引用格式：金浩，朱文博，段志奎，等. 基于注意力机制的 TDNN-LSTM 模型及应用[J]. 声学技术，2021，**40**(4)：508-514. [JIN Hao, ZHU Wenbo, DUAN Zhikui, et al. Attention mechanism based TDNN-LSTM model and its application[J]. Technical Acoustics, 2021, **40**(4): 508-514.] **DOI**: 10.16300/j.cnki.1000-3630.2021.04.011

基于注意力机制的 TDNN-LSTM 模型及应用

金 浩，朱文博，段志奎，陈建文，李艾园
(佛山科学技术学院，广东佛山 528000)

摘要：在大数据规模下，基于深度学习的语音识别技术已经相当成熟，但在小样本资源下，由于特征信息的关联性有限，模型的上下文信息建模能力不足从而导致识别率不高。针对此问题，提出了一种嵌入注意力机制层(Attention Mechanism)的时延神经网络(Time Delay Neural Network, TDNN)结合长短时记忆递归(Long Short Term Memory, LSTM)神经网络的时序预测声学模型，即 TLSTM-Attention，有效地融合了具有重要信息的粗细粒度特征以提高上下文信息建模能力。通过速度扰动技术扩增数据，结合说话人声道信息特征以及无词格最大互信息训练准则，选取不同输入特征、模型结构及节点个数进行对比实验。实验结果表明，该模型相比于基线模型，词错误率降低了 3.37 个百分点。

关键词：小样本；注意力机制；时延神经网络；长短时记忆递归网络
中图分类号：H107 **文献标志码**：A **文章编号**：1000-3630(2021)-04-0508-07

Attention mechanism based TDNN-LSTM model and its application

JIN Hao, ZHU Wenbo, DUAN Zhikui, CHEN Jianwen, LI Aiyuan
(*Foshan University*, *Foshan* 528000, *Guangdong*, *China*)

Abstract: With the development of big data, speech recognition technology based on deep learning has been quite mature, but under small sample resources, due to the limited relevance of feature information, the modeling ability of contextual information of the model is insufficient, which leads to low recognition rate. To solve this problem, a timing prediction acoustic model (named TLSTM-Attention), which consists of a time delay neural network (TDNN) embedded by attention mechanism layer (Attention) and a long and short time memory (LSTM) recurrent neural network, is proposed in this paper. This model can effectively fuse the coarse and fine particle features with important information to improve the modeling ability of context information. By using the velocity perturbation technique to amplify the data and combining the speaker's channel information features and the lattice-free maximum mutual information training criteria, and by selecting different input features, model structures and numbers of nodes, a series of comparative experiments are conducted. The experimental results show that compared with the baseline model, the word error rate of the model is reduced by 3.77 percentage points.

Key words: small sample; attention mechanism; time delay neural network (TDNN); long and short time memory recurrent network

0 引 言

在近十几年中，深度学习技术一直保持着飞速发展的状态，极大地推动了语音识别技术的不断发展。在大数据条件下，无论是传统语音识别技术、基于深度学习的语音识别技术，还是端到端语音识别技术、都已经相当成熟，各种商业化产品也相应落地实现。但在小样本数据下，由于系统对时序数据的上下文建模能力不足，导致语音识别效果仍不理想。为解决此问题，研究者们主要从丰富数据特征及优化建模方法等方向做了相应的研究。

在丰富数据特征方面，Saon 等[1]引入了身份认证矢量(Identity Authentication Vector, IVA) i-vector，它能够有效表征说话人和信道信息，并能提高低资源条件下语音识别的准确率[2]；Ghahremani 等[3]提出一种结合 i-vector 特征的音调提取算法，被证明能够丰富语音数据特征，提高模型上下文建模能力；Gupta 等将基于 i-vector 矢量的说话人自适应算

收稿日期：2020-11-08；**修回日期**：2021-01-23
基金项目：广东省基础与应用基础研究基金项目支持-粤佛联合基金项目支持(2019A1515110273)
作者简介：金浩(1995－)，男，辽宁铁岭人，硕士研究生，研究方向为语音识别。
通信作者：朱文博，E-mail: zhuwenbo@fosu.edu.cn

法成功应用在广播音频转录上[4]，得到了良好的识别率。

在优化建模方法方面，有研究者提出了不同于传统高斯混合建模(Gaussian Mixture Model, GMM)的深度神经网络建模方法，如时延神经网络[5](Time Delay Nerual Network, TDNN)、长短时记忆网络[6](Long Short Term Memory, LSTM)以及端到端[7]等基于深度学习的建模方法。但由于训练数据匮乏，时序特征重要程度的差异性在模型上难以体现，导致模型对时序数据的上下文建模能力仍不足。例如时延神经网络在对帧级特征信息进行时序拼接时，如果不能区分重要信息和非重要信息，则容易出现无效信息被重复计算和有效信息丢失的问题[8]。并且对 LSTM 来说，虽然其对长距离时序数据有一定的信息挖掘能力，但是当输入的时序数据包含的无效信息过长，训练模型时则会出现不稳定性和梯度消失的问题，导致模型捕捉时序依赖能力降低[9]。

由于注意力模型[10]具有使模型能够在有限资源下关注最有效的信息的优点，所以被广泛应用于机器翻译、图像识别等各种不同类型的深度学习任务中，具有较大的研发潜力。近年来，注意力机制开始被用于语音识别领域，Povey 等[11]和 Carrasco 等[12]提出一种受限的自我注意力机制层并应用于语音识别领域，有效提高了英语的语音识别率。有研究者提出了一种含有注意力模块的卷积神经网络，成功用在语音情感识别上，并取得了不错的效果[13]。Yang 等结合注意力机制能够关注有效信息的优点，提出了一种应用在情感分类上的注意力特征增强网络[14]。

因此，本文通过联合 TDNN 和 LSTM 声学模型并嵌入注意力机制，借助速度扰乱技术扩增数据同时引入说话人声道信息特征，并结合基于区分性训练的无词格的最大互信息训练准则来训练模型。针对小样本马来西亚方言数据集进行实验，深入分析不同输入特征、隐藏节点个数以及注意力结构对模型效果的影响。实验表明，本文提出的基于注意力机制的 TDNN-LSTM 混合模型整体表现良好，相比于基线模型词错率降低了 3.37 个百分点。

1　基于注意力机制的 TDNN-LSTM 模型架构

本文提出了一种基于注意力机制的 TDNN-LSTM 混合声学模型，即 TLSTM-Attention 模型，如图 1 所示。利用注意力机制处理特征重要度的差异，有效结合粗细粒度特征，充分提高 LSTM 捕捉时序特征依赖的能力，并结合无词格最大互信息训练准则[15](Lattice Free Maximum Mutual Information, LFMMI)对模型进行训练，以增强模型上下文的建模能力。

1.1　模型整体架构

TLSTM-Attention 模型共有 8 层结构组成，主要由时延神经网络模块、长短时记忆网络模块以及注意力模块三个部分组成。采用时延神经网络模块和长短时记忆网络模块以及注意力模块的交叉连接。该模型整体架构如图 1 所示，TDNN 模块对原始输入数据进行时序拼接，以多尺度方式提取更丰富的局部短序列特征。注意力层对多尺度特征进行差异性筛选，既能增强有效信息的利用率，又能减少计算参数、精简模型。LSTM 以注意力层抽取出带有重要程度差异性的粗粒度特征作为输入，再度抽取具有长依赖关系的细粒度特征，实现粗细粒度特征有效融合，能够在一定程度上避免因 LSTM 层步长过长，造成记忆丢失和梯度弥散的问题。最后结合注意力机制能够关注有效信息的优点，用于对输出结果进行分类以及预测。

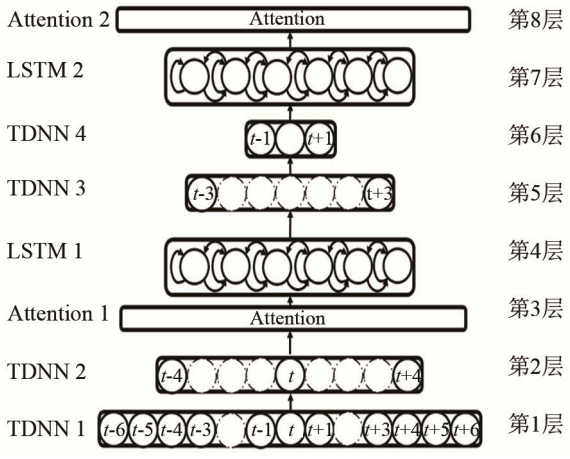

图 1　TLSTM-Attention 模型架构

Fig.1　Structural diagram of TLSTM-Attention model

1.2　时延神经网络模块

1.2.1　时延神经网络原理

时延神经网络是一种多层的前馈神经网络，网络结构如图 2 所示。与传统前馈神经网络采用全连接的层连方式不同，TDNN 将每层的输出都与前后若干时刻的输出拼接起来，相较于传统只能处理帧窗口中固定长度信息的前馈神经网络，TDNN 的输出不仅与当前时刻有关，还与前后若干时刻有关，因此能够有效描述上下层节点之间的时序关系，并且表现出更强的数据上下文信息建模能力和能够

适应动态时域特征变化的优势。每层隐藏层都可以和任意时刻输出进行拼接，体现了 TDNN 可以对更长的历史信息进行建模的能力。但是这也意味着 TDNN 在每一个时间步长，隐藏层的激活函数都会被计算一次，并且 TDNN 相邻节点之间的变化很小，可能包含了大量的无效信息，在训练的过程中容易出现反复计算且保留无效信息的问题。

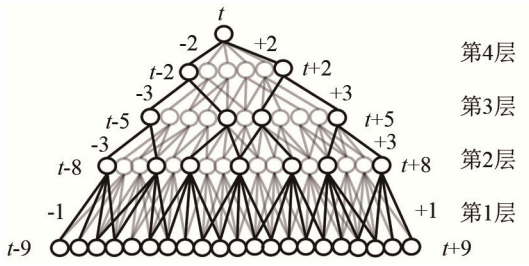

图 2 时延神经网络结构

Fig.2 The structure of time delay neural network

1.2.2 时延神经网络模块设计

TLSTM-Attention 模型共包含 4 个 TDNN 层，分别命名为 TDNN 1,2,3,4。TDNN 中通过设置每层参数来表示每一层输出拼接的时间步长以及依赖关系。使用{–*m*, *n*}表示将当前帧的历史第 *m* 帧、当前帧的未来第 *n* 帧和当前帧拼接在一起作为下一个网络层的输入，0 表示最后一层没有拼接的输入。假设 *t* 表示当前帧，在 TDNN 1 层，模型将原始数据的时序信号转换成特定的帧级特征向量作为输入，将帧进行{*t*–2, *t*–1, 0, *t*+1, *t*+2}时序拼接，处理后作为下一个隐藏层的输入。在 TDNN2 层，将上一层拼接后的帧进行{*t*–3, *t*–2, *t*–1,0, *t*+1, *t*+2, *t*+3}拼接，并将学习到的过去 5 帧及未来 5 帧的信息分类后作为注意力层的输入。在 TDNN 3 处，将对处理后赋予了注意力特性的帧级特征信息进行{*t*–3, *t*–2, *t*–1, 0, *t*+1, *t*+2, *t*+3}拼接，作为下一层的输入，在 TDNN 4 处，将帧进行{*t*–1, 0, *t*+1}拼接，拼接后的时序特征包含了过去及未来的 9 帧信息，作为下一个隐藏层的输入。

1.3 注意力层模块

1.3.1 注意力机制原理

注意力机制(Attention Mechanism)被认为是一种资源分配的机制，在深度神经网络的结构设计中，注意力机制所关注的资源就是权重参数。注意力机制总体可分为硬注意力机制与软注意力机制。硬注意力机制的核心是通过直接限制输入来达到聚焦有效信息的能力，但是对于时序数据的特性，直接限制输入则意味着数据完整性的缺失，将直接导致模型的上下文建模能力不足。与硬注意力机制不同，软注意力机制通过对特征信息进行注意力打分，并将其作为特征信息的权重参数，从而实现对特征信息差异性的关注。对于具有时序信息的语音数据，其中的特征信息包含的重要程度存在差异，重要的显著特征往往会包含更多的关联信息，对建模的影响程度更大。基于上述原理，本文将软注意力机制引入 TDNN-LSTM 模型中，为所有输入特征逐个加权进行打分，将归一化的平均打分作为特征的权重参数，有效地实现了粗细粒度特征的结合。

1.3.2 注意力层模块设计

TLSTM-Attention 模型嵌入了两层注意力层，分别设在整体结构的第三层和第八层。第一层注意力层，由前端 TDNN 2 网络进行时序拼接后的输出，作为注意力层的输入。首先计算每个帧级特征的标量分数 e_t，其表达式为

$$\boldsymbol{e}_t=\boldsymbol{v}^{\mathrm{T}}\boldsymbol{F}(\boldsymbol{W}\boldsymbol{h}_t+\boldsymbol{b})+\boldsymbol{k} \tag{1}$$

其中：$\boldsymbol{h}_t$ 为前端 TDNN 网络的输出，$\boldsymbol{v}^{\mathrm{T}}$ 为转移概率参数矩阵，$\boldsymbol{W}$ 为帧级特征的权重，$\boldsymbol{b}$ 为特征输出偏置项，$\boldsymbol{k}$ 为特征标量分数偏置项，$F(\cdot)$ 为 ReLU 激活函数。为减少异常数据影响，将得到的标量分数 $\boldsymbol{e}_t$ 进行归一化处理得到 α_t，其表达式为

$$\alpha_t=\frac{\exp(\boldsymbol{e}_t)}{\sum_t^{\mathrm{T}}\exp(\boldsymbol{e}_t)} \tag{2}$$

式中：T 为单个音频输入时长。归一化后的分数作为池化层的权重参数并计算平均权重向量 $\tilde{\boldsymbol{\mu}}$，其表达式为

$$\tilde{\boldsymbol{\mu}}=\sum_t^T\alpha_t\boldsymbol{h}_t \tag{3}$$

计算得到的平均权重向量系数与帧级特征信息结合，赋予模型关注重要度更高的特征，更好地实现时间序列的粗粒度特征的提取以及对 LSTM 输入信息的优化。在模型输出前的注意力层，将包含 18 帧的帧级特征信息，简化分类及预测，有效地精简模型并提高模型训练速度。

1.4 长短时记忆网络模块

1.4.1 长短时记忆网络原理

长短时记忆网络是由循环神经网络(Recurrent Neural Network, RNN)衍生而来的时序卷积神经网络，并在隐藏层的内部作了改进，增加了三个特殊的门控结构，通过权重参数的更新来选择有效的历史信息进行传递，实现对重要信息的保留和非重要信息的过滤，内部结构如图 3 所示。相较于 RNN 能更好地从输入数据学习，获得更好的上下文建模能力并能够挖掘时间序列中的时序变化规律。

其中 $\boldsymbol{x}_t$ 为 t 时刻的输入，$\boldsymbol{l}_t$ 为 t 时刻的输出，c 为长短时记忆单元信息的状态，维持信息的传递，i 代表输入门，决定当前信息 $\boldsymbol{x}_t$ 保留多少信息给 $\boldsymbol{c}_t$；f 代表遗忘门，遗忘门结构根据具有注意力特性的特征信息，决定保存多少前一时刻的单元状态 $\boldsymbol{c}_{t-1}$；o 代表输出门，决定 t–1 时刻的隐层状态有多少传递至当前状态的输出 $\boldsymbol{l}_t$。

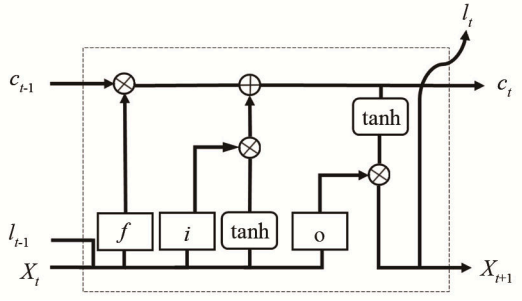

图 3　长短时记忆递归网络内部结构图
Fig.3　Internal structure of LSTM recurrent network

1.4.2 长短时记忆网络模块设计

LSTM 模块设计如图 4 所示，模型整体包含两层 LSTM，分别为 LSTM 1、LSTM 2。经过注意力层处理后的平均权重向量与特征信息结合得到 $\boldsymbol{x}_t$，作为 LSTM 1 层的输入。通过 LSTM 特有门控结构处理，对赋有注意力特征的时序特征进行长序列依赖发掘，进一步增强模型上下文信息的建模能力。设 $\sigma(\cdot)$ 表示门控 sigmoid 激活函数，$\boldsymbol{W}_x$ 为与输入层连接的权重参数矩阵，$\boldsymbol{W}_c$ 为与记忆单元连接的权重参数矩阵，上述流程对应公式为

$$\chi_t=\tilde{\boldsymbol{\mu}}\cdot\boldsymbol{h}_t \tag{4}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{x}_{t-1}+\boldsymbol{W}_{ci}\boldsymbol{c}_{t-1}+\boldsymbol{b}_i) \tag{5}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{l}_{t-1}+\boldsymbol{W}_{cf}\boldsymbol{c}_{t-1}+\boldsymbol{b}_f) \tag{6}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\tanh(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{l}_{t-1}) \tag{7}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{l}_{t-1}+\boldsymbol{W}_{co}\boldsymbol{c}_t+\boldsymbol{b}_o) \tag{8}$$

$$\boldsymbol{l}_t=\boldsymbol{o}_t\tanh(\boldsymbol{c}_t) \tag{9}$$

LSTM 1 通过学习前端 TDNN 网络模块的 11 帧赋予了注意力特性的特征，能够充分利用有效信息的权重比，对特征信息进行精准分类。并且通过 TDNN 4 层对特征数据进行时序拼接后，LSTM 2 层至少可以学习到上下文相关的 9 帧历史信息及 9 帧未来信息，整体提高模型上下文建模能力以及预测分类能力。

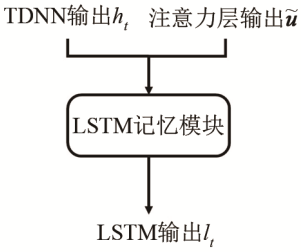

图 4　LSTM 模块设计结构
Fig.4　Structure of LSTM module

1.5　模型训练准则

本实验采用基于区分性训练的改进无词格最大互信息准则(Lattice Free Maximum Mutual Information, LFMMI)，建模单元如图 5 所示。改进的 LFMMI 准则由于降低神经网络对齐后的输出帧率，帧移从 10 ms 增加为 30 ms，因此音素状态数从 3 降为 1，用 s_p 表示，另外加上了一个用于自旋可重复 0 次或多次的空白状态 s_b。这样对于 1 帧的声学特征就要遍历整个隐马尔科夫模型(Hidden Markov Model, HMM)，相较于传统的 LFMMI[16]中 HMM 在音素状态级别建模，改进的 LFMMI，在音素级别建模，直接计算出相应的最大互信息(Maximum Mutual Information, MMI)和所有正确路径和混淆路径的后验概率。

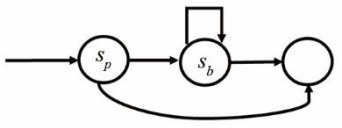

图 5　改进的 Lattice-free MMI 建模单元
Fig.5　Improved lattice-free MMI modeling unit

相比于标准语音识别系统，采用隐马尔科夫状态图(Hidden Markov, H)、音素上下文(Phone Context, C)、发音词典(Pronunciation Lexicon, L)、语言模型(Grammer Model, G)四部分有限状态转换器(Finite State Transducer, FST)组合成 HCLG 静态解码网络。改进的 LFMMI 针对小样本数据在音素级别建模，用音素语言模型(Phone Grammer Model, PGM)来代替词语言模型(Word Grammer Model, WGM)。由于小样本条件下音素个数比词个数少很多，因此 PGM 产生的 FST 图很小，最后得到的 HCP 解码网络也会小很多，P 代表 PGM，真正做到纯序列区分性训练，可以动态更新 MMI 部分的统计量并且减少模型训练时间。

2　实验设置

2.1　实验数据

实验采用的是由 Sarah Samson Juan 和 Laurent Besacier 收集的开源伊班语(IBAN)语料库。伊班语是婆罗洲的一种语言，并且是马来语和波利尼西亚语的一个分支，主要在马来西亚、加里曼丹和文莱等地普及。该语料库是由 23 个说话人录制完

成的，采样率设为 16 kHz，每个采样点进行 16 bit 量化，声道为单声道。该语料库总时长大约有 8 h，共包含 3 132 句伊班语语音数据，每句话时长约为 9 s。实验中随机选择 17 个说话人的语音数据作为训练集，6 个说话人的语音数据作为测试集。发音词典包含大概 3.7 万个单词。本文从网上的新闻演讲收集了大约 104 万个单词的文本进行 3 元语言模型训练。

2.2 语音识别系统搭建及性能指标

为避免语料库不足而产生过拟合的问题，本实验在训练集采用速度扰乱技术进行数据扩增[17]。为保证音频质量，语速调整应保持在 0.85 倍和 1.25 倍之间，因此本实验将扭曲因子参数设置为 0.9 和 1.1。每次训练期间会随机根据扭曲因子的参数，生成不同量的扭曲训练数据扩充训练集。同时由于采用速度扰乱技术后信号长度发生了变化，需要使用 GMM-HMM 系统对生成数据对齐，并将对齐后的低精度声学特征额外加入音量扰动以提取高精度声学特征，以 40 维梅尔频率倒谱系数(Mel Frequency Cepstrum Coefficient, MFCC)作为基础特征参数，同时添加说话人声道信息特征用于声学模型训练。将深度神经网络(Deep Neural Networks, DNN)模型作为基线模型，使用基于加权有限状态转换器(Weight Finite State Transducer, WFST)作为系统解码器，以 KALDI[18]为平台搭建了一个马来西亚方言语音识别系统。

每组实验在测试集上运行 3 次，以 3 次实验的平均词错误率为最终实验结果。词错误率的计算方法为

$$R_{\mathrm{WE}}=\frac{S+D+I}{T}\times 100\% \tag{10}$$

式中：S 代表替换错误词数，D 代表删除错误词数，I 代表插入错误词数，T 为句子中的总词数。R_{WE} 结果越小，表示识别性能越好。

3 实验结果及分析

3.1 不同神经网络的比较实验

本实验将 TLSTM-Attention 模型与 4 种模型进行对比实验：(1) DNN 模型包含六个隐藏层，一个输入层，一个输出层，每层节点数为 2 048 个，激活函数为 tanh。固定 15 帧上下文窗口，每帧提取 40 维 MFCC 特征，共计 600 维特征向量作为网络输入。(2) TDNN 声学模型包含六个隐藏层，一个输入层，一个输出层。每个隐藏层包含 256 个节点，激活函数为 tanh，分别采用{0}，{–1，1}，{–1，1}，{–3，3}，{–3，3}，{–3，3}配置进行时序拼接，其中{0}表示不进行时序拼接，{–1，1}表示对当前时刻的前后各一帧拼接。固定 5 帧上下文窗口，每帧提取 40 维 MFCC 特征，共计 200 维特征向量作为网络输入。(3) LSTM 声学模型包含六个隐藏层，一个输入层，一个输出层。每个隐藏层包含 256 个节点，包含 5 帧历史信息和 5 帧未来信息，后三个隐藏层为常规隐藏层，激活函数为 tanh。固定 3 帧上下文窗口，共计 120 维特征向量作为网络输入。(4) TDNN-LSTM 包含六个隐藏层，一个输入层，一个输出层。第一个隐藏层为包含 256 个节点的 TDNN，固定 5 帧上下文窗口，每帧提取 40 维 MFCC 特征，共计 200 维特征向量。第 2、4 和 6 隐藏层为包含 256 个节点的 LSTM，模块包含 5 帧历史信息和 5 帧未来信息。第三层和第五层是 TDNN 隐层，配置信息为{–3，3}。

表 1 为马来西亚方言在不同神经网络的声学模型的识别结果。从实验结果可以看出，TDNN-LSTM-Attention 得到的识别性能明显优于基线 DNN 模型，R_{WE} 从 18.20%下降到 15.06%，实验表明，基于 TDNN-LSTM-Attention 的声学模型能够有效提高模型上下文建模能力。

表 1 不同神经网络的词错误率对比结果

Table 1 Comparison of word error rates between different neural networks

模型	R_{WE}/%
DNN	18.20
TDNN	17.16
LSTM	17.36
TDNN-LSTM	17.01
TDNN-LSTM-Attention	15.06

3.2 基于注意力机制的 TDNN-LSTM 模型的不同结构比较实验

3.2.1 不同隐层个数和节点数的比较实验

在本实验中，分别对 TDNN 和 LSTM 神经网络不同隐藏层个数和节点数进行对比试验，其配置信息如表 2 所示。实验中分别设置隐藏层个数为 3、4、5 和 6，每个隐藏层包含 256 个节点。当隐藏层个数为 3 时，第 2 层为 LSTM 隐藏层；当隐藏层个数为 4 时，第 3 为 LSTM 隐藏层；当隐藏层个数为 5 时，第 3 层和第 5 层为 LSTM 隐藏层。当隐藏层个数为 6 时，第 3 层、第 6 层为 LSTM 隐藏层，其余层均为 TDNN 隐藏层。例如，使用 TDNN-LSTM-6-2 表示 TDNN-LSTM 包含 6 个隐藏层，对当前时刻前后两帧进行降采样。

表 2　不同隐层个数和节点数的词错误率对比结果
Table 2　Comparative of word error rates for different numbers of hidden layers and nodes

隐层个数	降采样节点数	R_{WE}/%
3	{−1，1}	17.16
4	{−1，1}	17.27
4	{−3，3}	17.15
5	{−3，3}	17.12
5	{−2，2}	17.05
6	{−2，2}	17.11
6	{−3，3}	17.29

实验结果如表 2 所示，其中 TDNN-LSTM 隐层数为 5 时，TDNN 降采样节点配置为{–2，2}的网络结构得到的实验结果最好，单词错误率为 17.05%，与基线 DNN 模型相比降低 1.15 个百分点。实验表明，随着隐藏层个数增加隐藏层节点数增加，单词错误率明显降低。这是因为随着层数和节点数的增加，将使 TDNN-LSTM 在训练过程中可以获得更多固定长度的时间上下文关联信息。

3.2.2　不同注意力层结构的比较实验

本实验以上面实验中表现最好的 TDNN-LSTM-5-2 模型为基准，模型基础结构不变，对注意力层的个数以及位置结构进行对比实验。实验中分别设置注意力层数为 1、2 及 3。当注意力层个数为 1 时，注意力层有两个位置结构，1-3 表示模型有 1 个注意力层结构，且位于该模型第 3 层；1-6 表示模型 1 个注意力层结构，且位于该模型第 6 层。当注意力层个数为 2 时，注意力层分别位于模型的第 3、8 层，用 2-3-8 表示。当注意力层个数为 3 时，注意力层分别位于模型的第 3、6、8 层，用 3-3-6-8 表示。

实验结果如表 3 所示，当注意力层个数为 2 时，即 Attention2-3-8 网络结构得到的实验结果最好，单词错误率为 14.83%，与基线 DNN 模型相比相对降低 3.37 个百分点。实验表明，适当嵌入注意层能够有效提高识别效果。这是因为模型中的注意力层能够关注特征的差异性，有效结合粗细粒度特征，但当注意层增加时模型将会过多的关注信息差异性，造成数据的原始性缺失进而导致识别率不佳。

表 3　注意力层的层数和位置不同的词错误率对比结果
Table 3　Comparison of word error rates for different layer numbers and positions of attention layers

注意力层数	注意力层位置结构	R_{WE}/%
1	Attention1-3	14.92
1	Attention1-6	14.95
2	Attention2-3-8	14.83
3	Attention3-3-6-8	15.07

3.3　基于 TLSTM-Attention 不同特征的比较实验

本实验以 13 维 MFCC 作为模型输入的基础特征，将基础特征进行二阶差分处理得到 26 维差分特征和 1 维的音高特征组合得到 40 维 MFCC，同时添加 100 维的 i-vector 特征作为附带特征。提取特征后对特征计算倒谱均值并在模型训练时动态进行归一化处理，减少异常特征信息数据对模型训练的影响。训练所用模型为 TDNN-LSTM-5-2-Attention2-3-8 模型，实验结果如表 4 所示。

表 4　不同声学特征的 TLSTM-Attention 模型词错误率对比结果
Table 4　Comparison of word error rates for TLSTM-Attention model with different acoustic features

声学特征	R_{WE}/%
13 维 MFCC	15.73
40 维 MFCC	14.83
40 维 MFCC+i-vector	14.52

表 4 的实验结果显示，对于基础特征来说，高维的 MFCC 能够更好地拟合基于注意力机制的 TDNN-LSTM 模型，并且基于 40 维的 MFCC 特征和 i-vector 特征组合的多输入特征，使得神经网络可以获取不同说话人特点和信道信息进行训练，比单输入特征在测试集上取得更好的识别率。能够在更长时序的语音序列建模，充分挖掘了上下文信息，从而提高模型的鲁棒性。

4　结 论

本文针对小样本资源下，模型上下文能力不足的问题，以基于注意力机制的 TDNN-LSTM 的模型为核心构建了一个马来语方言的语音识别系统，同时添加说话人声道信息特征，结合 LFFMI 训练准则，让模型在有限资源下充分对音素进行建模。实验结果表明，相比于 DNN 基线模型，基于注意力机制的 TDNN-LSTM 模型可以有效提高上下文建模能力，并且由于添加了说话人声道信息特征，在特征层面克服了用说话人无关的语音特征进行声学模型训练的不足。另外，本文的主要任务是从提高上下文建模能力角度来提高低资源下的语音识别效果，对于如何更有效提高小样本资源下语音识别的效果仍需要继续深入研究和探讨。

参　考　文　献

[1] SAON G, SOLTAU H, NAHAMOO D, et al. Speaker adaptation of neural network acoustic models using i-vectors[C]//2013 IEEE Workshop on Automatic Speech Recognition and Understanding.

Olomouc, Czech Republic. IEEE, 2013: 55-59.

[2] 蔡猛. 低资源条件下基于深度神经网络的语音识别声学建模研究[D]. 北京: 清华大学, 2015.

[3] GHAHREMANI P, BABAALI B, POVEY D, et al. A pitch extraction algorithm tuned for automatic speech recognition[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy. IEEE, 2014: 2494-2498.

[4] GUPTA V, KENNY P, OUELLET P, et al. I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy. IEEE, 2014: 6334-6338.

[5] WAIBEL A, HANAZAWA T, HINTON G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, **37**(3): 328-339.

[6] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[EB/OL]. 2014: arXiv: 1402.1128[cs.NE]. https://arxiv.org/abs/1402.1128

[7] BAHDANAU D, CHOROWSKI J, SERDYUK D, et al. End-to-end attention-based large vocabulary speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China. IEEE, 2016: 4945-4949.

[8] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural ne-twork architecture for efficient modeling of long temporal contexts[C]//Interspeech, 2015.

[9] LAI G K, CHANG W C, YANG Y M, et al. Modeling long- and short-term temporal patterns with deep neural networks[EB/OL]. 2017: arXiv: 1703.07015[cs.LG]. https://arxiv.org/abs/1703.07015

[10] VASWANI A , SHAZEER N, PARMAR N, et al. Attention Is All You Need[C]//Advances in Neural Information Processing Sys-tems. 2017: 5998-6008.

[11] POVEY D, HADIAN H, GHAHREMANI P, et al. A time-restricted self-attention layer for ASR[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada. IEEE, 2018: 5874-5878.

[12] CARRASCO M, BARBOT A. Spatial attention alters visual appearance[J]. Current Opinion in Psychology, 2019, **29**: 56-64.

[13] MIRSAMADI S, BARSOUM E, ZHANG C. Automatic speech emotion recognition using recurrent neural networks with local attention[C]//2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017: 2227-2231.

[14] YANG M, QU Q, CHEN X J, et al. Feature-enhanced attention network for target-dependent sentiment classification[J]. Neurocomputing, 2018, **307**: 91-97.

[15] POVEY D, PEDDINTI V, GALVEZ D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]//Interspeech 2016. ISCA, 2016.

[16] HADIAN H, SAMETI H, POVEY D, et al. End-to-end speech recognition using lattice-free MMI[C]//Interspeech 2018. ISCA: ISCA, 2018.

[17] KO T, PEDDINTI V, POVEY D, et al. Audio augmentation for speech recognition[C]//Interspeech, 2015.

[18] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]//Automatic Speech Recognition and Understanding(ICASRU), 2011 IEEE International Conf-erence on.IEEE, 2011: 1-4.