识别率高于更换前的模型，其中韵律短语的F值提升了4.9%，其他韵律边界的识别率也得到了一定的改善，韵律边界识别F-Score平均值相对提高了2%，这代表着韵母结构特征对提高模型识别率的有效性。

关键词

汉语短语，韵母时长，循环神经网络，韵律边界

LSTM Recognition Model of Chinese Prosody Boundary Based on Vowel Structure

Xinxiang Wei, Yizhi Wu, Wenming Gao

College of Information Science and Technology, Donghua University, Shanghai
Email: wxx.email@qq.com, yz_wu@dhu.edu.cn, gaowm219@163.com

Received: Mar. 26th, 2021; accepted: Apr. 21st, 2021; published: Apr. 28th, 2021

Abstract

With the popularity of speech synthesis, people's requirements for the naturalness and accuracy

文章引用: 魏新享, 吴怡之, 高文明. 基于韵母结构的 LSTM 汉语韵律边界识别[J]. 计算机科学与应用, 2021, 11(4): 1081-1088. DOI: 10.12677/csa.2021.114111

of synthesized speech are increasing, and the prosody boundary plays an important role in these two indicators. In people's communication, the part of pause between sentences is the boundary of prosody. How to improve the recognition rate of prosodic boundaries is still an important research content in the current academic circles. Based on the existing research theories, this paper proposes an LSTM Chinese prosody boundary recognition method based on the vowel structure. This method first extracts features from the corpus, then uses the structural features of the finals to normalize the duration of the finals, and finally uses the resulting feature data set to train the LSTM model to obtain a prosody boundary recognition model with a higher recognition rate. The results show that the recognition rate of the model that replaces the vowel duration with the normalized duration is higher than that of the model before the replacement. The F value of prosodic phrases is increased by 4.9%, and the recognition rate of other prosodic boundaries has also been improved. The average value of the recognition F-Score is relatively increased by 2%, which represents the effectiveness of the vowel structure characteristics in improving the model recognition rate.

Keywords

Chinese Phrase, Vowel Duration, Cyclic Neural Network, Prosodic Boundary

1. 引言

人们在日常生活中会利用很多方式进行信息的获取以及交流，语言交流就是其中最直接、有效的一个方式。语言所携带的信息分为字面的文字信息和韵律信息，韵律信息在声学上反映出来是语音中除音质之外的音高、音强、音长及语句停顿方面的变化，在人们主观听觉上反映出来的是抑扬顿挫的特征[1]。其中，在语句间停顿的部分即是韵律边界，目前比较主流的韵律结构划分方法是将韵律边界划分为 4 个级别，分别为 B0、B1、B2 和 B3，其中 B0 为韵律词边界，B1 为韵律短语边界、B2 为语调短语边界，B3 为句末边界。本文也采用这种划分方法，以“亚硝酸盐味微咸，易溶于水。”这句话为例，对其进行韵律结构划分如图 1。其中 PW (Prosody Word)代表韵律词对应 B0，PP (Prosody Phrase)代表韵律短语对应 B1，IP (Intonation Phrase)代表语调短语对应 B2，U (Utterance)代表话段对应 B3。正确的划分语音段的韵律结构可以显著的提升机器的语音识别率以及合成语音的自然度等。近年来，随着人机交互、自然语音理解、计算机辅助学习等需求的急剧上升，韵律边界的自动识别作为其中必不可少的一个环节自然也引起了许多学者的重视。

在本文中，韵律边界检测模型主要分为三部分，分别是语音信号预处理模块、特征提取模块以及边界检测模块。其中，模型识别准确率的一个重要制约因素就是特征的选择，许多学者对此开展了大量的研究工作。对于英语，有的研究者提取声学和词汇特征，包括语音部分(POS)标记，能量峰值以及基频(F0)的轮廓和斜率进行模型构建以检测音高和短语边界。有的研究者则使用单词持续时间，单词后的静音，强度和音高轮廓特征进行建模，并得出结论，静音是最有效的功能[2]。也有研究者从纯文本特征着手对短语中断的发生和类型进行建模。对于汉语，有的学者利用时长、基频及能量特征结合调核特征构建韵律边界识别模型，在 SY-CART 模型下取得了 78.3%的正确率[3]。但是，这些关于提取特征的研究是在音节级别上的，并且忽略了音节的内部结构，导致韵律短语边界的识别率较低。

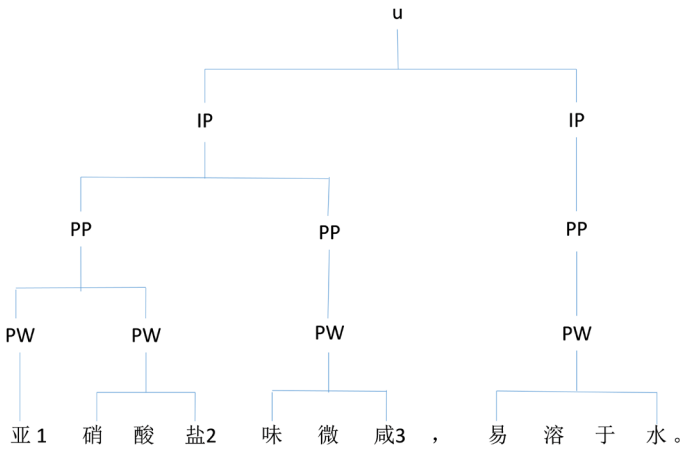

Figure 1. Examples of Chinese prosody boundary types
图 1. 汉语韵律边界类型示例

在学术界，对于话语的韵律结构和韵律单元内音节的持续时长之间的关系，早已有很多学者做出了大量研究，特别是在英语方面。Wightman 等提出了人们在发音时，通常会对短语尾部的音节时长进行延长，其研究内容聚焦在短语发音末尾处的时长变化[4]。通过类型学，Beckman 在当前研究基础上，指出对高层韵律单元末尾的时长进行延长的现象在世界上各地语言中都普遍存在，并推断该现象是所有语言共有的特点，与之同时提出的是，对于不同类型的语言，韵律单元其实对时长的影响也是不同的，也就是说该影响是随语言类型的改变而发生变化的[5]。

在汉语中，音节时长与韵律边界也是具有着很大的关联。汉语属于规范的单音节语言，其一个音节是由声母和韵母构成的[6]。已有研究表明，韵律边界和韵母时长的关系很大，与声母时长没有太大关联。所以在现在的韵律边界研究中，普遍会将韵母时长作为一个重要特征[7]。然而不同的韵母结构也会导致韵母的时长有所变化，这就导致直接利用韵母时长作为特征得到的模型其识别率仍不是很理想。故本文考虑从韵母结构出发，获取韵母结构特征，然后利用韵母结构特征对韵母时长进行归一化，从而平衡韵母结构不同对韵母时长带来的影响，最后将这些特征与统计模型相结合对输入语流进行韵律事件的识别。

2. 韵母结构模型

2.1. 韵母结构与韵母时长

汉语属于规范的单音节语言，其大多数词素由单个音节组成。在汉语发音中不存在初始或最终辅音簇。一个音节按照其词素形状可以划分为 CV，CVN，CVG，CGV，CGVG，及 CGVN，在这里 C 代表辅音，G 代表滑音/j，w，ɥ/，N 代表/n，ŋ/，V 代表/i，y，u，ə，a/。在汉语中，一个音节通常由声母和韵母构成，声母位于字始，且由辅音 C 充当。韵母位于声母 C 之后，通常由韵头、韵腹及韵尾组成，汉语中的韵母共计 39 个，其按词素形状可划分为：V，VN，VG，GV，GVG，GVN [8]。韵律边界识别模型通常将其持续时长即韵母时长作为一个重要特征。为了更清楚的分析韵律边界、韵母结构及韵母时长之间的关系，本文利用开放中文语料库，进行了相关统计和分析。边界与韵母时长数据分布如表 1 与图 2，根据柱状图可以发现韵母时长随着边界的延长，会有一定程度的延长，尤其是在 B0 到 B1 中间出现了比较明显的延长，这说明韵母时长对于 B0 和 B1 的区分可以提供重要线索。汉字的持续时间与韵母结构的关系如图 3，由图可以发现，对于韵母的持续时间，V 结构韵母时间最短，其时长比其他韵母结构的平均时长短了 37.2 毫秒。GVG 和 GVN 结构韵母时间最长，VN、VG 及 GV 结构的韵母时长居中，这

表明韵母结构的长度即韵头、韵腹及韵尾的齐全度会比较明显的影响到韵母时长。

Table 1. The distribution of vowel duration at each boundary
表 1. 各边界的韵母时长分布情况

韵律边界	最小时长(秒)	最大时长(秒)	均值(秒)	方差(秒)	样本数
B0	0.026693	0.32267	0.128164	0.001527	34758
B1	0.053571	0.418968	0.204478	0.001888	12357
B2	0.083599	0.520997	0.21877	0.001809	8611
B3	0.082275	0.417873	0.23635	0.00219	8733

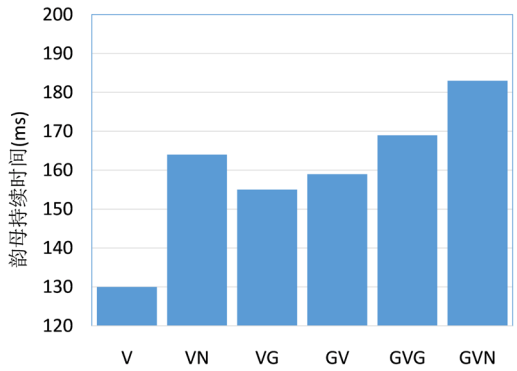

Figure 2. Relationship between prosodic boundary and vowel duration
图 2. 韵律边界与韵母时长关系图

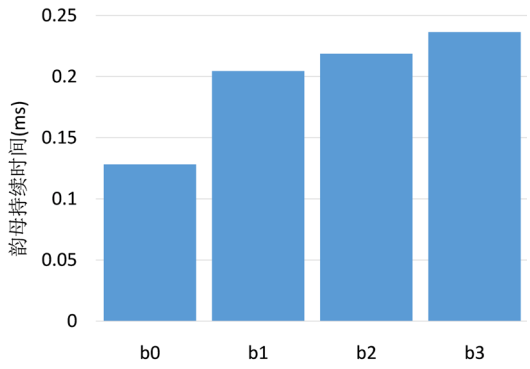

Figure 3. Relationship between prosody boundary and vowel duration
图 3. 韵律边界与韵母时长关系图

2.2. 基于韵母结构的时长归一化模型

在大多数的韵律边界识别研究中，韵母时长通常会被用来作为一个重要的声学特征。但是韵母时长不仅受韵律边界的影响，其还会受到韵母结构类型不同所带来的影响[9]。故韵母结构结合韵母持续时长可以为韵律边界识别提供重要线索。

一般的，韵母时长除了与韵母结构和韵律边界相关外，其还会受到语速的意向，所以直接进行韵母时长的绝对值比较是没有意义的，本文基于韵母结构提出如公式(1)的韵母时长归一化模型，

$$\overline{d(i)}=\frac{d(i)-\mu}{\sigma} \tag{1}$$

由实验可知道，不同的韵母结构会导致韵母时长的变化，为了排除这种影响本文引入一个比例因子α以平衡这些因素对均值μ和标准差σ的影响，具体如公式 2 和公式 3：

$$\hat{\mu}=\alpha\mu \tag{2}$$

$$\hat{\sigma}=\alpha\sigma \tag{3}$$

比例因子α 可由公式 4 算出：

$$\alpha=\frac{1}{M}\sum_{i=1}^{m}\frac{d(i)}{\mu} \tag{4}$$

由统计分析可知，韵母结构总体上分为 V，VN，VG，GV，GVG，GVN 六种类型，为了排除韵母结构的影响，对于$k\in\{\mathrm{V},\mathrm{VN},\mathrm{VG},\mathrm{GV},\mathrm{GVG},\mathrm{GVN}\}$，首先根据式 4 分别计算出各自的比例因子$\alpha_k$，然后由式 5 计算出最终的韵母时长$\overline{d_k(i)}$。

$$\overline{d_k(i)}=\frac{d_k(i)-\hat{\mu}_k}{\hat{\sigma}_k} \tag{5}$$

其中，$\hat{\mu}_k$和$\hat{\sigma}_k$由第 k 类韵母结构的比例因子α_k代入式得到。

3. 基于韵母结构和 LSTM 的韵律边界识别

3.1. 基于 LSTM 的韵律边界识别

韵律边界识别的过程是，首先对输入语音进行数据预处理以及特征的提取得到数据集(Y，X)。Y 是标签序列，X 是特征集，对于 X，其任意第 j 个样本 x(j)都是一个 D 维的向量，该向量包括文本特征，基频特征，相关时长特征，韵母结构特征，声调以及韵母时长特征。得到数据集(Y，X)后将其输入到 LSTM 网络模型中去训练，整体过程见图 4，最终得到一个能够对输入语音序列进行边界识别的网络模型。

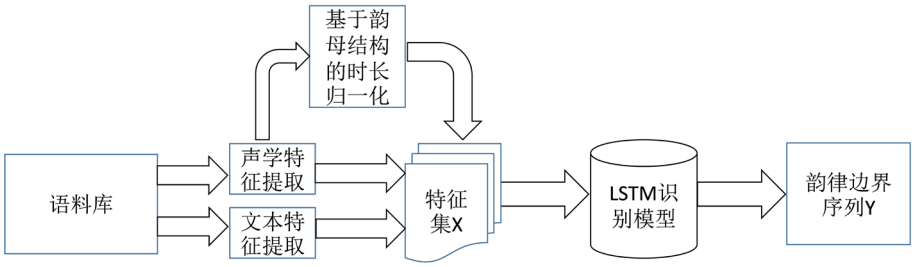

Figure 4. Identify the model flow
图4. 识别模型流程图

3.2. 特征选取

在语音韵律识别建模中，特征的选取是非常关键的一步，其选取的结果将直接影响到整个实验结果。本文所选取的特征分为文本特征和声学特征，其总体特征集关系如图 5 所示。文本特征分为词性特征和词长特征，声学特征分为基频、声母时长、静音时长、声调及基于韵母分类的归一化韵母时长特征。

3.3. LSTM 模型

有了数据集后，就需要选定合适的模型进行建模，本文选取的是 LSTM 模型。LSTM (Long Short-Term

Memory)是一种特殊的卷积神经网络(RNN)，能够学习长期的规律，其网络结构如图 6。它是由 Hochreiter 与 Schmidhuber 在 1997 年首先提出的，并且在后来的工作中被许多人精炼和推广[10]。它在序列建模领域应用得非常好，现在已被广泛的使用。LSTM 最大的优点就是于权重值更新其不存在梯度消失的缺陷。

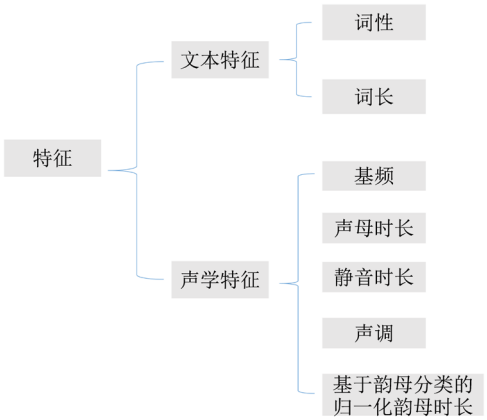

Figure 5. Feature set relationship diagram
图 5. 特征集关系图

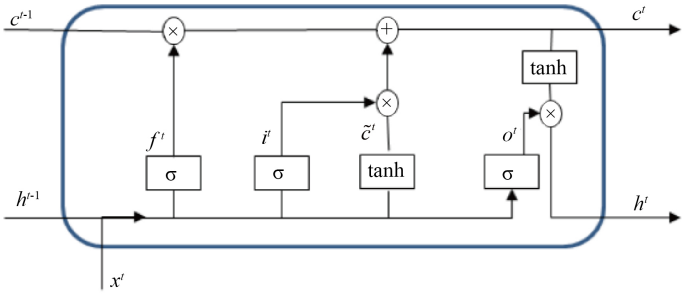

Figure 6. Long short-term memory network structure
图 6. 长短期记忆网络结构

LSTM 具有三个门，分别是输入门、输出门、遗忘门，LSTM 正是利用了这种门机制解决了依耐性缺陷。遗忘门通过权重矩阵 W_f 和 U_f 进行计算[11]。

除此之外，为了了解需要加入多少新信息到当前记忆单元中，我们需要让输入的 x^t 和 h^{t-1} 决定新状态信息 $\tilde{c}^t$ 的更新程度，

$$i^t = \sigma\left(W_i x^t + U_i h^{t-1} + b_i\right) \tag{6}$$

$$\tilde{c}^t = \tan h\left(W_{\tilde{c}} x^t + U_{\tilde{c}} h^{t-1} + b_{\tilde{c}}\right) \tag{7}$$

其中，W_i 和 U_i 是输入门计算的权重矩阵，$W_{\tilde{c}}$ 和 $U_{\tilde{c}}$ 是记忆单元状态计算的权重矩阵，b_i 和 $b_{\tilde{c}}$ 是偏置向量。然后把 c^{t-1} 和 $\tilde{c}^t$ 两部分信息筛选合并到一起，即得到当前记忆层的状态信息 c^t，

$$c^t = \left(f^t \times c^{t-1} + i^t \times \tilde{c}^t\right) \tag{8}$$

接下来根据当前状态信息 c^t 求解出隐藏层输出 h^t，这也需要进行筛选输出，即通过输出门参数 o^t 控制，

$$o^t = \sigma\left(W_o x^t + U_o h^{t-1} + b_o\right) \tag{9}$$

W_o 和 U_o 是输出门计算的权重矩阵，b_o 是偏置向量。

由以上步骤即求解出隐藏层输出 h^t，进而可利用传统的RNN的方法进一步改善参数。

3.4. 评价指标

在本次实验中，考虑到汉语韵律边界评测模型是一个 5 分类的模型，本次模型的评估选用的是多分类 F1 方法。本文通过下面四个指标来将模型评价体系量化确定下来，它们分别是：准确率(Accuracy)、精确率(Precision)、召回率(Recall)以及 F-score。定义如下：

$$\text{准确率} = \frac{\text{预测正确样本数}}{\text{预测样本数}} \times 100\% \tag{10}$$

$$\text{精确率} = \frac{\text{正确识别为该类韵律边界的样本数}}{\text{预测为该类韵律边界的样本数}} \times 100\% \tag{11}$$

$$\text{召回率} = \frac{\text{正确识别为该类韵律边界的样本数}}{\text{实际为该类韵律边界的样本数}} \times 100\% \tag{12}$$

$$F\text{值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{13}$$

4. 实验结果与分析

4.1. 试验配置

本次研究中采用的数据是中文标准女声音库，在此语料库中语言类型为标准普通话，每句平均字数在 16 字左右。语料库中的所有语料覆盖领域有文学作品、聊天对话、实事新闻即悦乐科技等等，它尽可能的全面的覆盖了类型、音调、音节音子及韵律等。该语料库依据合成语音该标注标准对音库进行了韵律层级标注、文本音字校对、语音的边界切分等标注[12]。本次实验环境：1. win10。2. python3.6.4。3. pip20.0.2。4. tensorflow1.14.0。在数据集上，本文将 10000 句实验语料按照一定的比例分为 3 组，训练集占比 70%，验证集和测试集分别占 15%。由于这些数据集得服从同分布的条件，本实验在划分它们时采取均匀随机抽样的方式以达到这个三类数据不存在交集的情况。本次实验采用的网络模型是 LSTM 模型，其超参数配置如表 2。

Table 2. Hyperparameter configuration
表 2. 超参数配置

网络模型	num_layer	num_hidden	batch_size	time_step	dropout	learningrate
LSTM	1	64	128	3	0.5	0.001

4.2. 实验结果分析

本文选取词性、词长、基频、声母时长、静音时长、声调作为基础特征模板(Bft, basic features template)，采用 LSTM 网络模型作为建模模型，进行韵律边界识别模型的构建。为了分析归一化韵母时长特征的有效性，本实验采用特征组合的方式进行实验，实验结果如表 3 所示。横向对比来看，两组实验的韵律短语边界识别率都低于其他韵律边界，这是由于语料库的韵律短语边界数据不够充分导致的。纵向对比来看，将普通的韵母归一化时长更换成基于韵母分类的韵母归一化时长后，韵律短语的识别率得到了一定4.1%的提升，其他边界的识别率也得到了一定的改善。整体来看，韵母时长第一类组合平均 F 值为 71.75，第二类组合为 73.375，平均 F 值提高了接近 2%，这表明基于韵母结构的韵母时长特征可以在一定程度上

改善韵律边界检测的性能。

Table 3. Phrase boundary detection results under different feature combinations
表 3. 不同特征组合下短语边界检测结果

特征组合	All	韵律词	韵律短语	语调短语	句末
	Acc (%)	F-score (P, R) (%)	F-score (P, R) (%)	F-score (P, R) (%)	F-score (P, R) (%)
Bft + 归一化时长	78.3	68.7 (66.2, 71.4)	47.1 (53.6, 42.0)	88.2 (80.8, 98.2)	83.0 (91.5, 76.0)
Bft + 基于韵母分类的归一化时长	79.0	69.3 (67.5, 71.4)	51.2 (54.2, 48.5)	88.9 (81.4, 98.0)	84.1 (91.8, 77.6)

5. 结论

本文提出了基于 LSTM 网络模型使用基于韵母分类的时长归一化声学特征进行汉语韵律边界识别方法，并通过实验验证了该特征对提高韵律边界识别率的有效性。在以后的工作中，将进一步研究不同的网络模型对韵律边界识别率的影响，以进一步提高边界识别模型的识别率。

参考文献

[1] 王洪君. 汉语的韵律词与韵律短语[J]. 中国语文, 2000(6): 525-536+575.

[2] Soto, V., Cooper, E., Rosenberg, A., *et al*. (2013) Cross-Language Phrase Boundary Detection. 2013 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, 26-31 May 2013, 8460-8464. https://doi.org/10.1109/ICASSP.2013.6639316

[3] 林举, 解焱陆, 张劲松, 张微. 基于声调核参数及 DNN 建模的韵律边界检测研究[J]. 中文信息学报, 2016, 30(6): 35-39+48.

[4] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., *et al*. (1992) Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. *The Journal of the Acoustical Society of America*, **91**, 1707-1717. https://doi.org/10.1121/1.402450

[5] Beckman, M.E. and Pierrehumbert, J.B. (1986) Intonational Structure in Japanese and English. *Phonology*, **3**, 255-309. https://doi.org/10.1017/S095267570000066X

[6] 梅晓, 熊子瑜. 普通话韵律结构对声韵母时长影响的分析[J]. 中文信息学报, 2010, 24(4): 96-103.

[7] 曹剑芬. 音段延长的不同类型及其韵律价值[J]. 南京师范大学文学院学报, 2005(4): 160-167.

[8] Wu, F. and Kenstowicz, M. (2015) Duration Reflexes of Syllable Structure in Mandarin. *Lingua*, **164**, 87-99. https://doi.org/10.1016/j.lingua.2015.06.010

[9] 王孟杰, 孟子厚. 基于语音参数的普通话韵母区别特征[J]. 声学技术, 2011, 30(1): 88-92.

[10] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Sak, H., Senior, A. and Beaufays, F. (2014) Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. *Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, 14-18 September 2014, 338-342.

[12] 标贝科技. 中文合成语音数据库[EB]. https://www.data-baker.com/open_source.html, 2021-03-19.