基于 Deep Speech 与多层 LSTM 的儿童朗读语音评价模型

郑纯军[1,2] 贾 宁[2]

(大连海事大学 辽宁 大连 116023)[1] (大连东软信息学院 辽宁 大连 116023)[2]

摘 要 现代人大多忽略了朗读的重要性,然而对于 5～12 岁的儿童,朗读不仅是学习过程中必备的技能,还是陶冶情操的有效手段。由于朗读语音信号的特征与评价标准之间存在着非线性关系,递归神经网络虽然适用于时间序列的预测,但是对长时间跨度的预测效果有限。基于此,根据儿童朗读语音特点及其评价体系,设计了一种基于 Deep-Speech 与三层长短期记忆(Long Short-Term Memory,LSTM)神经网络相结合的模型。首先,在添加注意力机制的基础上,提出朗读语音评价的准确性和流利性度量,以频谱图作为特征提取的输入,其中,朗读评价的准确性采用改进后的 Deep Speech 以提高音素识别的准确率,流利性评价将频谱图送至三层 LSTM 模型中以呈现时间序列的影响;然后,将结果送入注意力机制进行权重调节;最终,将计算的总评价结果用于儿童朗读语音的评分。使用"出口成章"软件提供的儿童朗读语料库和 TensorFlow 平台进行实验。结果表明,与传统的模型相比,此模型不仅可以精确判断朗读的正确性和朗读的流利性,而且其评价模型获得的评分结果较准确。

关键词 频谱图,长短期记忆网络,注意力机制,DeepSpeech,朗读语音评价模型

中图法分类号 TP183 **文献标识码** A

Children's Reading Speech Evaluation Model Based on Deep Speech and Multi-layer LSTM

ZHENG Chun-jun[1,2] JIA Ning[2]

(Dalian Maritime University, Dalian, Liaoning 116023, China)[1]

(Dalian Neusoft University of Information, Dalian, Liaoning 116023, China)[2]

Abstract Most modern people ignore the importance of reading. However, for children aged 5～12, reading aloud is not only an essential skill in the learning process, but also an effective means of cultivating sentiment. Since there is a non-linear relationship between the characteristics of the spoken speech signal and the evaluation criteria, the recurrent neural network is suitable for time series prediction, but its prediction effect is limited for long-term span. According to the characteristics of children's spoken speech and its evaluation system, a new model combining Deep Speech and three-layer LSTM (Long Short-Term Memory) neural network was designed. Firstly, on the basis of adding attention mechanism, the accuracy and fluency measure of speech evaluation are put forward, and the spectrum map is used as the input of feature extraction. Among them, the accuracy of reading uses the new version of Deep Speech to improve the accuracy of phoneme recognition. For fluency evaluation, the spectrogram is sent to the three-layer LSTM model to present the effects of the time series. Then, the results are sent to the attention mechanism for weight adjustment, and finally the total evaluation results are used for the evaluation of children's spoken speech. The experiment uses the children's reading corpus, which is provided by the "export chapter" software, and the experimental environment uses the TensorFlow platform. The experimental results show that compared with the traditional model, this model can accurately judge the correctness of spoken speech and the fluency of reading aloud, and the scoring results obtained by its evaluation model are more accurate.

Keywords Spectrogram, Long Short-Term Memory, Attention mechanism, DeepSpeech, Evaluation of spoken speech models

1 引言

人类语音包含不同类型的信息。鉴于声道结构的多样性,每个人的语音数据具有该说话者的特定信息。对于人类来说,从复杂语音数据中提取说话者,特定信息是微不足道的,但对于计算机来说,则是一项具有挑战性的任务。随着人工智能的发展,深度学习于 2009 年被引入语音识别领域,短短几年时间已被广泛应用于语音识别、说话者识别、文字识别、情感识别等相关领域。针对差异性及其表达形式,学者在领域内均取得了显著的成果,但鲜有学者针对未成年人的朗读语音进行深入研究,本文专注于儿童语音在朗读场景下的研究。

本文涉及的儿童年龄均为 5～12 岁,大部分儿童处于小学的学习阶段,而朗读是每名儿童必备的一项学习技能。小

本文受辽宁省自然科学基金项目(20180551068)资助。

郑纯军(1976－),男,副教授,CCF 会员,主要研究方向为大数据分析、语音情感分析,E-mail:zhengchunjun@neusoft.edu.cn(通信作者)。

学语文大纲指出：阅读是小学语文教学的基本环节。在这一个环节中，朗读是最重要、最经常的训练[1]。由此可以看出，在小学语文教学中朗读的重要性，由于学生上语言课次数与时间的限制，教师很难对每个学生的朗读效果做出系统的训练与评价。

通过引入深度学习技术，设计混合多种传统模型、注意力机制及语音识别技术，使用机器自动对朗读语音进行评价，为学生的课后朗读训练提供了一个良好的思路。

本文使用“出口成章”软件提供的儿童朗读语音语料库，提取大量朗读数据固定的若干特征，建立适用于朗读语音评价的模型，并设计朗读效果评判标准，最终通过实验验证了此模型的有效性与规范性。

2　相关工作

现有的科学研究主要集中在语音识别、说话者识别、语音情感分析和自然语言处理等领域，针对儿童朗读语音这一空白领域，本文工作的基本前提是从朗读语音中提取有评判价值的若干个特定特征，在此基础上，适当的选择有针对性的模型。

近年来，针对语音信号的分析主要有 3 种方法：1）从原始音频文件中提取信号特征[2]，捕获最原始的声学特征；2）直接在原始音频波形上运行深度学习模型[3]；3）利用自动语音识别（Automatic Speech Recognition，ASR）技术将语音转换为文本，然后采用传统的基于文本的分析系统[4]。目前，以上 3 种方法处于平行探索中，鲜有交叉研究领域存在。

2.1　原始声学特征提取技术

第一种语音分析方法需要将语音信号转换为计算机能够处理的语音特征向量，即低级描述符（Low-Level Descriptor，LLD）；大致可归纳为 3 种类型：韵律特征、谱特征和音质特征[5]。韵律特征与句法、语篇、信息结构等密切相关，主要包括音高、共振峰、时长、基频、能量等；谱特征是原始信号在声道中激发后而产生的测量特征，目前，常见的提取方法有：线性预测系数（Linear Prediction Coefficient，LPC）、梅尔频率倒谱系数（Mel-frequency Cepstral Coefficients，MFCC）[6] 和线性预测倒谱系数（Linear Prediction Cepstrum Coefficient，LPCC）等[7]；音质特征主要包括呼吸声音、喉化音、音素、词边界、明亮度等[8]。

5～12 岁儿童的言语声学和语言特征与成人不同[9]。例如，儿童的语音特征是音高较高[10]，共振峰发生在较高频率[11]。Saeid 等[12]通过实验证明，带宽对儿童言语的朗读语音评价有很大的影响力。

2.2　频谱图

第二种语音分析方法的核心思想是保留语音信号的完整特征，只将其转换为原始音频波形图，即语音频谱图（频谱图），该图由处理接收到的时域信号后衍生，相关的频谱图有振幅图、能量图、对数（log）能量图等。

Li 等[13]采用频谱图，将其与卷积神经网络（Convolutional Neural Networks，CNN）和注意力机制融合，用两个不同的卷积核分别提取时域特征和频域特征，将频谱图保存成图像直接进行归一化处理，其语音情感评价准确度较高。

Badshah 等[14]采用频谱图作为输入，其大量工作集中在切割语音的预处理环节，针对不同的切割时长、频率分辨率及模型，通过实验评估了频谱图对噪声的免疫能力。

2.3　递归神经网络模型与注意力机制

由于单个语音信号中提取的信息必须依赖于其上下文的内容，因此，一般将语音特征馈入递归神经网络（Recurrent Neural Networks，RNN）等深度学习模型中。例如，Etienne 等[15]从原始光谱图中提取高级特征，融合 CNN 和 LSTM 架构，设计了一个用于识别语音情感的神经网络，并使用了 IEMOCAP 数据集验证其有效性。Nicholas 等[16]融合了声谱图和一个三层 LSTM，在对比分析数据是否去噪的基础上，判断模型对于噪声数据的鲁棒性。Kang 等[17]利用门控循环单元（Gated Recurrent Unit，GRU）对语音情感进行识别，在加入噪声的基础上达到与 LSTM 相当的结果，但其可以应用在嵌入式设备上。

根据现有场景中语音信号的复杂程度，可以使用注意力模型对有效特征进行分析与提取，达到最佳的偏重效果。例如，Huang 等[18]将 CNN 与注意力机制相结合，实现了对音乐中富有情绪化内容的高亮。Mirsamadi 等[19]将本地注意力机制加入 RNN 中，通过集中提取与情感相关的短时帧级声学特征，以实现对说话者情绪的自动识别。

2.4　ASR 技术

第三种语音分析方法将语音转换为文本，通过识别音频中说话者所说的每个单词，将它们改变为单词嵌入，其中使用了自然语言处理中的一些技术，如词频-逆向文件频率（TF-IDF）和词袋（BOW）模型[20]。结果并不总是准确的，因为情绪检测的准确性取决于能否在口语中可靠地检测准确的发音[21]。此外，当语音被转换到文本时，一些情绪相关的信号特征也丢失，导致情绪分类的准确性降低。目前，在汉语识别领域，国内很多公司对外开放了语音识别的接口，例如科大讯飞、云知声、百度等，但少有公司将语音识别引擎开源，Deep-Speech 是百度开发的开源实现库，它使用复杂和前沿的机器学习技术创建了语音到文本的引擎[22]，方便开发人员对特定领域的子任务进行迁移。

2.5　朗读语音评价标准

本文主要从准确性、流利性两个维度来建立朗读语音评价标准。准确性主要从朗读者发音的音素相似性进行衡量，流利性主要从语音表达的流畅、连贯性等方面进行评价。文中准确性设定比重范围为 0.6～1，流利性比重范围为 0～0.4。根据此评价体系，建立朗读语音评价度量方法，假定准确性分值为 A，流利性分值为 F，朗读语音总分为 R，W_1 和 W_2 分别是准确性和流利性的系数，则有公式如下：

$$R=A*W_1+F*W_2 \tag{1}$$

其中，$\sum_{i=1,2} W_i=1$，假设 W_1 与 A 线性相关，且 W_1 的取值区间为[0.6，1]，则存在式(2)的关系：

$$R=-\lambda A^2+\lambda A\times F+A \tag{2}$$

其中，$\lambda=1-W_1$。

3　Deep Speech＋LSTM 的模型

文中设计了一种基于 Deep Speech 与 LSTM 相结合的朗读语音评价模型。该模型将朗读语音波形文件转换成频谱图作为模型的统一输入。

当 $\lambda=0.4$ 时，网络模型的整体框架图如图 1 所示。

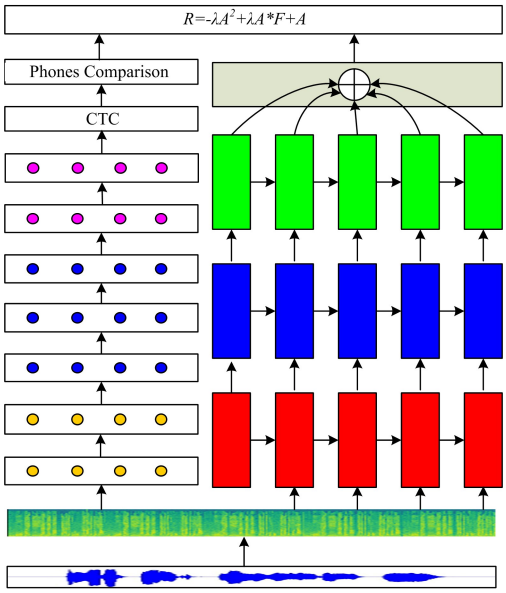

图 1 Deep Speech+LSTM 神经网络框架图

Deep Speech 分支模型首先用于将图谱输入到卷积神经网络，依次经过 GRU 网络层、多层的全卷积网络，然后进入 CTC(Connectionist Temporal Classification)，得到音素序列，与训练样本进行相似度比较，得到 0～1 之间相似度的值。

流利性评价模型是将图谱输入三层 LSTM 网络，再进入注意力机制(Attention)层，得到流利性评价结果。最后将准确性与流利性结合，得出最后的评价结果。

3.1 Deep Speech 模型

Deep Speech 模型为百度开源的基于 PaddlePaddle 的 Deepspeech2 神经网络模型，其功能包括特征提取、数据增强、模型训练、语言模型、解码模块等，强大且简单易用。它基于 LSTM-CTC (Connectionist Temporal Classification) 的端对端语音识别技术，将机器学习领域的 LSTM 建模与 CTC 训练引入传统的语音识别框架里。

3.1.1 频谱图输入

频谱图的横坐标是时间，纵坐标是频率，坐标点值为语音数据能量。由于是采用二维平面来表达三维信息，所以能量值的大小是通过颜色的深浅来表示，颜色深，表示该点的语音能量越强。

本文将格式为. wav 的朗读语音文件，以 20 ms/帧，通过汉明窗口完成加窗操作，每一帧通过使用快速傅里叶变换来计算各个频率的能量值，以 10 ms 为步长进行滑动，分别产生每帧的频谱图；最后按照时间顺序，将所有的频谱图拼接在一起，对应为这段朗读语音频谱图，作为本文两个神经网络模型的输入。频谱图的过程如图 2 所示。

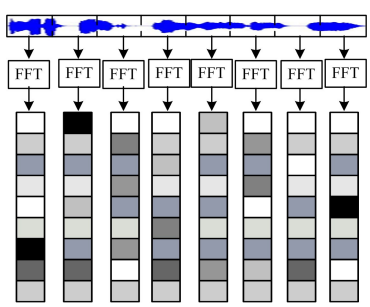

图 2 语音频谱生成过程图

3.1.2 GRU

GRU 是的一种加强版的 LSTM 网络，相比于 LSTM 网络，其结构更简单，是当前的一种流行网络。由于 GRU 是 LSTM 的变体，因此也可以解决 RNN 网络中的长依赖问题。与 LSTM 不同的是，GRU 模型中只有两个门：更新门和重置门。GRU 模型结构如图 3 所示。

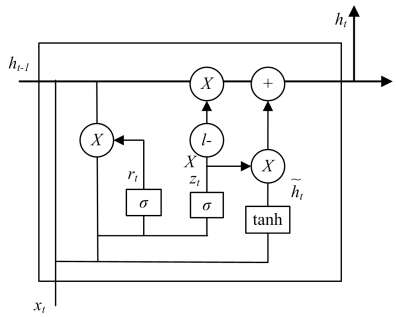

图 3 GRU 模型结构

图中的 zt 和 rt 分别表示更新门和重置门。更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大，说明前一时刻的状态信息被带入越多。重置门控制有多少前一状态信息被写入到当前的候选集 $\tilde{h}_t$ 上，重置门的值越小，前一状态的信息被写入的越少。

3.1.3 CTC

CTC(Connectionist Temporal Classification)算法，可以理解为基于神经网络的时序类分类，通过输入序列 x 得到输出序列 y，如可以获得输出序列的分布 $p(I/x)$，选择其中概率最大的那一个作为输出序列，如式(3)所示。

$$h(x)=\arg\max_{I\in L\leqslant T} p(I|x) \tag{3}$$

许多实际的序列学习任务可能会包含噪声，即没有事先对齐序列化数据。CTC 是计算一种损失值，主要的优点是可以自动对齐没有对齐的数据。

3.1.4 音素相似度判断

本文通过将语音转换成中文的 phones 序列与正确的 phones 序列进行相似度比较，来衡量朗读的准确性。

本文构建了一个相似度比较矩阵，该矩阵由 23 个声母与 24 个韵母构成，因涉及发音相似度的衡量，所以未考虑韵母的声调，设计的相似度比较矩阵 A 如表 1 所列。

表 1 相似度比较矩阵

	b	p	m	f	…	ing	ong
b	0	10	24	100	…	100	100
p	10	0	18	30	…	100	100
m	24	18	0	16	…	100	100
f	100	30	16	0	…	100	100
…	…	…	…		…	…	…
ing	100	100	100	100	…	0	30
ong	100	100	100	100	…	30	0

声韵母之间的相似度用 0～100 之间的数来表示，数字越大相似度越低，朗读者通过模型识别出音素后到混淆矩阵进行查找，识别出的声韵母与样本的声韵母对应位置为 1，其他位置为 0，生成矩阵 B，音素准确度矩阵 $C=A\cdot B$，C 中所有元素之和为音素的准确性值，一段语音所有音素准确性的平均值为整段语音的准确性评价值。

3.2 流利性评价模型

流利性评价模型采用三层 LSTM 网络，频谱图作为输

入，每层 LSTM 的隐藏节点设置为 256，第三层输出连接一个注意力层，获得最后的流利性评价结果。

3.2.1　LSTM

传统的 RNN 神经网络容易产生梯度消失或梯度爆炸的问题，因此在此基础上，Hochreiter 和 Schmidhuber 提出了 LSTM 模型，增加了存储长期有效数据的单元，从而克服了梯度问题，提升了预测的能力。

LSTM 作为更强大的 RNN 神经网络模型，在 LSTM 中引入了长时间信息效性的机制，这些信息有选择性的被控制并保存下来。LSTM 采用的策略是在每个神经元内部增加：输入门、输出门和忘记门。选用误差函数反馈权重，通过忘记门决定记忆单位是否被清除。默认的 LSTM 结构如式(4)所示。

$$\begin{aligned} f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i) \\ \widetilde{C}_t &= \tanh(W_{\widetilde{c}}[h_{t-1}, x_t] + b_{\widetilde{c}}) \\ C_t &= f_t * C_{t-1} + i_t * \widetilde{C}_t \\ o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \tag{4}$$

其中，W_f，W_i，$W_{\widetilde{c}}$，W_o 是权重参数；b_f，b_i，$b_{\widetilde{c}}$，b_o 是偏置；x_t 作为输入序列，结合上一个隐藏层 h_{t-1} 的状态，通过激活函数构成忘记门 f_t。输入门层 i_t 和输出门 o_t 也由 x_t 和 h_{t-1} 计算得到。忘记门 f_t 与前单元状态 C_{t-1} 联合以确定是否丢弃信息。

3.2.2　注意力机制

考虑到人类大脑对事物的感知是一个有选择性的集中注意力的过程，这种注意力机制可以被应用于深度学习领域，注意力可以被描述为用于分配有限信息处理能力的“选择机制”，它有助于快速分析目标数据，配合信息筛选和权重设置机制，提升模型的计算能力。

对于输入 x 的序列中的每个向量 x_i，可以按照式(5)计算注意力权重 α_i：

$$\alpha_i = \frac{\exp(f(x_i))}{\sum_j \exp(f(x_j))} \tag{5}$$

其中，$f(x_i)$ 是评分函数。

注意力层的输出，即 $attentive\_x$，是输入序列的加权之和。如式(6)所示。

$$attentive\_x = \sum_i \alpha_i x_i \tag{6}$$

4　实验验证

4.1　数据集合

本文使用“出口成章”软件的儿童朗读语音语料库中的数据来测试语音朗读评价模型的效果。研究人员招募了 400 名志愿者(5 岁－12 岁，平均年龄 9 岁，男、女各 50%)，每名志愿者在安静的环境下大声朗读指定的文字内容，及时记录他们的多种语音信号。同时，邀请 6 名播音专业的专家在聆听志愿者语音数据后对其流利性进行评分，评分范围为[0,5]的整数，当 70%专家的打分一致时，保留此样本的得分信息，否则要求专家在第二轮重复评分，第三轮无结果则直接丢弃此样本数据。

4.2　模型参数设置

涉及的模型参数主要与 LSTM 和注意力机制有关。其中，模型采用单向三层 LSTM，Batch_size 为 150，最大 Epochs 为 10000，学习速率为 0.001，Dropout 为 0.5。

初始化权值方法为 RandomUniform，神经元激活函数为 Tanh，优化器为 Adam，损失函数为均方误差。

4.3　结果分析

本文使用 Tensorflow 框架进行网络模型结构的搭建，从而对儿童朗读语音进行评价。以原始 Deep Speech 模型＋传统的 LSTM 模型作为基线，对照模型 1：原始 Deep Speech 模型＋双层 LSTM，模型 2：原始 Deep Speech 模型＋三层 LSTM＋注意力机制，模型 3：改进 Deep Speech 模型＋三层 LSTM，模型 4：当前系统(改进 Deep Speech 模型＋三层 LSTM＋注意力机制)，分别计算各模型预测结果与真实打分之间的相对误差，计算公式如式(7)所示：

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - t_i)^2} \tag{7}$$

其中，σ 代表了打分结果与真实结果之间误差的均方根，y_i 为模型打分值，t_i 为真实值。

图 4 为语音 ID 为 1001－1005 各个模型的打分数据。由图 4 可知，模型 4 的打分值与真实值之间的差异相对最小。

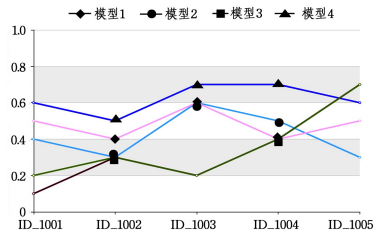

图 4　预测值与专家打分的统计数据

实验验证后，不同的模型预测结果的均方根误差如表 2 所列。

表 2　实验结果

模型	均方差误差/%
基线：原始 Deep Speech 模型＋传统的 LSTM	18.55
原始 Deep Speech 模型＋双层 LSTM	15.52
原始 Deep Speech 模型＋三层 LSTM＋注意力机制	15.3
改进 Deep Speech 模型＋三层 LSTM	12.25
当前系统(改进 Deep Speech 模型＋三层 LSTM＋注意力机制)	11.89

由表 2 可知，在所有模型中本文提出策略的表现最佳，拥有最优的准确率。其次是改进 Deep Speech 模型＋三层 LSTM 模型，采用原始 Deep Speech 模型的准确率均较低，而基线的准确率最差。由此可以确定改进的 Deep Speech 模型对结果产生了较大的影响，同时注意力机制也在一定程度上提升了准确率，此结论与朗读语音评价标准基本吻合。

结束语　针对儿童语音朗读评价问题，基于“出口成章”软件提供的儿童朗读语音语料库，本文设计改进后的 Deep Speech 与 LSTM 神经网络相结合的模型，同时增加了注意力机制对语音识别和频谱图通道分别进行特征提取，利用朗读语音评价模型，形成一套完整的回归问题来解决方案。

通过实验验证，本文提出的模型准确度较高，且均方误差值易收敛。

未来将进一步整合儿童朗读语音信息，获得更权威的数据，并自己设立儿童朗读语音数据库，对外提供相应的接口，同时，进一步优化现有的神经网络模型，以降低模型打分与真实值之间的误差。

(下转第 148 页)

RMB-USD exchange rate forecasting based on ARMA model [J]. Journal of Science of Teachers College & University, 2018 (4).

[7] LUO X. The Research of the Fluctuation Rules of USD/RMB Exchange Rate Series Based on GARCH Model[J]. Application of Statistics & Management, 2009, 28(2): 295-300.

[8] 朱可飞. 基于小波分析的人民币汇率预测方法研究[D]. 杭州: 浙江工商大学, 2014.

[9] SHIN T, HAN I. Optimal Signal Multi-Resolution by Genetic Algorithms to Support Artificial Neural Network Models for Financial Forecasting[C]// International Conference on Information Intelligence and Systems, 1999. IEEE, 2000: 586-593.

[10] CAO D Z, PANG S L, BAI Y H. Forecasting exchange rate using support vector machines[C]// International Conference on Machine Learning and Cybernetics. IEEE, 2005: 3448-34526.

[11] HUANG W, LAI K K, NAKAMORI Y, et al. Forecasting foreign exchange rates with artificial neural networks: a review[J]. International Journal of Information Technology & Decision Making, 2004, 3(1): 145-165.

[12] MAJHI R, PANDA G, SAHOO G. Efficient prediction of exchange rates with low complexity artificial neural network models[J]. Expert Systems with Applications, 2009, 36(1): 181-189.

[13] 王晴, 朱家明. KNN 算法在汇率预测中的应用及改进[J]. 兰州文理学院学报(自然科学版), 2017, 31(3): 27-31.

[14] WUTHRICH B, PERMUNETILLEKE D, LEUNG S, et al Daily prediction of major stock indices from textual www data[J]. HKIE Transaction, 1998, 5: 151-166.

[15] LAVRENKO V, T SCHMILlM, LAWRIE D, et al. Mining of concurrent text and time series [C]// KDD-2000 workshop on Text Mining. 2000: 37-44.

[16] 赵丽丽, 赵茜倩, 杨娟, 等. 财经新闻对中国股价影响的定量分析[J]. 山东大学学报, 2012, 47: 70-75.

[17] 金雪军, 祝宇, 杨晓兰. 网络媒体对股票市场的影响——以东方财富网股吧为例的实证研究[J]. 新闻与传播研究, 2013(12): 36-51.

[18] 佟瑞鹏, 谢贝贝, 安宇. 黑天鹅事件定义及分类的探讨[J]. 中国公共安全(学术版), 2017(2).

[19] 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010, 21(8): 1834-1848.

(上接第 111 页)

参 考 文 献

[1] 孙丽妍. 如何培养小学生的语文朗读能力[J]. 语文建设, 2018, 12: 97.

[2] BERTIN-MAHIEUX T, ELLIS D P W. Large-scale cover song recognition using hashed chroma landmarks[C]// 2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA). NY, USA: IEEE, 2011: 117-120.

[3] OORD A V D, DIELEMAN S, ZEN H G, et al. Wavenet: A generative model for raw audio[C]// arXiv: 1609. 03499. 2016.

[4] EZZAT S, EL GAYAR N, GHANEM M M. Sentiment analysis of call centre audio conversations using text classification[J]. International Journal of Computer Information Systems and Industrial Management Applications, 2012, 4(1): 619-627.

[5] 韩文静, 李海峰, 阮华斌, 等. 语音情感识别研究进展综述[J]. 软件学报, 2014, 25(1): 37-50.

[6] TRABELSI I, AYED D B. On the Use of Different Feature Extraction Methods for Linear and Non Linear kernels[J]. Computer Science, 2014.

[7] PALO H K, MOHANTY M N, CHANDRA M. Computational Vision and Robotics[C]// Advances in Intelligent Systems and Computing. 2015: 63-70.

[8] RODDY C. Emotion recognition in human-computer interaction [J]. Signal Processing Magazine, IEEE, 2001, 18(1): 32-80.

[9] GEROSA M, LEE S, GIULIANI D, et al. Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition[C]// Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2006(ICASSP 2006). 2006: 1393-1396.

[10] YILDIRIM S, NARAYANAN S, BOYD D, et al. Acoustic analysis of preschool children's speech[C]// Proc. 15th ICPhS. 2003: 949-952.

[11] LI, RUSSELL M J. An analysis of the causes of increased error rates in children's speech recognition[C]// Seventh International Conference on Spoken Language Processing. 2002.

[12] AFAVI S, NAJAFIAN M, HANANI A, et al. Speaker recognition for children's speech[J]. arXiv: 1609. 07498, 2016.

[13] LI P C, SONG Y, MCLOUGHLIN I V, et al. An Attention Pooling Based Representation Learning Method for Speech Emotion Recognition[C]// Interspeech. 2018: 3087-3091

[14] BADSHAH A M, AHMAD J, RAHIM N, et al. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network[C]// International Conference on Platform Technology & Service. IEEE, 2017.

[15] ETIENNE C, FIDANZA G, PETROVSKII A, et al. CNN+ LSTM Architecture for Speech Emotion Recognition with Data Augmentation[J]. Computer Science, 2018.

[16] CUMMINS N, AMIRIPARIAN S, HAGERER G. An Image-based Deep Spectrum Feature Representation for the Recognition of Emotional Speech[C]// ACM on Multimedia Conference. 2017.

[17] KANG J, ZHANG W Q, LIU J. Gated recurrent units based hybrid acoustic models for robust speech recognition[C]// 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). IEEE, 2016.

[18] HUANG Y S, CHOU S Y, YANG Y H. Pop Music Highlighter: Marking the Emotion Keypoints[J]. arXiv: 1802. 10495, 2018.

[19] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention[C]// ICASSP. IEEE, 2017.

[20] PASSALIS N, TEFAS A. Neural bag-of-features learning[J]. Pattern Recognition, 2017, 64: 277-294.

[21] KAUSHIK L, SANGWAN A, HANSEN J H. Automatic audio sentiment extraction using keyword spotting[C]// Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[22] AMODEI D, ANUBHAI R, BATTENBERG E, et al. Deep Speech 2: End-to-End Speech Recognition in English and Mandarin[J]. Computer Science, 2015.