

引用格式: 孙杰, 王宏, 吾守尔·斯拉木. 结合注意力机制和因果卷积网络的维吾尔语方言识别[J]. 声学技术, 2020, 39(6): 697-703. [SUN Jie, WANG Hong, Wushouer Silamu. The Uyghur dialect recognition based on attention mechanism and causal convolution networks[J]. Technical Acoustics, 39(6): 697-703.] DOI: 10.16300/j.cnki.1000-3630.2020.06.008

结合注意力机制和因果卷积网络的 维吾尔语方言识别

孙 杰^{1,2}, 王 宏², 吾守尔·斯拉木^{1,2}

(1. 新疆大学信息科学与工程学院, 新疆乌鲁木齐 830046; 2. 昌吉学院, 新疆昌吉 831100)

摘要: 针对传统 x-vector 模型生成方言语音段级表示时, 未考虑不同帧级特征对方言辨识作用不一致的问题, 以及维吾尔语的黏着性特点, 提出结合注意力机制和因果卷积网络的维吾尔语方言识别方法。首先使用多层因果卷积网络实现方言语音序列建模, 然后采用空洞卷积核增大感受野扩展采样范围, 最后使用注意力池化获取方言语音段级特征。维吾尔语方言识别实验结果表明, 所提方法较标准 x-vector 模型方言识别的识别准确率提升了 23.19 个百分点。

关键词: 注意力机制; 因果卷积网络; 空洞卷积; 维吾尔语方言; 识别

中图分类号: H107

文献标识码: A

文章编号: 1000-3630(2020)-06-0697-07

The Uyghur dialect recognition based on attention mechanism and causal convolution networks

SUN Jie^{1,2}, WANG Hong², Wushouer Silamu^{1,2}

(1. College of Information Science and Engineering, Xinjiang University, Urumqi 830046, Xinjiang, China;
2. Changji University, Changji 831100, Xinjiang, China)

Abstract: Considering that different frame features have different effects on dialect recognition when the traditional x-vector model is used to generate segment representation of dialect speech, and that Uighur language is an agglutinative language, a recognition method of Uighur dialect based on attention mechanism and causal convolution network is proposed. First, the multi-layer causal volume network is used to model the speech sequence, then the dilated convolution kernel is used to expand the sampling range of the receptive field, and finally the attention pooling is used to obtain the speech segment features. The experimental results of Uyghur dialect recognition show that the accuracy of the proposed method is 23.19 percentage higher than that of the standard x-vector model.

Key words: attention mechanism; causal convolution networks; dilated convolution; Uyghur dialect; recognition

0 引 言

方言识别亦称方言分类, 属于语种识别的范畴。方言作为特定共同语的地方变体, 具有“互相通话”功能^[1], 在语言学上具有很大的相似性, 因此方言识别要比语种识别更具挑战性^[2]。

现代维吾尔语划分为 3 个方言区: 中心方言、和田方言与罗布方言。中心方言包括伊犁、乌鲁木齐、吐鲁番、哈密、喀什和塔里木土语; 和田方言由和田、墨玉、洛浦、皮山、策勒、于田和民丰七

个土语组成; 罗布方言主要是现今若羌县境内的罗布人所操土语, 由于地理位置闭塞, 保留较多古语。目前, 关于维吾尔语方言识别的研究较少, 仅文献 [3] 提出了基于长短时记忆神经网络-统一背景空间 (Long Short Term Memory-Universal Background Model, LSTM-UBM) 的维吾尔语方言识别方法。很多研究者从语言学的角度对维吾尔语方言进行了辨识: 依据动词后是否缀接-mix 判断南部方言与北部方言^[4]; 把是否存在元音的唇部和谐作为区别罗布方言和中心方言的标准^[5-7], 但是这些都属于“口耳之学”, 很难用计算机进行处理。

主流的方言识别技术是建立在高斯通用背景模型 (Gaussian Mixture Model-Universal Background Model, GMM-UBM)^[8] 和联合因子分析技术 (Joint Factor Analysis, JFA)^[9] 上的全变量子空间建模方法 (Total Variability, TV), 它用一个低维度 (通常是 400

收稿日期: 2020-04-16; 修回日期: 2020-05-29

基金项目: 国家自然科学基金(U1435215、U1603262、61433012、201704041014)

作者简介: 孙杰(1976—), 男, 安徽阜阳人, 博士研究生, 副教授, 研究方向为方言识别、语音识别和说话人识别。

通讯作者: 吾守尔·斯拉木, E-mail: wushouer@xju.edu.cn

维或 600 维)的 i-vector 矢量表征方言^[10], 取得较好识别效果, 但是 i-vector 对训练和测试方言语音的时长、噪声和信道差异都很敏感, 对训练数据的要求较为严苛。随着神经网络在说话人识别方面取得的巨大成功, 研究者从特征域和模型域分别提出了深度瓶颈特征 (Deep Bottleneck Feature, DBF)^[11-12]和神经网络通用背景^[13-14]的方言识别 TV 模型。由于使用区分性的 DNN 网络获取不同方言语音的音素差异, 剔除了与音素无关的噪声干扰, 因而提取的方言语种 i-vector 更具鉴别性, 其识别性能好于传统的 GMM-UBM 生成性模型, 但是模型训练需要大量的标注语料, 对于方言识别而言代价较大。近期, 基于词嵌入技术的神经网络在自然语言处理方面取得良好效果^[15], 受此启发 Snyder 等学者提出了 x-vector 模型^[16-18], 其实质是一种端到端 (End-to-End) 的方言识别模型, 相关实验表明长时语音条件下的方言识别准确率高于 i-vector, 且与 DNN-UBM 相当。然而, x-vector 模型用池化层将帧级别特征转换为句子级特征时, 对语音段的帧特征计算了简单算数平均数, 即对不同帧采用相同的权重, 但是, 实际语音中每帧信号对方言语种的辨识贡献度是不一致的。

本文在对维吾尔语方言进行识别时, 做了两方面的创新工作: 一是在 x-vector 模型的池化层引入了注意力机制, 对引起方言差异的语音帧在计算段级特征时给予较大的权重; 二是采用因果卷积网络获取维吾尔语方言语音帧的因果关系, 实验结果表明, 融合了两种技术的 x-vector 系统的方言识别效果进一步提升。

1 注意力机制

1.1 注意力机制涵义

注意力机制实质是模仿人类观察物体时大脑视觉系统处理信息的方式, 即将有限的注意力放在众多信息中的重要区域, 挑选出关键信息, 抑制或忽略其他无关信息^[19-20]。方言语种识别任务中应用注意力机制的目标是挑选出与当前任务最为相关和最为关键的信息, 进而增强识别效果。

1.2 多头注意力模型

多头注意力 (Multi-head Attention) 机制^[21]使用多个查询状态 $\mathbf{Q}=\{q_1, q_2, \dots, q_M\}$, 并行地从输入特征中遴选出多个关键信息, 用不同的视角观察不同区域的信息, 并将每个单头注意力进行拼接, 最终构成多头注意力的输出值, 其模型如图 1 所示。

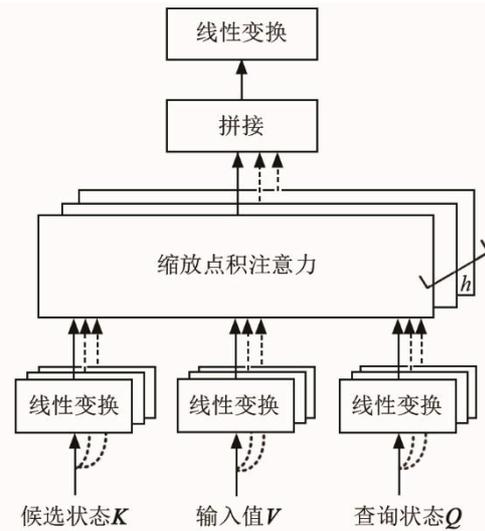


图 1 多头注意力模型
Fig.1 Multi-head attention model

计算多头注意力时, 首先对查询状态 \mathbf{Q} 、候选状态 \mathbf{K} 和输入值 \mathbf{V} 进行线性变换, 其变换表达式为

$$H_i(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \text{att}(\mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V, \mathbf{Q}\mathbf{W}_i^Q) \quad (1)$$

其中: $\text{att}(\cdot)$ 表示注意力得分计算函数; \mathbf{W}_i^K 、 \mathbf{W}_i^V 和 \mathbf{W}_i^Q 表示第 i 个输入的线性变换矩阵, 每个头的线性变换参数不共享, 也即每次对 \mathbf{Q} 、 \mathbf{K} 和 \mathbf{V} 进行线性变换的参数都不一样, 目的是为获取不同的注意力。然后再将每个头值输入缩放点积注意力模块, 计算各自的注意力, 并将所有输出进行拼接, 其表达式为

$$\text{att}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = H_1 \oplus H_2 \oplus \dots \oplus H_h \quad (2)$$

其中: h 表示注意力的计算次数, 属于超参数。简单拼接后得到的多头注意力内部结构松散, 对其实施线性变换可以使最终得到的多头注意力更加紧凑。另外, 每个单头注意力张成一个特征子空间, 多头注意力机制的优势就是从不同注意力张成的多个子空间中学习到互为补充的有用信息。

1.3 自注意力模型

自注意力机制^[22]是对多头注意力技术的进一步改进, 它更加注重内部信息的学习, 充分挖掘输入数据各部分的依赖性关系, 适合提取语音段内部各帧之间的相互关系。自注意力机制本质就是用输入特征 $\mathbf{X}=\{X_1, X_2, \dots, X_N\}$ 同时表示 \mathbf{Q} 、 \mathbf{K} 和 \mathbf{V} , 并且令 $\mathbf{Q}=\mathbf{K}=\mathbf{V}=\mathbf{X}$, 进而达到自我关注和自我挖掘。自注意力机制经常与多头注意力机制相结合使用, 其结合公式为

$$\text{multi_H}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \text{multi_H}(\mathbf{X}, \mathbf{X}, \mathbf{X}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_i\mathbf{W}_i^T\mathbf{X}^T}{\sqrt{d_x}}\right)\mathbf{X} \oplus \dots \oplus \text{softmax}\left(\frac{\mathbf{X}\mathbf{W}_h\mathbf{W}_h^T\mathbf{X}^T}{\sqrt{d_x}}\right)\mathbf{X} \quad (3)$$

其中： d_x 表示输入特征的维度，在式(3)中主要作用是防止过拟合； $\text{softmax}(\cdot)$ 为归一化指数函数，具体计算公式为 $\alpha(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}}$ 。

2 因果卷积

2.1 因果卷积网络

卷积神经网络(Convolutional Neural Networks, CNN)通常是在空间维度处理信息，例如图像识别中对像素信息的处理，因此使用 CNN 处理语音信号时，通常将语音信号转换为语谱图的形式再进行处理。为使 CNN 直接处理时序特征的语音信号，可以使用一维卷积网络，并通过增加卷积层数，同时配合一定的门控激活函数，实现对时序信号的“因果卷积”处理，门控激活为

$$y = \tanh(W_f * x) \cdot \sigma(W_g * x) \tag{4}$$

其中： x 、 y 分别表示神经元的输入和输出； W_f 、 W_g 分别表示卷积权重系数； $*$ 代表卷积操作； $\sigma(\cdot)$ 表示 sigmoid 函数。这种多层的一维卷积网络称之为因果卷积网络(Causal Convolution Networks, CCN)^[23]。输入层的序列数据通过因果卷积网络映射为标记序列，即 $f_{\text{CNN}}: X^{N+1} \rightarrow Y^{N+1}$ ，从而实现序列数据建模。

2.2 空洞卷积

因果卷积通过增加网络层数以及增大卷积核的尺寸实现长时序预测，同时也带来梯度弥散、模型复杂和拟合效果不佳等问题，针对此问题通过引入空洞卷积(Dilated Convolution)^[24-25]采样的方式进一步优化因果卷积网络。所谓空洞卷积就是在卷积核中加入空洞，从而增大感受野，扩展了观察数据的范围。空洞卷积采样可以表示为

$$F(s) = (x *_{d} f)(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i} \tag{5}$$

其中： s 表示输入序列的长度； f 为卷积核； d 为空洞因子； $*$ 表示卷积操作； k 为卷积核尺寸； $s-d \cdot i$ 卷积的历史跨度； $*_{d}$ 表示带有 d 个空洞因子的卷积操作。

3 结合注意力和因果卷积的方言识别模型

3.1 方言识别模型

首先，尽管基于 x-vector 模型的语种识别系统

取得了一定的识别效果，但是对维吾尔语这种黏着语而言，构成词语的词干和词缀的作用不同，与词干、词缀对应的所有帧的权重应该也不相同。其次，不同方言、同一个音素会有不同的音位变体，这些音位变体会引起语音的较大差异，因此在计算均值时可以为对应的帧特征分配更大的权重。另外，维吾尔语方言语音变化呈现出显著的先后关系。由于这三方面原因，本文使用自注意力机制和因果卷积网络对传统的 x-vector 语种识别模型进行改进，图 2 描绘了该系统架构，其中 diac_i 为第 i 类方言的缩写。该模型首先使用因果卷积网络提取方言语音的帧级特征，使用注意力模块计算帧级特征对应的权重，其次结合获取的权值在池化层对语音段计算加权统计信息，后使用一维卷积层获取方言的 x-vector 辨别矢量，最后使用 softmax 层输出方言种类的后验概率 P 。图 2 中，1,1 CONV1D@128 表示卷积核为 1×1 、个数为 128 的一维卷积。

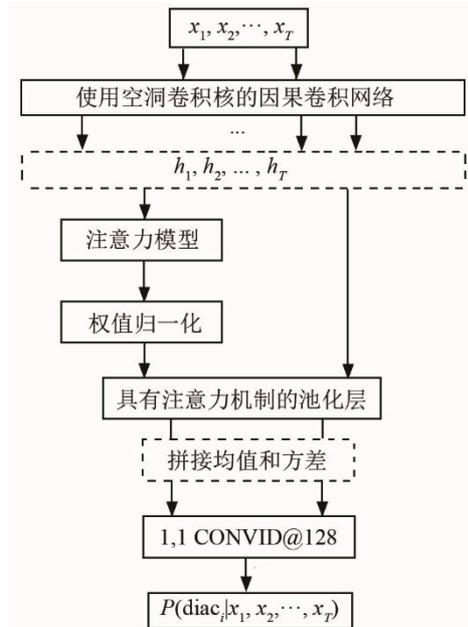


图 2 结合注意力和因果卷积的方言识别模型
Fig.2 Dialect recognition model combining attention mechanism and causal convolution networks

维吾尔语方言语音结构中元音和谐对辨识不同方言具有重要作用，元音和谐现象在维吾尔语中很常见，并且元音和谐发生在音素与音素之间，表现为前一个音素中的音位影响后一个音素中元音的发音。通常一个音素对应一个或几个语音帧，因此可以认为语音中前后帧之间具有较强的因果关系。图 3 为使用带有空洞卷积核的因果网络提取和田方言语音特征的过程示意图，音频语义为“vRvmqigE bardigan poyiz Kaysi wogzaldin maN-do(去乌鲁木齐的火车从哪个车站发车)?”

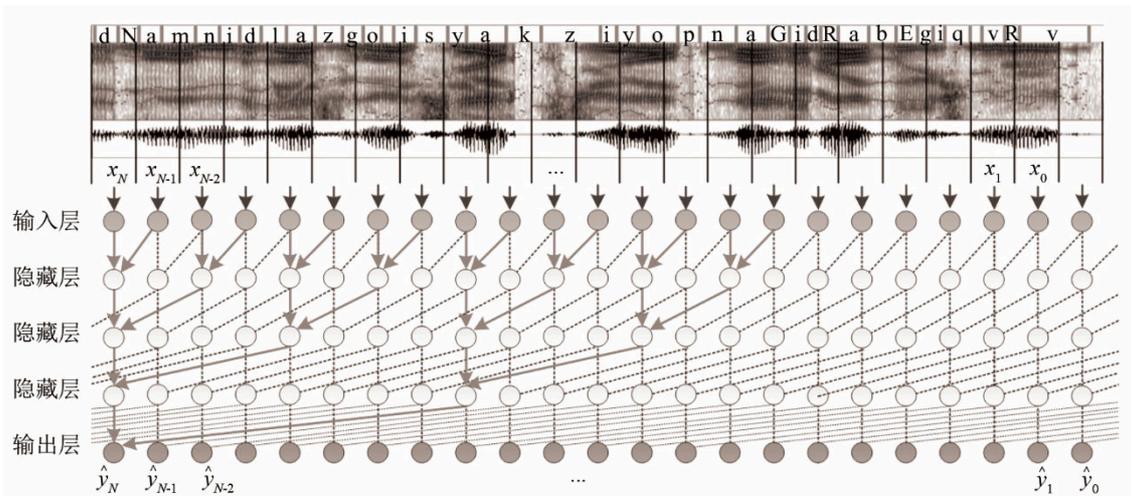


图3 因果网络提取和田方言语音特征示意图

Fig.3 Diagram of extracting speech features of Hotan dialect by causal networks

在“乌鲁木齐”(标注为vRvmqi)一词中,前元音/v/和前元音/i/发生和谐,根据黏着语的特性:(词根不断缀接其它音素),可以认为维吾尔语方言语音每一帧之间都具有因果关系。从生成模型的角度,这一段语音信号帧的联合概率可以表示为

$$p(x) = \prod_{t=0}^T p(x_t | x_0, \dots, x_{t-1}) \quad (6)$$

其中: $\mathbf{x} = \{x_0, \dots, x_T\}$ 表示语音段的帧信号,而使用空洞卷积核的因果卷积网络,通过考虑历史语音帧的因果卷积及门控激活函数的点积运算,可近似计算式(6)的联合概率。

3.2 结合注意力机制的池化层

使用注意力机制的维吾尔语方言识别模型中,采用了一个受限玻尔兹曼机计算查询状态和候选状态的相似性,其计算公式为

$$A = \text{softmax}[f(\mathbf{H}^T \mathbf{W})] \quad (7)$$

其中: $A = [\alpha_1, \alpha_2, \dots, \alpha_T]$ 表示方言语音帧注意力权重矩阵; $\mathbf{H} = [h_1, h_2, \dots, h_T]$ 表示由因果卷积网络隐藏层的输出值组成的矩阵,它同时作为注意力网络的输入值,其维度为 $d_h \times T$,而 d_h 是 h_t 的维度; \mathbf{W} 为受限玻尔兹曼机的权值矩阵, $f(\cdot)$ 是 ReLU 激活函数。通过式(7)即可得到方言帧级特征对应的权值,然后池化层就可以计算加权统计量,计算公式为

$$\boldsymbol{\mu} = \sum_{t=1}^T \alpha_t \mathbf{h}_t, \quad t=1, \dots, T \quad (8)$$

$$\boldsymbol{\sigma} = \sqrt{\sum_{t=1}^T \alpha_t \mathbf{h}_t \cdot \mathbf{h}_t - \boldsymbol{\mu} \cdot \boldsymbol{\mu}}, \quad t=1, \dots, T \quad (9)$$

另外,为从维吾尔语方言语音帧特征构成的不同

子空间中提取信息,注意力模块使用了多头注意力机制,平行地对因果卷积网络的输出帧特征重复计算注意力值,因此得到多组方言语音段的均值和标准差,所以需要对其进行拼接形成方言语音段的最终表示。

4 实验

4.1 方言数据集和评测指标

本文研究的维吾尔语方言识别目前在国际和国内均未有公开的标准测试数据集。清华大学公开的维吾尔语语音数据集 THUYG-20 只提供了说话人信息和文本标注信息^[26],并没有说明方言语种类别,因此只能用于维吾尔语说话识别和自动语音识别任务。本文实验使用的方言语种数据集是由新疆大学多语种信息技术重点实验室创建,三种方言语料均为手机录音的朗读式语句,采样频率为 16 KHz,采样位数 16 bits,语音时长为 5~30 s,保存格式为 WAV 类型。其中中心方言与和田方言男女发音人各为 41 人,每人朗读 120 句,而罗布方言女性发音人比例略大于男性发音人,分别为 49 人和 33 人,每人朗读 120 句,三种方言的语料各有 9 840 句。

方言和语种识别性能评测中经常也会使用方言语种识别正确率^[27]作为评测指标,即被正确分类语音段的百分比 P_{acc} :

$$P_{\text{acc}} = \frac{N_c}{N_t} \times 100 \quad (10)$$

其中: N_t 表示测试方言语音段的总数; N_c 表示被正确分类的语音段总数。

4.2 系统设计

为验证本文所提算法的有效性，按照文献[16]中的配置搭建基于 TDNN 的 x-vector 方言识别基线系统，称之为 TDNN-xvec。为探索因果卷积网络提取方言语音帧级特征对方言识别效果的影响，设计一个基于因果卷积的 x-vector 方言识别系统，其卷积层同样设置为 5 层，第一层至第五层空洞因子分别设置为 $d=1$ 、 $d=2$ 、 $d=4$ 、 $d=8$ 和 $d=16$ ，为了保持输入序列和标注序列的一致性，所有卷积层的滤波器数量均相同，并称为 CCN-xvec。另外，为验证结合注意力机制和因果卷积网络维吾尔语方言的识别效果，在 CCN-xvec 系统中加入注意力模块，获取权重系数的受限玻尔兹曼机的输入和输出神经元数量均与因果卷积层的输出帧数保持一致，将该系统称之为 CCN-att-xvec。最后，对基线 TDNN-xvec 方言识别系统加入注意力模块，将该系统称 TDNN-att-xvec。

4.3 模型训练

结合注意力机制和因果卷积网络的 x-vector 方言模型训练流程如图 4 所示，为充分利用有限方言语料，并增强模型的稳定性和可靠性，首先在从维吾尔语三种方言语料中挑选训练集和测试集时，采用十折交叉验证法，同时保证训练集数据不出现在验证集中。然后采用 G.723.1 技术规范^[28]对语音进行端点检测和倒谱均值减处理，分帧后每帧提取 30 维 MFCC 系数，同时计算其一阶和二阶差分系数，考虑到基线系统 TDNN-xvec 的第一层组合了当前时刻的前后两帧 $\{t-2, t-1, t, t+1, t+2\}$ 作为输入，CCN-att-xvec 同样使用 5 帧共计 450 维参数作为 CCN 的输入。采用有监督方式对神经网络训练，训练目标是最小化负对数似然函数，损失函数使用交叉熵函数。采用反向传播和梯度下降算法更新网络参数，参数更新公式为

$$v_t = \beta_1 v_{t-1} + (1 - \beta_1) d\theta_t \tag{11}$$

$$m_t = \beta_2 m_{t-1} + (1 - \beta_2) d\theta_t^2 \tag{12}$$

$$\theta_t = \theta_{t-1} - \eta \frac{v_t}{\sqrt{m_t + \varepsilon}} \tag{13}$$

式(11)~(13)中： v_t 、 m_t 和 θ_t 分别表示 t 时刻的冲量、

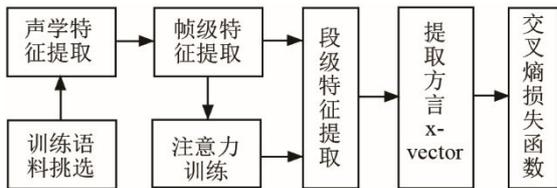


图 4 方言识别系统训练过程
Fig.4 Training process of dialect recognition system

光滑系数和网络参数； β_1 和 β_2 为超参数； η 为学习率； ε 为保持数值稳定的参数，初始学习率设置为 0.01，共计迭代 40 个周期。

4.4 实验结果

4.4.1 实验一

实验一对比了不同滤波器数量时的 TDNN-xvec 和 CCN-xvec 模型的方言识别性能。将 TDNN-xvec 和 CCN-xvec 模型中卷积层中卷积核数量分别设置为 64、128、256、512，实验中所有网络的卷积核尺度固定为 7。图 5 为方言识别结果，从图中可以看出，TDNN-xvec 方言识别系统随着卷积核数量的增加，识别正确率不断降低，两者之间呈现明显的负相关性。而 CCN-xvec 方言识别系统卷积核从 64 个增加到 128 个时，识别正确率最高，正确率为 85.80%，继续增加卷积核数量，方言识别正确率缓慢降低。值得注意的是，具有不同卷积核数量的 CCN-xvec 模型的方言识别正确率都高于对应的 TDNN-xvec 模型。这说明在 x-vector 框架下，因果卷积网络比 TDNN 更能够提取到具有辨别性的维吾尔语方言语音帧级特征。

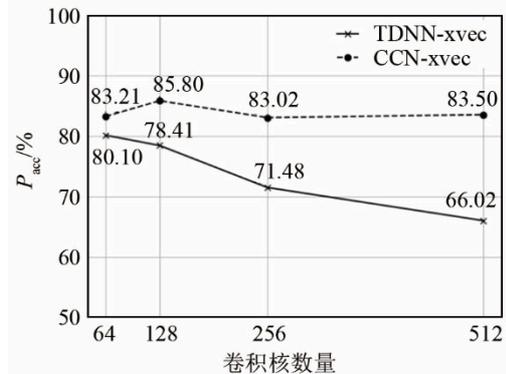


图 5 不同尺度卷积核模型的方言识别正确率
Fig.5 Correctness rate of dialect recognition based on the convolution kernel model with different scales

4.4.2 实验二

实验二对比了添加注意力机制后 TDNN-att-xvec 和 CCN-att-xvec 模型的方言识别性能。实验时将计算帧权重的受限玻尔兹曼机的神经元与卷积核的数量设置为一致，加入注意机制后模型的维吾尔语方言识别结果如图 6 所示。从实验结果来看，一个明显的结论就是 CCN-att-xvec 系统的识别性能始终优于 TDNN-att-xvec 系统的识别性能，并且两个模型在卷积核数量为 128 个时性能最优。另外将实验二与实验一进行对比可以发现两点：(1) 加入注意力机制的 CCN-att-xvec 比没有融合注意力机制的 CCN-xvec 识别正确率总体上有一定程度提升，识

别正确率最大提升 6.19 个百分点,说明注意机制与因果卷积网络结合有助于提高维吾尔语方言识别率;(2) 加入注意力机制的 TDNN-xvec 系统在卷积核数量为 64 时,识别正确率低于未使用注意力机制的系统,而在卷积核数量为 128、256 和 512 个时方言识别正确率又都高于未使用注意机制的系统。形成这种结果的原因是:对于 TDNN-xvec 系统而言,当训练语料数量一定、语音段长度一定时,滤波器数量的增多意味着网络最终输出的帧级特征维度越大,这些特征中存在大量重叠的上下文信息,这些信息简单拼合在一起会相互干扰,滤波器数量越多干扰越大,识别正确率降低程度越大,而使用注意力机制后,就相当于对这些高维冗余信息进行了主成分分析(Principal Component Analysis, PCA)^[29-30],做了降维处理,维度越大反而提取到的有用信息越多,识别效果越好。

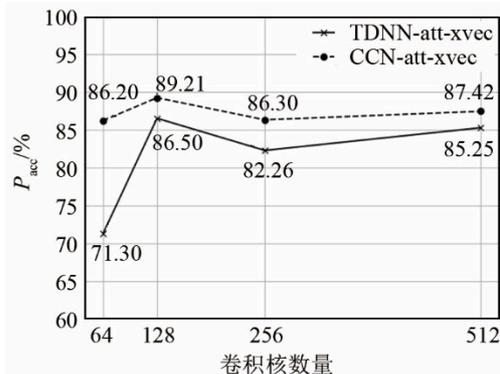


图 6 加入注意力机制后模型的方言识别性能

Fig.6 Dialect recognition performance of the model with attention mechanism

4.4.3 实验三

实验三为验证本文所提方法对其他方言识别的有效性,分别使用 TDNN-xvec 和 CCN-att-xvec 模型对长沙话、南昌话和上海话(简称湘、赣、吴)三种方言进行识别,识别结果如图 7 所示。三种方言数据来自科大讯飞方言挑战赛公开的部分方言语料,每种方言训练数据为 6 600 条,同样使用十折交叉验证法划分训练集与测试集,且保证训练集中无测试集中的发音人语料。图 7 中 TDNN-xvec 和 CCN-att-xvec 分别表示两种模型对长沙话、南昌话和上海话的识别结果。从识别结果可以看出,在不同卷积核个数情况下 CCN-att-xvec 模型方言识别正确率均比 TDNN-xvec 模型的要高。说明注意力机制的因果卷积网络相对传统 x-vector 模型,不仅对维吾尔语有较高的识别正确率,而且对汉语方言也有较好识别性能,模型有较强的泛化性。通过与实验二进行对比,可以发现 CNN-att-xvec 模型

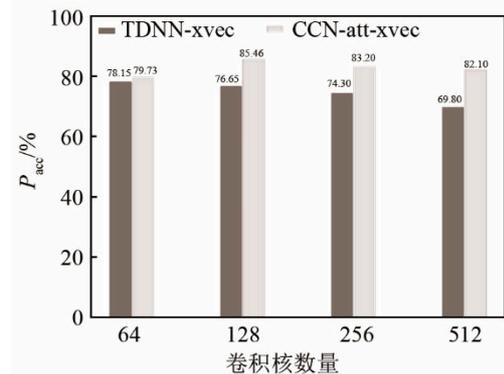


图 7 融合注意力机制模型的湘、赣、吴方言识别正确率

Fig.7 Recognition rate of Xiang, Gan and Wu dialects based on attention mechanism model

对湘、赣、吴和上海话三种方言的识别正确率略低于维吾尔语三种方言的正确率,这可能和维吾尔语的黏着性有关。

5 结 论

本文提出了结合注意力机制和因果卷积网络的 x-vector 维吾尔语方言识别模型。实验结果表明,使用空洞卷积采样技术的因果卷积网络提取的维吾尔语方言帧级特征比 TDNN 的更具辨识性,且加入注意机制后,基于 TDNN 的 x-vector 模型和基于 CCN 的 x-vector 模型方言识别性能均有相当程度的提升,特别是后者的维吾尔语方言识别正确率比标准 x-vector 模型最高提升了 23.19 个百分点。

参 考 文 献

- [1] 李小凡, 项梦冰. 汉语方言学基础教程[M]. 北京: 北京大学出版社, 2010.
- [2] SHON S, ALI A, GLASS J. Convolutional neural networks and language embeddings for end-to-end dialect recognition[J]. Odyssey 2018 The Speaker and Language Recognition Workshop. 2018.
- [3] 孙杰, 吾守尔·斯拉木, 热依曼·吐尔逊, 等. 维吾尔语方言识别及相关声学分析[J]. 声学学报, 2019, 44(6): 1083-1092.
SUN Jie, Wushouer Silamu, Reyiman Turson, et al. Acoustic analysis and language recognition of Uyghur[J]. Acta Acustica, 2019, 44(6): 1083-1092.
- [4] 阿米娜·阿帕鲁娃. 论现代维吾尔语方言及民族文学语言的基础方言和标准音[J]. 民族语文, 1980, 6(2): 24-30.
- [5] 高士杰. 维吾尔语和田方言的主要特点[J]. 中央民族学院学报, 1984, 34(2): 69-77.
- [6] 司马义·阿不都热依木. 濒危维吾尔语方言个案研究——以罗布泊方言的调查分析为例[J]. 语文学刊, 2018, 38(5): 37-46.
Esmael Abdurehim, Endangered Uyghur dialect: a case study of the lopnor dialect[J]. Journal of Language and Literature Studies, 2018, 38(5): 37-46.
- [7] 司马义·阿不都热依木. 维吾尔语罗布泊方言中音变现象的音系学分析[J]. 语言与翻译, 2017, 130(2): 32-42.
Esmael Abdurehim. Analysis on the phonological processes in the lopnor dialect of Uyghur[J]. Language and Translation, 2017,

- 130(2): 32-42.
- [8] MAY T, Van de PAR S, KOHLRAUSCH A. Noise-robust speaker recognition combining missing data techniques and UBM[J]. *Proceeding of IEEE Transactions on Audio, Speech, and Language Processing*, 2012, 20(1): 108-121.
- [9] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. *Proceeding of IEEE Transactions on Audio, Speech, and Language Processing*, 2011, 19(4): 788-798.
- [10] BOULKENAFET Z, BENGHERABI M, HARIZI F, et al. Forensic evidence reporting using GMM-UBM, JFA and I-vector methods: Application to Algerian Arabic dialect[C]//2013 8th International Symposium on Image and Signal Processing and Analysis, IEEE, 2013: 404-409.
- [11] SONG Y, HONG X H, JIANG B, et al. Deep bottleneck network based i-vector representation for language identification[C]//Inter-speech 2015 16th Annual Conference of the International Speech Communication Association, 2015.
- [12] 洪新海, 宋彦, 蒋兵, 等. 采用DBN的TV改进方法在语种识别中的应用[J]. *信号处理*, 2015, 31(9): 1152-1158.
HONG Xinhai, SONG Yan, JIANG Bing, et al. Improved total variability modeling method using deep bottleneck network for language identification[J]. *Journal of Signal Processing*, 2015, 31(9): 1152-1158.
- [13] KENNY P, GUPTA V, STAFYLAKIS T, et al. Deep neural networks for extracting baum-welch statistics for speaker recognition[C]//Proceedings of Odyssey: The Speaker and Language Recognition Workshop. Joensuu, Finland: International Speech Communication Association (ISCA), 2014: 293-298.
- [14] RICHARDSON F, REYNOLDS D, DEHAK N. Deep neural network approaches to speaker and language recognition[J]. *IEEE Signal Processing Letters*, 2015, 22(10): 1671-1675.
- [15] ALMUHAREB A, ALSANIE W, AL-THUBAITY A. Arabic word segmentation with long short-term memory neural networks and word embedding[J]. *IEEE Access*, 2019, 7(2): 12879-12887.
- [16] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: robust DNN embeddings for speaker recognition[C]//Proceeding of IEEE International Conference on Acoustics, Speech and Signal Processing, 2018: 5329-5333.
- [17] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Proceeding of INTERSPEECH, 2017: 999-1003.
- [18] XU L, DAS R K, YILMAZ E, et al. Generative x-vectors for text-independent speaker verification[C]//Proc of Spoken Language Technology Workshop(SLT 2018). Athens, Greece, 2018: 1014-1020.
- [19] FIRAT O, CHO K, BENGIO Y. Multi-way, multilingual neural machine translation with a shared attention mechanism[C]//Association for Computational Linguistics, 2016: 866-875.
- [20] CHOROWSKI J K, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]//Advances in neural information processing systems, 2015: 577-585.
- [21] TAO C, GAO S, SHANG M, et al. Get the point of my utterance! learning towards effective responses with multi-head attention mechanism[C]//IJCAI, 2018: 4418-4424.
- [22] SHEN T, ZHOU T, LONG G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- [23] CHANG S Y, LI B, SIMKO G, et al. Temporal modeling using dilated convolution and gating for voice-activity-detection[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5549-5553.
- [24] ZHANG Q, CUI Z, NIU X, et al. Image segmentation with pyramid dilated convolution based on ResNet and U-Net[C]//International Conference on Neural Information Processing, Springer, Cham, 2017: 364-372.
- [25] SHI W, JIANG F, ZHAO D. Single image super-resolution with dilated convolution based multi-scale information learning inception module[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 977-981.
- [26] ROUZI A, SHI Y, ZHIYONG Z, et al. THUYG-20: A free uyghur speech database[J]. *Journal of Tsinghua University (Science and Technology)*, 2017, 57(2): 182-187.
- [27] BAHARI M H, SAEIDI R, HAMME H V, et al. Accent recognition using i-vector, Gaussian mean supervector and Gaussian posterior probability supervector for spontaneous telephone speech[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 7344-7348.
- [28] International Telecommunications Union-Telecommunication Standardization Sector (ITU-T). Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s: ITU-T G.723.1-2006[S/OL]. [2006-05]. http://www.ptsn.net.cn/standard/std_query/show-itut-3119-1.htm.
- [29] CANDÈS E J, LI X, MA Y, et al. Robust principal component analysis[J]. *Journal of the ACM*, 2011, 58(3): 1-37.
- [30] BRO R, SMILDE A K. Principal component analysis[J]. *Analytical Methods*, 2014, 6(9): 2812-2831.