文章编号:1007-1423(2014)14-0003-05 DOI:10.3969/j.issn.1007-1423.2014.14.001

语音识别中基于 i-vector 的说话人归一化研究

李亚琦,黄浩

(新疆大学信息科学与工程学院,乌鲁木齐 830046)

摘要:

i-vector 是反映说话人声学差异的一种重要特征,在目前的说话人识别和说话人验证中显示了有效性。将 i-vector 应用于语音识别中的说话人的声学特征归一化,对训练数据提取 i-vector 并利用 LBG 算法进行无监督聚类,然后对各类分别训练最大似然线性变换并使用说话人自适应训练来实现说话人的归一化。将变换后的特征用于训练和识别,实验表明该方法能够提高语音识别的性能。

关键词:

说话人识别; i-vector; 最大似然线性变换; 特征提取; 说话人归一化; LBG 算法

基金项目:

国家自然科学基金资助项目(No.61365005、No.60965002)

0 引言

说话人归一化的主要目的就是消除人际随机差异,提高恒定参数,即滤掉个人特征,获得具有语言学意义的信息。其另一个作用是消减录音时的发音风格(正式、差异、紧张等)的差异。归一化一般有两个步骤,一是对坐标平移,二是对频域压缩或扩大。语音特征的冲击特征受噪声环境的影响,归一化方法在语音识别系统当中用来补偿环境噪声不匹配的影响,进而再提高系统的识别率。

一般常用的特征归一化方法[1]主要有倒谱均值归一化(Cepstral Mean Normalization, CMN)[1]和倒谱方差归一化(Cepstral Variance Normalization,CVN)[2],两者都属于对倒谱矩的归正方法。使带噪语音特征参数的概率密度函数与纯净语音的概率密度函数更为接近,消减测试语料和训练语料环境的不匹配度。其中,CMN 是对一阶矩做归一化,CVN 是在 CMN 的基础上再对二阶矩进行归一化。这两种都是常用的方法,CMN 在倒谱域中消除了包含大部分信道失真的直流分量,而 CVN 通过对方差的进一步归一化,使得带噪语音信号和纯净语音信号的概率密度函数的差异减小。我们将 CVN 与 CMN 共同使用的方法称为均值方差归一化。这些说话人归一化的方法已经在语音识别任务当中显示了其有效性。同时说话人归一化技术也被认为是改善说话人识别系统的识别率的一种非常有效的途径。

说话人识别(Speaker Recognition,SR)又称话者识别,是从说话人的一段语音中分析和提取出说话人的个性特征,自动鉴别说话人身份的一项技术。说话人识别的主要途径就是提取出反映说话人的声道特性的差异特征进行比对。近几年来,研究者们提出了一系列以 Gauss 混合模型-通用背景模型(GMM-UBM)为基础的说话人建模方法[3],使得说话人识别系统的性能有了突飞猛进的提高。在美国国家标准技术局组织的评测中占主导地位的 Gauss 混合模型超矢量-支持向量机和联合因子分析。

2011 年,Dehak N 和 Kenny P 在分析联合因子的基础上,提出了可区分性向量 i-vector 模型。传统的联合因子分析 JFA 模型是基于说话人和信道因素这两个明显的空间:说话人空间被定义为特征语音矩阵 V,信道空间被定义为特征信道矩阵 U。然后提出基于单一的空间方法来取代这两个部分空间,我们将这种新的空间定义称之为“总体变化子空间”[4],它同时包含了说

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

话人和信道两种变化。因此定义的总体变换子空间包含了特征矢量和最大的总体变化协方差矩阵的特征值。在该种模型中，定义说话人影响和信道影响在GMM超矢量空间中是没有明显区别的。在美国国家标准技术局组织的评测中占主导地位的Gauss混合模型超矢量-支持向量机(GSV-SVM)、联合因子分析(JFA)以及i-vector（identity vector）等说话人建模技术均以GMM-UBM系统为基础，其中i-vector系统的性能最好，它可以反映人与人之间发音的差异，成为目前国内外研究的主流方法[5]。

对于语音识别任务来说，用于识别的特征应该尽量去除说话人的声学差异，由于i-vector可以很好地反映说话人声学差异。本文着重对于将i-vector携带的差异信息用于语音识别，利用LBG算法对提取的i-vector进行无监督聚类，对各类分别训练最大似然线性变换，来对说话人的差异进行归一化。将变换后的特征用于训练和识别，实验表明该方法能明显提高语音识别的性能。

1 基于i-vector的说话人聚类

1.1 i-vector提取方法

给定有I条训练语音数据[6] $y=\{Y_i \mid i=1,2,3,\cdots,I\}$，其中 $Y_i=\{y_1^{(i)},y_2^{(i)},\cdots,y_{T_i}^{(i)}\}$ 是一列来自第 i 个训练语音片段的 D 维特征矢量，在本文中是 D=39 维的MFCC特征序列。首先采用最大似然准则训练高斯混合模型GMM，通常称为统一背景模型(UBM)：

$$p(y)=\sum_{K=1}^{K} C_k N(y;m_k,R_k) \tag{1}$$

其中 C_k 是第 k 个混合高斯的系数，$N(y;m_k,R_k)$ 表示均值 m_k 是 D 维矢量，$D\times D$ 维对角协方差矩阵 R_k 的高斯分布。我们用 M_0 表示将 M_k 连接起来的 $(D\cdot K)$ 维的超矢量，R_0 表示 $(D\cdot K)\times(D\cdot K)$ 的块对角矩阵，R_K 作为它的第 K 个块矩阵。我们用 $\Omega=\{c_k,m_k,R_k \mid k=1,2,3,\cdots,k\}$ 来表示UBM-GMM参数。给定一个语音片段 Y_i，用 $(D\cdot K)$ 维低维随机超矢量 $M(i)$ 描述文本无关的描述说话人的差异：

$$M(i)=M_0+Tw(i) \tag{2}$$

$w(i)$ 是一个 F 低维随机矢量符合标准的正态分布 $N(\cdot;0,\mathrm{I})$，T 是 $(D\cdot K)\times F$ 维的描述总体变化的矩阵，对于给定的 Y_i，Ω 和 T，i-vector可以由上式得到，从而解决如下的问题：

$$\hat{w}(i)=\arg\max \prod_{t=1}^{T}\prod_{k=1}^{K} N(y_t^{(i)};M_K(i),R_K)^{p(k\mid y_t^{(i)},\Omega)} p(w(i)) \tag{3}$$

这里的 $M_K(i)$ 是 $M(i)$ 的第 K 个 D 维低维矢量，其中：

$$p(k\mid y_t^{(i)},\Omega)=\frac{c_k N(y_t^{(i)};M_K(i),R_K}{\sum_{l=1}^{K} c_l N(y_l^{(i)});M_l(i),R_l)} \tag{4}$$

上述问题的解决给i-vector的提取公式就可以得到：

$$\hat{w}(i)=I^{-1}(i)T^T R_0^{-1}\Gamma y(i) \tag{5}$$

其中：

$$l(i)=I+T^T\Gamma(i)R_0^{-1}T \tag{6}$$

$\Gamma(i)$ 是一个 $(D\cdot K)\times(D\cdot K)$ 的块对角矩阵，$\gamma_k(i)I_{D\times D}$ 作为它的第 K 个块成分；$\Gamma_y(i)$ 是一个 $(D\cdot K)$ 的低维超矢量，$\Gamma_{y,k}(i)$ 作为它的第 k 个 D 低维矢量。Baum-Welch统计量 $\gamma_k(i)$ 和 $\Gamma_{y,k}(i)$ 通过下式计算可得：

$$\gamma_k(i)=\sum_{t=1}^{T_i} p(k\mid y_t^{(i)},\Omega) \tag{7}$$

$$\Gamma_{y,k}(i)=\sum_{t=1}^{T_i} p(k\mid y_t^{(i)},\Omega)(y_i^{(i)}-m_k) \tag{8}$$

从上面i-vector $\hat{w}(i)$ 的求解过程来看，最终归结于总体变化矩阵 T 可以通过下面的步骤估算得到[7]：

①初始化

在 $[Th_1,Th_2]$ 中随机地选取 T，设定T中每个成分的初始值，这里的 Th_1 和 Th_2 是两个控制参数。对于每个训练的语音片段，利用公式(7)和(8)计算其相应的Baum-Welch统计量。

②设定 E 值

对于每个训练的语音片段 Y_i，用充足的数据和当前对 T 的估计，计算 $w(i)$ 的期望值，计算的方法如下：

$$E[w(i)]=I^{-1}(\mathrm{i})T^T R_0^{-1}\Gamma y(i) \tag{9}$$

$$E[w(i)w^T(i)]=E[w(i)]E[w^T(i)]+I^{-1}(i) \tag{10}$$

③设定 M 值

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

采用一个方程更新总体变化矩阵 T：

$$\sum_{t=1}^{I}\Gamma(i)TE[w(i)w^{T}(i)]=\sum_{i=1}^{I}\Gamma_{y}(i)E[w^{T}(i)] \qquad (11)$$

④重复或者中止：反复步骤②到③，找到迭代次数的固定值或者直到目标函数收敛。

1.2 基于 i-vector 的说话人识别

本文从训练语句中 78871 条语音数据随机抽取 1000 条语句进行说话人分类实验，即判断测试是由哪一个说话人实验输出。每段语音片段都有一一对应的说话人编号。

①对所有的训练语句提取特征矢量，得到一个 78871×70 维的矩阵，其中每一行对应的是每条语句的特征矢量，本文的迭代次数设置为 n=10，多次迭代使得所提取的特征矢量更加收敛，使得后续的系统识别性能更佳。

②在上述得到的 78871 个特征矢量中用随机数选取 1000 个语音片段作为测试数据，即得到一个 1000×70 维的矩阵。

③利用余弦距离计算公式测试语音数据（1000×70 维矩阵）和训练语音数据（77871×70 维矩阵）的值，得到一个余弦距离值的矩阵（77871×1000 维矩阵）。

④索引③所得到的余弦距离值矩阵的每一行的最大值，检验最大值所对应的测试语音数据和训练语音数据是否为同一个说话人编号。若是同一个说话人，则计数增加 1，直到 1000 个测试数据检验完毕，得到总的正确计数。

⑤通过④所得到的正确识别计数/1000，可以得到本实验的说话人的识别率。

1.3 基于 LBG 的 i-vector 聚类

LBG 距离矢量算法对 i-vector 提取的说话人特征矢量进行分类，相较于一般的 LBG 聚类的分类（2 类、4 类、8 类和 16 类）[6]，本文的研究重点在于聚成两大类，发现这两个大类正好是男女性别的声学特征。该发现对于说话人识别归一化的研究起到了促进作用。

LBG 算法是个不断迭代的算法，以平均失真 D 的相对误差小于预先设定的某一门限阈值 ε 作为算法的收敛准则，其基本思想是在每次迭代时都用最小距离准则，对训练样本重新分类，使每次迭代后总的量化失真减小。其主要步骤如下[8]：

（1）设置量化失真阈值 δ、初始量化失真 D(0)=∞、最大迭代次数 MAX 以及码字初值；

（2）设迭代初值 m=1；

（3）码字为中心，根据最近领域准则将 X 分成 L 类；

（4）计算量化失真；

（5）计算量化失真改进量的相对值；

（6）计算新码字（聚类中心）；

（7）判断量化失真改进量的相对值是否小于设置量化失真阈值，若是，则输出码字作为码本；若否，判断迭代次数，当 m<MAX 时，则 m=m+1，转到（4）。

本文将用余弦距离打分的方法来替换 LBG 算法中平均量化误差方法。用目标说话人的 i-vectors 和测试说话人的 i-vectors 的余弦内核的价值作为判定打分[4]，使得语音识别中的说话人识别的系统性能更佳。

$$score(W_{target},W_{test})=\frac{<W_{target},W_{test}>}{||W_{target}||\ ||W_{test}||}\ \underset{<}{\geq}\ \theta \qquad (12)$$

这个类中心的值与阈值 θ 对比做出最终的决定。它的优势在于不必须要目标说话人的录入。余弦内核作为说话人识别的判定打分使得过程更快速和简化[9]。测试的语音 i-vectors 与训练的 i-vectors 的余弦距离的值最大，则表明两个 i-vectors 是属于同一个说话人。

2 基于 i-vector 聚类说话人归一化

2.1 基于说话人自适应训练的归一化方法

本文采用说话人自适应训练对具有先验类别特征语音数据进行归一化。说话人自适应训练（Speaker Adapted Training，SAT）[10-11]利用特定说话人数据对说话人无关（Speaker Independent，SI）码本进行改造，其目的是得到说话人自适应（Speaker Adapted，SA）码本以提升识别性能。基于模型变换的转化模型参数，通常利用均值和高斯参数的协方差矩阵。在自适应训练过程中，给定一系列变换就可以建立规范模型，且每种变换都代表了特有的声学环境。通过域类特殊变换和识别使用，使这个模型自适应。广泛使用的技术[12]包括最大似然线性回归（MLLR）[13]、分类自适应训练（CAT）和约束最大似然线性回归（CMLLR）。约束最大似然线性回归（CMLLR）[14]变换参数是语音识别系统中常用的说话人模型调整技术，它可以实现说话人无关模型到说话

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

人相关模型的线性调整，也可以实现说话人环境间的线性调整，模型的参数调整经由仿射变换进行。更多关于 CMLLR 方法以及利用 CMLLR 进行说话人自适应训练的方法参见文献[15]。

2.2 语音识别中基于 i-vector 聚类的说话人归一化

基于 i-vector 聚类的说话人归一化的思想是，对于 i-vector:

①利用训练数据建立说话人无关的声学模型；

②对所有的语音数据提取 i-vector，并建立 i-vector 和语音数据的标号；

③用 LBG 算法进行对 i-vector 进行聚类，根据聚类的结果，对所有的语音数据进行标记聚类的类别；

④利用①中的说话人无关模型，以及语音数据的类别进行自适应训练，得到归一化的声学模型以及每个类别的线性变换；

⑤通过 i-vector 的判定测试的数据属于哪类，然后用该类的最大似然线性变换成新的矢量；

⑥用新的矢量进行说话人的识别。

3 实验与结果

本文采用 863 汉语连续语音识别数据库的带调音节输出实验来验证所提出方法的有效性，带调音节输出能够考虑纯粹的声学特性。因为实验着重讨论声学模型的识别结果，在识别过程中没有使用语言模型。训练语料包含 100 个说话人发声的 100 小时的 86271 条语句，语音数据采样率为 16bit/16kHz。谱观察向量采用 39 维，包括倒谱均值归一化的 12 阶美尔频率倒谱系数(MFCC)、归一化对数能量及其一阶、二阶导数。语音识别测试语句包括另外 10 个说话人的 5972 条语音数据。

本文 i-vector 的提取方法实验中，UBM-GMM 模型中的高斯成分 K=512，i-vector 的维数 F=50，i-vector 的特征提取的谱观察向量维数采用 D=39。我们从 86271 条训练语句中随机抽取 7000 条语句作为开发集进行说话人分类实验，即判断测试是由哪一个说话人实验输出，对 7000 条集内测试语句的说话人分类正确率为 98.01%。说明 i-vector 能有效反映说话人的差异。接下来进行说话人聚类实验，我们将训练集中提取的 i-vector 聚为两类，发现聚类后的结果正好与说话人的性别相吻合。这说明两类 i-vector 聚类反映的是说话人的性别声道的差异。我们对开发集 7000 条测试语音数据进行性别分类实验，对该测试语音提取 i-vector，对每条测试语句的 i-vector 计算与上述两类 i-vector 余弦得分最大值，并将测试语句说话人性别信息与两类性别信息比对得到性别测试结果，对于开发集 7000 条集内测试数据性别测试正确率为 98.32%。从性别分类实验来看，i-vector 的两类聚类已经能够十分准确地反映男女说话人的声道性别差异。

语音识别声学模型采用 HTK 的 HERest 训练得到，HMM 采用 3 状态 8 个混合高斯单音子模型，最大似然训练模型的基线带调音节输出识别结果为 65.80%。在 2 类 i-vector 聚类完成之后，根据训练集的类别标号以及基线声学模型进行说话自适应训练得到归一化模型以及两类的线性变换。在进行测试时，根据测试语音的 i-vector 找到相应变换，将原有特征变换为新特征进行识别，连续语音识别结果为 66.78%。上述结果表明，基于 i-vector 的声学特征归一化能够有效提高语音识别有效性。

表1 实验结果总结

实验名称	正确率(%)
说话人分类	98.01
说话人性别分类	98.32
语音识别	65.80
归一化语音识别	66.78

4 结语

本文将 i-vector 应用于语音识别中的说话人的声学特征归一化。通过对训练数据提取特征 i-vector，并利用 LBG 算法进行无监督聚类，然后对各类分别训练最大似然线性变换并使用说话人自适应训练来实现说话人的归一化。将变换后的特征用于说话人的训练和识别，通过实验表明基于 i-vector 的方法能够有效地提高语音识别的性能。我们将训练集中提取的 i-vector 聚为两类，发现聚类后的结果正好与说话人的性别相吻合。这说明两类 i-vector 聚类反映的是说话人的性别声道的差异。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

参考文献：
[1]A. Acero，X. Huang. Augmented Cepstral Normalizationfor Robust Speech Recognition[C]. Proc. of IEEE Automatic Speech Recognition Workshop. Snow- bird, Utah, USA: 1995
[2]P. Jain，H. Hermansky. Improved Mean and Variance Normalization for Robust Speech Recognition[C]. Proceedings of 2001 IEEE International Conference on Acoustics, Acoustics and Signal Processing. Salt Lake City, Utah, USA: 2001
[3]Reynolds D,Quatieri T,Dunn R.Speaker Verification Usinadapted Gaussian Mixture[J]. Digital Signal Processing,2000,10:19~41
[4]N.Dehak,P.J.Kenny,R.Dehak,P.Dumouchel,andP. Ouellet. Front-end Factor Analysis for Speaker Verification[J]. IEEE Transaction on Audio,Speech,and language Processing，2011，19(4):788~798
[5]何亮，栗志意，蔡猛等.集合分类中的鉴别式局部信息距离保持映射[J].清华大学学报(自然科学版)，2011,51(7):1010~1016
[6]Y.Zhang,J.Xu,Z.-J.Yan,Q.Huo. A i-vector Based Approach to Training Data Clustering for Improved Speech Recongnition. Proc. Interspeech-2100
[7]张文林，张卫强，刘加，李弼程，屈丹. 自动化学报，2011,37(12):14951502)
[8]J.Xu,Y.Zhang,Z.-J.Yan,and Q. Huo,An i-vector Based Approach to Acoustic Sniffing for Irrelevant Variability Normalization Based Acoustic Model Training and Speech Recognition". Proc. Interspeech-2011：1701~1704
[9]J.Xu,Y.Zhang,Z.-J.Yan, Q.Huo. A New i-vecter Approach and Its Application to Irrelevent Variability Normalization Basced Acoustic Model Training, MLSP-2011,Beijing,china,6pages
[10]孙圣和，陆哲明. 矢量量化技术及应用[M]. 北京：北京科学出版社，2002
[11]Lee C-H , Lin C-H , Juang B-H. A Study on Speaker Adaptation of the Parameters of Continuous Density Hidden Markov Models[J]. IEEE Transactions on Signal Processing, 1991, 39(4): 806~814
[12] 吕萍，王作英，陆大琻. 基于最大似然模型插值的快速说话人自适应算法[J]. 中文信息学报，2002，16(1):49~3
[13]M. J. F. Gales, P. C. Woodland. Mean and Variance Adaptation within the MLLR Framework. Computer Speech and Language, 1996，10(4)：249~264
[14]Marc Ferras, Cheung Chi Leung, Claude Barrasland Jean-Luc Gauvain, Constrained MLLR for Speaker Recognition. Proc. ICASSP 2007：53~56
[15]Kai Yu. Adaptive Training for Large Vocabulary Continuous Speech Recognition. PhD Thesis, Cambridge University，2006.7

作者简介：
李亚琦(1988-)，男，湖北嘉鱼人，硕士研究生，研究方向为语音识别技术
黄浩(1976-)，男，新疆乌鲁木齐人，博士，副教授，研究方向为自动语音识别技术、口语理解与人机交互技术
收稿日期：2014-04-03　　修稿日期：2014-05-06

Research on Speaker Normalization Based on i-vector in Speech Recognition

LI Ya-qi，HUANG Hao

(Department of Information Science and Engineering, Xinjiang University, Urumqi 830046)

Abstract:
i-vector is an important feature which reflects differences of acoustic characteristics between speakers, and has shown effectiveness in speaker identification and speaker verification. Applies the i-vector method to speaker normalization in speech recognition: extracts the i-vectors of training data and carries out unsupervised clustering using the LBG algorithm. Then performs speaker adaptive training using the cluster information. Speech recognition experiments show that this method can consistantly improve the performance.
Keywords:
Speech Recognition; i-vector; Maximum Likelihood Linear Transforms; Feature Extractor; LBG Algorithm

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net