DOI:10.7544/issn1000-1239.2019.20180471
56(5):1082-1091,2019

基于降噪自动编码器的语种特征补偿方法

苗晓晓[1,2] 徐 及[1,2] 王 剑[1]
[1](中国科学院声学研究所语言声学与内容理解重点实验室 北京 100190)
[2](中国科学院大学 北京 100190)
(miaoxiaoxiao@hccl.ioa.ac.cn)

Denoising Autoencoder-Based Language Feature Compensation

Miao Xiaoxiao[1,2], Xu Ji[1,2], and Wang Jian[1]
[1](*Key Laboratory of Speech Acoustics and Content Understanding*, *Institute of Acoustics*, *Chinese Academy of Sciences*, *Beijing* 100190)
[2](*University of Chinese Academy of Sciences*, *Beijing* 100190)

Abstract Language identification (LID) accuracy is often significantly reduced when the duration of the test data and the training data are mismatched. This paper proposes a method to compensate language features using a denoising autoencoder (DAE). Use of denoising autoencoder-based language feature compensation can map language features from variable length utterances into a fixed length representation. Therefore the problem of length mismatch and unbalanced phoneme distribution can be mitigated. The algorithm first converts the speech signal to low level acoustic features by framing and transforming, and then estimates its i-vector and phonetic vector. These two vectors are then concatenated and fed into the DAE-based language feature compensation processing unit. The compensated i-vector from the output of the DAE, and the original i-vector, are presented to the back-end classifier to obtain two score vectors. These two score vectors are finally fused at a score level to obtain a final result. Tests on NIST-LRE07 demonstrate that this feature compensation method improves identification performance over various test speech durations. Compared with traditional LID systems, the performance for 30 s test utterances improves by 3.16%, while the performance for 10 s test utterances improves by 2.90%. Compared with the end-to-end LID system, the performance on 3 s test utterances is increased by 3.21%.

Key words language identification (LID); i-vector; phoneme vector; feature compensation; denoising autoencoder (DAE)

摘 要 在语种识别中,当训练语音与测试语音长度失配时,系统的识别性能会出现严重下降.基于降噪自动编码器(denoising auto-encoder, DAE)的方法对不同长度测试语音的语种特征进行补偿,把不同长度的语音特征都映射为固定长度的语音特征,一定程度上解决了长度失配和音素分配不平衡的问题.具体分为4个环节:1)语音信号经过分帧、变换得到底层声学特征;2)提取语音信号的原始i-vector,同时计算其音素向量;3)对原始i-vector和音素向量进行拼接,送入基于DAE的语种特征补偿处理单元得到补偿后的i-vector;4)将补偿后的i-vector和原始i-vector分别送入后端分类器得到2个分数向量,

收稿日期:2018-06-27;**修回日期**:2018-12-10
基金项目:国家重点研发计划项目(2016YFB0801203,2016YFB0801200)
This work was supported by the National Key Research and Development Program of China (2016YFB0801203, 2016YFB0801200).
通信作者:徐及(xuji@hccl.ioa.ac.cn)

并将其在得分域融合后进行判决.在 NIST-LRE07 上的实验结果表明:所提出的语种特征补偿算法在各种测试语音时长上的识别性能均有提升.相比传统的语种识别系统,测试语音时长为 30 s 时性能相对提升 3.16%,测试语音时长为 10 s 时性能相对提升 2.90%.相比端到端语种识别系统,测试语音时长为 3 s 时性能相对提升 3.21%.

关键词 语种识别;i-vector;音素向量;特征补偿;降噪自动编码器

中图法分类号 TN912.3; TP18

语种识别(language identification, LID)是指自动判定给定语音段,从该语音信号中提取各语种的差异信息,判断语言种类的过程[1].人类的听觉系统是世界上最准确的语种识别系统.经过短时间的训练学习,人们能够快速准确地判定语种类别.即使是不熟悉的语种,人们也能对其语种与所知道的语种做出一个粗略的判断.语种识别的目标是将这种能力赋予计算机使语种分类自动化.根据测试集语种类别,可以将语种识别任务分成闭集语种识别和开集语种识别.闭集的训练集语种类别包含所有测试集语种类别,而开集的训练集并没有包含所有测试集语种类别[1].近年来,语种识别技术在这 2 个任务上已取得了长足的进步.然而仍有很多不足,尤其是面对短时语音段语种识别、高混淆度的语言识别、大量集外语种任务时.本文主要针对闭集语种识别问题展开研究.

传统的语种识别技术可分为基于音素层特征和基于声学层特征.基于音素层特征的语种识别技术是将音素层特征作为识别依据,主要考虑了不同语种有着不同的音素集合,音节和音素出现的频率有很大差异,以及音节和音素的组合大不相同等因素.一般通过音素识别器,先将语音信号解码为音素序列或者音素网格.在建模阶段,通常为每个语种建立 N 元文法(Ngram)模型或者 N 元文法统计量模型[2].常用的方法有音素识别后接 N 元文法模型(phoneme recognizer followed by language model, PRLM)[2]、并行音素识别器后接语言模型(parallel phone recognition followed by language modeling, PPRLM)[3-4]和并行音素识别器后接向量空间模型(parallel phoneme recognizer followed by vector space model, PPRVSM)[5]等.基于声学层特征的语种识别技术依赖于声学层特征.通过对语音信号分帧、变换提取声学层特征,采用概率统计或鉴别性方法对其建模.常用的声学层特征有美尔频率倒谱系数(Mel-frequency cepstral coefficient, MFCC)[6]、感知线性预测系数(perceptual linear predictive, PLP)[7]、滑动差分倒谱(shifted delta cepstrum, SDC)[8]等.主流系统有混合高斯模型-全局背景模型(Gaussian mixture model-universal back-ground model, GMM-UBM)[9]、高斯超向量-支持向量机(GMM super vector-support vector machines, GSV-SVM)[10]和基于全差异空间的(total variability, TV) i-vector 系统[11]等.

近几年,深度神经网络(deep neural networks, DNNs)[12]模型在语种识别任务上得到快速发展.一方面从前端特征提取层面,蒋兵等人[13]利用 DNN 强大的特征抽取能力,提取了深度瓶颈特征(deep bottleneck feature, DBF);另一方面从模型域出发,Lei 等人[14]提出了基于 DNN 的 TV 建模策略.

此外也出现了基于深度学习的端对端语种识别系统,摒弃了传统的语种识别系统框架.2014 年 Google 的研究人员[15]将特征提取、特征变换和分类器融于一个神经网络模型中,这是端对端系统首次被成功应用于语种识别任务.随后有研究人员在此基础上发掘了不同神经网络的优势,包括延时神经网络(time-delay neural network, TDNN)[16]、长短时记忆递归神经网络(long short term memory-recurrent neural network, LSTM-RNN)[17].2016 年 Geng 等人[18]利用注意力机制模型(attention-based model),结合 LSTM-RNN 搭建了端对端语种识别系统,也取得了不错的语种识别性能.2018 年 Jin 等人[19]提出了基于 LID-net、LID-bnet 的端对端语种模型,并利用 LID-net 模型提取统计量,对比原有的 DNN-TV 系统,从网络中间层中获取 LID-senone 特征,证明了这个特征更具有语种区分性.同年 Cai 等人[20-22]提出了一种基于可学习的字典编码层的端对端系统,从底层声学特征直接学习语种类别信息,摒弃了声学模型,也取得了较优的识别性能.

Vincent 等人[23]于 2008 年提出降噪自动编码器(denoising auto-encoder, DAE).该神经网络用于抑制输入信号中的噪声因子,能够有效加强系统的鲁棒性,广泛应用于语音增强[24]、语音信号去混响[25]等

领域. 2015 年 Yamamoto 等人[26]首次将 DAE 应用于短时说话人识别,2017 年 Yang 等人[27]在此基础上,改进并提出了基于 DAE 的短时说话人 i-vector 补偿技术,取得了较好的说话人识别性能.

目前,将降噪自动编码器用于语种识别领域的研究鲜见报道. 本文基于国内外关于语种识别和降噪自动编码器的研究,实现了基于降噪自动编码器的语种特征补偿方法. 本文的主要贡献有 2 个方面:

1) 目前的语种识别系统在训练语音与测试语音长度匹配的情况下具有较高的识别率,而当长度失配时,其性能也随之下降. 为了解决这个问题,本文提出一种基于降噪自动编码器的语种特征补偿方法,将不同长度的语音特征都映射为固定长度的语音特征,一定程度上解决了长度失配问题.

2) 语音学的研究[28]表明,在世界范围内,几乎没有任何语音拥有相同的音素集合,即使有些语言共用同一套音素体系,音素出现的频率也有所差别. 因此在语种识别中,音素作为重要的特征,可以被用来有效区分语种. 而事实上,测试语音小于 10 s 或者更短时,语音的音素分布严重不均衡[29]. 在这种情况下语种特征的提取也是不可信的,极大地影响了识别性能. 基于 DAE 的语种特征补偿算法,将短时语种特征映射到长时语种特征空间,以得到音素分布更为平衡的短时语音段表示,缓解了短时测试语音音素分布不平衡的问题.

1 相关工作

本文搭建了目前国际主流的 2 个语种识别系统:基于全差异空间的语种识别系统和基于端对端神经网络的语种识别系统,并以此作为实验的基线系统. 二者建模方法有所差异,前者是传统的生成式的建模方法,将 GMM 均值超矢量映射成低维向量,再利用语种标签信息训练得到后端分类器. 而后者在整个训练过程中直接使用深度神经网络进行语种识别,这也是当前语种识别方向的研究热点.

1.1 基于全差异空间的语种识别系统

基于全差异空间的方法首先被成功应用于说话人识别任务[30-31],随后被迁移到语种识别领域[32-33],成为语种识别领域的主流方法之一. 基本框架如图 1 所示:

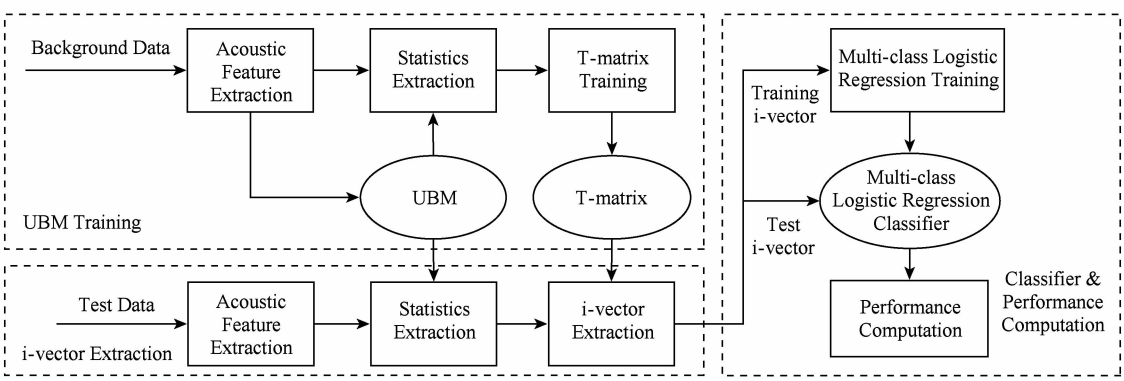

Fig. 1 structure of the GMM i-vector LID system

图 1 GMM i-vector 语种识别系统框图

TV 系统通过定义一个低维空间,该空间不区分捕获到的差异信息是否与语种、说话人和信道信息相关,而是将 GMM 超向量[10]中的语种、说话人变化空间和信道变化空间合并为全差异空间来进行建模. 因为强制分离空间的话有可能会因为分离的不正确而丢失重要的信息,这些信息在后端建模中是无法弥补的. 为避免有关语种的有效信息丢失,假设所有的空间都被合并成一个统一的空间,由全差异矩阵 $\boldsymbol{T}$ 来构建. 直观上来讲,全差异空间会将高维的 GMM 超向量 $\boldsymbol{M}$ 映射成一个低维向量. 假设 $\boldsymbol{M}$ 能被分解为

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

其中,$\boldsymbol{m}$ 是与语种和信道等无关的超向量,即 UBM 的均值超向量. 因为 UBM 是由背景数据训练得到的,无任何先验信息,因此可以认为 $\boldsymbol{m}$ 与语种、说话人、信道等无关. $\boldsymbol{T}$ 为全差异矩阵,$\boldsymbol{w}$ 是全差异因子,也叫 i-vector,是对 GMM 超向量的低维表示. i-vector 作为模型中的隐含变量,满足高斯分布,其后验分布:

$$\boldsymbol{w}(u)=(\boldsymbol{I}+\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{N}(u)\boldsymbol{T})^{-1}\boldsymbol{T}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{F}(u), \tag{2}$$

其中,$\boldsymbol{\Sigma}$ 是对角协方差矩阵,定义了 GMM 超向量中未被 TV 空间描述的噪声部分. $\boldsymbol{N}(u)$ 和 $\boldsymbol{F}(u)$ 每句

训练语音在 UBM 上的 Baum-Welch 统计量:

$$\boldsymbol{N}_c(u)=\sum_{t=1}^{L}p(c\mid\boldsymbol{u}_t),\tag{3}$$

$$\boldsymbol{F}_c(u)=\sum_{t=1}^{L}p(c\mid\boldsymbol{u}_t)(\boldsymbol{u}_t-\boldsymbol{m}_c),\tag{4}$$

其中,$p(c|\cdot)$表示在 UBM 的第 c 个高斯上的后验概率;$\boldsymbol{u}_t$ 表示训练语音 u 的第 t 帧特征,共 L 帧;$\boldsymbol{m}_c$ 表示在 UBM 的第 c 个高斯上的均值向量.

换一种角度理解,TV 方法是将每段语音当作独立的个体,认为每段都属于不同的语种,再根据各语音段与 UBM 均值超向量的差异性得到具有语种区分性的低维因子表示,也可以被认为是一种将语音信号特征映射到一个低维空间的概率主成分分析.

从以上的描述中,可以看出在估计全差异矩阵和提取 i-vector 的过程中,并没有用到语种类别信息,因此还需要对 i-vector 进行区分性训练,可以使用逻辑回归作为后端分类器,从而得到包含更多语种信息的 i-vector.

1.2 基于端对端神经网络的语种识别系统

近年来,深度学习在语音信号处理领域中得到了快速的发展,基于端对端神经网络的语种识别系统也逐渐成为主流.它摒弃了传统的全差异空间建模的方法,在训练过程中将特征提取、变换以及后端分类器融于一个神经网络中,引入语种标号进行区分性训练.相比传统的 TV 系统,端对端网络有 2 个优势.1)运用了区分性的建模方法,在整个网络的训练过程中,企图寻找不同类别之间的最佳分类面,并没有侧重去拟合数据的分布.2)直接利用语种标签信息,不断优化网络参数,使得所提特征更具语种特性.而 TV 建模仅仅在后端分类器的训练过程中引入了语种标签信息.由于研究时间过短,端对端网络还有许多可改进的空间,是当前的研究热点之一.

本文实现了文献[20-21]提出的一种新的端对端语种识别系统,它结合了卷积神经网络(convolutional neural network, CNN)在帧级特征上强大的建模能力和时域平均池化层(temporal average pooling, TAP)将帧级特征转换到句级特征的池化能力.该系统被称为 CNN-TAP,基本框架如图 2 所示.

具体来说,CNN-TAP 包含一个基于卷积层的前端特征提取器,设语音信号的声学特征为 $\boldsymbol{x}_1$,$\boldsymbol{x}_2$,…,$\boldsymbol{x}_T$,共 T 帧.其中 $\boldsymbol{x}_T$ 表示第 T 帧特征经过非线性变换映射后,得到的特征包含更多语种信息;接着利用池化层得到句级特征表示,经过 TAP 层后,

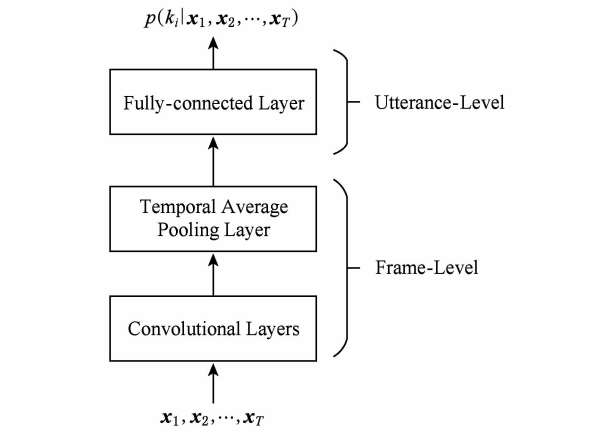

Fig. 2 Structure of the CNN-TAP end-to-end LID system

图 2 CNN-TAP 端到端语种识别系统框图

不同长度的输入语音得到了固定维度的句级向量表示;最后得到类别后验概率 $p(k_i\mid\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T)$,$k_i$ 表示第 i 个类别.

2 语种特征修复

本文所提出的语种特征补偿方法框架如图 3 所示.首先从语音信号分帧、变换得到底层声学特征;之后利用 1.1 节描述的基线系统提取原始 i-vector;同时按照 2.1 节所提出的方法计算音素向量;将拼接后的 i-vector 和音素向量,送入基于 DAE 的语种特征补偿处理单元映射得到补偿后的 i-vector,详见 2.2 节;最后将补偿后的 i-vector 和原始 i-vector 分别送入后端分类器得到分数向量,并将其在得分域融合.

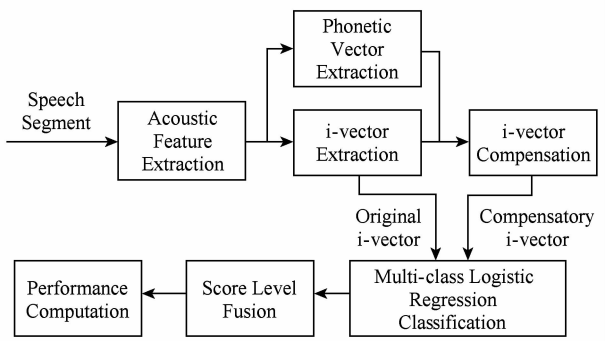

Fig. 3 Structure of DAE based i-vector feature compensation

图 3 基于 DAE 的 i-vector 补偿结构框图

2.1 音素向量

语言学研究表明:不同音素集、不同音素的组合及出现的频率表征了不同类别的语言[28].显而易见,音素信息在语种识别中有很重要的作用.本文尝

试用较为简单的方法提取语音的音素信息作为辅助信息,增加不同语种之间的区分度.

音素信息的提取过程为:先利用高斯混合模型计算语音信号每帧的后验概率,再将一句话所有帧的后验概率求和取平均作为新的特征,这里称为音素向量,计算为

$$p_c(u)=\frac{1}{L}\sum_{t=1}^{L}p(c\mid \boldsymbol{u}_t), \tag{5}$$

其中,$p_c(u)$是训练语音 u 在 UBM 的第 c 个高斯上的后验概率;t 表示帧数,共 L 帧.

2.2 基于降噪自动编码器的 i-vector 补偿方法

降噪自动编码器是自动编码器(auto-encoder, AE)的改良版,旨在用被破坏的输入数据重构出原始未被破坏的数据,从而提取、编码出具有鲁棒性的特征.在语音信号处理领域,DAE 被用于加强语音信号,包括去噪、去混响等.

本文提出基于降噪自动编码器的 i-vector 补偿方法,结构框图如图 4 所示.DAE 补偿网络的特征向量由语音的 i-vector $\boldsymbol{w}(u)$和音素向量 $\boldsymbol{p}(u)$拼接组成.输入短时语音语种特征向量 $\boldsymbol{x}_1(u)=[\boldsymbol{w}_1(u),\boldsymbol{p}_1(u)]$,标签为长时语音语种特征向量 $\boldsymbol{x}_L(u)=[\boldsymbol{w}_L(u),\boldsymbol{p}_L(u)]$.隐层之间的前向传递过程为

$$\boldsymbol{x}_{i+1}(u)=g(\boldsymbol{W}_{(i,i+1)}\boldsymbol{x}_i(u)+\boldsymbol{b}_{(i,i+1)}), \tag{6}$$

其中,$g()$为非线性函数,$\boldsymbol{W}_{(i,i+1)}$和 $\boldsymbol{b}_{(i,i+1)}$为前一隐层与后一隐层之间的权重参数和偏置参数,最后输出层为补偿语种特征 $\boldsymbol{x}_N(u)=[\boldsymbol{w}_N(u),\boldsymbol{p}_N(u)]$,称为补偿向量.$\boldsymbol{w}_N(u)$为补偿后的 i-vector,$\boldsymbol{p}_N(u)$为补偿后的音素向量.

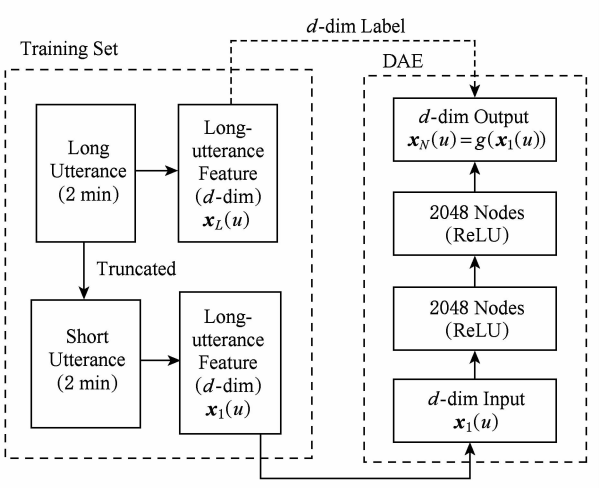

Fig. 4 The diagram for language feature compensation

图 4 语种特征补偿方法图示

该网络的训练是一种有监督的训练.主要思想是输入被干扰的向量,重构出原始输入.本文实验中输入的是短时语音语种特征,目标向量为长时语音语种特征.训练数据的准备分 3 个步骤:

1) 提取训练语音 S_i 的 i-vector 和音素向量,拼接得到目标向量 $\boldsymbol{x}(S_i)$.训练语音的长度范围是 0~2 min;

2) 将训练语音 S_i 分别切成 3 s,10 s 和 30 s,对每个语音段 $s_{i,j}$ 提取 i-vector 和音素向量,拼接构成短时向量 $\boldsymbol{x}(s_{i,j})$;

3) 最终的训练数据对为$(\boldsymbol{x}(S_i),\boldsymbol{x}(s_{i,j}))$.通过最小化目标函数对 DAE 网络进行参数优化.目标函数可以是目标向量和补偿向量之间的均方误差.

2.3 融合策略

每条语音经过语种特征补偿处理单元后得到 2 个 i-vector:原始 i-vector 和补偿后的 i-vector.为了充分利用这 2 个特征,本文在模型后端采用 2 种融合策略:i-vector 融合和得分融合.

i-vector 融合方法是将语音信号的原始 i-vector 和补偿后的 i-vector 进行线性融合后,再送入后端分类器得到最终的判决结果:

$$\boldsymbol{w}^{\mathrm{f}}(u)=(1-\alpha)\boldsymbol{w}(u)+\alpha\boldsymbol{w}^{\mathrm{comp}}(u),0\leqslant\alpha\leqslant1, \tag{7}$$

其中,$\boldsymbol{w}(u)$是原始 i-vector,$\boldsymbol{w}^{\mathrm{comp}}(u)$是它的补偿 i-vector.

得分融合方法是直接在 TV 模型输出的得分上进行线性融合:

$$s^{\mathrm{f}}(u)=(1-\alpha)s(\boldsymbol{w}(u))+\alpha s(\boldsymbol{w}^{\mathrm{comp}}(u)),0\leqslant\alpha\leqslant1, \tag{8}$$

其中,$s()$为分数后端计算得分的函数.

3 语种特征补偿实验

3.1 实验设置

1) 测试集.从 20 世纪 90 年代起,美国国家标准技术研究院(National Institute of Standards and Technology, NIST)组织的语种识别评测比赛(language recognition evaluation, LRE),为语种识别的研究提供了统一的、公共的数据集.本实验采用美国国家标准技术局在 2007 年闭集条件下的语种识别评测数据集[34].这个测试集包含 14 个语种:阿拉伯语(Arabic, AR)、孟加拉语(Bengali, BE)、英语(English, EN)、波斯语(Farsi, FA)、俄语(Russian, RU)、德语(German, GE)、印地语(Hindustani, HI)、日语(Japanese, JA)、韩语(Korean, KO)、中文(Chinese, CH)、西班牙语(Spanish, SP)、泰米尔语(Tamil, TA)、泰国语(Thai, TH)和越南语(Vietnamese, VI).其中中文除了包含普通话,还有

闽南语、吴方言和粤语.英语除了美国英语外,还包括印度英语.而印地语包括了北印度语和乌尔都语.NIST07 的测试数据按照语音的时长分为 30 s,10 s,3 s 这 3 种,每种包括 2 158 条 14 个语种的语音.

2) 训练集.训练语料主要使用 CallFriend 数据库[35].该数据库包含 12 个语种:阿拉伯语、英语、波斯语、法语、德语、印地语、日语、韩语、普通话、西班牙语、泰米尔语和越南语.LRE2007 的测试数据还有 2 个 CallFriend 数据库不包含的语种、方言,因此,训练数据在 CallFriend 数据库的基础上还添加了 LRE2003,LRE2005,LRE2007 开发集.除此之外还包含 2008 年说话人识别评测比赛(speaker recognition evaluation, SRE)[36]的训练数据,为该测试提供补充训练数据:孟加拉语,俄语、泰国语、闽南语、吴方言、粤语、阿拉伯语和乌尔都语.训练集的数据分布严重不平衡,英语、中文占所有语料的 57.06%.

语种识别的测试标准主要采用 NIST-LRE07 测试标准平均代价(average cost, C_avg)[34]和错误率(error rate, ER)来评价.另外本文在 4.2 节中还将提到虚警率、漏警率[34].这些指标从不同角度反映了语种识别系统性能的好坏,它们都是越小越好.C_avg 的定义为

$$C\_avg = \frac{1}{N_1}\Big(\sum_{L_t} C_{\text{miss}} P_{\text{target}} P_{\text{miss}}(L_{\text{target}}) + C_{\text{fa}} P_{\text{out-of-set}} P_{\text{fa}}(L_{\text{target}}, L_{\text{non-target}}) + \sum_{L_n} C_{\text{fa}} P_{\text{out-of-set}} P_{\text{fa}}(L_{\text{target}}, L_{\text{non-target}})\Big), \quad (9)$$

$$P_{\text{out-target}} = (1 - P_{\text{target}} - P_{\text{out-of-set}})/(N_1 - 1) \quad (10)$$

其中,N_1 为集合中语种总数;L_t 和 L_n 分别表示目标语种和非目标语种;$P_{\text{miss}}(L_t)$表示目标语种为 L_t 时的漏检率;$P_{\text{fa}}(L_t, L_n)$是目标语种为 L_t 时的虚警率;C_{miss}和 C_{fa}分别是漏检和虚警的惩罚因子;P_{target}为目标语种的先验概率;$P_{\text{non-target}}$为非目标语种的先验概率;$P_{\text{out-of-set}}$为集外语种的先验概率.本论文实验只考虑闭集测试的情况,因此 $P_{\text{out-of-set}}=0$. NIST-LRE07 设定 $C_{\text{miss}}=C_{\text{fa}}=1$,$P_{\text{target}}=0.5$.

本文采用 2 个基线系统:基线 1 是基于 TV 空间的 i-vector 系统,分为前端声学特征提取和后端统计建模 2 部分.前端底层声学特征采用基于 PLP 线性扩展的 SDC 特征,后端建模利用 TV 建模方法.TV 建模中高斯混合数为 512,i-vector 维数为 600,采用逻辑回归分类器.基线 2 是基于端对端深度神经网络的语种识别系统.与文献[20-21]相似,卷积神经网络采用深度残差网络 resnet34,采用交叉熵(cross entropy)准则进行训练,采用随机梯度下降法(stochastic gradient descent, SGD)更新网络参数.Mini-batch 大小设为 128,每个 Mini-batch 的帧长范围为[200,1000].共 60 个 epoch,学习率为 0.1.

随机选择 5 000 条训练数据训练高斯混合数为 32 的 UBM,提取 32 维的音素向量.基于降噪自动编码器的网络输入是 632 维向量,由 600 维 i-vector 和 32 维音素向量组成.网络共有 4 层,隐层节点数是 2048,网络的结构是 632-2048-2048-632.采用最小均方误差准则进行训练.经过对比 2 种融合策略,实验均采用得分融合策略,$\alpha=0.1$.关于融合系数 α 的选取,将在 4.2 节实验结果中给出验证.

3.2 实验结果

为了验证本文所提算法的有效性,本节将描述多组实验.首先以音素向量作为唯一变量,验证音素向量补偿的效果.表 1 列出了 i-vector 直接通过 DAE 网络映射以及音素向量和 i-vector 拼接后映射,在不同时长测试语音下的评价指标 ER 和 C_avg 的变化趋势.表 1 中的 Direct Mapping 表示提取出的 i-vector 直接经过 DAE 网络映射.而 Concatenate Mapping 表示送入 DAE 网络的特征是音素向量和 i-vector 的拼接向量.2 组实验的唯一变量是音素向量.DAE 网络由 3 s 的短时语种特征和 2 min 的长时语种特征对训练得到.实验后端均采用逻辑回归分类器.映射后的 i-vector 和原始 ivector 按照 0.1∶0.9 在得分域融合.

Table 1 The Results of Phonetic Vector Compensation

表 1 音素向量补偿性能对比 %

Different Systems	Test Utterance Length/s					
	30		10		3	
	ER	C_avg	ER	C_avg	ER	C_avg
Baseline 1	8.85	5.67	22.38	13.88	48.93	29.95
Baseline 2	8.53	5.31	21.59	12.94	50.32	29.98
Direct Mapping	8.80	5.65	22.15	13.83	48.84	29.91
Concatenate Mapping	8.78	5.64	22.10	13.75	48.70	29.87

从表 1 可以看出,相比基线 1,i-vector 直接映射后的识别性能在 3 种测试时长上均有提升,尤其是在测试语音时长为 10 s 和 3 s 的短时情况,因为本组实验的 DAE 网络训练数据由 3 s 和 2 min 语音对构成,短时测试语音与模型更匹配,性能提升更多.

相比基线 2,i-vector 直接映射和拼接映射后识别性能在 3 s 的测试时长上也有所提升. 音素向量和 i-vector 合并后,语种识别性能进一步得到改善,说明音素向量有一定的补偿效果.

下面验证补偿网络的有效性. 针对不同的测试条件,在训练阶段,长时的训练语料被切割成时长分别为 30 s,10 s 和 3 s 的短时语音段,并组成 3 种时长的短时语音语种训练集合,分别学习对应的补偿网络. 表 2 列出了针对不同测试时长的训练数据分别训练相应的补偿系统,在不同时长测试语音下的评价指标 *ER* 和 *C_avg* 的变化情况. 表 2 中的 30 s compensation 表示补偿网络的训练数据是 30 s 和 2 min 的训练对.

Table 2 Performance Comparison on Baseline and Compensation System

表 2 基线系统和补偿系统性能对比 %

Different Systems	Test Utterance Length/s					
	30		10		3	
	ER	*C_avg*	*ER*	*C_avg*	*ER*	*C_avg*
Baseline 1	8.85	5.67	22.38	13.88	48.93	29.95
Baseline 2	8.53	5.31	21.59	12.94	50.32	29.98
30 s compensation	8.57	5.59	22.15	13.77	48.84	29.90
10 s compensation	8.67	5.61	21.73	13.61	48.75	29.89
3 s compensation	8.78	5.64	22.10	13.75	48.70	29.87

从表 2 可以看出,本文提出的补偿算法在各种测试时长上的识别性能都有提高. 尤其是在测试语音时长和训练语音时长完全匹配的情况下,性能提升最大. 相比基线 1,测试语音时长为 30 s,经过 30 s 补偿网络映射后,错误率相对降低了 3.16%;测试语音时长为 10 s 的错误率经过 10 s 补偿网络映射后相对降低了 2.90%;测试语音时长为 3 s 的错误率,经过 3 s 补偿网络映射后相较基线提升不大. 这是因为使用 3 s 时长的短时训练样本由于语音段过短使其在提取 i-vector 时包含的语种相关信息非常容易受到影响,直接用于网络训练可能导致模型的估计不准确. 相比基线 2,补偿网络在测试语音时长为 30 s 和 10 s 的性能没有提升,但基本可以与基线持平;测试语音时长为 3 s 的错误率经过 3 s 补偿网络映射后相对降低了 3.21%. 总的来说,针对不同测试时长,利用时长互相匹配的训练数据分别训练相应的补偿系统能进一步提升补偿网络系统的性能.

Vincent 等人[23]证明了 DAE 网络的可以对加噪数据进行降噪,得到更加鲁棒的不变性特征,获得输入的更有效表达. 文献[22]提到一般情况下,高维的数据都处于一个较低维的流形曲面上,被干扰向量的分布稀疏,远离低维曲面. 而 DAE 网络可以将稀疏的特征映射到更为紧实的低维曲面上. 这与本文所描述的基于 DAE 的补偿网络的主要思想是一致的,短时特征可以被当做是长时特征经过噪声干扰后的数据. 补偿网络得输入是短时语音的 i-vector,通过学习隐含特征,可以有效重构出长时语音的 i-vector.

图 5 显示出 30 s,10 s,3 s 测试时长的 *C_avg* 随补偿 i-vector 权重 α 变化的情况,不同测试时长语音分别经过对应时长的补偿网络,采用式(8)的得分融合策略.

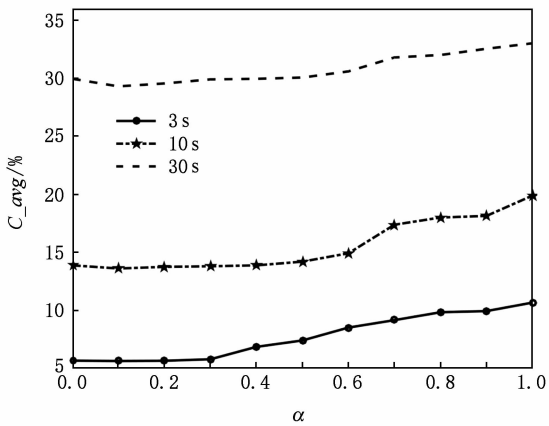

Fig. 5 The variation of *C_avg* with respect to score fusion coefficient

图 5 *C_avg* 随得分融合系数变化图

从图 5 中可以看出,α 取 0.1～0.2 时融合系统的性能稍好于基线 1,$\alpha=0.1$ 时在 3 种测试时长下性能最优. 这说明基于 DAE 的补偿网络提供了额外的语言信息,与原始 i-vector 语言信息互补.

混淆矩阵可以很直观地反映出各语种的虚警率、漏警率以及各语种间的混淆程度. 本文统计了测试语音在 3 s 补偿系统上的得分所构成的混淆矩阵. 如图 6 所示,纵轴代表真实类别,横轴代表预测类别. ACC 代表该语种识别精度. 从图 6 可以看出,各语种大部分样本都能保证分类正确,彼此间不存在较大的干扰. 但是某些类别之间混淆程度较大导致系统的虚警率和漏警率升高. 3.1 节中提到训练数据分布不均衡,其中 MA 所占比例最大,它们的识别精度分别为 71.94%,84.50%,虚警率为 31.02% 和 22.97%. 这是因为这部分语料比重大,在训练阶段系统会侧重优化这些语种的模型,使得最终模型的预测结果有偏差. 还有一些语种即使训练语料所

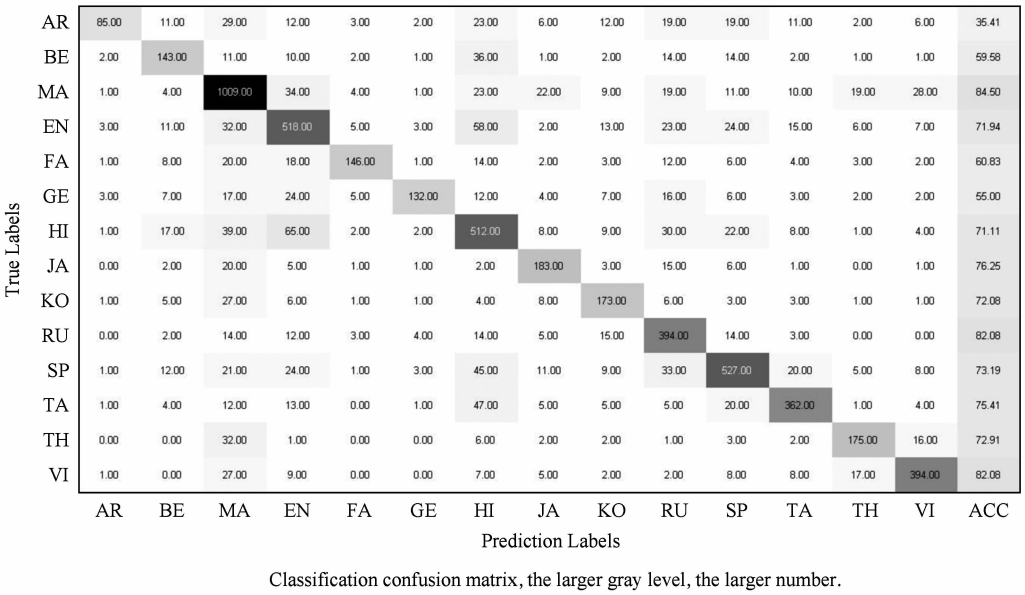

Fig. 6 Classification confusion matrix

图 6 分类混淆矩阵

占比例不大,却仍存在很高的虚警率,这说明简单的依赖底层声学特征是无法准确地给出易混语种的判别结果.从分析结果可以看出本文提出的模型虽然较基线系统有一定的性能提升,但还存在很大的提升空间.

4 总 结

语种识别系统在训练语音和测试语音长度不匹配时,性能会出现大幅下滑,严重制约语种识别技术在实际中的应用.本文提出基于降噪自动编码器的语种特征补偿方法来解决这个问题.在 NIST-LRE07 上的实验结果表明,所提出的语种特征补偿算法在 30 s,10 s 和 3 s 这 3 种测试条件下均可以获得不同程度的性能提升.对实验细节的进一步分析表明:目前音素向量提取过程较为简单,并没有充分挖掘出语音的音素分布信息,未能起到较大作用,后期仍有进一步改良的空间.

参 考 文 献

[1] Li Haizhou, Ma Bin, Lee K. Spoken language recognition: From fundamentals to practice [J]. Proceedings of the IEEE, 2013, 101(5): 1136-1159

[2] Zissman M A. Comparison of four approaches to automatic language identification of telephone speech [J]. IEEE Transactions on Speech and Audio Processing, 1996, 4(1): 31

[3] Yan Yonghong, Barnard E. An approach to automatic language identification based on language-dependent phone recognition [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 1995: 3511-3514

[4] Yan Yonghong, Barnard E, Cole R A. Development of an approach to automatic language identification based on phone recognition [J]. Computer Speech & Language, 1996, 10(1): 37-54

[5] Li Haizhou, Ma Bin, Lee C. A vector space modeling approach to spoken language identification [J]. IEEE Transactions on Audio, Speech & Language Processing, 2007, 15(1): 271-284

[6] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1980, 28(4): 65-74

[7] Hermansk H. Perceptual linear predictive (PLP) analysis of speech [J]. The Journal of the Acoustical Society of America, 1990, 87(4): 1738-1752

[8] Torres-Carrasquillo P A, Singer E, Kohler M A, et al. Approaches to language identification using Gaussian mixture models and shifted delta cepstral features [C] //Proc of the 7th Int Conf on Spoken Language Processing. Piscataway, NJ: IEEE, 2002: 89-92

[9] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models [J]. Digital Signal Processing, 2000, 10(1/2/3): 19-41

[10] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speakers verification [J]. IEEE Signal Processing Letters, 2006, 13(5): 308-311

[11] Dehak N, Torres-Carrasquillo P A, Reynolds D A, et al. Language recognition via i-vectors and dimensionality reduction [C] //Proc of the 12th Annual Conf of the Int Speech Communication Association. Baixas, Florence: International Speech and Communication Association, 2011: 857-860

[12] Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313(5786): 504-507

[13] Jiang Bing, Song Yan, Wei Si, et al. Performance evaluation of deep bottleneck features for spoken language identification [C] //Proc of the 9th Int Symp on Chinese Spoken Language Processing. Piscataway, NJ: IEEE, 2014: 143-147

[14] Lei Yun, Scheffer N, Ferrer L, et al. A novel scheme for speaker recognition using a phonetically-aware deep neural network [C] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2014: 1695-1699

[15] Lopez-Moreno I, Gonzalez-Dominguez J, Plchot O, et al. Automatic language identification using deep neural networks [C] //Proc of the 39th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2014: 5337-5341

[16] Garcia-Romero D, McCree A. Stacked long-term TDNN for spoken language recognition [C] //Proc of the 17th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2016: 3226-3230

[17] Gonzalez-Dominguez J, Lopez-Moreno I, Sak H, et al. Automatic language identification using long short-term memory recurrent neural networks [C] //Proc of the 15th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2014: 2155-2159

[18] Geng Wang, Wang Wenfu, Zhao Yuanyuan, et al. End-to-end language identification using attention-based recurrent neural networks [C] //Proc of the 17th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2016: 2944-2948

[19] Jin Ma, Song Yan, McLoughlin I. LID-senones and their statistics for language identification [J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2018, 26(1): 171-183

[20] Cai Weicheng, Cai Zexin, Liu Wenbo, et al. Insights into end-to-end learning scheme for language identification [C/OL] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018 [2018-05-20]. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Insights+into+end-to-end+learning+scheme+for+language+identification&btnG

[21] Cai Weicheng, Cai Zexin, Zhang Xiang, et al. A novel learnable dictionary encoding layer for end-to-end language identification [C/OL] //Proc of IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2018 [2018-05-20]. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=A+novel+learnable+dictionary+encoding+layer+for+end-to-end+language+identification+&btnG

[22] Cai Weicheng, Chen Jinkun, Li Ming. Exploring the encoding layer and loss function in end-to-end speaker and language recognition system [C/OL] //Proc of Speaker Odyssey. 2018 [2018-06-01]. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Exploring+the+Encoding+Layer+and+Loss+Function+in+End-to-End+Speaker+and+Language+Recognition+System+&btnG

[23] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders [C] //Proc of the 25th Int Conf on Machine learning. New York: ACM, 2008: 1096-1103

[24] Lu Xugang, Tao Y, Matsuda S, et al. Speech enhancement based on deep denoising autoencoder [C] //Proc of the 14th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2013: 436-440

[25] Ishii T, Komiyama H, Shinozaki T, et al. Reverberant speech recognition based on denoising autoencoder [C] //Proc of the 14th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2013: 3512-3516

[26] Yamamoto H, Koshinaka T. Denoising autoencoder-based speaker feature restoration for utterances of short duration [C] //Proc of the 16th Annual Conf of the Int Speech Communication Association. Dresden, Germany: International Speech and Communication Association, 2015: 1052-1056

[27] Yang I, Heo H, Yoon S, et al. Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using deep neural network [C] //Proc of the 42nd IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 5490-5494

[28] Muthusamy Y K. A segmental approach to automatic language identification [D]. Portland, Oregon, USA: Oregon Health & Science University, 1993

[29] Hasan T, Saeidi R, Hansen J H L, et al. Duration mismatch compensation for i-vector based speaker recognition systems [C] //Proc of the 38th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2013: 7663-7667

[30] Dehak N, Kenny P, Dehak R, et al. Frontend factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech & Language Processing, 2011, 19(4): 788-798

[31] Dehak N, Dehak R, Kenny P, et al. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification [C] //Proc of the 10th Annual Conf of the Int Speech Communication Association. Brighton, United Kingdom: International Speech and Communication Association, 2009: 1559-1562

[32] Dehak N, Torres-Carrasquillo P A, Reynolds D A, et al. Language recognition via i-vectors and dimensionality reduction [C] //Proc of the 12th Annual Conf of the Int Speech Communication Association. Baixas, France: International Speech and Communication Association, 2011: 857-860

[33] Yang Jinchao, Zhang Xiang, Suo Hongbin, et al. Language recognition with language total variability [C] //Proc of the 2011 Int Conf on Innovative Computing and Cloud Computing. New York: ACM, 2011: 6-9

[34] Martin A F, Le A N. NIST 2007 language recognition evaluation [S/OL]. Gaithersburg, Maryland: National Institute of Standards and Technology. [2017-09-20]. https://catalog.ldc.upenn.edu/docs/LDC2009S04/LRE07EvalPlan-v8b-1.pdf

[35] CallFriend Corpus. Linguistic data consortium [S/OL]. [2017-09-20]. http://www.ldc.upenn/ldc/about/callfriend.html

[36] National Institute of Standards and Technology. The NIST Year 2008 Speaker Recognition Evaluation Plan [S/OL]. Gaithersburg, Maryland: National Institute of Standards and Technology. [2017-10-20]. https://www.nist.gov/sites/default/files/documents/2017/09/26/sre08_evalplan_release4.pdf

Miao Xiaoxiao, born in 1994. PhD candidate. Her main research interests include language identification, speaker recognition and deep learning.

Xu Ji, born in 1986. PhD, associate professor. His main research interests include speech recognition, deep learning and ocean acoustics.

Wang Jian, born in 1977. Master, associate professor. His main research interests include speech signal processing.