ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 850－857
DOI:10. 16337/j. 1004-9037. 2020. 05. 006

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

基于内容的x-vector文本相关SV研究

陈亚峰，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)

摘　要: x-vector系统将一段不定长的语音通过神经网络映射成固定维的矢量来表征说话人信息,该系统在文本无关的说话人确认(Speaker verification, SV)任务中取得了优异的性能。本文将其应用到文本相关的SV任务中,在x-vector模型选择上,采用残差神经网络以获得更有区分性的x-vector;在包含多字符的语句中,对每个字训练一个残差神经网络;在提取过程中,每一字单独提取一个x-vector并单独进行说话人判决,最后将多个判决得分进行融合后给出最终的识别结果。实验是在数据库RSR2015 Part Ⅲ上进行的,提出的方法在男性和女性测试集上等错误率分别有15.34%、19.7%的下降。

关键词: 说话人确认;文本相关;深度神经网络;声纹特征

中图分类号: TN912.3　　**文献标志码:** A

Content-Dependent x-vector for Text-Dependent Speaker Verification

CHEN Yafeng, GUO Wu

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

Abstract: The x-vector system maps a variable-length speech to a fixed-dimensional speaker embeddings via neural networks, and performs well in text-independent speaker verification. Here, it is applied to the text-dependent speaker verification and different x-vectors are extracted according to different contents in one sentence. In model selection, deep residual network (DRN) is used to obtain more discriminative x-vector. For a sentence with multiple words, word-dependent DRNs are trained to extract word-dependent x-vectors, which are separately fed to different backend classifiers. Finally, multiple scores are fused to obtain the final verification results. Experiments on Part Ⅲ of the RSR2015 dataset show that the proposed method can achieve equal error rate (EER) reduction of 15.34% and 19.7% for male and female, respectively.

Key words: speaker verification (SV); text-dependent; deep neural network; speaker characteristics

引　　言

说话人确认(Speaker verification, SV)是判断一段测试语音与其所声明身份是否一致的过程。SV又分为文本相关的SV(Text-dependent SV)和文本无关的SV(Text-independent SV)。从目前的SV技术水平来看,相对于文本无关的SV的低准确率,文本相关的SV将内容与声纹特征结合起来,有效地提高了识别准确率,从而在商业应用中获得了广泛的应用[1]。

收稿日期:2019-12-05;**修订日期:**2020-04-25

近几年来,基于因子分析的全变量(Total variability, TV)系统的i-vector[2]算法一直是文本无关的SV中主流的方法。首先通过将一段语料映射到一个低维的子空间中,得到表征该说话人的特征矢量i-vector,再进行低维空间的信道补偿算法和得分判决算法以获得更优的SV性能。该方法在大数据集上训练和测试取得了不错效果,同样也被用于文本相关的SV中[3]。

文本相关的SV一般将语音内容限制为:(1)固定短语;(2)一组预定义短语;(3)特定的随机内容组合。典型的如数字串,在测试中由系统生成需要判断的语音内容并进行声纹确认[4]。经典的文本相关SV应用中,注册和验证通常都使用一个固定短语或者一组预定义短语,如果文本信息被泄露,则安全性会大大降低。在文献[5]中,针对随机内容组合的这种应用,提出基于音素的i-vector系统,其对语料中的22个音素分别建模,提取每个音素的i-vector,再结合后端算法判决得分。在文献[6]中,在对10数字分别建模的i-vector系统的基础上,提出了一种新的后端信道补偿算法,进一步提高了识别的准确率。

随着深度学习的在图像、自然语言处理以及语音识别[7]等领域上取得优异的效果,其强大的特征提取能力可以帮助声纹系统获得更具有说话人区分性的信息。因此,基于深度神经网络的SV方法被广泛使用,最主流的算法是提取出表征说话人特征矢量x-vector[8],再结合后端处理算法进行信道补偿和得分判决。但针对文本相关的SV任务,存在因训练数据过少导致深度神经网络的过拟合问题,提取出来的x-vector区分性不够。本文中,采用不同的网络结构以及网络预训练等策略解决该问题。

本文针对文本内容为随机数字序列的SV任务提出并构建了一个基于内容建模的x-vector系统。首先,利用语音识别模型将语料分割成不同的内容(10个数字),然后分别针对每个数字微调一个预训练好的深度神经网络,得到不同数字的特征提取器,使用这些特征提取器提取对应内容(数字)的x-vector。后端处理算法也分别针对不同的内容(数字)单独训练,最后将测试语料中各个内容(数字)的得分求和的平均计算最终得分。实验在RSR2015数据库上进行,由于Part Ⅲ语料内容是数字串,因此后面的描述中用“数字”来代表内容。若语料内容不是数字,本文提出的方法依旧适用:先语音识别文本内容,再分词建模提取说话人特征,最后运用后端算法计算最终得分。从结果上来看,提出的算法可以明显提升系统性能。

1 x-Vector/PLDA

1.1 x-vector

x-vector系统所使用的深度神经网络结构主要分为帧处理层(Frame-level layers)、池化层(Pooling layer)、段处理层(Segment-level layer)3部分[9],如图1所示。

(1) 帧处理层

传统的x-vector系统帧处理层由5层时延神经网络(Time delay neural network, TDNN)[10]组成,以帧为单位对低层输入特征进行非线性映射获得帧级别的表示。对于一段输入语料$\boldsymbol{X}=\{x_1,x_2,\cdots,x_T\}$($T$为帧数),那么每一帧处理层输出为

$$\boldsymbol{f}_t^i=f(\boldsymbol{w}^{\mathrm{T}}*\boldsymbol{x}_t^c+\boldsymbol{b}) \tag{1}$$

式中,$\boldsymbol{f}_t^i$为第i层第t帧输出矢量,$\boldsymbol{x}_t^c$为第t帧附近的输入特征拼接起来的矢量,来学习第t帧附近的信息,$\boldsymbol{w}$和$\boldsymbol{b}$为权重矩阵和偏置矢量,f为非线性激活函数,这里取ReLU函数。

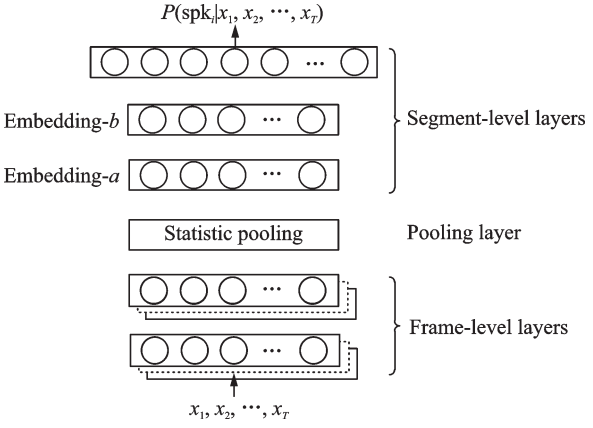

图1 深度神经网络示意图

Fig.1 Diagram of the deep neural network

(2) 池化层

图1网络中的池化层为统计池化层(Statistics pooling layer),是由帧处理层的输出分别计算均值和标准差拼接而成,以此得到一段语料的统计特性作为表示。统计池化层的输出为

$$\boldsymbol{s}=\text{concat}\left[\text{mean}(\boldsymbol{f}_1,\cdots,\boldsymbol{f}_t);\text{std}(\boldsymbol{f}_1,\cdots,\boldsymbol{f}_t)\right] \tag{2}$$

式中,$\boldsymbol{s}$表示统计池化层输出矢量,$\boldsymbol{f}_t$表示帧处理层的第t帧输出矢量,mean表示对所有帧求均值,std表示对所有帧求标准差。

(3) 段处理层

段处理层用统计池化层的输出$\boldsymbol{s}$作为输入,通过若干层前向DNN网络提取段级别的矢量来表征说话人。x-vector中采用2层DNN,表示为

$$\boldsymbol{\sigma}^l=f(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}+\boldsymbol{b}) \tag{3}$$

式中$\boldsymbol{\sigma}^l$为段处理层第l层输出。

段级别的矢量经过段处理层的进一步处理,最后通过Softmax分类器进行分类,来预测目标说话人的类别以进行区分性训练。训练采用交叉熵损失函数[11]

$$E=-\sum_{n=1}^{N}\sum_{k=1}^{K}d_{nk}\ln\left(P(\text{spk}_r|\boldsymbol{x}_{1:T}^n)\right) \tag{4}$$

式中,第n段语料如果属于第k个说话人,则d_{nk}为1,否则为0;$P(\text{spk}_r|\boldsymbol{x}_{1:T}^n)$为Softmax分类器对给定第$n$段语料特征的预测输出。神经网络利用批量BP算法进行训练,更新网络参数。

(4) x-vector提取

在段处理层中,第1,2层的输出都可以用来作为一段语料的低维矢量表示,一般采用第1层的线性部分输出(Embedding-a)作为最终的说话人矢量表示,即x-vector,其优异的性能在近些年获得了广泛应用。

1.2 概率线性判别式分析

在获得语音的低维矢量表示之后,采用当前主流的后端判别概率线性判别式分析(Probabilities linear discriminant analysis, PLDA)进行SV。给定一条语料u,PLDA模型可以写成

$$\boldsymbol{\omega}(\boldsymbol{u})=\boldsymbol{\mu}+\boldsymbol{V}\boldsymbol{y}(\boldsymbol{u})+\varepsilon(\boldsymbol{u}) \tag{5}$$

式中,$\boldsymbol{\mu}$是所有数据x-vector的均值,$\boldsymbol{V}$是载荷矩阵,它的每一列是说话人子空间的基。$\boldsymbol{y}(\boldsymbol{u})$是$\boldsymbol{\omega}(\boldsymbol{u})$映射在说话人子空间的隐变量,$\varepsilon(\boldsymbol{u})$是残余噪声项。高斯PLDA模型是建立在观测值服从高斯分布这一基础上建立的。但是在实际应用中,x-vector分布是不满足高斯条件的,为了提高PLDA算法效果,需要对x-vector做如下均值方差归一化[12]

$$\boldsymbol{\omega}(\boldsymbol{u})\leftarrow\frac{\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\omega}(\boldsymbol{u})-\boldsymbol{\mu})}{\left\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\omega}(\boldsymbol{u})-\boldsymbol{\mu})\right\|} \tag{6}$$

式中,$\boldsymbol{\Sigma}$是所有训练数据x-vector协方差矩阵。在测试阶段,利用PLDA模型计算两端语料相似度得分。假设H_1表示两段语料来自于同一个说话人,假设H_0表示2段语料来自于不同的说话人,2段语料对应的x-vector分别为$\boldsymbol{\omega}(\boldsymbol{u}_1)$和$\boldsymbol{\omega}(\boldsymbol{u}_2)$,那么最终的似然度得分计算如下

$$s(\boldsymbol{u}_1,\boldsymbol{u}_2)=\log\frac{P(\boldsymbol{\omega}(\boldsymbol{u}_1),\boldsymbol{\omega}(\boldsymbol{u}_2)|H_1)}{P(\boldsymbol{\omega}(\boldsymbol{u}_1),\boldsymbol{\omega}(\boldsymbol{u}_2)|H_0)P(\boldsymbol{\omega}(\boldsymbol{u}_1),\boldsymbol{\omega}(\boldsymbol{u}_2)|H_0)} \tag{7}$$

2 基于内容的x-vector / PLDA模型

在文本相关的SV中,内容是很重要的一个区分性信息。前面所述的x-vector系统都是对一段语音进行统一的矢量提取,没有考虑内容对x-vector的影响。本文针对这种情况采用不同的数字分别训练

残差神经网络并分别提取x-vector。基于内容的x-vector系统包含训练阶段和测试阶段,图2为说话人识别流程图。在说话人模型注册阶段,首先进行数据预处理:提取训练语料的30维梅尔频率倒谱系数(Mel frequency cepstral coefficient, MFCC)特征,并利用端点检测算法除去静音帧;再进行语料切分:利用语音识别模型将每条语料切割成若干数字,由于声音信噪比高,采用高斯混合模型与隐马尔科夫模型(Gaussian mixture models and hidden markov model, GMM-HMM)模型已经能够获得很好的语音识别准确率;利用训练好的深度神经网络模型分别提取每个数字的x-vector,完成模型的注册。在测试阶段,数据预处理与训练阶段相同,提取注册语料和测试语料中各数字相应的x-vector,使用线性判别式分析(Linear discriminant analysis, LDA)、PLDA后端信道补偿算法获取各数字的得分,最后将测试语料各数字得分求和平均计算最终得分。

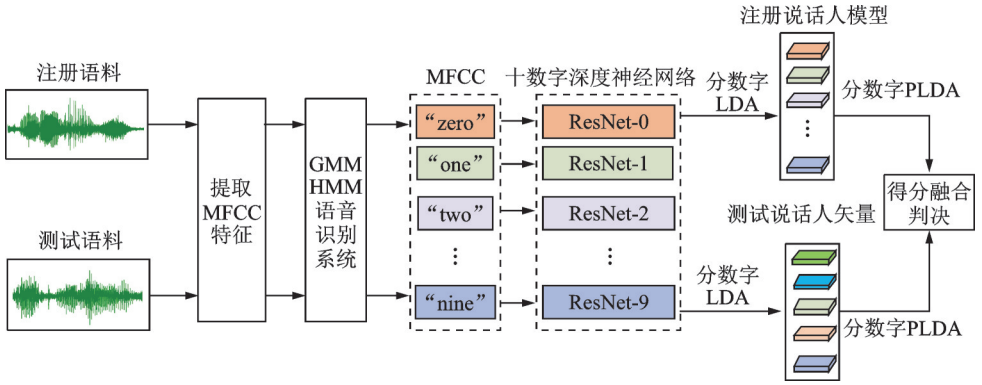

图2 基于内容的x-vector系统流程图

Fig.2 Block diagram of digit-dependent x-vector system

图2中提取x-vector神经网络首先使用大量数据预训练得到一个初始网络,然后用训练集的每个不同的数字来训练得到10个与数字相关的神经网络。由于x-vector是与数字相关的,因此也用训练集的不同数字的x-vector来单独训练LDA、PLDA模型。

2.1 语音内容切分

在训练阶段,本文中采用RSR2015 Part Ⅲ中的bkg+dev中Part Ⅲ数据训练GMM-HMM模型,采用单音素声学模型,每个音素采用三状态建模,用得到的语音识别模型对语音做识别后,并对每个数字做强制对齐可以得到其起始和终止时间,并以此作为后续说话人训练和测试实验所用。

2.2 x-vector模型选择与训练

传统的x-vector系统在帧处理层采用的是TDNN结构,这种结构在数据量较大时,性能优异,但声纹表征的提取能力依旧不足。本文采用改进的34层深度残差网络(ResNet-34)[12]代替TDNN,如图3所示。与标准ResNet-34相比,除去第1个卷积层后的池化层,并修改各个卷积层中各卷积核大小,具体参数见图3。其以帧为单位对低层输入特征进行非线性映射获得帧级别的表示。与普通DNN不同的是,ResNet-34引入残差学习模块[13],解决了随着网络层数的加深,准确率不升反降的问题。池化层为统计池化层,段处理层由2个具有512节点数的fc层和softmax层组成,输出节点为对应的目标说话人。

传统的x-vector系统是训练一个神经网络,既然采用每个字来分别建模,最好的方式是每个不同的字都建立一个神经网络分别提取x-vector,这样具有更高的区分性。不可回避的是,采用每个字建立一个x-vector的提取网络面临着数据不足的问题。为解决训练数据过少导致深度神经网络的过拟合问题,采用网络预训练策略。首先用大量数据训练一个相对稳健的模型,本文中使用Voxceleb中的开发

集[14]和Voxceleb2中的开发集[15]训练一个深度神经网络，然后将输出节点替换，固定除最后一个隐层外的所有底层参数，然后分别用各个字的特征语料训练最后一个隐层参数，待网络收敛后，只固定BN(Batch normalization)层[16]参数，重新训练网络至收敛。

当10个数字的深度神经网络训练完毕，将注册语料和测试语料中的各数字特征作为网络输入来提取对应数字的x-vector，再分别进行信道补偿和得分判决。

2.3 基于数字的PLDA模型

当基于字的深度神经网络训练完毕，将注册语料和测试语料中的各个字特征作为网络输入来提取对应数字的x-vector。在得到每个数字的表征说话人的矢量x-vector之后，采用基于字的PLDA模型。

给定一条语料x，基于字的PLDA模型如下

$$\boldsymbol{\omega}_d(\boldsymbol{u})=\boldsymbol{\mu}_d+\boldsymbol{V}_d\boldsymbol{y}_d(\boldsymbol{u})+\boldsymbol{\varepsilon}_d(\boldsymbol{u}) \tag{8}$$

与式(5)不同的是，上式所有变量都是针对特定数字d，$\{\boldsymbol{\mu}_d,\boldsymbol{V}_d,\boldsymbol{\Sigma}_d\}$这些参数都是由其对应数字的归一化的x-vector训练，归一化过程如式(9)所示

$$\boldsymbol{\omega}_d(\boldsymbol{u})\leftarrow\frac{\boldsymbol{\Sigma}_d^{-1/2}(\boldsymbol{\omega}_d(\boldsymbol{u})-\boldsymbol{\mu}_d)}{\left\|\boldsymbol{\Sigma}_d^{-1/2}(\boldsymbol{\omega}_d(\boldsymbol{u})-\boldsymbol{\mu}_d)\right\|} \tag{9}$$

式中，$\boldsymbol{\mu}_d$和$\boldsymbol{\Sigma}_d$是所有关于数字d的x-vector的均值和协方差矩阵。

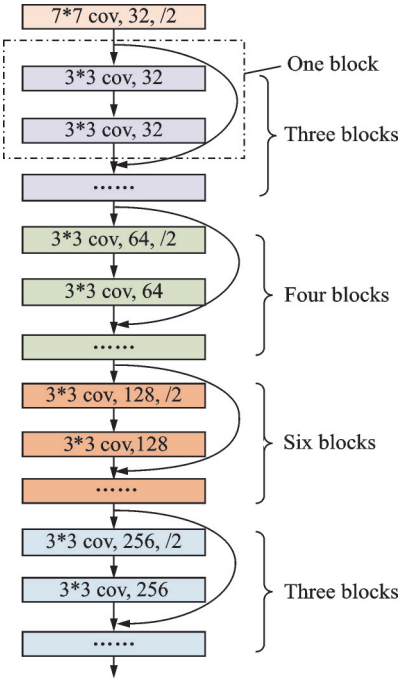

图3 深度残差网络

Fig.3 A deep residual network

然后，对注册语料和测试语料中的各个数字进行得分判决，如式(10)所示

$$s_d(\boldsymbol{u}_1,\boldsymbol{u}_2)=\log\frac{P(\boldsymbol{\omega}_d(\boldsymbol{u}_1),\boldsymbol{\omega}_d(\boldsymbol{u}_2)|H_1)}{P(\boldsymbol{\omega}_d(\boldsymbol{u}_1),\boldsymbol{\omega}_d(\boldsymbol{u}_2)|H_0)P(\boldsymbol{\omega}_d(\boldsymbol{u}_1),\boldsymbol{\omega}_d(\boldsymbol{u}_2)|H_0)} \tag{10}$$

式中，假设H_1表示对于数字d，$\boldsymbol{\omega}_d(\boldsymbol{u}_1),\boldsymbol{\omega}_d(\boldsymbol{u}_2)$是来自于相同的说话人，假设$H_0$表示对于数字$d$，它们来自不同的说话人。最后，将不同数字的得分进行合并，统计出每条语料的判决得分如下

$$s(\boldsymbol{u}_{\text{test}},\boldsymbol{u}_{\text{enroll}})=\frac{1}{|D_x|}\sum_{d\in D_x}s_d(\boldsymbol{u}_{\text{test}},\boldsymbol{u}_{\text{enroll}}) \tag{11}$$

式中，D_x表示测试语料中含有的数字集合，$|D_x|$表示测试语料中包含数字的个数，$s_d(\boldsymbol{u}_{\text{test}},\boldsymbol{u}_{\text{enroll}})$的计算如式(10)所示。式(11)是不同数字得分求和平均的过程。

3 仿真实验与结果分析

3.1 实验数据与评价指标

本次实验是RSR2015数据库上进行的，RSR2015是一个针对文本相关的SV任务的英文数据库，其中包含300个说话人(男157人，女143人)。按照不同说话人进行分类，分为：background(bkg)，development(dev)和evaluation(eval)[17]，具体如表1所示。按照语料内容分为Part Ⅰ、Part Ⅱ、Part Ⅲ这3部分。Part Ⅰ语料是固定短语，Part Ⅱ语料是家用电器控制命令，Part Ⅲ语料是随机数字串

表1 RSR2015数据库分类

Table 1 Partitioning of RSR2015 人

类别	男	女
Background	50	47
Development	50	47
Evaluation	57	49

语音。本文在Part Ⅲ进行SV实验。Part Ⅲ语料中,10数字串平均时长5.19 s,5数字串平均时长3.06 s,除去静音帧后,其有效时长分别为2.07,1.09 s。

本文采用等错误率(Equal error rate,EER)和最小错误代价函数(Minimal detection cost function,MinDCF(p-target)=0.01)作为评价指标[18]。

3.2 实验系统

在RSR2015数据集上分性别训练和测试,注册和测试数据全部来自Eval,10数字串的语料注册,5数字串的语料测试[5]。构建了914 688个测试:其中男性有526 167个,目标说话人的个数有9 231个,非目标说话人的个数有516 936个,女性有388 521个,目标说话人有7 929个,非目标说话人有380 592个。除本文提出的算法之外,另外还采用了7个主流的系统进行对比。

GMM i-vector系统:UBM模型分性别训练,UBM模型的训练数据分别是男性说话人和女性说话人的bkg+dev数据。UBM高斯数为1 024,对应i-vector系统的T矩阵训练数据同UBM模型,i-vector取400维。后端LDA算法进行信道补偿,将维度降到128,再利用PLDA算法得分判决。LDA、PLDA模型训练数据为bkg+dev中Part Ⅲ语音。

基于内容的i-vector系统:这是文献[8]提出的一种算法。GMM-HMM模型也是分性别训练,训练数据分别是男性说话人和女性说话人的bkg+dev中Part Ⅲ数据。UBM按不同数字建模,训练数据与GMM-HMM模型相同,UBM高斯数为16,对应i-vector系统的T矩阵训练数据同UBM模型,i-vector取100维,LDA降维至60,PLDA得分判决。LDA、PLDA模型训练数据与GMM i-vector系统相同。

RSR-TN-xvector系统:5层TDNN作为帧处理层,各层节点数分别为512,512,512,512,1 536。池化层为统计池化层;段处理层由2个具有512节点数的fc层和softmax层组成。整个神经网络训练数据是RSR2015 Part Ⅲ中bkg+dev中所有语音数据。

RSR-RN-xvector系统:改进的ResNet-34网络作为帧处理层,如图3所示;其他所有配置与TN-xvector相同。

TN-xvector系统:网络结构和RSR-TN-xvector系统完全相同。网络训练数据是Voxceleb中的开发集和Voxceleb2中的开发集,x-vector取512维,LDA、PLDA配置和GMM i-vector系统相同。

基于内容的TN-xvector系统:网络结构与TN-xvector系统相同,深度神经网络预训练数据与TN-xvector系统训练数据相同。当网络预训练完成后,将RSR2015 Part Ⅲ中bkg+dev中Part Ⅲ数据作为训练数据重新对网络进行微调。LDA、PLDA配置和GMM i-vector系统相同。

RN-xvector系统:改进的ResNet-34网络作为帧处理层,如图3所示;其他所有配置与TN-xvector相同。

基于内容的RN-xvector系统:网络结构与RN-xvector系统相同,其他所有配置与基于内容的TN-xvector系统相同。

3.3 实验结果与分析

表2列出了6个系统在测试集上的实验结果。

由RSR-TN-xvector和RSR-RN-xvector系统在男女测试集上性能可知,仅使用RSR2015数据库中的数据训练深度神经网络,说话人识别性能会大幅降低。因网络参数过多,而训练数据不足导致出现过拟合,参数量越大,过拟合现象越严重,识别率越低。故本文采用预训练等策略提高识别性能。

传统的x-vector系统帧处理层为5层TDNN构成,但特征提取能力与ResNet-34相比依旧不足。RN-xvector系统相较于TN-xvector系统在男性和女性测试集上EER分别相对提升31.79%、22.31%,MinDCF相对提升15.57%、26.81%。表明网络层次的加深,会进一步增强声纹特征的提取能力,说明

表2 Part Ⅲ 测试集实验结果

Table 2 Experimental results on test set of Part Ⅲ

实验系统	男性测试集		女性测试集	
	EER/%	MinDCF	EER/%	MinDCF
GMM i-vector	2.968	0.339 1	4.402	0.474 8
基于内容的i-vector	2.253	0.226 4	2.510	0.255 2
RSR-TN-xvector	12.77	0.885 9	14.20	0.916 5
RSR-RN-xvector	16.76	0.946 6	18.56	0.973 0
TN-xvector	3.001	0.348 7	3.052	0.398 3
基于内容的TN-xvector	2.351	0.256 2	2.421	0.261 7
RN-xvector	2.047	0.294 4	2.371	0.291 5
基于内容的RN-xvector	1.733	0.244 1	1.904	0.219 6

帧处理层的替换对于提取声学特征中的说话人信息有一定的作用。

文本内容作为辅助信息的应用在文本相关的SV实验中也取得了一定的效果。基于内容的i-vector系统相较于GMM i-vector系统在男性和女性测试集上EER分布分别提升24.09%、42.95%，MinDCF相对提升33.24%、46.25%。基于内容的TN-xvector系统相较于TN-xvector系统在男性和女性测试集上EER分布分别提升21.66%、20.67%，MinDCF相对提升26.53%、34.3%。基于内容的RN-xvector系统相较于RN-xvector系统在男性和女性测试集上EER分别提升15.34%、19.7%，MinDCF相对提升17.08%、24.66%。充分验证了内容建模的有效性，体现了基于内容的说话人信息提取的鲁棒性。

RN-xvector系统是将x-vector系统应用到文本相关的SV任务中，并且使用性能更优的ResNet-34替换传统的TDNN网络，并针对文本内容分别建模。相较于其他7个主流系统，在男性和女性测试集上都获得了一致的性能提升。

4 结束语

本文提出并构建了基于内容的x-vector系统，该系统针对一句话中的不同字分别利用深度神经网络进行前端建模，取代了传统方法中对整句话的建模。在RSR2015数据集Part Ⅲ上的SV实验结果表明：基于内容的x-vector系统相对于x-vector系统在测试集上的性能有很大提升，说明了本文所提出方法的有效性。进一步与基于内容的i-vector系统相比，性能提升更加明显。下一步准备改进传统的后端信道补偿算法以及得分规整来进一步提高实验性能。

References:

[1] PODDAR A, SAHIDULLAH M, SAHA G. Speaker verification with short utterances: A review of challenges, trends and opportunities[J]. IET Biometrics, 2017, 7(2): 91-101.

[2] DEHAK N, KENNY P J, DEHAK R, et al. Front-end factor analysis for speaker verification[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2010, 19(4): 788-798.

[3] DEBNATH S, SONI B, BARUAH U, et al. Text-dependent speaker verification system: A review[C]//Proceedings of 2015 IEEE 9th International Conference on Intelligent Systems and Control (ISCO). [S.l.]: IEEE, 2015: 1-7.

[4] MAGHSOODI N, SAMETI H, ZEINALI H, et al. Speaker recognition with random digit strings using uncertainty normalized HMM-based i-vectors[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(11): 1815-1825.

[5] CHEN L, LEE K A, MA B, et al. Phone-centric local variability vector for text-constrained speaker verification[C]//

Proceedings of Sixteenth Annual Conference of the International Speech Communication Association. [S.l.]: [s.n.], 2015: 229-233.

[6] CHEN P, QUO W, HU G. Digit-dependent local i-vector for text-prompted speaker verification with random digit sequences [C]//Proceedings of 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP). [S.l.]: IEEE, 2016: 1-5.

[7] 戴礼荣, 张仕良, 黄智颖. 基于深度学习的语音识别技术现状与展望[J]. 数据采集与处理, 2017, 32(2): 221-231.
DAI Lirong, ZHANG Shiliang, HUANG Zhiying. Deep learning for speech recognition: Review of state-of-the-arts technologies and prospects[J]. Journal of Data Acquisition & Processing, 2017, 32(2): 221-231.

[8] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2018: 5329-5333.

[9] SNYDER D, GARCIA-ROMERO D, POVEY D, et al. Deep neural network embeddings for text-independent speaker verification[C]//Proceedings of Interspeech. [S.l.]: [s.n.], 2017: 999-1003.

[10] PEDDINTI V, POVEY D, KHUDANPUR S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Proceedings of Interspeech. Dresden, Germany: [s.n.], 2015: 3214-3218.

[11] ZHANG X, TRMAL J, POVEY D, et al. Improving deep neural network acoustic models using generalized maxout networks [C]//Proceedings of 2014 IEEE International Conference onAcoustics, Speech and Signal Processing (ICASSP). [S.l.]: IEEE, 2014: 215-219.

[12] BOUSQUET P M, LARCHER A, MATROUF D, et al. Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis[C]//Proceedings of Odyssey Speaker and Language Recognition Workshop. Singapore: [s.n.], 2012: 157-164.

[13] 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. [S.l.]: IEEE, 2016: 770-778.

[14] NAGRANI A, CHUNG J S, ZISSERMAN A. VoxCeleb: A large-scale speaker identification dataset[J]. Telephony, 2017, 3: 33-39.

[15] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep speaker recognition[C]//Proceedings of Interspeech. Hyderabad, India: [s.n.], 2018:1086-1090.

[16] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]//Proceedings of International Conference on Machine Learning. Liue, France: [s.n.], 2015: 448-456.

[17] LARCHER A, LEE K A, MA B, et al. Text-dependent speaker verification: Classifiers, databases and RSR2015[J]. Speech Communication, 2014, 60: 56-77.

[18] SADJADI S O, KHEYRKHAH T, TONG A, et al. The 2016NIST speaker recognition evaluation[C]//Proceedings of Interspeech. Stockholm, Sweden: IEEE, 2017: 1353-1357.

作者简介:

陈亚峰(1997-),男,硕士研究生,研究方向:声纹识别,E-mail: yfchen97@mail.ustc.edu.cn。

郭武(1973-),男,博士,副教授,研究方向:语音信号处理, E-mail:guowu@ustc.edu.cn。

(编辑:张彤)