ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 34,No. 2,Mar. 2019,pp. 322-330
DOI:10. 16337/j. 1004-9037. 2019. 02. 015

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:+86-025-84892742

基于卷积神经网络的语种识别系统

金 马 宋 彦 戴礼荣

(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)

摘 要: 从给定语音中提取有效语音段表示是语种识别的关键点。近年来深度学习在语种识别应用中有重要的进展,通过深度神经网络可以提取音素相关特征,并有效提升系统性能。基于深度学习的端对端语种识别系统也表现出其优异的识别性能。本文针对语种识别任务提出了基于卷积神经网络的端对端语种识别系统,利用神经网络强大的特征提取能力及区分性建模能力,提取具有语种区分性的基本单元,再通过池化层得到有效语音段表示,最后输入全连接层得到识别结果。实验表明,在NIST LRE 2009数据集上,相比于现阶段国际主流语种识别系统,提出的系统在30 s,10 s和3 s等语音段上错误率分别相对下降了1.35%,12.79%和29.84%,且平均错误代价在3种时长上均相对下降30%以上。

关键词: 语种识别;卷积神经网络;语音段表示;语种区分性基本单元;端对端机制

中图分类号: TN912.34 **文献标志码:** A

Language Identification Based on Convolutional Neural Network

Jin Ma, Song Yan, Dai Lirong

(National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China)

Abstract: A key problem of language identification (LID) is how to design effective representations which are specific to language information. Recent advances in deep neural networks (DNNs) have led to significant improvements in language identification. The acoustic feature extracted from a structured DNN which is discriminative to phoneme or tri-phone states can significantly improve the performance. End-to-end schemes also show its strong capability of modelling in recent years. A novel end-to-end convolutional neural network (CNN) LID system is proposed, called language identification network (LID-net), taking advantage of neural networks (NNs) with the capability in feature extraction and discriminative modelling, which can extract units that discriminant to languages, and we call them LID-senones, thus can extract effective utterance representation with pooling layer. Evaluations on NIST LRE 2009 show improved performance compared to current state-of-the-art deep bottleneck feature with total variability (DBF-TV) method, can achieve 1.35%, 12.79% and 29.84% relative equal error rate (EER) improvement on 30, 10 and 3 s utterances and receive over 30% relative gain in C_{avg} on all durations.

Key words: language identification; convolutional neural network; utterance representation; language identification (LID)-senone; end-to-end scheme

---

基金项目: 国家自然科学基金(U1613211)资助项目。
收稿日期: 2017-03-22;**修订日期:** 2017-05-25

引　言

语种识别是指利用计算机自动判定语音片段所属语言种类的过程。据统计,全世界已查明的语言数量为7 099种[1],而中国的56个民族就有80多种彼此不能通话的语言和地区方言[2]。随着全球国际化的日益加深,如何在全国甚至全世界范围内进行无障碍交流成为不能忽略的问题。面对如此庞大的语言体系,一个人能掌握的语言种类非常有限。因此,自动语种识别技术的重要性和地位也显得愈发重要。语种信息属于语音信号中的弱信息,不像内容信息可以直接通过识别结果反映,需要通过对语音中的底层信息加以组合、建模和分析才能够得到。如何对语音的底层声学信息进行有效特征提取和统计建模,从而得到有效的语音段表示一直是语种识别的关键问题。

到目前为止,全变量因子(i-vector)一直是主流语种识别系统中语音段的表示,主要原因是i-vector比较紧凑且包含充足的信息量,可以得到性能优异的语种识别系统[3-4],其最初主要应用于声纹识别领域[5-6]。然而,i-vector是通过无监督的全变量子空间建模算法(Total variability, TV)得到,因此通常需要线性判别分析(Linear discriminant analysis, LDA)和类内协方差规整(Within-class covariance normalization, WCCN)等区分性训练做进一步噪声补偿处理,才能得到较好的识别结果。通常情况下,i-vector提取算法、噪声补偿算法和得分计算算法统称为TV模型后端。

近期许多深度学习技术,包括深度神经网络(Deep neural networks, DNNs)在语音信号处理中有着广泛的应用[7-9]。在语种识别系统中,由于DNN的区分性建模能力,将其与TV系统的结合,使得语种识别性能有了进一步提升。在前端特征建模方面,前期工作中提出了深瓶颈特征(Deep Bottleneck feature, DBF),并搭建了基于DBF的TV系统(DBF-TV)[10-11]。DBF是通过自动语音识别系统中音素识别器深度瓶颈网络(Deep bottleneck network, DBN)得到,相比于传统的声学特征,DBF能够去除与音素无关的信息,包括说话人信息、信道信息和背景噪声。而在后端建模方面,文献[12-14]利用DNN对单音素(Monophone)或者三音子状态(Tri-phone states)进行统计建模,从而提升TV系统的性能。无论是前端帧级特征提取或是后端建模,DNN都凸显了其强大的区分性建模能力。然而上述这些方法都是在单音素或者三音子状态上进行映射,并不是直接对语种的差异性信息进行建模,因此在高混淆度及短时语音上的识别性能会有较大幅度的衰减。

最近也有学者提出了端对端网络的语种识别系统,这类网络摒弃了传统的TV框架,充分利用神经网络的区分性建模能力,直接对语种差异性信息进行建模,取得了较好的识别性能。文献[15]使用DNN进行语种识别,然而网络的输入层受到维数的限制,只包含21帧的底层声学参数信息,限制了网络对语种信息的建模能力。文献[16-18]利用长短时记忆递归神经网络(Long short term memory-recurrent neural network, LSTM-RNN)对语音信号进行建模。由于其独特的结构设计,LSTM-RNN适合处理和预测时间序列中间隔和延迟很长的信息,并一定程度上解决了传统RNN模型的梯度消失和梯度爆炸问题,但是LSTM-RNN模型复杂度高,训练时间长。

本文提出了一种基于卷积神经网络(Convolutional neural network, CNN)的端对端语种识别系统,该网络结合了DNN在前端特征的建模能力和CNN从帧级特征到段级特征的映射能力(由于DNN的全连接层也可以用卷积的形式进行表达,因此DNN的全连接层和CNN中的卷积层都可看作卷积层),从底层声学特征直接得到语种标号,称为语种识别网络(Language identification network, LID-net)。该网络直接对语种的差异性信息进行建模,可以得到带有语种区分性的基本单元(Language identification senone, LID-senone),在语音识别中,带有音素区分性的基本单元被定义为senone,因此带有语种区分性的基本单元被称为LID-senone。并利用不同语种在LID-senone统计量上的分布差异性进行语种识别。同时,在前端特征建模过程中利用了音素识别器的DBN网络,缓解了LID-net训练时出现的过拟合问

题。实验结果表明,相比于目前国际主流的DBF-TV系统,该网络在不同时长上的性能评价指标中均有提升。

1 DBF-TV基线系统

本文采用的基线系统为现阶段国际主流的DBF-TV系统,充分利用了DBF特征对三音子状态的描述能力和TV算法对声学特征统计建模的优势,其流程图如图1所示。

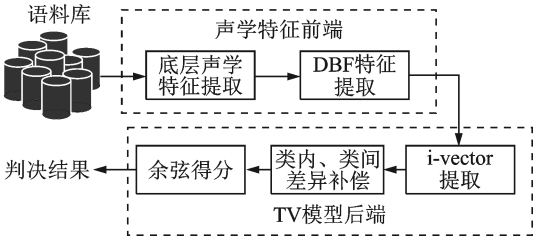

图1 DBF-TV系统流程图

Fig.1 Framework of DBF-TV system

DBF-TV系统可以分为两个部分:声学特征前端和TV模型后端。声学特征前端主要进行DBF特征的提取,利用DBN对底层声学特征进行映射,从而得到TV模型后端所需的语音学特征;TV模型后端主要利用语音学特征进行统计建模得到语音段表示,并在i-vector因子上进行区分性建模补偿,计算余弦得分,最终得到判决结果。

1.1 声学特征前端

由于底层声学特征包含着十分丰富的信息,而一些无关的噪声会对语种系统的识别性能造成影响,在声学特征前端需要尽量消除无关噪声。文献[19-22]证明提升声学特征对音素描述的准确程度会使得语种识别性能得到提高,因此使用自动语音识别中基于三音子状态的音素识别器DBN来提取DBF特征。该DBN的输入是底层声学特征,输出是所对应的三音子状态,网络中设置一层节点数目较少的隐层,称为瓶颈(Bottleneck, BN)层,DBF的具体提取过程可参考文献[10]。由于该网络在底层声学特征和三音子状态之间建立了信息提取关系,因此可以有效去除与音素无关的噪声。在提取到DBF特征后,使用TV模型后端进行处理。

1.2 TV模型后端

TV是建立在高斯混合模型(Gaussian mixture model, GMM)均值超矢量上的统计建模算法,将语音段信息的所有差异通过一个统一的空间进行描述,称之为全差异空间$\boldsymbol{T}$。TV的数学模型表示为

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{w} \tag{1}$$

式中:$\boldsymbol{M}$是每段语音的GMM均值超矢量,$\boldsymbol{m}$表示语种无关的通用背景模型(Universal background model, UBM)均值超矢量,$\boldsymbol{T}$表示描述全差异空间的投影载荷矩阵,$\boldsymbol{w}$表示均值超矢量$\boldsymbol{M}$在载荷矩阵空间下对应的低维因子表示,其后验概率均值称为i-vector,并服从均值为0、方差为$\boldsymbol{I}$的高斯分布。由于GMM均值超矢量往往有数十万维,而i-vector维数通常控制在数百,它的优势体现在把GMM均值超矢量尽可能无损压缩到一个低维的矢量空间中。$\boldsymbol{T}$空间的建立极大地减少了需要估计的参数,缓解了GMM模型中协方差矩阵估计不准的问题。然而,全差异空间建模中并没有利用到训练数据的标记信息,因此在提取i-vector后需要使用区分性模型进行噪声补偿。在经过补偿后,可以直接计算两段语音向量的余弦相似度得到判决结果,余弦距离计算为

$$\mathrm{DIS}(X,Y)=\cos\theta=\frac{\boldsymbol{x}\cdot\boldsymbol{y}}{\|\boldsymbol{x}\|\cdot\|\boldsymbol{y}\|} \tag{2}$$

2 基于卷积神经网络的语种识别系统

2.1 语种识别网络框架

尽管DBF-TV通过DNN引入了三音子状态的区分性信息,但由于这些信息并不直接在语种的差

异性上建模,在易混淆及短时语音的统计量建模上仍然会造成比较大的偏差。因此提出了一个基于语种识别任务的卷积神经网络,称为LID-net,网络的结构如图2所示。整个网络包括DNN层、卷积层、池化层和全连接层,利用神经网络(Neural network, NN)的区分性建模能力将底层声学特征经过复杂的非线性变换映射到带有语种区分性的基本单元LID-senone。然后,通过池化层对LID-senone进行简单加权平均来代替传统TV生成式的学习方法,从而得到语音段表示。最后,通过全连接层得到每个语音段在每个语种类别上的后验概率。

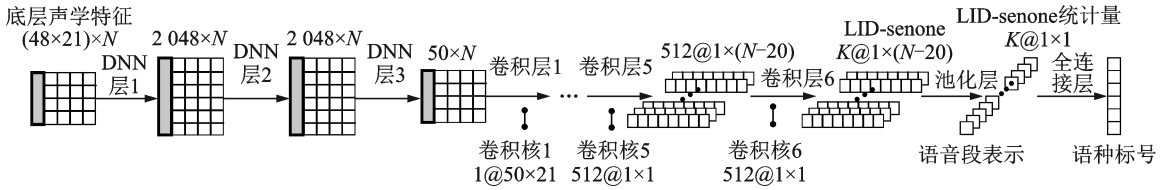

图2　语种识别卷积神经网络系统框架

Fig.2　Structure of language identification convolutional neural network

网络的参数配置如表1所示。其中输入或卷积核尺寸1@50×N表示输入或卷积核的高度是50,宽度是N,通道数是1。网络使用了两次扩帧处理。第1次在DNN层1扩帧,使用了固定的10-1-10扩帧窗,即包括当前帧及其前后各扩展的10帧,共计21帧来表示当前帧的语音信号,这在基于DNN的语音信号处理中非常常见。经过DNN层后,可以得到语种相关的50维BN特征;第2次在卷积层1扩帧,扩帧数由卷积核的大小控制,例如卷积核尺寸为1@50×21,表示使用10-1-10的扩帧窗。在池化层,语音帧级信息被直接池化到段级的语音段表示,由于需要从不定长的语音特征池化到固定长度的语音段表示,网络使用了文献[23]提出的空间金字塔池化层(Spatial pyramid pooling, SPP)来代替传统的池化层。经过SPP层,K@1×N的帧级特征被池化到固定长度的语音段级矢量,从而可以使用全连接层直接进行分类。

表1　语种识别卷积神经网络参数配置表

Tab.1　Configuration of language identification convolutional neural network

层数	名称	输入尺寸	配置
1	DNN层1	(48×21)×N	连接数:(48×21)×2 048
2	DNN层2	2 048×N	连接数:2 048×2 048
3	DNN层3	2 048×N	连接数:2 048×50
4	卷积层1	1@50×N	卷积核尺寸:1@50×21
5～9	卷积层2～6	512@1×(N−20)	卷积核尺寸:512@1×1
10	池化层	K@1×(N−20)	池化尺寸:1×(N−20)
11	全连接层	K@1×1	卷积核尺寸:K@1×1

DNN层1～DNN层3的作用是特征转换,由于三音子状态的统计量信息可用来进行语种识别,因此在特征转换时借助了三音子状态的信息(见2.3节)。经过DNN层3之后,可以得到低维的BN特征表达,不同于传统的底层声学特征,譬如梅尔频率倒谱系数(Mel frequency cepstrum coefficient, MFCC)或者感知线性预测(Perceptual linear predictive, PLP)系数,它具有更鲁棒的语种区分性,因此称为语种相关特征。卷积层和池化层可看作语音段向量表示的提取器。语种相关特征经过若干卷积层映射到具有语种区分性的基本单元LID-senone,其统计量信息会比三音子状态携有更多的语种区分性信

息,更适合进行语种识别。一段语音中,LID-senone通过池化层可以得到LID-senone统计量信息,由于LID-senone统计量是一个固定长度的向量表示,因此可以直接送入全连接层进行语种分类。

2.2 LID-net结构分析

CNN在计算机视觉领域通常会使用5×5等小卷积核对特征映射图进行卷积操作[24],这是因为图像维度之间的相关性非常高。但是在LID-net上,通过观察图3,语种相关特征的协方差矩阵除了对角线上的值不为零,其他值几乎都为零,表明该特征维间相关性非常弱,在特征上使用小卷积核没有意义。因此卷积核1的尺寸覆盖了语种相关特征所有维度及部分时域,在经过卷积层1后,特征的维数变成了512@1×(*N*−20)。

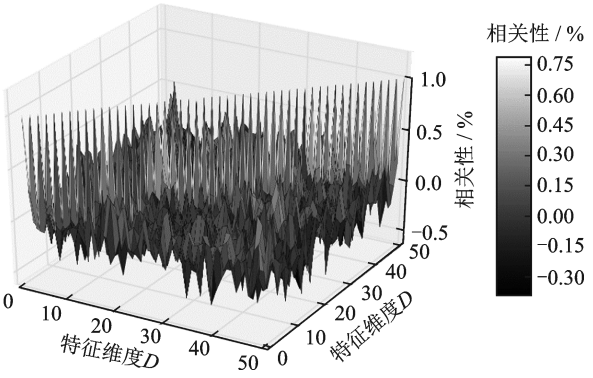

图3 语种相关特征协方差矩阵图

Fig.3 Covariance matrix of language dependent feature

为了验证LID-senone及其统计量的合理性,搭建了只包含1层卷积层的LID-net。由于通过前向计算得到的LID-senone的统计量被送入全连接层用于语种识别,因此LID-senone统计量具有语种区分性。如图4所示,采集了4段语音的LID-senone统计量,4段语音均属于波斯语种类,其中2段语音来自波斯语(Farsi),另外2段语音来自达里语(Dari)。为了便于分析,随机选取了35个LID-senone统计量。可以看出,相同语种的统计量分布更加相似,而不同语种的统计量分布显示出了较大的差异。

此外,假如每一帧特征都有相对应的LID-senone,那么它在所有LID-senone上激活值组成的矢量就具有稀疏性。图5展示了1段语音中某4帧LID-senone的激活情况,为了方便观察,激活值都做了Soft-max归一化处理。从图5(a)和(b)可以发现某1个LID-senone被激活;图5(c)展示了LID-senone的转

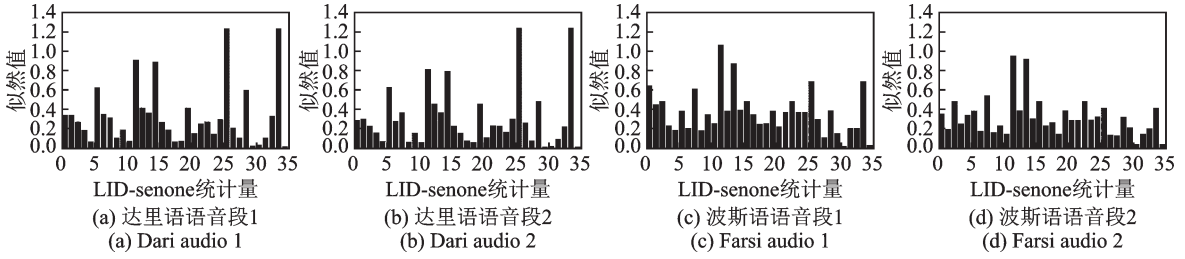

图4 LID-senone统计量

Fig.4 Statistics of LID-senone

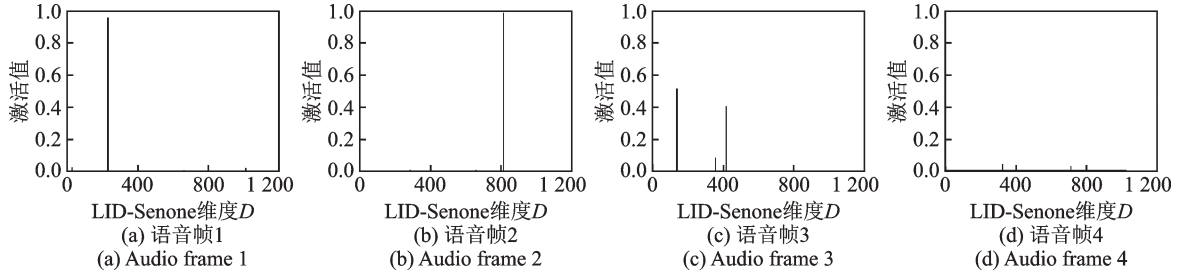

图5 LID-senone的激活值示意图

Fig.5 Activation values of LID-senone

换帧,有不止一个LID-senone被激活;(d)是非语音帧,没有LID-senone被激活。

根据上述实验分析认为LID-senone的存在是合理的,每1帧特征都有与其相对应的LID-senone,并且其统计量信息具有语种区分性。

2.3 LID-net训练步骤

在自动语音识别中使用的DNN网络,每一帧声学特征都会被映射到对应的三音子状态上,然而对于语种识别,1段语音才对应1个语种标号。由于网络的参数规模比较大而训练语料不足,非常容易出现过拟合的情况,而LID-net大部分参数都集中在DNN层,因此使用迁移学习方法[25]利用音素识别网络对DNN层的参数进行初始化来缓解过拟合问题。如前所述,音素信息也有助于语种识别分类,并且音素分类器在每1帧上都映射到相对应的三音子状态,因此先训练1个音素识别器,然后把部分参数作为LID-net中DNN层的参数初始化。

具体来说,首先利用公开语音数据集SwitchBoard[26]训练1个7层DBN网络的音素识别器,输入特征是48维,前后各扩10帧,网络输出的状态数是3 020,并带有1个50维的BN层,网络的结构是48×21—2 048—2 048—50—2 048—2 048—3 020。网络完成训练后,把音素识别器网络的前3层参数作为LID-net中DNN层的初始化参数,再对LID-net进行训练。训练时,需要把DNN层参数的学习率降低,通常设为正常学习率的1/10。网络训练完成后,输出每个语种后验概率直接使用余弦距离计算得分。

3 实验结果及分析

3.1 数据集及性能评价指标

为了验证提出网络的有效性,在NIST LRE 2009公开标准数据集上进行相关实验。该数据集包含23个语种,每个语种的训练数据均包含两种信道:电话语音信道(Conversational telephone speech, CTS)和美国之声(Voice of America, VOA)窄带广播信道,训练数据在经过语音端点检测(Voice activity detection, VAD)和切分后,时长约1 100 h。此外,还包括15 000条验证数据和30 000余条测试数据,其中包含3种时长的语音,30 s和10 s的长时语音及3 s的短时语音,每个时长都需要训练对应的语种识别网络。

评价指标采用NIST评测使用的等错误率(Equal error rate, EER)和平均错误代价(Average decision cost function, C_{avg})。其中EER是当虚警率(False alarm, FA)和漏警率(Miss alarm, MA)相等时二者的值,C_{avg}的计算方式为

$$C_{avg}=\frac{1}{N_L}\sum_{L_T}\left\{C_{Miss}\cdot P_{Target}\cdot P_{Miss}(L_T)+\sum_{L_N}\{C_{FA}\cdot P_{Non\text{-}Target}\cdot P_{FA}(L_N,L_T)\}\right\} \tag{3}$$

式中N_L表示所有待识语种的数目;L_T和L_N分别表示目标语种和非目标语种;C_{Miss}和C_{FA}表示漏判决和错误判决一条语音的代价;P_{Target}和$P_{Non\text{-}Target}$分别表示目标语种和非目标语种的先验概率。根据NIST LRE 2009测试标准,计算时$C_{Miss}=C_{FA}=1$,$P_{Target}=0.5$,$P_{Non\text{-}Target}=(1-P_{Target})/(N_L-1)$[27]。

3.2 相关系统定义

本文采用的语种识别系统具体描述如下。

系统1(LID-net):本文提出的系统。底层声学特征采用48维PLP特征(13维特征+3维基音频率特征,并计算一阶和二阶差分),使用2.3节所述方法初始化DNN层。识别网络包含6个卷积层,其中前5层卷积通道数为512,第6层卷积通道数由32变化到512作为实验对比。网络训练迭代15轮,初始学习率设为0.05,每迭代5次学习率降10倍。网络输出矢量直接使用余弦距离计算得分。

系统2(DBF-TV):本文搭建的基线系统。底层声学特征与系统1一致,并使用1.1节和2.3节描述

的DBN网络提取50维DBF特征。TV系统采用期望最大化(Expectation maximization, EM)算法迭代5轮训练1个秩为400的T矩阵,然后提取i-vector,再使用LDA和WCCN进行噪声补偿后使用余弦距离计算得分,作为对照实验使用。

系统3(DBF-TV-Ferrer):文献[14]采用的语种识别DBF-TV系统,作为对照实验使用。

3.3 实验结果及分析

实验1 卷积核尺寸对系统性能影响

为了构建一个较为合理的语种区分性单元LID-senone,对卷积核1的尺寸进行了相关实验,尺寸从1@50×1变换到1@50×26,每次尺寸长度增加5。为了使LID-senone在混淆语种中也能有较好的区分性,在NIST LRE 2009数据集中选取了6个最易混淆的语种进行本次实验,分为3个方言对,分别是:达里语(Dari)和波斯语(Farsi)、俄语(Russian)和乌克兰语(Ukrainian)、印地语(Hindi)和乌尔都语(Urdu)[28]。网络只包含1个卷积层,通道数是1 024。系统的性能由EER(%)进行评价,性能对比如图6所示。

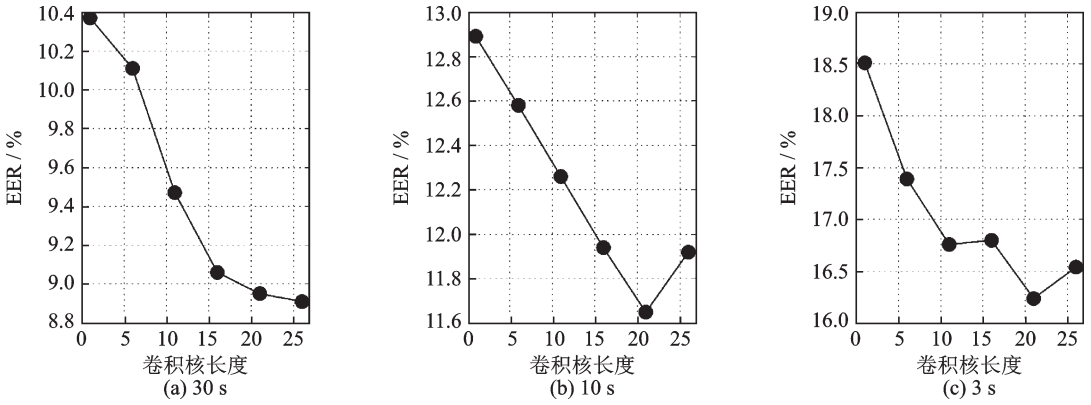

图6 不同卷积核尺寸的系统性能对比图

Fig.6 Performance on different convolutional filter sizes

从图6可以看出,随着卷积核的长度增长,3种时长下语种识别的性能都在逐步提升。其中10 s和3 s网络在卷积核尺寸为1@50×21时性能最优;而对于30 s网络,卷积核尺寸在1@50×26时比在1@50×21时只有轻微提升。因此整体来说,卷积核1尺寸设置在1@50×21比较合适,相当于在卷积层1进行了10-1-10的扩帧,即每21帧语种相关特征可以表示1个LID-senone。再考虑到DNN层固定的10-1-10扩帧窗,相当于每41帧底层声学特征可以较好地表示1个LID-senone。在语音识别领域,通常认为11帧或15帧底层声学特征可以较好地对三音子信息进行建模,但对于LID-senone明显需要更长时的信息。在下面的实验中,卷积核1的尺寸统一配置为1@50×21。

实验2 卷积神经网络语种识别系统及对照系统性能对比

为了验证本文提出的端对端语种识别系统的有效性,将该系统与3.2节的系统2和系统3进行性能对比。LID-net第6层卷积层的通道数由32变化到512。所有系统的性能由EER(%)和C_{avg}(%)进行评价,如表2所示。其中每个时长中性能最优系统的数值用粗体表示。可以发现,LID-net在不同时长上的指标均好于两个基线系统DBF-TV及DBF-TV-Ferrer。总体来说,语音的时长越短,系统性能提升的幅度越大。相较于DBF-TV系统,LID-net系统的EER在30 s, 10 s和3 s时长上分别相对下降了1.35%, 12.79%和29.84%;C_{avg}在3种时长上分别相对下降了32.73%, 31.77%和32.49%。这是因为经过CNN的区分性建模,不同语种的得分分布更具有区分性,从而实现更优的系统性能。因此LID-net这种端对端模型可以很好地对语种信息进行建模,尤其在短时语音上比生成性模型有更大优势。实验

表 2　不同语种识别系统性能对比

Tab.2　Performance comparison on different language identification systems

系统名称	第6个卷积层通道数	3 s		10 s		30 s	
		EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
LID-net	32	7.67	6.02	2.74	1.54	1.49	1.05
	64	7.76	5.99	2.92	1.64	1.54	0.75
	128	7.58	6.15	2.89	2.00	1.55	0.91
	256	**7.57**	**5.05**	**2.66**	**1.46**	**1.46**	1.21
	512	7.79	6.64	2.81	1.49	1.50	**0.74**
DBF-TV	N/A	10.79	7.48	3.05	2.14	1.48	1.10
DBF-TV-Ferrer[14]	N/A	N/A	6.82	N/A	1.98	N/A	1.15

结果表明,LID-net的第6个卷积层通道数需要合理设置,通道数太小会对系统性能造成影响,而太大则会导致过拟合。

4　结束语

本文针对语种识别任务提出了一个基于卷积神经网络的端对端语种识别系统,网络层数较深但比较直观。通过DNN层可以得到语种相关特征;再通过卷积层得到LID-senone;最后通过池化层得到LID-senone统计量并送入分类器得到识别结果。由于音素信息有助于语种识别,而DNN层有大量的待训练参数。因此先训练1个音素分类器,并把部分参数作为DNN层初始化参数,然后训练LID-net网络。对比现阶段主流的DBF-TV系统,EER在30 s,10 s和3 s时长上分别相对下降了1.35%,12.79%和29.84%,C_{avg}在所有时长上均相对下降了30%左右。然而,LID-net中池化层只是对LID-senone做了简单的加权平均,造成了大量的信息损失。因此,如何充分利用LID-senone的信息将是未来工作的重点。

参考文献:

[1] Lewis M P, Simons G F, Fennig C D. Ethnologue: Languages of the world [M]. 18th ed. Dallas: SIL International Publications, 2009.

[2] 吴伟平, 李兆麟. 语言学与华语二语教学 [M]. 香港: 香港大学出版社, 2009: 99.
Wu Weiping, Li Zhaolin. Chinese as a second language teaching and research society [M]. Hong Kong, China: Hong Kong University Press, 2009: 99.

[3] Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data [J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(3): 345-354.

[4] Martınez D, Plchot O, Burget L, et al. Language recognition in ivectors space [C]//Proceedings of the 15th Annual Conference of the International Speech Communication Associations (Interspeech). Firenze, Italy: International Speech Communication Association (ISCA), 2011: 861-864.

[5] Kenny P, Boulianne G, Dumouchel P. Eigenvoice modeling with sparse training data [J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(3): 345-354.

[6] Kenny P, Boulianne G, Ouellet P, et al. Speaker and session variability in GMM-based speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1448-1460.

[7] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition [C]//Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China:IEEE, 2016: 4960-4964.

[8] Matějka P, Glembek O, Novotný O, et al. Analysis of DNN approaches to speaker identification [C]//Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Shanghai, China:IEEE, 2016: 5100-5104.

[9] Zeinali H, Burget L, Sameti H, et al. Deep neural networks and hidden markov Models in i-vector-based text-dependent

speaker verification [C]//Proceedings of Odyssey: The Speaker and Language Recognition Workshop. Bilbao, Spain: International Speech Communication Association (ISCA), 2016: 24-30.
[10] Jiang B, Song Y, Wei S, et al. Deep bottleneck features for spoken language identification [J]. Plos One, 2014, 9(7): e100795.
[11] Song Y, Jiang B, Bao Y B, et al. I-vector representation based on bottleneck features for language identification[J]. Electronics Letters, 2013, 49(24): 1569-1570.
[12] Lei Y, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network [C]// Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 1695-1699.
[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] Ferrer L, Lei Y, McLaren M, et al. Study of senone-based deep neural network approaches for spoken language recognition [J]. IEEE Transactions on Audio, Speech and Language Processing, 2016, 24(1): 105-116.
[15] Lopez M I, Gonzalez D J, Plchot O, et al. Automatic language identification using deep neural networks[C]//Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy: IEEE, 2014: 5337-5341.
[16] Gonzalez D J, Lopez M I, Sak H, et al. Automatic language identification using long short-term memory recurrent neural networks[C]//Proceedings of the 15th Annual Conference of the International Speech Communication Associations (Interspeech). Singapore: International Speech Communication Association (ISCA), 2014: 2155-2159.
[17] Zazo R, Lozano D A, Gonzalez D J, et al. Language identification in short utterances using long short-term memory (LSTM) recurrent neural networks [J]. Plos One, 2016, 11(1): e0146917.
[18] Gelly G, Gauvain J L, Le V B, et al. A divide-and-conquer approach for language identification based on recurrent neural networks[C]//Proceedings of the 17th Annual Conference of the International Speech Communication Associations (Interspeech). San Francisco, USA: International Speech Communication Association (ISCA), 2016: 3231-3235.
[19] Ambikairajah E, Li H, Wang L, et al. Language identification: A tutorial [J]. IEEE Circuits and Systems Magazine, 2011, 11 (2): 82-108.
[20] Diez M, Varona A, Penagarikano M, et al. On the use of phone log-likelihood ratios as features in spoken language recognition [C]// Proceedings of Spoken Language Technology IEEE Workshop (SLT). Miami, USA:IEEE, 2012: 274-279.
[21] Siniscalchi S M, Reed J, Svendsen T, et al. Universal attribute characterization of spoken languages for automatic spoken language recognition [J]. Computer Speech & Language, 2013, 27(1): 209-227.
[22] McLoughlin I. Applied speech and audio processing: With Matlab examples [M]. Cambridge: Cambridge University Press, 2009.
[23] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [C]// Proceedings of European Conference on Computer Vision (ECCV). Zurich, Switzerland: Springer International Publishing, 2014: 346-361.
[24] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, USA:IEEE, 2016: 770-778.
[25] Pan S J, Yang Q. A survey on transfer learning [J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.
[26] Godfrey J J, Holliman E C, McDaniel J. SWITCHBOARD: Telephone speech corpus for research and development [C]// Proceedings of the 15th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). San Francisco, USA:[s.n.], 1992: 517-520.
[27] Martin A F, Greenberg C S. The 2009 NIST language recognition evaluation [C]// Proceedings of Odyssey: The Speaker and Language Recognition Workshop. Brno, Czechoslovakia: International Speech Communication Association (ISCA), 2010: 30.
[28] Jiang B, Song Y, Wei S, et al. Task-aware deep bottleneck features for spoken language identification [C]// Proceedings of the 15th Annual Conference of the International Speech Communication Associations (Interspeech). Singapore: International Speech Communication Association (ISCA), 2014: 3012-3016.

作者简介:

金马(1990-),硕士研究生,研究方向:语种识别和模式识别,E-mail: jinma525@mail.ustc.edu.cn。

宋彦(1972-),男,副教授,研究方向:语种识别及音、视频分析和检索,E-mail: songy@ustc.edu.cn。

戴礼荣(1962-),男,教授,研究方向:语音识别和信号处理,E-mail: lrdai@ustc.edu.cn。

(编辑:陈珺)