基于拼音约束联合学习的汉语语音识别

梁仁凤[1,2]，余正涛[1,2*]，高盛祥[1,2]，黄于欣[1,2]，郭军军[1,2]，许树理[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
{liangrenfeng3，ztyu}@hotmail.com，gaoshengxiang.yn@foxmail.com
{huangyuxin2004，guojjgb}@163.com，xushulitony@outlook.com

摘要

当前的语音识别模型在英语、法语等表音文字中已经取得很好的效果。然而，汉语是一种典型的表意文字，汉字与语音没有直接的对应关系，但拼音作为汉字读音的标注符号，与汉字存在相互转换的内在联系。因此，在汉语语音识别中利用拼音作为解码约束，引入一种更接近语音的归纳偏置。基于多任务学习框架，提出一种基于拼音约束联合学习的汉语语音识别方法，以端到端的汉字语音识别为主任务，以拼音语音识别为辅助任务，通过共享编码器，同时利用汉字与拼音识别结果作为监督信号，增强编码器对汉语语音的表达能力。实验结果表明，相比基线模型，提出方法取得更优的识别效果，词错误率WER降低了2.24个百分点。

关键词： 端到端；汉语语音识别；联合学习；拼音

Chinese Speech Recognition Based on Pinyin Constraint Joint Learning

Renfeng Liang, Zhengtao Yu, Shengxiang Gao, Yuxin Huang, Junjun Guo, Shuli Xu
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China
{liangrenfeng3, ztyu}@hotmail.com, gaoshengxiang.yn@foxmail.com
{huangyuxin2004, guojjgb}@163.com, xushulitony@outlook.com

Abstract

Current speech recognition models have achieved good results in phonetic language such as English and French. However, Chinese is a typical ideographic writing, and there is no direct correspondence between Chinese characters and phonetics, but Pinyin, as a mark of the pronunciation of Chinese characters, has an internal connection with Chinese characters. Therefore, Pinyin is used as a decoding constraint in Chinese speech recognition, and an inductive bias closer to speech is introduced. Based on a multi-task learning framework, a Chinese speech recognition method based on pinyin constraint joint learning is proposed. The end-to-end Chinese character speech recognition is the main task, and the pinyin speech recognition is the auxiliary task. By sharing the encoder, the Chinese characters and the pinyin recognition are used at the same time. The result is used as a supervision signal to enhance the ability of the encoder to express Chinese speech. Experimental results show that compared with the baseline model, the proposed method achieves better recognition results, and the word error rate WER is reduced by 2.24 percentage points

*通信作者：余正涛 ztyu@hotmail.com
项目基金：国家自然科学基金（61732005，61761026）;云南省高新技术产业专项（201606）

Keywords: End-to-end , Chinese speech recognition , Joint learning , Pinyin

1 引言

自动语音识别（Automatic Speech Recognition，ASR）是把语音中包含的词汇内容转换为计算机可理解的文本。随着深度学习的快速发展，ASR系统主要分为两类：传统混合系统和当前主流的端到端模型。传统混合系统 (Sainath et al., 2013)基于深度神经网络隐马尔可夫模型（Deep Neural Networks - Hidden Markov Models ，DNN-HMM）对声学模型建模、使用发音字典将音素序列转换为词、再通过一个语言模型将词序列映射为句子，系统训练时，这些声学、发音和语言组件有不同的激活函数，通常单独训练和优化。为了弥补传统混合系统的不足，当前流行的端到端模型 (Liu et al., 2019)将传统混合系统折叠为一个单一的神经网络，去除传统框架中所有中间步骤和独立子任务，输入语音特征，直接输出源语言文本，具有容易训练、模型简单和联合优化的优势，取得很好的效果。当前端到端模型流行的方法主要有连接时序分类算法 (Graves et al., 2006)（Connectionist Temporal Classification，CTC）CTC和使用CTC与注意力机制的混合方法 (Moritz et al., 2019)。CTC不需要对训练语料预先分段和后处理输出标签，然而，CTC基于条件独立假设训练ASR模型，缺乏对输入序列间上下文关系的建模。对此，注意力对齐机制 (Bahdanau et al., 2014)第一次使用到基于序列到序列结构的语音识别模型中 (Chorowski et al., 2014)，但由于过度灵敏的关注对齐方式应用到真实的语音识别场景中表现出比较差的效果。于是， (Kim et al., 2017)结合CTC和注意力机制的优势提出基于两者的混合语音识别模型， (Moritz et al., 2019)和 (Sarl et al., 2020)在混合模型的改进取得不错的效果。

综上所述，端到端的模型主要在英语、法语等表音文字的语音识别中取得很好的效果，然而，汉语是一种典型的表意文字，每一个汉字表示个别词或词素的形体，不与语音直接发生联系，当前端到端的模型对汉字的识别存在一些不足。 (Chan et al., 2016)对汉字识别的研究工作中表明模型对汉字的识别收敛速度较慢。拼音作为汉字的读音标注文字，直接表示汉字语音，拼音与汉字存在内在转换关系，基于音节（拼音）的研究工作 (Zhou et al., 2018)持续至今。将语音特征识别为音节单元 (Qu et al., 2017)、再通过一个转换模型将拼音变换为汉字 (Liu et al., 2015)的级联模型存在错误传播，为了避免这种问题， (Chan et al., 2016)提出汉字-拼音识别模型，只在训练时使用拼音帮助对汉字的识别，但是这种方法识别字符错误率（Character Error Rate，CER）达到59.3，对此， (Zhou et al., 2018)提出基于Transformer (Vaswani et al., 2017)的贪婪级联解码器模型，取得相对较好的效果。

基于以上研究工作，在汉语语音识别中，引入拼音作为对汉字解码的约束，能够促使模型学习更好的语音特征。在汉语中，对汉字的识别类似于语音翻译（Spoken Language Translation，ST） (Di Gangi et al., 2019)，对拼音的识别可以视为对汉语的语音识别。在ST领域， (Weiss et al., 2017)提出将语音识别和语音翻译联合学习可以有效提高模型翻译性能。从 (Weiss et al., 2017)的研究工作中受到启发，在多任务学习框架下Caruana (1997)，提出基于拼音约束联合学习的汉语语音识别方法，在汉语语音识别中引入拼音语音识别任务作为辅助任务联合训练，共同学习，相互促进。在AISHELL-1 (Bu et al., 2017)中文训练语料上，词错误率WER值比基线模型降低2.24个百分点。

2 基于拼音约束联合学习的汉语语音识别方法

模型共享一个编码器，拼音语音识别和汉语语音识别分别有一个解码器，训练时，模型的交叉熵是两个解码器分别计算损失后正则求和；反向传播时，编码器的参数被两个任务同时更新，达到两个任务共同促进，相互增强的效果。模型结合 (Weiss et al., 2017)研究工作和 (Kim et al., 2017)提出的混合模型，并对其做了进一步改进。具体结构概览图如图1所示，包括三个部分：共享编码器、拼音语音识别和基于拼音约束联合学习的汉字识别，本节分别将从1.1、1.2、1.3节对以上部分进行介绍。

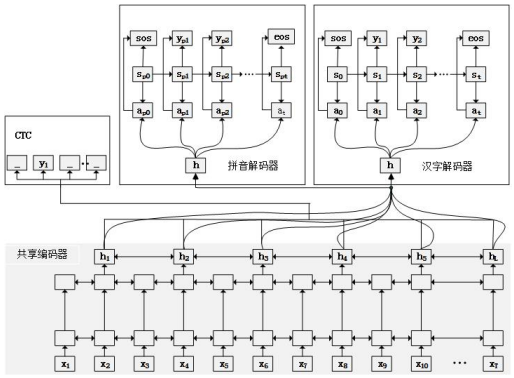

图1　基于拼音约束联合学习的汉语语音识别结构图

2.1　共享编码器

模型共享一个编码器，编码器采用双向长短期记忆网络（Long Short Term Memory networks, LSTM），双向LSTM结构见图1中的共享编码器部分。共享编码器将语音信号特征$x = (x_1, x_2, ..., x_T)$作为输入，使用VGG对x抽取特征转为高纬的隐表征，输出为$h = (h_1, h_2, ..., h_L)$。这里T表示语音特征的帧索引，L为对语音特征下采样后的帧索引$(L \leq T)$。编码器的编码过程表示为

$$h = Encoder\,(x) \qquad (1)$$

2.2　拼音语音识别

拼音语音识别模型采用当前流行的基于注意力机制的编码器-解码器框架，编码器采用2.1节介绍的共享编码器结构。解码器采用单向LSTM，见图1中的拼音解码器部分。解码器以共享编码器的输出h作为输入，基于时刻t前的输出标签序列，得到每一个t时刻预测拼音p标签y_{pt}的概率分布：

$$P\,(y_p|h) = \prod_t P\left(y_{pt}|h, y_{p(1:t-1)}\right) \qquad (2)$$

$$y_{pt} = LSTM\left(h, y_{p(1:t-1)}\right) \qquad (3)$$

对于每一时间步t，基于所有的输入语音特征h和注意力机制权重$a_{t,l}$产生文本向量c_t：

$$c_t = \sum_l a_{t,l} h_l \qquad (4)$$

这里的$a_{t,l}$通过softmax层计算：

$$a_{t,l} = \frac{\exp\left(\gamma e_{t,l}\right)}{\sum_l \exp\left(\gamma e_{t,l}\right)} \qquad (5)$$

$$e_{t,l} = \omega^T \tanh\left(W s_{t-1} + V h_l + U f_{t,l} + b\right) \qquad (6)$$

$$f_t = F * a_{t-1} \qquad (7)$$

这里，训练参数有ω、W、V、U、F、γ是模型的锐化因子，l为帧索引，$*$表示一维卷积，f_t通过$*$与卷积参数F计算得出。

解码器使用c_t、t时刻前的输出标签$y_{p(t-1)}$和隐状态s_{t-1}生成当前时刻的隐状态s_t和预测拼音标签y_{pt}：

$$s_t = LSTM\,(s_{t-1}, y_{t-1}, c_t) \qquad (8)$$

$$y_{pt} = Generate\,(s_t, c_t) \qquad (9)$$

这里LSTM代表单向循环神经网络，Generate代表前馈网络。

结合公式（2），拼音语音识别的损失函数可以通过以下公式计算：

$$L_p\,(h, y_p) = -\ln P\,(y_p|h) \qquad (10)$$

这里拼音序列$y_p = (y_{p1}, y_{p2}, ...y_{pt})$，其中$t \leq T$。

2.3 基于拼音约束联合学习的汉字识别

基于共享编码器的输出h作为输入，汉字解码器同样以h作为输入，结合时刻t前的输出标签序列，通过简单的前馈网络和softmax激活函数，得到每一个时刻t预测汉字标签y_t的概率分布$P(y|h)$，基于$P(y|h)$，汉字语音识别交叉损失熵可以通过以下公式计算：

$$L(h,y) = -\ln P(y|h) \qquad (11)$$

这里汉字序列$y = (y_1, y_2, ...y_t)$。

在多任务学习框架下，提出模型的交叉损失熵通过拼音解码器和汉字解码分别计算损失后的正则求和联合训练。拼音语音识别作为辅助任务帮助模型增强对汉字的识别能力，与此同时，汉语语音识别作为主要任务促进模型对拼音监督信号的感知。反向传播时，通过共享编码器，能同时感知拼音和汉字的监督信号，编码器的参数被拼音语音识别和汉字语音识别同时更新。结合公式（10）、（11），基于拼音约束联合学习的汉字识别交叉熵损失函数表示为

$$L_{hybrid}(h,y) = \lambda L_p(h,y_p) + (1-\lambda) L(h,y) \qquad (12)$$

这里λ为模型可微调的超参数：$\lambda \in (0,1)$。

考虑CTC具有使模型快速收敛的优势，且不需要对输入、输出序列做一一标注和对齐，因此，提出的模型结合了CTC。通常情况下，CTC与循环神经网络（Recurrent Neural Network，RNN）结合，RNN作为编码器，把语音特征序列x转为高纬的隐状态h，该编码器过程如公式（1）。基于语音隐表帧h，CTC假设输出汉字标签之间条件独立，标签之间允许插入空白表示（-），求出标签序列任何一条路径$\pi = (\pi_1, \pi_2, ..., \pi_T)$的概率分布$p(\pi/h)$，由于多条路径序列可能只对应一条汉字标签序列，通过定义一个多对一的映射函数F（$\pi \in F(y)$）将路径序列映射到标签序列y，采用前后向算法有效求得标签序列的最大概率分布$p(y/h)$，基于$p(y/h)$，可以计算CTC的负对数似然函数L_{CTC}。本文模型结合CTC的交叉熵损失函数通过以下公式计算：

$$L(h,y) = (1-\lambda_1) L_{hybrid}(h,y) + \lambda_1 L_{CTC}(h,y) \qquad (13)$$

$$L_{CTC}(h,y) = -\ln(P(y|h)) \qquad (14)$$

$$P(h,y) = \sum_{\pi \in F(y')} P(\pi|h) \qquad (15)$$

这里λ_1为模型可微调的超参数：$\lambda_1 \in (0,1)$，y'为映射标签序列。

3 实验

3.1 数据设置

见表1，使用由希尔贝壳中文普通话开源的语音数据库AISHELL-1 (Bu et al., 2017)证明了本文方法的有效性。该训练语料包括200个说话者，其中训练集有120098条语音（约150个小时），验证集有14326条语音（约10个小时），测试集有7176条语音（约5个小时）。通过Torchaudio[1]工具，提取以上训练语料步长为10毫秒、窗口大小为25毫秒、维度为40的梅尔倒频谱filter-bank特征。

分类	时长	男声	女声
训练集	150	161	179
验证集	10	12	28
测试集	5	13	7

Table 1: 实验训练集AISHELL-1

[1]https://pypi.org/project/torchaudio/

3.2 评价指标

本文使用词错误率作为模型的评价指标，词错误率简称WER（Word Error Rate），将模型预测的输出序列与监督信号序列进行比较，计算WER的公式：

$$WER = 100 * \frac{S + D + I}{N} \qquad (16)$$

这里S、D、I 表示替换、删除和插入的字数，N为监督信号字序列的总字数。

3.3 参数设置

对于未登录字，使用特殊字符“UNK代替”，超参数均设置为0.2时模型效果最好 (Kim et al., 2017)，dropout设为0.25。模型采用Adadelta算法Zeiler (2012)进行优化，batch-size设置为16，共享编码器采用4层的卷积网络和5层的双向LSTM，双向LSTM每个方向有512个隐状态单元，两个解码器均是一个单层的有512个隐状态单元的LSTM，Attention机制使用location-aware attention (Chorowski et al., 2014)。在词嵌入层，每个字表征为256纬的向量。拼音的字表大小为1400，汉语的字表大小为4500。

3.4 基线模型

本文共选择了3个基线模型，分别在训练数据集AISHELL-1进行试验，得到WER评分。模型包括基于音节的贪婪级联解码模型、S2S结合CTC的混合模型（S2S+CTC）和级联模型。

贪婪级联解码模型 (Zhou et al., 2018)，是使用两个beam search级联解码的Transformer模型。

混合S2S+CTC语音识别系统 (Kim et al., 2017)，是一种利用CTC和基于Attention序列到序列两者优势的模型，是当前常用的语音识别系统。

级联系统:将汉语语音特征序列识别为拼音文本序列，再采用一个额外的语言模型将拼音文本转写为汉语文本,采用由Pinyin2Hanzi3[2]将拼音文本序列转变为汉语文本序列。

3.5 本文方法有效性分析

对比基线模型，在AISHELL-1数据集上，验证了本文方法的有效性。使用WER值作为模型的评价指标，见下表2。

根据表2的实验结果分析：相比S2S+CTC（拼音识别），S2S+CTC（汉字识别）的WER值在验证集上高4.93个百分点，在测试集上高5.04个百分点，这说明当前的端到端语音识别模型对表意文字的识别效果不佳。相比基线模型S2S+CTC（汉字识别），提出模型在验证集上的WER值低2.5个百分点，在测试集上的WER值低2.24个百分点，说明在当前的汉语语音识别中引入拼音语音识别作为辅助任务联合训练，增强了模型对汉字的识别能力。相比级联系统+CTC，提出模型在验证集上的WER值低1.31个百分点，在测试集上低1.05个百分点，说明在汉语语音识别中引入拼音语音识别任务，提出的方法避免了级联系统导致的错误传播问题，取得比级联系统更好的识别效果。相比贪婪级联解码模型，提出模型在验证集上的WER值低6.1个百分点，在测试集上的WER值低4.95个百分点，这说明提出的模型在汉语语音识别中引入拼音取得相对较好的效果。

模型	λ, λ_1	WER（dev）	WER（test）
贪婪级联解码模型	-	16.16	17.64
S2S+CTC（拼音识别）	0,0.2	7.63	9.89
S2S+CTC（汉字识别）	0,0.2	12.56	14.93
级联系统+CTC	0,0	11.37	13.74
提出模型	0.2,0.2	10.06	12.69

Table 2: 提出模型对比基线模型的实验结果

为了讨论拼音语音识别任务和CTC对汉字识别的影响，对提出的模型去除CTC结构进行消融性实验结果分析，且分别将级联系统和S2S+CTC模型均消去CTC结构。三个模型训练

[2]https://github.com/letiantian/Pinyin2Hanzi

时间基本一致。相比S2S-CTC（拼音识别），S2S-CTC（汉字识别）在验证集集上的WER值高6.23个百分点，在测试集上的WER值高6.45个百分点，说明当前的端到端语音识别系统对表意文字的识别效果不佳。相比基线模型S2S-CTC（汉字识别），提出模型-CTC在验证集的WER值低2.61个百分点，在测试集上低2.57个百分点；相比级联系统-CTC，提出模型-CTC在验证集上低1.5个百分点，在测试集上低2.31个百分点，说明提出模型在不受CTC影响下，引入拼音约束联合学习，增强了模型对语音特征的表达。

模型	λ, λ_1	WER（dev）	WER（test）
S2S-CTC（拼音识别）	0,0	10.57	12.57
S2S-CTC（汉字识别）	0,0	16.80	19.02
级联系统-CTC	0,0	15.69	18.76
提出模型-CTC	0.2,0	14.19	16.45

Table 3: 消融性实验结果分析

结论

由于汉字与语音没有直接的联系，拼音与汉字、语音具有内在关系，提出基于拼音约束联合学习的汉语语音识别方法，通过多任务学习框架，联合拼音语音识别、汉字语音识别任务共同学习，取得了更好的效果。进一步研究工作可以将拼音序列变换汉字序列视为一个机器翻译任务，通过共享解码器方式的联合学习等。

参考文献

Ashish Vaswani,Noam Shazeer,Niki Parmar,Jakob Uszkoreit,Llion Jones,Aidan N. Gomez ,Łukasz Kaiser,Lukaszkaiser,Illia Polosukhin 2017. Alternation. *Attention is All you Need*, Neural Information Processing Systems.

Alex Graves, Santiago Fernández, and Faustino Gomez. 2006. Alternation. *Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks*, International Conference on Machine Learning ACM.

Alexander H. Liu, Tzu-Wei Sung, Shun-Po Chuang, Hung-yi Lee, Lin-shan Lee 2019. Alternation. *Sequence-to-sequence Automatic Speech Recognition with Word Embedding Regularization and Fused Decoding*, arXiv.

Dzmitry Bahdanau, Kyunghyun Cho,and Yoshua Bengio 2014. Alternation. *Neural Machine Translation by Jointly Learning to Align and Translate*, Computer Science.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, Hao Zheng. 2017. Alternation. *Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline*, 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). IEEE, 2017: 1-5.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio 2014. Alternation. *End-to-end continuous speech recognition using attention-based recurrent NN: First results*, preprint arXiv.

Leda Sarl, Niko Moritz, Takaaki Hori,and Jonathan Le Roux 2020. Alternation. *Unsupervised Speaker Adaptation using Attention-based Speaker Memory for End-to-End ASR*, IEEE International Conference on Acoustics,Speech and Signal Processing.

Mattia A. Di Gangi, Matteo Negri and Marco Turchi 2019. Alternation. *Adapting Transformer to End-to-End Spoken Language Translation*, conference of the international speech communication association.

Matthew D. Zeiler 2012. *Adadelta: an adaptive learning rate method*, preprint arXiv.

Niko Moritz, Takaaki Hori,and Jonathan Le Roux 2019. Alternation. *Streaming end-to-end speech recognition with joint CTC-attention based models*, IEEE ASRU.

Niko Moritz, Takaaki Hori,and Jonathan Le Roux 2019. Alternation. *Triggered attention for end-to-end speech recognition*, in Proc. IEEE ICASSP, May2019, pp. 5666–5670.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu,and Zhifeng Chen 2017. Alternation. *Sequence-to-sequence models can directly translate foreign speech*, Proceedings of Interspeech.

Rich Caruana 1997. *Multitask learning. Machine Learning.* 28(1):41–75

Shiyu Zhou, Linhao Dong, Shuang Xu, and Bo Xu. 2018. Alternation. *Syllable based sequence-to-sequence speech recognition with thetransformerin mandarin chinese*, Interspeech.

Suyoun Kim,Takaaki Hori,and Shinji Watanabe 2017. Alternation. *Joint ctc-attention based endto-end speech recognition using multi-task learning*, International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Tara Sainath, Abdel-rahman Mohamed, Brian Kingsbury, and Bhuvana Ramabhadran. 2013. Alternation. *Deep Convolutional Neural Networks for LVCSR*, IEEE International Conference on Acoustics,Speech and Signal Processing.

William Chan,Ian Lane 2016. Alternation. *On online attention-based speech recognition and joint mandarin character-pinyin training*, Interspeech.

Yi Liu,Jing Hua,Xiangang Li,Tong Fu, and Xihong Wu 2015. Alternation. *Chinese syllable-tocharacter conversion with recurrent neural network based supervised sequence labelling*, Signal and Information Processing Association Annual Summit and Conference.

Zhongdi Qu,Parisa Haghani,and Eugene Weinstein 2017. Alternation. *Syllable-based acoustic modeling with ctc-smbr-lstm*, IEEE ASRU.