2018 年 10 月 第 10 期
Vol. 39 No. 10 2018

基于链接时序分类的日语语音识别

孙 健，郭 武

(中国科学技术大学 语音及语言信息处理国家工程实验室，合肥 230027)
E-mail: sjian17@ mail. ustc. edu. cn

摘 要： 目前，端到端的语音识别系统因其简洁性和高效性成为大规模连续语音识别的发展趋势. 本文将基于链接时序分类的端到端技术应用到日语语音识别上，考虑到日语中平假名、片假名和日语汉字多种书写形式的特性，通过在日语数据集上的实验，探讨了不同建模单元对识别性能的影响；进一步将音素信息应用到模型的初始网络训练中，改善语音识别系统性能，最终效果优于基于隐马尔可夫模型和双向长短时记忆网络的主流语音识别系统.

关 键 词： 语音识别；日语；链接时序分类；端到端
中图分类号： TP183 **文献标识码：** A **文 章 编 号：** 1000-1220(2018)10-2129-05

Towards Connectionist Temporal Classification Speech Recognition System for Japanese

SUN Jian, GUO Wu
(University of Science and Technology of China, National Engineering Laboratory for Speech and Language Information Processing, Hefei 230027, China)

Abstract: The end-to-end framework has become the state-of-the-art method in large vocabulary continuous speech recognition (LVCSR) because of its simplicity and efficiency. In this paper, the end-to-end technology based on Connectionist Temporal Classification (CTC) is applied to Japanese speech recognition. Considering the characteristic of various written forms among hiragana, katakana and kanji in Japanese, we discuss the impact of different modeling units on recognition performance through experiments on Japanese dataset. Then we combine phoneme information into the acoustic model to improve the performance. Experiments demonstrate the effectiveness of the proposed methods, which can achieve better performance than the mainstream speech recognition system based on Hidden Markov Model and Bi-directional long-short memory network.

Key words: automatic speech recognition; Japanese; end-to-end; connectionist temporal classification

1 引 言

随着深度学习的快速发展，神经网络取代混合高斯模型[1](Gaussian Mixture Model，GMM)，并与隐马尔可夫模型[2](Hidden Markov Model，HMM)相结合，对状态进行建模，使得大规模连续语音识别的性能获得了显著的提高. 近年来，循环神经网络(Recurrent Neural Networks，RNN)及其变体—长短时记忆网络[3](Long Short-Term Memory，LSTM)成功应用于语音识别，解决了普通的深度神经网络[4](Deep Neural Network，DNN)无法对语音信号时序特性建模的缺点，语音识别性能进一步提升. 但是以 HMM 为框架的识别算法对语言学知识的要求较高，包括上下文相关音素状态绑定，发音字典的准备等等，开发难度较大. 另外训练过程中，需要通过强制对齐，获得帧级标注，任务复杂程度较高，且忽略了语音序列内在特性，无法全局优化整个语音序列.

为解决上述问题，A. Graves 等人提出链接时序分类技术[5,6](Connectionist Temporal Classification，CTC)和端到端的识别系统，解决了输入和输出标签对应关系未知情况下的序列分类问题，全局优化语音序列. 与之前提到的混合模型不同，CTC 不需要隐马尔可夫模型，仅需要单独的神经网络即可完成整个语音识别任务. 在基于 CTC 的端到端系统中，将语音序列直接映射到标注序列所在的空间，消减了发音词典，语言模型等成分，极大地简化了语音识别的步骤[7,8].

本文探讨了日语语音识别的相关问题. 日语是日本国的官方语言，日语中主要使用的文字包括平假名(例如："あ")，片假名(例如："テ")，和日语汉字(例如："日本语"). 平假名包含了日语中所有的发音[9]，片假名用于书写外来词，拟声词，拟态词和一部分动、植物的名称，日语汉字用于表示实物的名称或动作. 日语中一般混合使用三种字体，其中平假名和片假名一一对应，由于假名同音歧义的现象比较严重，因此日语中汉字使用十分广泛，常用汉字有 2000 多个，而且所有的汉字均可通过假名表达. 虽然日语中的发音单元并不多，但书写单元种类繁多，表现形式复杂，因此以 CTC 技术为核心的端到端语音识别系统中，选择合适的建模单元能够对识别性能进一步优化. 本文首先采用字型(gra-phoneme)即全部的假名和常用汉字共 2795 个单元进行建模，实验结果与双向长短时记忆网络(BiLSTM-HMM)系统差距较大. 进一步，在已经知道日语语音学知识的情况下，将其结合到端到端识别系统中. 选择以音素为建模单元训练任务，系统性能得到提升.

在进行 CTC 的实验中，我们发现字型作为建模单元的神经网络输出的后验概率比较尖锐，随机初始化的网络容易陷入局部最优解，因此本文把以音素为建模单元的训练模型作为初始网络以提升前者模型的鲁棒性，使得识别性能大幅提高，此外我们将传统的语言模型与CTC相结合，系统效果获

收稿日期：2017-12-12 收修改稿日期：2018-02-01 基金项目：国家重点研发计划专项项目(2016YFB1001303)资助. 作者简介：孙 健，男，1995 年生，硕士研究生，研究方向为语音识别；郭 武，男，1973 年生，博士，副教授，研究方向为语音信号处理.

得明显的提升,超过当前主流的 BiLSTM-HMM 系统.

2 基于深度神经网络的声学模型

当前语音识别的主要方法是采用循环神经网络(RNN)及其变体和隐马尔可夫模型相结合进行声学模型的训练. 循环神经网络利用过去的信息,将上一时刻隐层输出输入到当前时刻的隐层中,保留了之前的信息,如图 1 所示. 语音信号作为一个时间序列,上下文依赖性较强,因此循环神经网络很快被应用于语音识别. 理论上 RNN 可以处理任意长的序列,但是由于梯度消失,导致 RNN 无法利用较远时刻的信息.

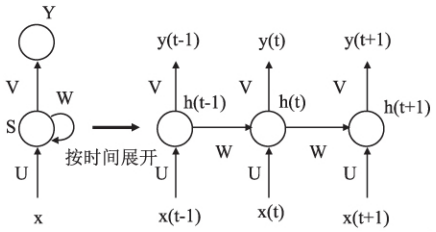

图 1 RNN 时间展开图

Fig. 1 Unfolded RNN structure

为解决这一问题,RNN 衍生出一种变体—长短时记忆网络[10](LSTM). 原始的 RNN 网络中,隐藏层只有一个状态,无法解决序列的长时依赖问题,所以在隐藏层节点中额外引入一个 cell 单元,cell 单元利用了门的概念,通过门的控制保留长时信息. Cell 单元包含"输入门","输出门"和"遗忘门",其中输入门决定当前语音信号如何保存到 cell 单元中,输出门决定 cell 单元状态如何作为隐藏层的输出,遗忘门决定上一时刻的 cell 单元状态如何保存到当前时刻的 cell 单元中.

图 2 展示了 cell 单元的工作原理,输入信号包括当前输入信息 x_t,上一时刻隐藏层的输出 $\boldsymbol{h}_{t-1}$,以及上一时刻的 cell 单元状态 $\boldsymbol{C}_t$,cell 单元的输出包括隐藏层当前状态 $\boldsymbol{h}_t$,以及 Cell 单元的当前状态 $\boldsymbol{C}_t$.

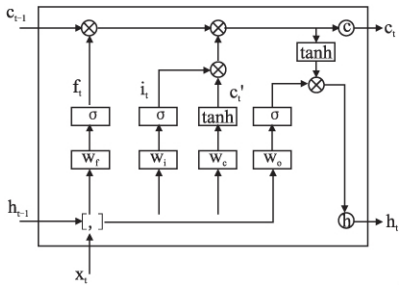

图 2 cell 结构

Fig. 2 Architecture of memory cell

具体计算见式(1)到式(6),其中 $\boldsymbol{f}_t$,$\boldsymbol{i}_t$,$\boldsymbol{o}_t$ 分别表示遗忘门,输入门和输出门,∘表示按元素乘运算.

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{C}'_t = \tanh(\boldsymbol{W}_c [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \circ \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \circ \boldsymbol{C}'_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = o_t \circ \tanh(\boldsymbol{C}_t) \tag{6}$$

长短时记忆网络利用 cell 结构解决了时间序列的长时依赖问题,有效使用了当前时刻之前的信息,然而在一些任务中,除了过去的信息,未来信息也很重要,因此双向长短时记忆网络[11] 登上舞台. 本文在后续实验中,采用双向长短时记忆网络完成 CTC 系统的搭建.

3 链接时序分类

3.1 链接时序分类简介

链接时序分类(CTC)主要用于处理时序分类任务,尤其是输入信号与目标标签对齐结果未知的情况. 链接时序分类技术在整个输入序列的任何一点都可以进行标签预测,解决了传统语音识别中需要强制对齐的问题. 通过链接时序分类技术进行神经网络训练的准则称为 CTC 准则.

对于语音识别,输入序列为 $\boldsymbol{x} = \{x_1, x_2, \cdots, x_T\}$,得到输出序列 $\boldsymbol{a} = \{a_1, a_2, \cdots, a_N\}$,CTC 准则中,集合中的每一个符号代表一个建模单元(例如字,字母,音节等),将输出标签集合 $\mathbf{A} = \{a_i\}$ 扩展为 $\mathbf{A}' = \mathbf{A} \cup \{blank\}$,*blank* 用来表示静音帧或者分割叠字,神经网络的 softmax 层增加一个额外节点表示输出 *blank* 的概率. 用 $y^t_{l_t}$ 表示第 t 帧语音对应标签为 l_t 的条件概率,则对于输入序列,可能的输出序列 $\boldsymbol{l}$ 对应的后验概率为:

$$P(\boldsymbol{l}|\boldsymbol{x}) = \prod_{t=1}^{T} y^t_{l_t} \tag{7}$$

基于集合 $\mathbf{A}'$ 得到序列 $\boldsymbol{l}$,需要映射到集合 $\mathbf{A}$ 所对应的空间,因此定义函数 $\mathrm{F}: \mathbf{A}'^{\mathrm{T}} \rightarrow \mathbf{A}^{\leqslant \mathrm{T}}$ 合并连续相同的标签,删除 *blank* 得到完整的输出结果,例如 $\mathrm{F}(-aa-a-b) = F(-a-a-bb)$ (aab). 记语音序列对应的真实标签序列为 $\boldsymbol{L}$,故 CTC 准则下要求最大化所有可能的路径 $\boldsymbol{l}$ 对应的后验概率之和,对应的损失函数为式(9).

$$P(\boldsymbol{L}|\boldsymbol{x}) = \sum_{l \in \mathrm{F}^{-1}(L)} P(\boldsymbol{l}|\boldsymbol{x}) \tag{8}$$

$$\mathrm{O} = -\ln P(\boldsymbol{L}|\boldsymbol{x}) \tag{9}$$

3.2 目标函数计算

目标函数 $P(\boldsymbol{L}|\boldsymbol{x})$ 可由动态规划算法进行解决. 考虑到输出路径中的 *blank*,修改标签序列 $\boldsymbol{L}$ 为 $\boldsymbol{L}'$,在路径 $\boldsymbol{L}$ 的开始,结束和每两个连续的标签之间增加 *blank*,若 $\boldsymbol{L}$ 的长度为 M,则 $\boldsymbol{L}'$ 的长度为 2M + 1. 定义前向变量 $\alpha(t, m)$ 表示 t 时刻输出的标签为 $\boldsymbol{L}'$ 路径中第 m 个符号的后验概率. 在计算前向概率之前,对于任意序列 $\boldsymbol{s}$,用 $\boldsymbol{s}_{p:q}$ 表示 $\boldsymbol{s}$ 的子序列 $\{s_p, s_{p+1}, \cdots, s_q\}$,定义集合 $\mathbf{V}(t, m) = \{\boldsymbol{l} \in \mathbf{A}'^{t}: \mathrm{F}(\boldsymbol{l}) = \boldsymbol{L}_{1:m/2}, l_t \rightarrow L'_m\}$. 则前向概率如下:

$$\alpha(t, m) = \sum_{l \in \mathbf{V}(t,m)} \prod_{i=1}^{t} y^i_{l_i} \tag{10}$$

根据上述公式,目标函数可表示为时刻 T 输出 *blank* 或者没有输出 *blank* 的前向概率之和,即:

$$P(\boldsymbol{L}|\boldsymbol{x}) = \alpha(T, 2M+1) + \alpha(T, 2M) \tag{11}$$

所有正确的路径必须起始于 *blank*,或者是 $\boldsymbol{L}$ 的第一个输出标签.

$$\alpha(1, 1) = y^1_{blank} \tag{12}$$

$$\alpha(1, 2) = y^1_{L_1} \tag{13}$$

$$\alpha(1,m) = 0, \forall m > 2 \tag{14}$$

故前向概率的迭代形式如下:

$$\alpha(t,m) = y^t_{L'_m} \sum_{i=f(u)}^{m} \alpha(t-1,i) \tag{15}$$

其中

$$f(m) = \begin{cases} m-1 & \text{若 } L'_m = blank \text{ 或者 } L'_{m-2} = L'_m \\ m-2 & \text{其他} \end{cases} \tag{16}$$

在每一个时间点都要考虑是否有足够的时长来完成剩余序列,故前向概率需要满足下式:

$$\alpha(t,m) = 0 \quad \forall m < 2M - 2(T-t) \tag{17}$$

同理定义后向概率 $\beta(t,m)$ 表示满足前向概率 $\alpha(t,m)$,且从 $t+1$ 时刻开始到输出序列 L 结束的所有可能路径的概率之和,计算过程与前向概率类似,不再赘述.

因此,在训练样本集合 $S=\{x,L\}$ 上的神经网络的损失函数可表示为:

$$O(S) = -\ln \prod_{(x,L)\in S} P(L|x) = -\sum_{(x,L)\in S} \ln P(L|x) \tag{18}$$

$$O(x,L) = -\ln P(L|x) = -\ln \sum_{m=1}^{|L'|} \alpha(t,m)\beta(t,m) \tag{19}$$

3.3　基于 CTC 的日语识别单元选择

日语拥有复杂的书写系统,主要包括平假名、片假名和日语汉字三种文字系统,同时也可以以日语罗马字转写为拉丁字母. 日语汉字的读音复杂,大多包含音读(音読)和训读(训読)两类,音读将古代汉语读音日语化,训读保留汉字含义,采用日语固有读音方法,通常使用平假名和片假名为日语汉字注音(见图 3). 罗马字多用于商标和招牌,文章中一般很少使用.

文本标注: 次官は十二から
假名注音: じかんはじゅうにから

图 3　假名注音方式

Fig. 3　Pronunciation in kana

现在语音识别有发音字典,在图 4 中,左边是我们图 3 中的假名以及汉字分词后的词单元,右侧则是在经典的语音识别中用到的所谓音素,根据每个字的发音组成进行处理.

次	ji
官	ka N
は	ha
十二	ju: ni
か	ka
ら	ra

图 4　发音字典 左侧是假名,汉字以及两者构成的词语,右侧是音素组成

Fig. 4　Pronunciation dictionary on the left side of the pronunciation dictionary are kana, kanji and the words formed by the two. The right side is composed of phonemes

在经典的 DNN-HMM 或者 LSTM-HMM 框架下,将图 4 中的音素的绑定三音子单元(tri-phone)中的状态作为神经网络的建模单元. 另一方面,采用 CTC 建模的策略下,可以直接忽略图 4 所示的词典,直接采用图 3 所示的标注来进行模型训练. 本文中,采用两种策略实现 CTC 的日语识别,首先是直接采用假名(平假名、片假名)和汉字作为声学建模的输出单元,也就是最常用的日语分词都不再采用,直接根据字型(gra-phoneme)来做输出单元,而不考虑这些字到底是单音节还是多音节,或者根本都无法组成一个音节,本文的策略对非拉丁字母的端对端识别具有一定的参考; 第二种策略是,既然有日语分词和日语词典,我们将其应用到 CTC 的声学建模中,也就是把一句话的 gra-phoneme 拆解成以单音素(mono-phone)为单元的音素串,但是采用 CTC 的优化准则来训练声学模型,从训练语句一句话的角度来优化模型参数.

3.4　CTC 模型训练参数初始化

相对英语和汉语这两种世界广泛应用的语言而言,日语识别语料还是相对较少,如何在语料较少的情况下训练一个稳健的声学模型也是一个很重要的研究点. 对于深度学习而言,首先采用相对好的参数来初始化模型参数,避免陷入局部最优解,是目前最常用的一种策略. 最常用的初始化策略是采用大语种(如英汉)的模型参数作为初始值. 故本文在搭建 BiLSTM-HMM 系统时,采用 300 小时的 switchboard 英语数据集的训练结果作为初始网络,增强系统鲁棒性.

正像汉语一样,日语中也存在着多发音字现象,而直接把字形拿来建模是无法考虑这种情况的; 另外,不同字的字频分布也很不均匀,在基于日语字的 CTC 系统的训练过程中,我们发现随机初始化的 CTC 相较于传统的 HMM 模型,神经网络输出的后验概率比较尖锐,训练过程不稳定,容易过早收敛,陷入局部最优解. 考虑到已经有日语的音素信息的词典,而这种音素信息相对而言比较可靠,以音素为建模单元的模型相对而言稳健性更好,将其作为以字为建模单元的初始网络,从而可以避免陷入局部最优解的不足.

4　实验结果及分析

4.1　实验数据集和实验平台

本文在 King-ASR-117 日语数据集上进行实验. 该数据库收集了安静环境下 122945 条语音数据,长达 145.2 小时,所有语音数据均为 16KHz 采样率、16bit、单通道的格式. 在实验中我们挑选了 ~106.2k 条语音数据(~123h)作为训练集,~5.4k 条语音数据(~6.21h)作为开发集,~2.5k 条语音数据(~2.88h)作为测试集. 本文以 Kaldi[12] 和 Eesen[13] 作为实验平台,比较了基于隐马尔可夫模型的语音识别系统和基于链接分类技术的端到端系统的识别效果.

4.2　基于 HMM 的语音识别系统

实验中将 39 维梅尔频率倒谱系数(MFCC 特征)作为 GMM-HMM 混合系统的输入信号,在 GMM-HMM 系统中,通过高斯分裂和决策树聚类最终绑定状态数目为 12970,用得到的模型对训练数据做强制对齐得到帧级标签,作为后续神经网络的训练数据.

在 LSTM-HMM 训练中,采用 108 维 filterbank 特征进行训练. 当前帧利用之前发生的 40 帧信号获得过去信息,同时在输入语音帧和输出标签中加入一定时延得到一部分之后的信息. 网络共有 3 层隐藏层,隐层节点为 1024,输出维度仍然是 12970.

为更好地利用上下文信息,我们采用双向长短时记忆网络,BiLSTM-HMM 系统与 LSTM-HMM 结构基本一致,只是在当前帧的前后各使用了 40 帧语音数据. 另外,我们采用英语的

SwitchBoard 作为初始网络来提高系统的识别正确率.

4.3 基于 CTC 的语音识别系统

基于链接时序分类的语音识别系统,摈弃了隐马尔可夫模型,直接从输入序列映射到输出序列. 本实验中采取 3 层隐藏层,每层 1024 个隐藏节点的 BiLSTM 网络,和 108 维 filter-bank 特征进行声学模型的训练. 我们训练了两个 CTC 的基本系统,一个是以字作为神经网络输出节点,也就是基于 gra-phoneme 的系统,另外一个是以音素为神经网络输出节点,也就是基于 phoneme 的系统. 在这两个基本系统的基础上,用后者的训练得到的网络来作为初始网络,再重新优化以字作为输出节点的系统,得到识别性能更优的系统.

在搭建以日字为建模单元的 CTC 系统时,经统计,数据集中假名、汉字共有 2794 个单元,测试集中包含少量的集外英语词汇,故添加 1 个 blank 单元,利用 2795 个建模单元进行实验. 实验采取多句并行的方法,利用 GPU,加快训练速度. 在实验过程中,我们发现了不同单元词频差异很大,如图 5. 这种字频差异很大会导致模型相对鲁棒性较差.

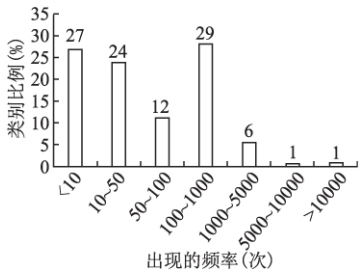

图 5 以字为建模单元,以频率对不同建模单元分类

Fig. 5 Gra-phoneme as modeling unit, frequency of different modeling unit

考虑到资源稀疏性对实验结果的影响,我们利用发音词典,以音素为建模单元进行实验,数据集中共有音素 237 个,加入 *blank* 后,网络的输出节点为 238,统计各个音素出现的频率如图 6 所示,相对均衡性更好,训练得到的模型理应更稳健.

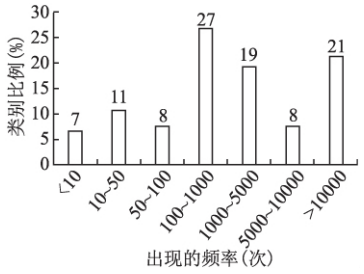

图 6 以音素为建模单元,以频率对不同建模单元分类

Fig. 6 Phoneme as modeling unit, frequency of different modeling unit

得到以音素为建模单元的 CTC 网络后,将其作为日字 CTC 训练系统的初始网络,增强系统的鲁棒性,避免训练过程陷入局部最优解.

由于基本的 CTC 方法不考虑语言学知识,严重影响识别效果[14-16],因此本文采用加权有限状态转换机[17](Weighted Finite-State Transducer, WFST)的方法,将语言模型,词典,标注符号打包在一起生成庞大的搜索网络进行解码,提高了解码效率和识别的准确率.

以上描述的多个系统的识别词错误率如表 1 所示.

4.4 实验分析

在隐马尔可夫模型的框架下,LSTM 的实验结果相较于 GMM 提高了 6.57 个百分点,可见神经网络对人类认知世界的拟合能力非常强大. 将 LSTM 替换为 BiLSTM,神经网络高效地利用上下文的信息,最终我们的基线系统词错误率为 16.22%.

表 1 实验结果

Table 1 Experimental results

实验任务	词错误率(WER%)
GMM-HMM	25.18%
LSTM-HMM	18.61%
BiLSTM-HMM + initial_net	16.22%
CTC-Gra-phoneme	17.80%
CTC-Phoneme	17.37%
CTC-Gra-phoneme + initial_net	15.53%

在基于链接时序分类的语音识别系统,我们首先以日字作为建模单元(CTC-Gra-phoneme),随机初始化网络模型,词错误率为 17.80%. 降低建模单元的颗粒度之后,利用音素作为建模单元(CTC-Phoneme),实验性能得到进一步提升,词错误率为 17.37%,但相对于 BiLSTM-HMM 仍有差距,这主要是传统的 HMM 建模单元采用的是三音子单元(tri-phone),而 CTC 采用的是单音子(mono-phone),区分性差一些. 但是,将 CTC-Phoneme 系统得到的网络作为初始模型添加到日字 CTC 系统中,最终词错误率为 15.53%,这也证明了将音素初始信息加入后,模型参数更加可靠稳健.

基于 CTC 的日字识别系统对日字进行建模,能够从语音空间直接映射到手写空间,针对日语中存在大量的同音歧义字,有比较好的建模能力. 在下例中,标注为日语汉字,括号中是日语汉字对应的假名,表明其发音,其中"次官"和"時間"假名注解相同. 可以看到 CTC 系统和 HMM 系统的识别结果发音相同,但是 HMM 系统的识别结果却不同于标注. 因此选择端到端的方式针对日语语音识别具有一定的合理性.

例:

标注: 次官(じかん) 空(から)

CTC 识别结果: 次官 空

HMM 识别结果: 時間(じかん) から

5 实验总结

本文研究了基于链接时序分类的端到端技术,在日语数据集上,根据日语文字的特点,搭建了完整的语音识别系统,通过实验比较了不同颗粒度建模单元对识别性能的影响,最终基于 CTC 的语音识别系统性能超越 BiLSTM-HMM 系统,证明了 CTC 技术在日语语音识别上的有效性,也验证了如果能够将音素信息结合到模型训练中,可以进一步提升性能.

References:

[1] Yu D, Deng L. Automatic speech recognition: a deep learning approach [M]. Springer Publishing Company, Incorporated, 2014: 13-20.

[2] Rabiner L R. A tutorial on hidden Markov models and selected ap-

plications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[3] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[4] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups [J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[5] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks [C]. Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006: 369-376.

[6] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks [C]. Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 6645-6649.

[7] Audhkhasi K, Ramabhadran B, Saon G, et al. Direct acoustics-to-word models for English conversational speech recognition [J]. arXiv preprint arXiv: 1703.07754, 2017.

[8] Müller M, Stüker S, Waibel A. Phonemic and graphemic multilingual CTC based speech recognition [J]. arXiv preprint arXiv: 1711.04564, 2017.

[9] Han Shao-xiang. The new version of sino-japan communication and primary standard Japanese(the first volume) [M]. Beijing: People's Education Press, 2005: 7-11.

[10] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. arXiv preprint arXiv: 1402.1128, 2014.

[11] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.

[12] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [C]. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, IEEE Signal Processing Society, 2011.

[13] Miao Y, Gowayyed M, Metze F. EESEN: end-to-end speech recognition using deep RNN models and WFST-based decoding [C]. Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on. IEEE, 2015: 167-174.

[14] Prabhavalkar R, Rao K, Sainath T N, et al. A comparison of sequence-to-sequence models for speech recognition [C]. Proc. of Interspeech, 2017.

[15] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition [C]. Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016: 4945-4949.

[16] Kim S, Hori T, Watanabe S. Joint CTC-attention based end-to-end speech recognition using multi-task learning [C]. Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017: 4835-4839.

[17] Hori T, Nakamura A. Speech recognition algorithms using weighted finite-state transducers [J]. Synthesis Lectures on Speech and Audio Processing, 2013, 9(1): 19-21.

附中文参考文献:

[9] 韩绍祥. 新版中日交流标准日本语初级(上) [M]. 北京: 人民教育出版社, 2005: 7-11.