doi: 10.13413/j.cnki.jdxblxb.2023080

基于注意力机制语谱图特征提取的语音识别

姜 囡[1], 庞永恒[1], 高 爽[2]
(1. 中国刑事警察学院 公安信息技术与情报学院, 沈阳 110854;
2. 东北大学 信息科学与工程学院, 沈阳 110819)

摘要: 针对连接时序分类模型需具有输出独立性的假设, 对语言模型的依赖性强且训练周期长的问题, 提出一种基于连接时序分类模型的语音识别方法. 首先, 基于传统声学模型的框架, 利用先验知识训练基于注意力机制的语谱图特征提取网络, 有效提高了语音特征的区分性和鲁棒性; 其次, 将语谱图特征提取网络拼接在连接时序分类模型的前端, 并减少模型中循环神经网络层数进行重新训练. 测试分析结果表明, 该改进模型缩短了训练时间, 有效提升了语音识别准确率.

关键词: 语音识别; CTC 模型; 循环神经网络; 注意力机制
中图分类号: TP391 **文献标志码**: A **文章编号**: 1671-5489(2024)02-0320-11

Speech Recognition Based on Attention Mechanism and Spectrogram Feature Extraction

JIANG Nan[1], PANG Yongheng[1], GAO Shuang[2]
(1. *School of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang* 110854, *China*;
2. *College of Information Science and Engineering, Northeastern University, Shenyang* 110819, *China*)

Abstract: Aiming at the problem that the connected temporal classification model needed to have output independence assumption, and there was strong dependence on language model and long training period, we proposed a speech recognition method based on connected temporal classification model. Firstly, based on the framework of traditional acoustic model, spectrogram feature extraction network based on attention mechanism was trained by using prior knowledge, which effectively improved the discrimination and robustness of speech features. Secondly, the spectrogram feature extraction network was spliced in the front of the connected temporal classification model, and the number of layers of the recurrent neural network in the model was reduced for retraining. The test analysis results show that the improved model shortens the training time, and effectively improves the accuracy of speech recognition.

Keywords: speech recognition; CTC model; recurrent neural network; attention mechanism

收稿日期: 2023-03-08.

第一作者简介: 姜 囡(1979—), 女, 汉族, 博士, 教授, 从事智能识别的研究, E-mail: zgxj_jiangnan@126.com.

基金项目: 教育部重点研究项目(批准号: E-AQGABQ20202710)、辽宁省自然科学基金(批准号: 2019-ZD-0168)、辽宁省科技厅联合开放基金机器人学国家重点实验室开放基金(批准号: 2020-KF-12-11)、中国刑事警察学院重大计划培育项目(批准号: 3242019010)、公安学科基础理论研究创新计划项目(批准号: 2022XKGJ0110)和证据科学教育部重点实验室(中国政法大学)开放基金(批准号: 2021KFKT09).

近年来，在信息处理、通信与电子系统、自动控制等领域相继出现了不同用途的语音识别系统，低信噪比下语音识别已显露强大的技术优势和生命力，但仍面临诸多问题需要完善. 环境噪音和杂音对语音识别的效果影响不容忽视；在多语言混合、无限词汇识别和自适应方面需要进一步改进，以达到不受口音、方言和特定人影响的要求[1-3].

传统基于 GMM-HMM(Gaussian mixture model-hidden Markov model)的语音识别框架得到广泛使用和研究[4-5]，提出了语音识别声学模型的区分性训练准则，有效提升了语音识别的准确率和识别速率. 基于声学模型的区分性训练(discriminative training，DT)准则和最大似然估计训练 GMM-HMM，根据区分性准则对模型参数进行更新，可以显著提升模型的性能[6-8]. 文献[9-10]将深层神经网络用于语音的声学建模，在音素识别 TIMIT(声学-音素连续语音语料库，The DARPA TIMIT acoustic-phonetic continuous speech corpus)任务上获得成功，但由于 TIMIT 是一个小词汇量的数据库，连续语音识别任务更关注的是词组和句子的正确率，识别效果并不理想. 文献[11]提出了在 CD-DNN-HMM 框架下利用混合带宽训练数据，提高宽带语音识别精度的策略，DNN(deep neural network)提供了使用任意特性的灵活性.

随着基于 GMM-HMM 的语音识别框架被打破，文献[12]考虑卷积提供了一种改进的混响建模能力，将 CNN(convolutional neural network)与全连接的 DNN 相结合，可以在参数小于 DNN 的特征向量中建立短时间相关性模型，从而更好地泛化未发现的测试环境. 文献[13]提出一种具有规则化自适应性的基于长短时记忆(long and short term memory，LSTM)循环神经网络的声学模型. 在混合模型中，DNN/CNN/RNN(recurrent neural network)与隐 Markov 模型(HMM)是分开优化的，又由于语音识别中存在序列识别问题，因此联合优化模型中的所有组件会有效提升识别率，因而端到端的处理方式被引入到语音识别系统中. 文献[13]提出了一种基于 Gammatone 频率倒谱系数(GFCC)的特征提取方法，弥补了 Mel 频率倒谱系数(MFCC)特征的局限性. 文献[14-18]提出了混合式 CTC (connected temporal classification)/注意端到端的语音识别技术，这是一种用于语音识别、文本识别等任务的深度学习模型，它可以直接从时序数据中学习并预测出整段序列的标签. CTC 模型的基本架构是一个由多个循环神经网络(RNN)单元组成的双向网络，该网络在输入时序数据后，可以输出整段序列的标签概率分布，通过反向传播算法进行训练，最终得到的模型可以对新的未知序列进行标签预测，有效利用了训练和译码的优点，采用多任务学习框架提高鲁棒性，可以实现快速收敛.

本文提出一种基于连接时序分类模型的语音识别方法. 充分利用先验知识，基于注意力机制训练语谱图特征提取网络，有效提升提取特征的区分性和鲁棒性. 最终保证模型训练时间有效降低，提升语音识别准确率.

1　端到端模型的语音识别框架

1.1　语音识别系统流程

语音识别系统包括声学特征提取、语言和声学模型、发音词典和解码器 5 个模块，如图 1 所示. 从语音信号中提取到的声学特征经过统计训练到声学模型，把该声学模型作为识别单元的模板，结合发音词典和语言模型，经过解码器处理得到识别结果.

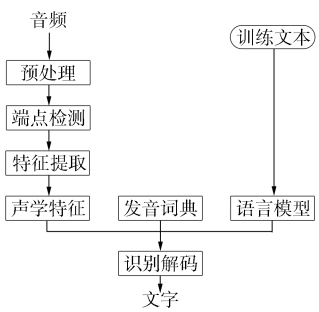


图 1　语音识别系统原理框架

Fig. 1　Principle framework of speech recognition system

1.2　基于 CTC 的端到端语音识别框架

CTC 的连续语音识别系统主要包括循环神经网络和 CTC 函数计算层，其中循环神经网络包含 4 层 LSTM 单元，如图 2 所示，其作用主要是提取输入特征中的有效时序信息，提升特征的区分性能和表达能力. CTC 函数计算层则可以直接计算输出序列后验概率.

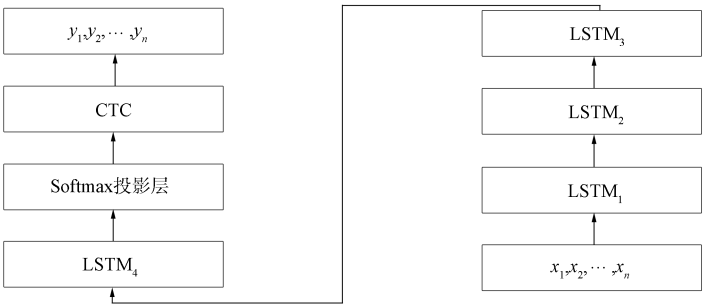


图 2 基于 CTC 的端到端语音识别系统框架

Fig. 2 End-to-end speech recognition system framework based on CTC

LSTM 模型是 RNN 的一种变形. RNN 存在梯度消失的问题，即后面时间节点对前面时间节点的感知力下降，网络深度过大无法训练. 图 3 为循环神经网络模型的结构. 由图 3 可见，X_t 是网络在 t 时刻的输入，经过网络结构处理得到 h_t，它可以进行输出，也可以传递给下一时刻. y_{t+1} 为 $t+1$ 时刻网络得到的输出结果，包含了当前时刻和所有历史时刻共同作用的结果，以此可以对时间序列建模. 其中隐含层状态计算公式为

$$h_t = f(\boldsymbol{W}_{hx}x_t + \boldsymbol{W}_{hh}h_{t-1}); \tag{1}$$

输出和隐含层状态的关系为

$$y_t = g(\boldsymbol{W}_{hy}h_t), \tag{2}$$

其中 $f(\cdot)$是隐含层激活函数，$g(\cdot)$是输出层激活函数，$\boldsymbol{W}_{hx}$，$\boldsymbol{W}_{hy}$和 $\boldsymbol{W}_{hh}$是权重矩阵.

LSTM 在 RNN 的基础上，增加了一个存储单元、一个输入门、一个输出门和一个遗忘门，如图 4 所示.

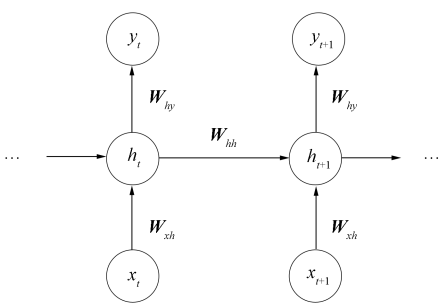


图 3 循环神经网络模型的结构

Fig. 3 Structure of recurrent neural network model

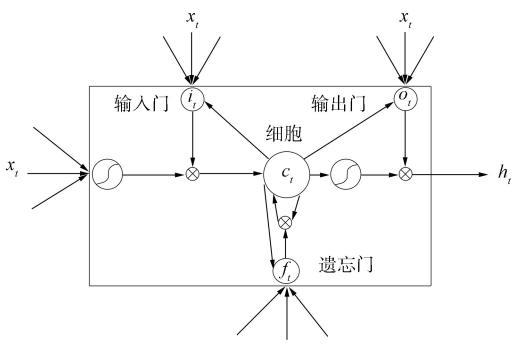


图 4 LSTM 模型的内部结构

Fig. 4 Internal structure of LSTM model

在 CTC 算法中，同一个实际音素序列可由基于帧的音素序列转换，多个基于帧的音素序列的后验概率相加可得到实际音素序列的后验概率. 通常 CTC 算法在进行计算前，都要在识别音素集合中加入空白符“—”和序列映射函数 Γ，其中空白符不但代表连续语音中的静音区间，也可作为两个音素之间的分隔符. 加入空白符后新的音素集合 $A=\{c_1,c_2,\cdots,c_m,—\}$元素数量为 $m+1$，因此需要把模型中 Softmax 层单元数调整至 $m+1$. 序列映射函数 Γ 定义为：$A_T \rightarrow A_{\leqslant T}$，即把基于帧的音素序列映射成长度较短的实际音素序列.

基于 CTC 和文本对齐的音频示例如图 5 所示. 由图 5 可见，分帧式方法需要做每个音素的标记，而 CTC 只是预测一系列的峰值和紧接着一些可能的空白符用于区分字母，分帧式的方法可能出现对齐错误，因为相邻两帧的标签概率分布图太接近，如在 dh，ax 有明显的重叠，而 CTC 却没有出现这种情况.

虽然基于 CTC 的端到端语音识别模型摒弃了传统 HMM 框架，输出可以达到音素甚至更小的状态，但其仍存在一些不足：

1）CTC 损失函数训练时间较长，优化空间相对复杂，损失函数的数值敏感较易出现波动，越接近优化目标，优化空间变得越小，易导致反优化；

2）有输出独立性的假设，对于语言模型的依赖性较强；

3）循环神经网络在时域上呈递归结构，不能同时处理多帧数据.

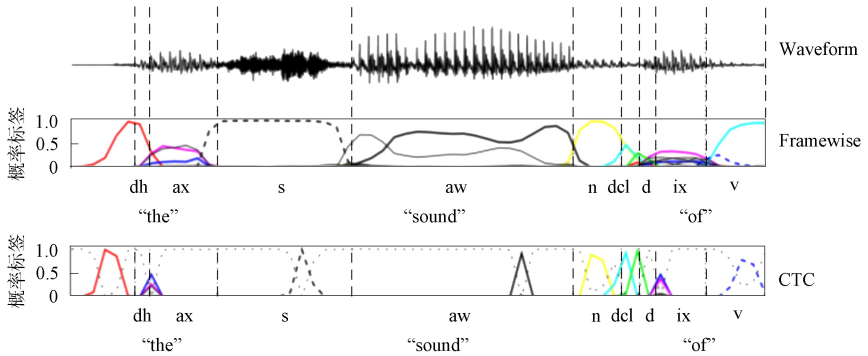


图 5　分帧式与 CTC 对语音信号进行文本对齐

Fig. 5　Text alignment of speech signals with framing and CTC

2　基于注意力机制语谱图特征提取模型

针对 CTC 的端到端语音识别模型的局限性，研究者已提出在端到端模型的前端增加一个基于注意力机制的语谱图特征提取模型. 基于语谱图提取声学特征相当于对原始语音非线性特征进行压缩变换，相比常规 MFCC 等特征维度更低，同时对背景噪声、信道和说话人的鲁棒性和区分度也更强. 利用端到端可以实现直接建模，将特征提取模型拼接到识别模型的前端，对融合后的识别模型进行联合训练，并微调参数，可有效提升识别准确率.

2.1　利用语谱图提取发音特征

语音帧长通常是 20～50 ms，为准确获取说话人音质、音色的信息，选择单个完整音节作为语谱图的最小单位，时间为 50～200 ms. 考虑通常自然发音的频率范围在 100～3 000 Hz，设置纵坐标的频率标尺范围. 下面用一段语音的语谱图进行分析.

图 6 为一个灰度窄带语谱图，灰度值深浅渐变的像素值可精确反映不同时间点频率分布的动态变化. 语音信号在低频处分布较多，高频处分布较少，且含有许多不同的花纹，如横杠、竖直条和乱纹等，其中横杠与时间轴平行，表示共振峰，深色条纹呈周期性均匀分布，从它对应的纵轴频率和宽度可确定相应的共振峰频率和带宽. 共振峰的个数、走向以及带宽都是重要的特征参数. 为更清晰直观地说明语谱图特征参数的优越性. 将语音信号的各特征参数进行对比，结果列于表 1.

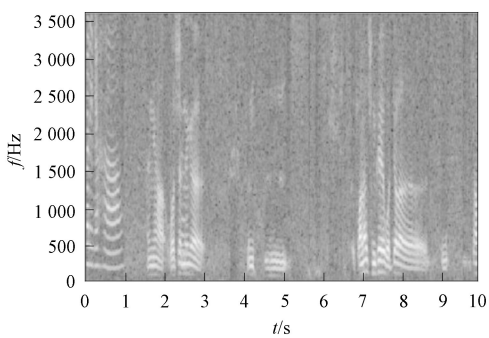


图 6　语音信号语谱图

Fig. 6　Spectrogram of speech signal

表 1　语音信号各特征参数对比

Table 1　Comparison of each characteristic parameter of speech signal

特征参数	特征提取容易程度	语音信息量	抗噪性和稳定性	特征参数	特征提取容易程度	语音信息量	抗噪性和稳定性
MFCC 参数	容易	多	弱	共振峰	困难	多	强
线性预测(LPCC)参数	容易	少	弱	语谱图	容易	多	强

现在大多数语音识别系统都采用 MFCC 参数，但它的抗噪性和稳定性较差，因此本文采用在语谱图上提取特征，具有较强的稳定性和鲁棒性，能进一步提升语音识别的准确率.

2.2 注意力机制模型

注意力机制模拟人类视觉机制忽略无关特征，将注意力分配在特征序列的部分相关区域，从特征序列中提取有效特征. 在序列到序列模型中，将特征序列$(h_1,h_2,\cdots,h_T)$映射成一个目标向量$\boldsymbol{ct}_o$，计算输出序列的元素y_o. 为便于后端模式分类，将序列中压缩后的重要信息变换为固定长度向量，再将$\boldsymbol{ct}_o$作为注意力子网络后的解码网络输入，依次计算出循环神经网络隐含层状态序列$(s_1,s_2,\cdots,s_O)$和输出序列$(y_1,y_2,\cdots,y_O)$. 注意力模型结构如图7所示.

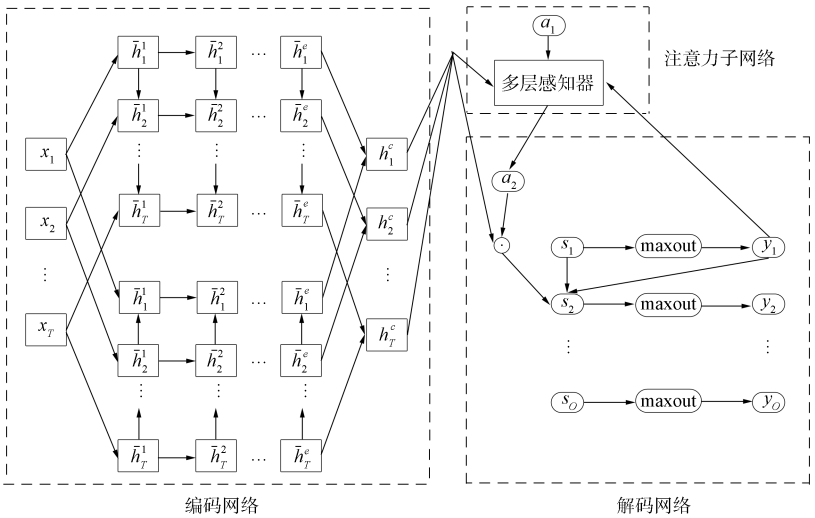


图7 注意力模型结构

Fig. 7 Structure of attention model

输出序列位置$o\in\{1,2,\cdots,O\}$所对应的目标向量为$\boldsymbol{ct}_o$，根据注意力机制其计算过程如下. 输出序列位置$o-1$的隐含层状态为s_{o-1}，首先计算其与t时刻的特征关联度：

$$e_{o,t}=a(s_{o-1},h_t), \tag{3}$$

其中$a(\cdot)$为只含一个隐含层的多层感知机. 注意力子网络表示为

$$e_{o,t}=\boldsymbol{\omega}^{\mathrm{T}}\tanh(\boldsymbol{W}[s_{o-1},h_t]+\boldsymbol{b}), \tag{4}$$

其中$\boldsymbol{W}$为输入层到隐含层的权重矩阵，$\boldsymbol{b}$为偏置向量，$\boldsymbol{\omega}$为隐含层到输出层的权重矩阵.

将对所有时刻特征的关联度进行指数归一化后的数值作为注意力系数：

$$\alpha_{o,t}=\frac{\exp\{e_{o,t}\}}{\sum_{t=1}^{T}\exp\{e_{o,t}\}}, \tag{5}$$

其中$\alpha_{o,t}$为权重，针对全部时刻的特征加权求和，利用注意力机制，得到输出序列位置o对应的目标向量$\boldsymbol{ct}_o$为

$$\boldsymbol{ct}_o=\sum_{t=1}^{T}\alpha_{o,t}h_t. \tag{6}$$

采用注意力机制计算目标向量的过程可记为

$$\boldsymbol{ct}_o=\mathrm{attention}((h_1,h_2,\cdots,h_T),s_{o-1}). \tag{7}$$

2.3 基于语谱图特征提取模型的训练

由于模型中注意力子网络和解码网络的连接为环状结构，所有变量均可导，因此可针对目标函数进行最优化训练模型参数.

数据集$\{(x_1,x_2,\cdots,x_{T_n}),(p_1,p_2,\cdots,p_{O_n})\mid n\in[1,N],\ n\in\mathbb{Z}\}$包含$N$张语谱图，利用梯度下降法求解下式中目标函数的最小值：

$$\hat{\boldsymbol{\theta}}=\arg\min_{\boldsymbol{\theta}}\frac{1}{N}\sum_{n-1}^{N}[-\log P(p_1^n,p_2^n,\cdots,p_{O_n}^n\mid x_1^n,x_2^n,\cdots,x_{T_n}^n,\boldsymbol{\theta})], \tag{8}$$

其中 n 为语谱图编号，$p_1^n,p_2^n,\cdots,p_{O_n}^n$ 为正确音素标记序列，$x_1^n,x_2^n,\cdots,x_T^n$ 为语音的特征序列，$\boldsymbol{\theta}$ 为模型中全部的权重矩阵和偏置向量.

考虑对序列进行建模，针对单段语音的序列计算方式为

$$P(p_1^n,p_2^n,\cdots,p_O^n\mid x_1^n,x_2^n,\cdots,x_T^n,q)=P(p_1^n\mid x_1^n,x_2^n,\cdots,x_{T_n}^n,q)P(p_2^n\mid x_1^n,x_2^n,\cdots,x_{T_n}^n,q)\times$$
$$P(p_O^n\mid x_1^n,x_2^n,\cdots,x_{T_n}^n,q)=$$
$$y_1^{p_1}y_2^{p_2}\cdots y_o^{p_o}\cdots y_{O_n}^{p_O}y_{O_{n+1}}^{\langle\mathrm{eos}\rangle},\tag{9}$$

其中 $y_o^{p_o}(o\in\{1,2,\cdots,O\})$表示模型解码网络生成输出序列中第 o 个输出向量中表示音素 p_o 出现的概率，$y_{O_{n+1}}^{\langle\mathrm{eos}\rangle}$ 表示输出序列第 O_{n+1}个输出向量中表示序列终止符〈eos〉出现概率的分量.

3　基于语谱图特征提取模型的语音识别

3.1　匹配相似度定义

引入更灵活的方式比较两个序列之间的相似度：最长公共子串(longest common substring，LCS)和编辑距离匹配法(edit distance matching).

3.2　注意力机制端到端语音识别设计

在语谱图利用注意力模型提取高维特征，在基于 CTC 的端到端模型上进行语音识别. 步骤如下：

1) 样本输入. 基于自适应门限的分形维数算法对音频进行语音端点检测，然后转化成灰度语谱图 $X_i(i=1,2,\cdots,n)$，作为系统的输入.

2) 训练基于注意力的模型. 首先引入不同的函数和计算机制，根据输入序列 X 和目标序列 P，计算两者的相似性或者相关性，点积为

$$\mathrm{Similarity}(X,P)=X\cdot P,\tag{10}$$

余弦相似性为

$$\mathrm{Similarity}(X,P)=\frac{X\cdot P}{\|X\|\cdot\|P\|},\tag{11}$$

MLP 网络为

$$\mathrm{MLP}(X,P).\tag{12}$$

引入 Softmax 的计算方式对第一步的得分进行值转换，采用如下公式计算：

$$a_i=\mathrm{Softmax}(\mathrm{Sim}_i)=\frac{\mathrm{e}^{\mathrm{Sim}_i}}{\sum_{j=1}^{L_x}\mathrm{e}^{\mathrm{Sim}_j}}.\tag{13}$$

计算特征 Value_i 对应的权重系数，然后加权求和即可得到 Attention 数值：

$$\mathrm{Attention}(X,P)=\sum_{i=1}^{L_X}a_i\cdot\mathrm{Value}_i.\tag{14}$$

3.3　注意力模型提取语谱图特征

由于注意力输出序列未知，同时考虑注意力模型中解码网络位置输出对计算结果的影响，基于带序列终止符的 BeamSearch 算法在注意力模型解码网络的输出中，将负概率最低的序列作为输出. 按照序列的长度逐步枚举解码结果，并将其存储在一个容量为 beam_size 序列集合中. 根据集合中元素，在进行每次搜索时序列输入和解码网络将序列长度拓展 1 位，能同时得到多个候选结果和对应的得分. 从中可筛选出最好的 beam_ size 个序列保留在集合中，根据此流程循环运行.

3.4　基于 CTC 端到端模型的训练和识别

在注意力模型中从语谱图上得到高维特征，输入到 LSTM 神经网络，继续经过一层全连接神经网络，将最后一个维度映射作为所有可能的类标. 同时数据会进入 CTC 模块和计算损失函数，经反向传播神经网络，预测输入数据所对应的序列.

综上，基于注意力机制的端到端语音识别算法的系统框图如图 8 所示.

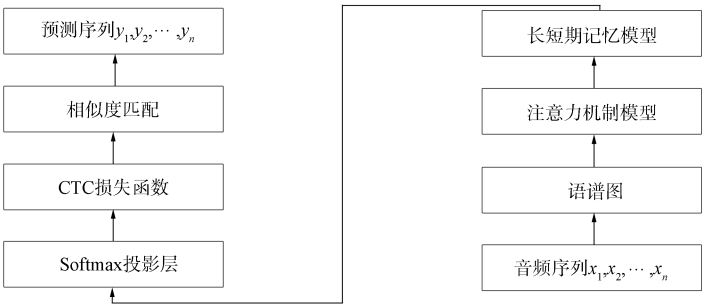


图8 基于注意力机制的端到端语音识别系统框图

Fig. 8 Flow chart of end-to-end speech recognition system based on attention mechanism

4 实验分析

4.1 实验配置

在PyCharm软件TensorFlow框架下进行语音识别实验. 使用的语料库是thchs-30语音库，该数据库设计的目的是作为863数据库的补充，尽可能提高中文发音的涵盖率.

thchs-30语音库是使用单个碳粒麦克风在安静的办公室环境下录制的，总时长约30 h，参与录音的人员为会说流利普通话的大学生，设置采样频率为16 kHz，其文本选取自大容量的新闻和文章，全部为女声. 表2列出了该数据集包含的全部内容. 其中开发集的作用是调整参数、选择特征，与训练集进行交叉验证. 训练和测试的目标数据可分为词和音素两类. 表3列出了训练thchs语料库的一些基本信息.

表2 thchs语料库包含的全部内容

Table 2 All contents contained in thchs corpus

数据集	音频时长/h	句子数/个	词数/个	数据集	音频时长/h	句子数/个	词数/个
训练集	25	10 000	198 252	测试集	6.15	2 495	49 085
开发集	2.14	893	17 743				

表3 thchs语料库的基本信息

Table 3 Basic information of thchs corpus

加载训练样本	样本数	词汇表大小	最长句子字数	最长语音/s
2022-10-18	8 911	2 666	75	673

在语音识别中，常用的评价指标除有识别的准确率，还有词错误率(word error rate，WER). 为保持识别的词序列和目标序列的一致性，需要进行替换、删除或插入等操作，错误率用WER表示，其计算公式为

$$\mathrm{WER}=\frac{\mathrm{S}+\mathrm{D}+\mathrm{I}}{N}\times 100\%, \tag{15}$$

其中S表示替换，D表示删除，I表示插入，N为单词数目.

语音识别结构如图9所示，分别提取MFCC特征，用CNN和Attention在语谱图上提取特征，再分别输入到HMM，RNN和LSTM+CTC模型中进行语音识别，对比分析，以验证改进算法的有效性.

4.2 基于MFCC特征的语音识别模型分析

选取thchs语料库中一段名为“A2_46.wav”的音频文件，该音频内容为“换句话说，一个氯原子只能和一个钠原子结合，而两个氯原子才能和一个镁原子化合.”音频时长为8 s，提取的时域波形如图10所示.

对音频进行端点检测，由于音频中含有多段短语音段，用红色竖线标记过于杂乱. 因此对检测出的短语音段，用红色覆盖，静音部分用灰色标记，得到的语音波形如图11所示.

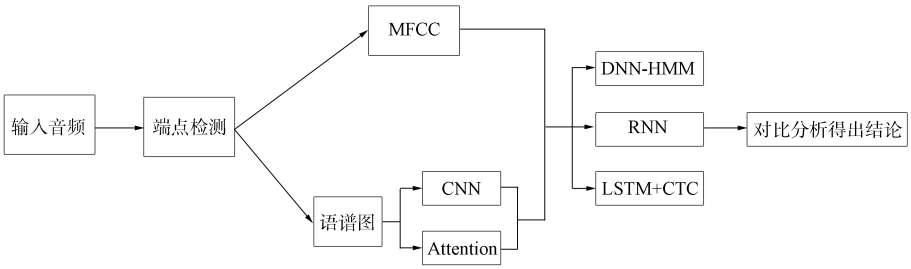


图 9　语音识别结构

Fig. 9　Structure of speech recognition

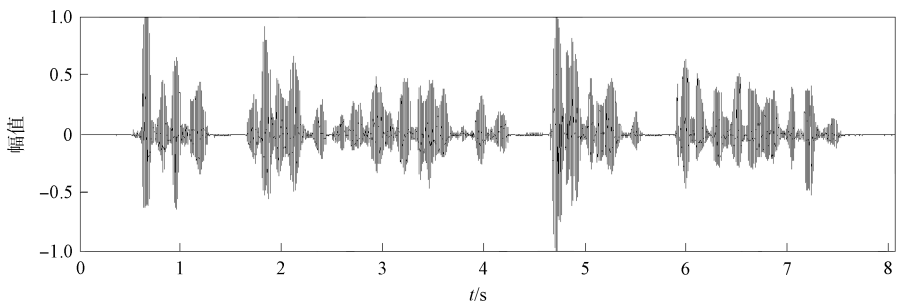


图 10　语音波形

Fig. 10　Speech waveform

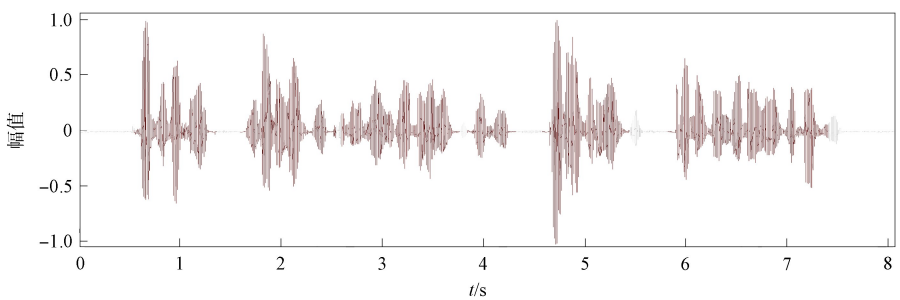


图 11　端点检测的语音波形

Fig. 11　Speech waveform for endpoint detection

提取特征维度为 12 维的语音 MFCC 特征，并对其逐帧提取. 各语音的长度不同，所得特征矩阵的大小也不同，例如当一段语音的帧数为 n，则对应提取的特征是一个 $n\times12$ 的特征矩阵，因此无法统一模型输入的尺度. 因此，计算全部帧的 MFCC 向量均值，并将其作为整段语音的 MFCC 特征，按列求取特征矩阵的平均值，归一化后可得到 1×12 的 MFCC 特征向量. 将 MFCC 特征分别输入到 DNN-HMM，RNN 和 LSTM＋CTC 三个模型中，对比平均迭代周期和识别错误率，以证明 LSTM＋CTC 模型的优越性.

采用三音素的状态作为 DNN 网络的标签信息. 整个网络包括 1 个输入层、6 个隐含层和 1 个输出层. Tanh 函数作为隐含层的激活函数，输出层分类用 Softmax 函数，以最小化交叉熵为目标函数，权重学习率为 0.001. 整个 RNN 网络包括 1 层输入层，输入层的节点数为 300；3 个隐含层每层有 2 048 个节点；一个输出层，对应 36 016 个节点. Sigmoid 函数作为隐含层的激活函数，输出层分类用 Softmax 函数. 采用最小化交叉熵作为目标函数，学习率初始值为 0.001.

针对 LSTM 单元训练神经网络的隐含层，设置状态维度为 300，LSTM 堆叠的层数为 4，隐藏节点为 40 个，1 个全连接层. CTC 为目标损失函数，学习率为 0.001.

表 4 列出了基于 MFCC 特征的语音识别模型对比结果. 由表 4 可见，LSTM＋CTC 语音识别模型在训练时间和识别性能上都优于其他模型，平均迭代周期降低到 37 min，识别的错误率仅为 7.5%.

表4 基于MFCC特征的语音识别模型对比结果

Table 4 Comparison results of speech recognition model based on MFCC feature

语音识别模型	平均迭代周期/min	词错误率(WER)/%
DNN-HMM	70	18.1
RNN	53	11.2
LSTM+CTC	37	7.5

4.3 基于语谱图特征的语音识别模型分析

继续采用“A2_46.wav”音频，转化成语谱图如图12所示. 图12为灰度窄带语谱图，通过灰度值渐变的像素值可精确反应不同时间点频率分布的动态变化情况. 图13为灰度统计直方图的实例及原图. 由图13可得到各种像素所占的比例，同时可对应语谱图中的频率信息. 因为语音信号频率本身具有高频部分不明显、低频处明显的特征，并且图中显示低频部分颜色较深，且颜色的覆盖率也相对较大，这与人类的发音特点一致.

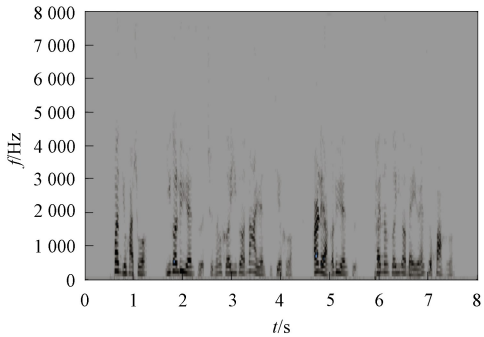


图12 “A2_46.wav”音频的语谱图

Fig. 12 Spectrogram of “A2_46.wav” audio

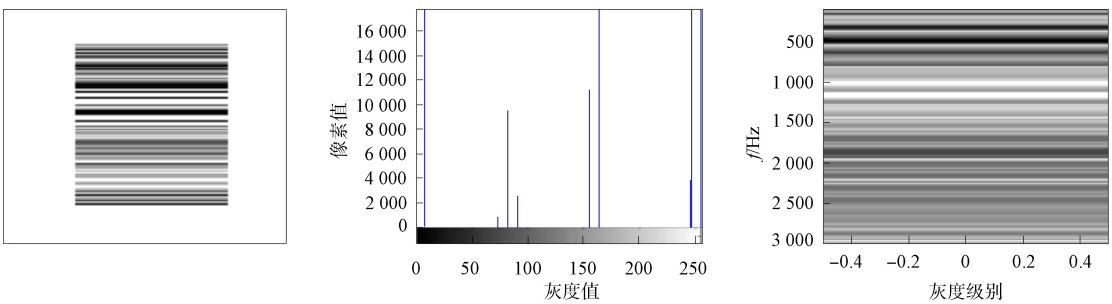


图13 灰度统计直方图的实例及原图

Fig. 13 Example of histogram of grayscale statistics and original drawings

将语谱图分别输入到CNN网络和注意力模型中，再拼接在LSTM+CTC模型中，对比并证明注意力机制的优越性. 针对CNN网络中的每个卷积层，采用3×3的小卷积核，并在多个卷积层后加上池化层，增强了CNN的表达能力. 但它与注意力网络相比，不能很好地表达语音的长时相关性，鲁棒性也较差. 图14为两种网络模型识别错误率的对比. 由图14可见，用CNN在语谱图上提取特征得到识别错误率平均约为4.9%，而基于注意力模型得到的识别错误率约维持在4.0%，识别效果得到了有效提升.

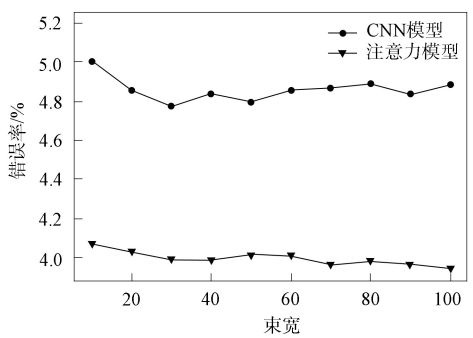


图14 两种网络模型识别错误率的对比

Fig. 14 Comparison of recognition error rates of two network models

4.4 实验结果对比分析

针对下面选取的9段中文音频文件分别进行识别分析验证. 进行端点检测后，基于注意力机制的语音识别模型的识别结果列于表5. 由表5可见，只有第4段和第7段语音识别结果出现明显的局部错误. 基于注意力机制的端到端的语音识别模型识别的错误率可降低到约5%. 表6统计了不同特征在不同模型下语音识别的词错误率.

表 5　9 段中文音频的识别结果

Table 5　Recognition results of 9 Chinese audio

序号	音频名称	音频内容	识别结果	识别准确度
1	A2_46.wav	换句话说，一个氯原子只能和一个钠原子结合，而二个氯原子才能和一个镁原子化合	换句话说 一个 氯 原子 只能 和 一个 钠 原子 结合 而 二个 氯 原子 才能和 一个 镁 原子 化合	内容准确
2	A2_135.wav	一定还有别样的生活，有城市、有音乐、有遥望远方事物的等待，有令人狂喜的发现	一定 还有 别样 的 生活有 城市 有 音乐 有 遥望远方 事物 的 等待 有 令人 狂喜 的 发现	内容准确
3	A4_221.wav	韩国的基本目标是射箭 3 块金牌，柔道 3 块金牌，羽毛球 2 块金牌以及举重等 12 块金牌	韩国 的 基本 目标 是 射箭 3 块 金牌 柔道 3 块 金牌 羽毛球 2 块 金牌 以及 举重 等 12 块 金牌	内容准确
4	A6_19.wav	现年 58 岁的切尔诺梅尔金是从 1992 年 12 月开始担任俄罗斯政府总理职务的	现年(5)18 岁(的) 切尔 诺 梅尔 金 是 从 1992 年 12 力 开战 担任 俄罗斯 政理 总 总 职务 的	内容部分准确
5	A7_53.wav	于是黑白颠倒，鸡鸭鱼肉受冷落，野菜窝头成了餐桌的宠儿	于是 黑白 颠倒 鸡鸭 鱼肉 受 冷落 野菜 窝头 成了 餐 桌上 的 宠儿	内容大致准确
6	A7_243.wav	有 353 户人家在洱海网箱养鱼 85 亩，产鱼 400 多吨，平均亩产 5 450 公斤	有 353 户 人家 在 洱海 网箱 养鱼 85 亩 产 鱼 400 多 吨 平均亩产 达到 5 450 公斤	内容大致准确
7	A8_249.wav	每日下午三四点钟，陪得志大海的游泳池畔太阳伞下坐坐，吹海风，欣赏游泳人的风姿	每日 下午(三四点) 钟 陪 得志 到 大海(的) 我 游泳池 畔 太阳伞 坐(坐) 求 吹吹(海风)，欣赏 游泳 人 的 风姿	内容部分准确
8	A12_91.wav	西安东、南、北城门前有许多宣传牌，挂一块一年要花 2 万元，我们也挂了一块	西 安东(南) 北 城(门)前门 有 许多 宣传牌 挂一块 牌 一年 要 花 2 万元 我们 也 挂了 一块	内容大致准确
9	A22_117.wav	我曾足迹遍布海内外，但我始终萦绕不能忘怀的却是这生我养我的中原故土	我 曾 足迹 遍 及 海内外 但 使 我 始终萦 不能 忘怀 的 却是 这 生 养 我 的 中原 故土	内容大致准确

表 6　不同特征在不同模型下语音识别的错误率

Table 6　Error rates of speech recognition for different features under different models　%

特征	DNN-HMM	RNN	LSTM＋CTC
MFCC	18.1	11.2	9.5
语谱图＋CNN	15.4	10.6	5.4
语谱图＋Attention	12.9	9.3	4.2

由表 6 可见：从特征的角度，用注意力机制提取语谱图得到的错误率要低于 MFCC 特征和利用 CNN 提取语谱图特征的错误率；从模型的角度，LSTM＋CTC 语音识别模型明显要优于 DNN-HMM 和 RNN 模型. 因此，基于注意力机制的 LSTM＋CTC 模型能有效降低识别错误率，缩短了训练时间.

综上所述，本文以 thchs 语料库为样本集，提取了 MFCC 特征，用 CNN 和 Attention 在语谱图上提取特征，通过输入到 HMM，RNN 和 LSTM＋CTC 模型的识别分析对比，基于注意力机制的 LSTM＋CTC 模型能同时弥补其他两种模型的缺陷，缩短了训练周期，并有效提高了语音识别的准确率.

参　考　文　献

[1] 王建荣，张句，路文焕. 机器人自身噪声环境下的自动语音识别 [J]. 清华大学学报(自然科学版)，2017，57(2)：153-157. (WANG J R，ZHANG J，LU W H. Automatic Speech Recognition with Robot Noise [J]. Journal of Tsinghua University (Science and Technology)，2017，57(2)：153-157.)

[2] 姚庆安，张鑫，刘力鸣，等. 融合注意力机制和多尺度特征的图像语义分割 [J]. 吉林大学学报(理学版)，2022，

60(6): 1383-1390. (YAO Q A, ZHANG X, LIU L M, et al. Image Semantic Segmentation Based on Fusing Attention Mechanism and Multi-scale Features [J]. Journal of Jilin University (Science Edition), 2022, 60(6): 1383-1390.)

[3] 苟鹏程, 宗群. 车载语音识别及控制系统的设计与实现 [J]. 计算机应用与软件, 2017, 34(5): 129-134. (GOU P C, ZONG Q. Design and Realization of Vehicle Speech Recognition and Control System [J]. Computer Applications and Software, 2017, 34(5): 129-134.)

[4] HAN L H, WANG B, DUAN S F. Development of Voice Activity Detection Technology [J]. Application Research of Computers, 2010, 27(4): 1220-1226.

[5] 金超, 龚铖, 李辉. 语音识别中神经网络声学模型的说话人自适应研究 [J]. 计算机应用与软件, 2018, 35(2): 200-205. (JIN C, GONG C, LI H. Speaker Adaptation Research of Neural Network a Coustic Model in Speech Recognition [J]. Computer Applications and Software, 2018, 35(2): 200-205.)

[6] 陶勇, 朱媛. 基于深度神经网络声学模型及其改进的语音识别研究 [J]. 吉林大学学报(理学版), 2021, 59(4): 539-546. (TAO Y, ZHU Y. Research on Speech Recognition Based on Deep Neural Network Acoustic Model and Its Improvement [J]. Journal of Jilin University (Science Edition), 2021, 59(4): 539-546.)

[7] CAO D Y, GAO X, GAO L. An Improved Endpoint Detection Algorithm Based on MFCC Cosine Value [J]. Wireless Personal Communications, 2017, 95(3): 2073-2090.

[8] KHELIFA O M, ELHADJ Y M, YAHYA M. Constructing Accurate and Robust HMM/GMM Models for an Arabic Speech Recognition System [J]. International Journal of Speech Technology, 2017, 20: 937-949.

[9] 陈爱月, 徐波, 申子健. 基于高斯混合模型及TIMIT语音库的说话人身份识别 [J]. 信息通信, 2017, 7: 51-52. (CHEN A Y, XU B, SHEN Z J. Speaker Recognition Based on Gaussian Mixture Model and TIMIT Speech Database [J]. Information and Communication, 2017, 7: 51-52.)

[10] LANJEWAR R B, MATHURKAR S, PATEL N. Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and *K*-Nearest Neighbor (KNN) Techniques [J]. Procedia Computer Science, 2015, 49(1): 50-57.

[11] CUI X D, JING H, CHIEN J T. Multi-view and Multi-objective Semi-supervised Learning for HMM-Based Automatic Speech Recognition [J]. IEEE Transactions on Audio Speech & Language Processing, 2012, 20(7): 1923-1935.

[12] MAAS A L, QI P, XIE Z, et al. Building DNN Acoustic Models for Large Vocabulary Speech Recognition [J]. Computer Speech & Language, 2017, 41(7): 195-213.

[13] 邵玉斌, 陈亮, 龙华, 等. 基于改进GFCC特征参数的广播音频语种识别 [J]. 吉林大学学报(理学版), 2022, 60(2): 417-424. (SHAO Y B, CHEN L, LONG H, et al. Broadcast Audio Language Identification Based on Improved GFCC Feature Parameters [J]. Journal of Jilin University (Science Edition), 2022, 60(2): 417-424.)

[14] HINTON G, DENG L, DONG Y, 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.

[15] YI J Y, WEN Z Q, TAO J H, et al. CTC Regularized Model Adaptation for Improving LSTM RNN Based Multi-accent Mandarin Speech Recognition [J]. Journal of Signal Processing Systems, 2017, 90(2): 1-13.

[16] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition [J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[17] 张宇, 张鹏远, 颜永红. 基于注意力LSTM和多任务学习的远场语音识别 [J]. 清华大学学报(自然科学版), 2018, 58(3): 249-253. (ZHANG Y, ZHANG P Y, YAN Y H. Long Short-Term Memory with Attention and Multitask Learning for Distant Speech Recognition [J]. Journal of Tsinghua University (Science and Technology), 2018, 58(3): 249-253.)

[18] 龙星延, 屈丹, 张文林. 结合瓶颈特征的注意力声学模型 [J]. 计算机科学, 2019, 46(1): 260-264. (LONG X Y, QU D, ZHANG W L. Attention-Based Acoustic Model with Bottleneck Features [J]. Journal of Computer Science, 2019, 46(1): 260-264.)

(责任编辑: 韩 啸)