·多媒体技术及应用· 文章编号:1000-3428(2018)05-0262-06 文献标志码:A 中图分类号:TN912

基于改进 i-vector 的说话人感知训练方法研究

梁玉龙,屈 丹,邱泽宇

(解放军信息工程大学 信息系统工程学院,郑州 450002)

摘 要: 基于辨识向量(i-vector)的说话人感知训练方法使用 MFCC 作为输入特征对 i-vector 进行提取,但 MFCC 较差的特征鲁棒性会影响该训练方法的识别性能。为此,提出一种基于改进 i-vector 的说话人感知训练方法。设计基于 SVD 的低维特征提取方法,用其提取的特征替代 MFCC 对表征能力更优的 i-vector 进行提取。实验结果表明,在捷克语语料库中,相对于 DNN-HMM 语音识别系统与原始基于 i-vector 的说话人感知训练方法,该方法的识别性能分别提升了 1.62% 与 1.52% ,在 WSJ 语料库中,该方法识别性能分别提升了 3.9% 和 1.48% 。

关键词: 说话人感知训练;辨识向量;深度神经网络;奇异值矩阵分解;瓶颈特征

中文引用格式: 梁玉龙,屈 丹,邱泽宇. 基于改进 i-vector 的说话人感知训练方法研究[J]. 计算机工程,2018,44(5):262-267.

英文引用格式: LIANG Yulong,QU Dan,QIU Zeyu. Research on Speaker Aware Training Method Based on Improved i-vector[J]. Computer Engineering,2018,44(5):262-267.

Research on Speaker Aware Training Method Based on Improved i-vector

LIANG Yulong,QU Dan,QIU Zeyu

(School of Information and Systems Engineering,PLA Information Engineering University,Zhengzhou 450002,China)

【Abstract】 The performance of speaker aware training method based on i-vector is poor because of using MFCC which has the relative poor robustness as the input feature for the extraction of the i-vector. To solve this problem,an improved i-vector based speaker aware training method is proposed. Firstly,a low dimensional feature extraction method based on SVD is proposed,and then the feature extracted by this method is used to replace the MFCC,which can extract better i-vector. Experimental results show that,in the Vystadial_cz corpus,compared with the DNN-HMM speech recognition system and the original i-vector based speaker aware training method,the recognition performance of this method is increased by 1.62% and 1.52% respectively,in the WSJ corpus,the recognition performance of this method is increased by 3.9% and 1.48% respectively.

【Key words】 speaker aware training; i-vector; Deep Neural Network (DNN); Singular Value Matrix Decomposition (SVMD); bottleneck feature

DOI: 10.19678/j.issn.1000-3428.0046946

0 概述

近年来,在连续语音识别应用中存在一个难以忽视的问题,即由训练数据与测试数据间的说话人不匹配导致的系统性能下降。虽然基于深度神经网络(Deep Neural Network,DNN)[1-5]的语音识别系统极大地提升了语音识别的性能,但在该类系统中仍然存在一个隐含假设:训练数据和测试数据服从相同的概率分布,该假设在实际中很难满足,主要原因是训练阶段难以获得与测试环境相匹配的数据,或匹配数据较少,通常不能对应用场景进行全覆盖,使得训练和测试的条件仍存在不匹配的问题。

可以使用说话人自适应技术解决模型和测试间说话人不匹配的问题,对此,许多研究机构已经做了大量关于 DNN 自适应方面的研究。这些方法中,文献[6-12]中基于辨识向量(i-vector)的说话人感知训练方法备受青睐,其基本思想是将 i-vector 和原始输入特征拼接后对 DNN 模型进行训练,该方法操作简单且容易与其他自适应方法兼容。上述文献主要关注纯净语音条件下的基于 i-vector 的说话人感知训练方法,文献[13-15]则研究噪声条件下基于 i-vector的自适应方法,研究结果显示基于 i-vector 的说话人感知训练方法同样适用于噪声条件。

虽然学者们针对基于 i-vector 的说话人感知训

基金项目: 国家自然科学基金(61673395,61403415);河南省自然科学基金(162300410331)。

作者简介: 梁玉龙(1991—),男,硕士研究生,主研方向为语音识别、机器学习;屈 丹,副教授、博士生导师;邱泽宇,硕士研究生。

收稿日期: 2017-04-25 **修回日期:** 2017-05-31 **E-mail:** yulonglianghb@163.com

练做了大量研究,但由于在获取 i-vector 的过程中常使用 MFCC 作为特征,MFCC 虽然具有较好的表征能力和一定的鲁棒性,但其低层特征表征能力有限,且在恶劣环境中的鲁棒性欠佳,导致用其提取的 i-vector表征能力受到影响。一些研究机构试图应用其他鲁棒性更强的特征代替 MFCC 特征来获取性能更优的 i-vector,其中优先考虑的是瓶颈(bottleneck)特征[16],该特征的表征能力和鲁棒性均优于 MFCC,因此,其受到各研究机构的普遍青睐,但由于在提取 bottleneck 特征时,在 DNN 结构中引入了 bottleneck 层,该策略降低了 DNN 的帧分类准确率,使得系统的识别性能受到一定的影响。

针对上述问题,本文提出一种基于改进 i-vector 的说话人感知训练方法,其主要特点是在获取i-vector 的过程中替换掉传统特征 MFCC。首先,训练一个与说话人无关的 DNN 模型;然后,应用奇异值矩阵分解(Singular Value Matrix Decomposition,SVMD)算法对 DNN 某一隐层的权值矩阵进行分解,用分解后的矩阵代替原始权值矩阵,并应用该网络提取低维特征;最后,应用该特征完成 i-vector 提取器的训练与 i-vector 的提取,进行说话人感知训练。

1　基于 i-vector 的说话人感知训练方法

将说话人信息输入到 DNN 后,DNN 能自动利用说话人信息对网络参数进行调整,该方法称为说话人感知训练[17]。

1.1　训练方法原理

说话人感知训练方法即从句子中估计说话人信息,然后将这些信息输入到网络中,通过 DNN 训练算法自动理解如何利用这些说话人信息完成模型参数的调整。图 1 所示为说话人感知训练过程示意图,DNN 的输入包括声学特征和说话人信息 2 个部分,其余部分与 DNN 模型相同。

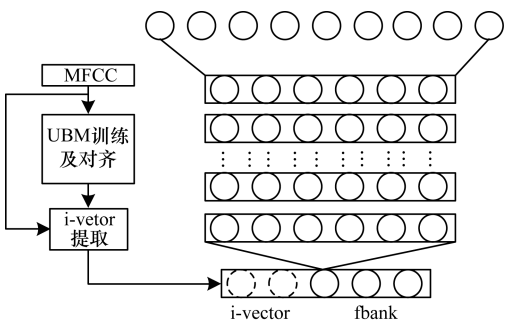

图 1　基于 i-vector 的说话人感知训练过程

当输入特征不包含说话人信息时,第一个隐层的激励为:

$$\boldsymbol{v}^1 = f(\boldsymbol{z}^1) = f(\boldsymbol{W}^1\boldsymbol{v}^0 + \boldsymbol{b}^1) \tag{1}$$

其中,$\boldsymbol{v}^0$ 表示输入声学特征向量,$\boldsymbol{W}^1$ 表示权值矩阵,$\boldsymbol{b}^1$ 表示偏置向量,$\boldsymbol{z}^1$ 表示输入声学特征向量的线性变换。当加入说话人信息后,式(1)变为:

$$\begin{aligned}\boldsymbol{v}^1_{\mathrm{SaT}} &= f(\boldsymbol{z}^1_{\mathrm{SaT}}) = f\left([\,\boldsymbol{W}^1_v \quad \boldsymbol{W}^1_s\,]\begin{bmatrix}\boldsymbol{v}^0\\ \boldsymbol{s}\end{bmatrix} + \boldsymbol{b}^1_{\mathrm{SaT}}\right) = \\ &f(\boldsymbol{W}^1_v\boldsymbol{v}^0 + \boldsymbol{W}^1_s\boldsymbol{s} + \boldsymbol{b}^1_{\mathrm{SaT}}) = \\ &f(\boldsymbol{W}^1_v\boldsymbol{v}^0 + (\boldsymbol{W}^1_s\boldsymbol{s} + \boldsymbol{b}^1_{\mathrm{SaT}}))\end{aligned} \tag{2}$$

其中,$\boldsymbol{s}$ 表示标志说话人的特征向量,$\boldsymbol{W}^1_v$ 和 $\boldsymbol{W}^1_s$ 分别表示声学特征和说话人信息相关的权值矩阵。由式(2)可知,传统 DNN 使用的偏置向量为 $\boldsymbol{b}^1$,而说话人感知训练的偏置向量为 $\boldsymbol{b}^1_s = \boldsymbol{W}^1_s\boldsymbol{s} + \boldsymbol{b}^1_{\mathrm{SaT}}$。

说话人感知训练的优点是其暗含、高效的自适应过程。由式(2)可以看出,说话人感知训练算法无需单独的自适应步骤,其自适应过程可以理解为对偏置项做的变换,该过程使得模型对不同的说话人都适用。如果能够可靠地将说话人信息估计出来,则说话人感知训练将在 DNN 自适应框架中具有优势。

1.2　i-vector 原理

i-vector 技术在说话人识别及说话人确认中作为说话人信息矢量被广泛应用,该技术之所以有如此广泛的应用,原因主要有以下 2 点:1) i-vector 表示了说话人特征中最重要的信息,且其值是低维的;2) i-vector 不仅可以用于 GMM 模型的自适应,也可以用于 DNN 模型的自适应。因此,i-vector 可以作为说话人自适应的一个理想工具。下文介绍 i-vector 的计算推导过程[17]。

i-vector 提取首先需要训练一个通用背景模型(Universal Background Model,UBM),UBM 是一个由 K 个对角协方差高斯组成的高斯混合模型,用来描述整个数据空间的分布,该模型可以表示为:

$$\boldsymbol{x}_t \sim \sum_{k=1}^{K} c_k N(\boldsymbol{x};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \tag{3}$$

其中,$\boldsymbol{x}_t \in \mathbb{R}^{D\times 1}$ 表示由 UBM 生成的声学特征矢量,c_k、$\boldsymbol{\mu}_k$、$\boldsymbol{\Sigma}_k$ 分别表示第 k 个高斯分布的混合权重、高斯均值、对角协方差矩阵。假设第 s 个说话人的声学特征为:

$$\boldsymbol{x}_t(s) \sim \sum_{k=1}^{K} \tilde{c_k} N(\boldsymbol{x};\boldsymbol{\mu}_k(s),\boldsymbol{\Sigma}_k) \tag{4}$$

其中,$\boldsymbol{\mu}_k(s)$ 表示第 s 个说话人从 UBM 自适应得到的属于第 k 个高斯分布的均值。进一步假设自适应后的说话人均值 s 与均值 $\boldsymbol{\mu}_k$ 存在如下关系:

$$\boldsymbol{\mu}_k(s) = \boldsymbol{\mu}_k + \boldsymbol{T}_k\boldsymbol{w}(s), 1 \leqslant k \leqslant K \tag{5}$$

其中,$\boldsymbol{T}_k$ 表示全变换空间矩阵,其包含 M 个基矢量,这些基矢量组成了高斯均值向量空间的一个子空间,该子空间包含整个均值向量空间最核心的部分,$\boldsymbol{w}(s)$ 表示第 s 个说话人的 i-vector。

i-vector 是一个隐含变量,如果假设 i-vector 满足均值为 0、方差为单位方差的高斯分布,且每一帧都属于某一固定的高斯分量,同时全变换空间矩阵

$\boldsymbol{T}$ 是已知的,则可以估计后验概率分布如下:

$$\mathrm{P}(\boldsymbol{w}|\{\boldsymbol{x}_t(s)\}) = N(\boldsymbol{w};\boldsymbol{L}^{-1}(s)\sum_{k=1}^{K}\boldsymbol{T}_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\theta}_k(s),\boldsymbol{L}^{-1}(s)) \tag{6}$$

其中,精度矩阵 $\boldsymbol{L}(s)\in\mathbb{R}^{M\times M}$ 表达式为:

$$\boldsymbol{L}(s) = \boldsymbol{I} + \sum_{k=1}^{K}\gamma_k(s)\boldsymbol{T}_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{T}_k \tag{7}$$

零阶与一阶统计量分别为:

$$\gamma_k(s) = \sum_{t=1}^{T}\gamma_{tk}(s) \tag{8}$$

$$\boldsymbol{\theta}_k(s) = \sum_{t=1}^{T}\gamma_{tk}(s)(\boldsymbol{x}_k(s) - \boldsymbol{\mu}_k(s)) \tag{9}$$

其中,$\gamma_{tk}(s)$ 是第 s 个说话人的第 t 帧特征序列属于第 k 个高斯分量的后验概率。i-vector 可以看作是变量 $\boldsymbol{W}$ 在最大后验概率(MAP)下的点估计:

$$\boldsymbol{w}(s) = \boldsymbol{L}^{-1}(s)\sum_{k=1}^{K}\boldsymbol{T}_k\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\theta}_k(s) \tag{10}$$

由式(10)可以看出,i-vector 就是后验分布的均值。

由于$\{\boldsymbol{T}_k|1\leqslant k\leqslant K\}$是未知的,因此需要使用期望最大化(Expectation Maximization,EM)算法从特定说话人的声学特征 $\{\boldsymbol{x}_t(s)\}$中,根据最大似然(Maximum Likelihood,ML)准则来进行估计。其中,EM 算法的 E(Expectation)步骤的辅助函数为:

$$Q(\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_K) = -\frac{1}{2}\sum_{s,t,k}\gamma_{tk}(s)[\lg|\boldsymbol{L}(s)| + (\boldsymbol{x}_t(s)-\boldsymbol{\mu}_k(s))^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t(s)-\boldsymbol{\mu}_k(s))] \tag{11}$$

式(11)等价于:

$$Q(\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_K) = -\frac{1}{2}\sum_{s,k}[\gamma_k(s)\lg|\boldsymbol{L}(s)| + \gamma_k(s)\mathrm{Tr}\{\boldsymbol{\Sigma}_k^{-1}\boldsymbol{T}_k\boldsymbol{w}(s)\boldsymbol{w}^{\mathrm{T}}(s)\boldsymbol{T}_k^{\mathrm{T}}\} - 2\mathrm{Tr}\{\boldsymbol{\Sigma}_k^{-1}\boldsymbol{T}_k\boldsymbol{w}(s)\boldsymbol{\theta}_k^{\mathrm{T}}(s)\}] + C \tag{12}$$

将式(12)对 $\boldsymbol{T}_k$ 求导后可以得到 EM 算法的 M(Maximization)步骤:

$$\boldsymbol{T}_k = \boldsymbol{C}_k\boldsymbol{A}_k^{-1}, 1\leqslant k\leqslant K \tag{13}$$

其中,式(14)与式(15)通过 E 步骤得到。

$$\boldsymbol{C}_k = \sum_s\boldsymbol{\theta}_k(s)\boldsymbol{w}^{\mathrm{T}}(s) \tag{14}$$

$$\boldsymbol{A}_k = \sum_s\gamma_k(s)[\boldsymbol{L}^{-1}(s) + \boldsymbol{w}(s)\boldsymbol{w}^{\mathrm{T}}(s)] \tag{15}$$

2 基于改进 i-vector 的说话人感知训练方法

2.1 改进的 i-vector 提取方法

传统的 i-vector 提取方法用 MFCC 作为输入特征,为使 i-vector 的鲁棒性更强,一些研究机构利用 bottleneck 特征代替 MFCC 特征,实现 i-vector 提取器的训练与 i-vector 的提取。但在提取 bottleneck 特征时,设置的 DNN 网络 bottleneck 层节点数远小于其他隐层节点数,导致系统的帧分类准确率受到影响,为此,本文提出应用基于 SVD 的低维特征提取方法得到低维特征,用其代替 MFCC 特征完成 i-vector提取器的训练与i-vector的提取。

目前研究 DNN 模型的矩阵分解方法主要关注神经网络的参数减少,如文献[18]提出的思想。这些方法分解 DNN 模型的权值,利用低秩分解或 SVD 减少神经网络无用参数的数量,但其重构的神经网络在识别精度上没有太大变化。基于 SVD 的低维特征提取方案如图 2 所示,该方法使用 SVD 对某一隐层的权值矩阵进行分解(该权值矩阵不包括偏移向量),将分解后得到的基矩阵代替原始矩阵,然后应用新的网络提取低维特征。

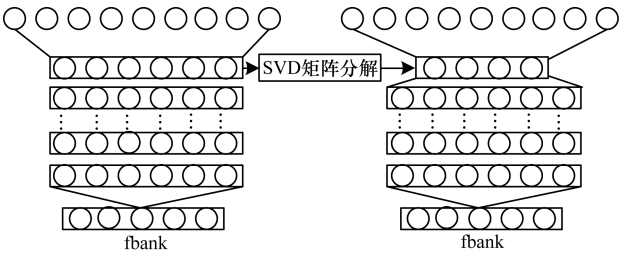

图 2 基于 SVD 的低维特征提取方法示意图

采用基于 SVD 的低维特征提取方法的原因有2 点:

1)因为无法直接对隐层的线性输出进行变换,所以需要使用间接方法,在计算 DNN 隐层的线性输出时,层与层间的权值矩阵作用于每一帧特征,因此,可以将权值矩阵看作是一种具有一定的整体分布特性的广义映射函数。

2)同一层的权值矩阵与偏置向量没有整体性联系,很难对偏移向量和权值矩阵同时进行操作,因此,在该特征层不设置偏移向量。

用 SVD 算法对权值矩阵进行分解的过程表示为:

$$\boldsymbol{A}_{m\times n} = \boldsymbol{U}_{m\times m}\boldsymbol{S}_{m\times n}\boldsymbol{V}_{n\times n}^{\mathrm{T}} \approx \boldsymbol{U}_{m\times k}\Sigma_{k\times k}\boldsymbol{V}_{k\times n} = \boldsymbol{U}_{m\times k}\boldsymbol{N}_{k\times n} \tag{16}$$

其中,$\boldsymbol{A}$ 为带分解矩阵,$\boldsymbol{U}$ 为一个 $m\times m$ 的 $\boldsymbol{U}$ 矩阵,矩阵 $\boldsymbol{U}$ 为一个 $m\times n$ 的对角矩阵且其对角线上的元素非负,$\boldsymbol{V}^{\mathrm{T}}$ 为 $\boldsymbol{V}$ 的转置,$\boldsymbol{S}$ 的对角线元素是矩阵 $\boldsymbol{A}$ 的奇异值,奇异值按降序排列,在这种情况下,对角矩阵 $\boldsymbol{S}$ 由 $\boldsymbol{A}$ 唯一确定。此时可以保存 k 个奇异值和 $\boldsymbol{A}$ 的近似矩阵 $\boldsymbol{U}_{m\times k}\boldsymbol{N}_{k\times n}$。

2.2 训练方法步骤

获取改进的 i-vector 后,将得到的改进 i-vector 与原始输入特征进行拼接,得到新的包含说话人信息的输入特征后,利用该特征对模型进行训练与识别。基于改进 i-vector 的说话人感知训练方法过程如图 3 所示。

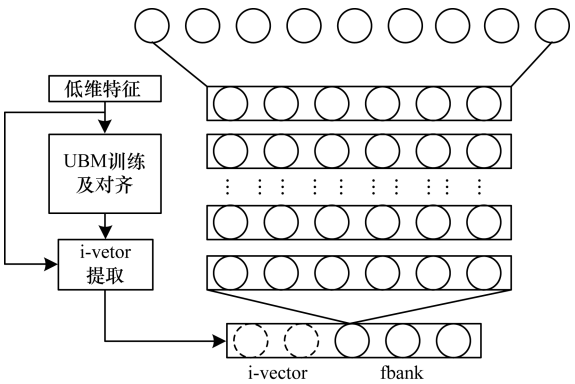

图 3　基于改进 i-vector 的说话人感知训练示意图

该训练方法的主要步骤如下:

1)训练数据模型 SI-DNN;

2)应用 SVD 对最后一层隐层权值矩阵进行分解,并用该结果代替原始权值矩阵;

3)应用网络提取新的低维特征;

4)应用低维特征进行 i-vector 的提取;

5)应用改进的 i-vector 进行说话人感知训练。

3　实验结果与分析

3.1　语料库简介

为验证本文所提特征的识别性能,采用如下2 种语料库进行测试:

1)WSJ 语料库,国际通用的英文语料库,数据由麦克风在安静环境下录制得来。训练集包含 WSJ 0 和 WSJ 1 两部分,共 81.3 h。其中,WSJ 0 包含84 个说话人,共 7 138 句,总时长为 15.1 h,WSJ 1 包含 200 个说话人,共 30 278 句,总时长为 66.2 h。测试集包括 Eval 92 和 Dev 93 两部分。本文使用 Dev 93 作为测试集,该部分包含 10 个说话人,共 503 句,总时长为 65 min。

2)Vystadial 2013 Czech data(Vystadial_cz),开源的捷克语语料库,总时长约 15 h,主要由 3 类数据组成:Call Friend 电话服务语音数据、Repeat After Me 语音数据和 Public Transport Info 口语对话系统语音数据。其中,训练数据集共 22 567 句,126 333 个词语,总时长为 15.25 h;测试集共 2 000 句,11 204 个词语,总时长为 1.22 h。

3.2　实验工具与评价指标

3.2.1　实验工具

实验使用的工具包括 2 个:开源工具包 Kaldi 和 PDNN + Kaldi。Kaldi 工具包主要实现数据准备、特征提取、语言模型和声学模型的训练与解码。PDNN 工具包主要实现 DNN 的搭建与训练。

3.2.2　评价指标

连续语音识别的结果一般为词序列,采用动态规划算法将识别结果与正确的标注序列对齐后进行比较,其中产生的错误类型分为 3 类:插入错误,删除错误,替代错误。插入错误是由于在 2 个相邻的标注间插入其他词所引起,删除错误是由于在识别结果中找不到与某个标注对应的词所引起,替代错误是由于识别得到的词与对应的标注不相符所引起。

假设某个测试集中标注的总个数为 N,插入错误个数为 I,删除错误个数为 D,代替错误个数为 R,则词错误率(WER)的定义如下:

$$WER = \frac{I + D + R}{N} \times 100\% \tag{17}$$

该评测指标越低,表明系统的识别性能越好。

3.3　基线系统

本文采用的基线系统为基于 i-vector 的说话人感知训练模型,将其命名为 DNN + i-vector 模型,由于实验中需要比较基于 SVD 提取的低维特征与 bottleneck 特征的性能,且这 2 个模型的训练都基于 GMM-HMM 模型,因此本节将给出这 3 个模型的具体参数设置。

1)GMM-HMM + LDA + MLLT + SAT 模型。输入特征为 13 维的 MFCC 特征,训练三音子 GMM 声学模型。首先,经过线性区分性分析(Linear Discriminant Analysis,LDA)将 9 帧拼接的特征降到 40 维;然后,采用特征空间最大似然线性回归(feature-space Maximum Likelihood Leaner Regression,fMLLR)进行特征归一化;最后,进行说话人自适应训练(Speaker Adaption Training,SAT)。对于 WSJ 语料库和 Vystadial_cz 语料库,采用的高斯混元数均为 9 000。

2)DNN-HMM/DNN-HMM + i-vector 模型。采用 DNN 对聚类后的三音子状态的似然度进行建模。以 WSJ 语料库的 DNN 模型为例,该模型包括6 个隐层,每个隐层包含 1 024 个节点,激活函数为 Sigmoid 函数。输入层包含 11 帧 40 维 fbank 特征,DNN 的输入节点为 440 个,输出层节点数为 GMM-HMM + LDA + MLLT + SAT 模型中绑定后的三音子状态数,有 3 415 个节点。用后向传播(Back Propogation,BP)算法对 DNN 进行训练,以 DNN 计算得到的预估计概率分布与实际概率分布间的交叉熵作为目标函数。在 BP 算法中,随机梯度下降法的 mini-batch 大小为 256。BP 过程所用的绑定状态标注由 GMM-HMM + LDA + MLLT + SAT 模型对训练集进行强制对齐得到。使用受限玻尔兹曼机(Restricted Boltzmann Machines,RBMs)对 DNN 参数初始化。最终的网络结构参数设置为“440-1024-1024-1024-1024-1024-1024-3415”。与 WSJ 语料库参数设置相

似,Vystadial_cz 语料库的网络结构参数设置为:“440-1024-1024-1024-1024-2125”。对于 DNN + i-vector 模型,只有输入需要拼接 100 维的i-vector,因此,其输入变为 540,其余设置相同。

3)BNF + GMM-HMM + LDA + MLLT 模型。首先,采用 DNN 模型进行 BNF 提取,然后将 BNF 输入到 GMM-HMM + LDA + MLLT 模型中,该模型由上述第一个模型 GMM-HMM + LDA + MLLT + SAT 中省略最后 SAT 训练所得。对于 BNF 提取网络而言,输入特征与 DNN 模型的输入特征相同。经过多次实验表明,对于 WSJ 语料库,相应的 bottleneck DNN 的网络结构参数设置为“440-1024-1024-1024-1024-40-1024-3415”时性能最佳,对于 Vystadial_cz 语料库,bottleneck DNN 设置为“440-1024-1024-40-1024-2125”时 bottleneck 特征的性能最佳。2 个语料库使用的声学模型均为 GMM-HMM + LDA + MLLT。

DNN 训练的学习速率初始值为 0.08,当相邻 2 轮训练的误差小于 0.2% 时,学习速率减半,当减半后相邻 2 轮的误差再次小于 0.2% 时训练停止(如果一直大于 0.2% ,则最多进行 8 次学习)。冲量值设为 0.5,mini-batch 尺寸设为 256。基线系统词错误率如表 1 所示。

表 1 基线系统词错误率 %

语音识别系统	*WER*	
	WSJ	Vystadial_cz
BNF + GMM-HMM + LDA + MLLT	7.25	50.25
DNN-HMM	6.93	48.07
DNN-HMM + i-vector	6.76	48.02

3.4 基于 SVD 的低维特征提取

基于 SVD 的低维特征提取步骤为:首先,初始化一个与说话人无关的 DNN 模型(SI-DNN);然后,对 DNN 基线系统某一层的权值矩阵应用 SVD 算法做矩阵分解;最后,用分解后的基矩阵替换原始权值矩阵。

应用该特征重新训练 GMM-HMM + LDA + MLLT 声学模型并解码。其中,影响识别性能的因素主要有 2 个:1)对 DNN 的哪一层权值矩阵进行分解;2)对权值矩阵分解多少维效果更优。根据这 2 个因素,本文分别做实验进行验证。实验结果如表 2和表 3 所示。

表 2 WSJ 语料库 DNN-SVD 词错误率结果

矩阵分解后的维数	*WER*/%	
	SVD-1	SVD-2
30	7.52	8.68
40	7.14	8.25
50	7.30	8.22
60	7.34	8.05

表 3 Vystadial_cz 语料库 DNN-SVD 词错误率结果

矩阵分解后的维数	*WER*/%	
	SVD-1	SVD-2
30	45.67	50.87
40	45.92	50.28
50	46.01	49.74
60	45.89	49.45

表 2 中“SVD-1”表示最后一层隐层的权值矩阵,“SVD-2”表示倒数第 2 层隐层的权值矩阵,词错误率表示由 DNN + 矩阵分解 + GMM-HMM + LDA + MLLT 组成的语音识别系统的词错误率。从表 2 的结果中可以看出,对于 WSJ 语料库,当使用 SVD 对最后一个隐层的权值矩阵做分解并取分解维数为 40 时,效果最好。由表 3 的结果可以看出,对于 Vystadial_cz 语料库,当使用 SVD 对最后一层隐层的权值矩阵做分解并取分解维数为 30 时,效果最好。

由上述结果可知,基于矩阵分解的方法克服了帧分类准确率下降的问题,与基线系统 BNF + GMM-HMM + LDA + MLLT 相比,其 WSJ 语料库的识别性能提升了 1.52% ,Vystadial_cz 语料库的识别性能提升了 9.11% 。由于矩阵分解的算法解决了低资源情况下的数据不充分训练问题,因此其在数据量较小的 Vystadial_cz 语料库上的识别性能提升得更高,在数据量相对充足的 WSJ 语料库上性能提升不明显。

DNN 通过每层的非线性变换将输入特征变得越来越抽象,鲁棒性也越来越强,因此,理论上由最后一层得到的特征表征能力会优于由倒数第 2 层得到的特征,在 WSJ 与 Vystadial_cz 语料库中的实验结果也证明了这一点。本文分析认为,分解尺寸的大小应该与数据量的多少有关,超出或少于某个范围,会导致特征表征稀疏或特征表示不充分,进而导致系统的识别性能下降。

3.5 基于改进 i-vector 的说话人自适应方法

基于改进 i-vector 的说话人自适应方法步骤为:首先,将 SVD-BN 特征代替原 MFCC 特征进行 i-vector提取器的训练与 i-vector 的提取,得到改进后的 i-vector;然后,将改进的 i-vector 代替原始 i-vector,与 DNN 的输入特征进行拼接后送入 DNN 进行训练与识别。该方法所用模型的其余参数设置与基线 DNN + i-vector 模型相同。实验结果如表 4 所示。

表 4 基于改进 i-vector 的说话人感知训练识别结果 %

语音识别系统	*WER*	
	WSJ	Vystadial_cz
DNN-HMM + (改进)i-vector	6.66	47.29

由表 1、表 4 可以看出,在 Vystadial_cz 语料库中,相对 DNN-HMM 语音识别系统,本文方法识别性能提升了1.62% ,相对原始基于 i-vector 的方法,本文方法识别性能提升了 1.52% 。在 WSJ 语料库的实验中,上述性能分别提升了3.9% 和 1.48% 。实验结果表明,改进的 i-vector 在提取时应用了基于 SVD 分解得到的低维特征,该特征克服了帧分类准确率下降的问题,因此,其鲁棒性与表征能力更优,使得到的 i-vector 包含更有用的说话人信息,最终使得整个识别系统的性能得到提升。

4　结束语

传统的 i-vector 提取方法主要应用 MFCC 作为输入特征。由于 MFCC 的鲁棒性与表征能力均较差,因此本文提出一种基于改进 i-vector 的说话人自适应方法,该方法在一定程度上克服了帧分类准确率下降的问题,由其提取的特征表现出了较好的鲁棒性。实验结果表明,相比原有基于 i-vector 的方法,该方法的系统识别性能较高。下一步将考虑应用更优的算法以获取更有效的特征表征,使系统更鲁棒、识别率更高。

参考文献

[1] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition:the shared views of four research groups[J]. IEEE Signal Processing Magazine,2012,29(6):82-97.

[2] DAHL G E,YU D,DENG L,et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio Speech and Language Processing,2012,20(1):30-42.

[3] 李传朋,秦品乐,张晋京. 基于深度卷积神经网络的图像去噪研究[J]. 计算机工程,2017,43(3):253-260.

[4] 梁玉龙,屈　丹,李　真,等. 基于卷积神经网络的维吾尔语语音识别[J]. 信息工程大学学报,2017,18(1):44-50.

[5] 秦楚雄,张连海. 低资源语音识别中融合多流特征的卷积神经网络声学建模方法[J]. 计算机应用,2016,36(9):2609-2615.

[6] LIAO H. Speaker adaptation of context dependent deep neural networks[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C. ,USA:IEEE Press,2013:7947-7951.

[7] SEIDE F,LI G,CHEN X,et al. Feature engineering in context-dependent deep neural networks for conversational speech transcription[C]//Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding. Washington D. C. ,USA:IEEE Press,2011:24-29.

[8] YAO K,YU D,SEIDE F,et al. Adaptation of context-dependent deep neural networks for automatic speech recognition[C]//Proceedings of 2012 IEEE Workshop on Spoken Language Technology. Washington D. C. , USA:IEEE Press,2012:366-369.

[9] HAMID O A, JIANG H. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition[EB/OL]. [2017-04-25]. http://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_1248.pdf.

[10] SELTZER M,YU D,WANG Y. An investigation of deep neural networks for noise robust speech recognition[C]// Proceedings of 2013 IEEE International Conference on Acoustics,Speech and Signal Processing. Washington D. C. , USA:IEEE Press,2013:7398-7402.

[11] YOSHIOKA T,RAGNI A,GALES M J. Investigation of unsupervised adaptation of DNN acoustic models with filterbank input[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C. ,USA:IEEE Press, 2014:6344-6348.

[12] DELCROIX M,KINOSHITA K,HORI T,et al. Context adaptive deep neural networks for fast acoustic model adaptation[C]//Proceedings of 2015 IEEE International Conference on Acoustics,Speech and Signal Processing. Washington D. C. ,USA:IEEE Press,2015:5270-5274.

[13] KARANASOU P, WANG Y, GALES M J F, et al. Adaptation of deep neural network acoustic models using factorized i-vectors[EB/OL]. [2017-04-20]. http://www.isca-speech.org/archive/archive_papers/interspeech_2014/i14_2180.pdf.

[14] SENIOR A, MORENO I L. Improving DNN speaker independence with i-vector inputs[C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C. ,USA: IEEE Press,2014:225-229.

[15] ROUVIER M, FAVRE B. Speaker adaptation of DNN-based ASR with i-vectors:does it actually adapt models to speakers? [EB/OL]. [2017-04-20]. http://pageperso.lif.univ-mrs.fr/~benoit.favre/papers/favre_interspeech 2014a.pdf.

[16] YU C, OGAWA A, DELCROIX M, et al. Robust i-vector extraction for neural network adaptation in noisy environment[EB/OL]. [2017-04-15]. http://www.isca-speech.org/archive/interspeech_2015/papers/i15_2854.pdf.

[17] SAON G, SOLTAU H, NAHAMOO D, et al. Speaker adaptation of neural network acoustic models using i-vectors[C]//Proceedings of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Washington D. C. ,USA:IEEE Press,2013:55-59.

[18] XUE S F,HAMID O A,JIANG H,et al. Fast adaptation of deep neural network based on discriminant codes for speech recognition[J]. IEEE/ACM Transactions on Audio,Speech and Language Processing,2014,22(12): 1713-1725.

编辑　吴云芳