第 34 卷 第 5 期 计 算 机 工 程 2008 年 3 月
Vol.34 No.5 Computer Engineering March 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)05—0196—02 文献标识码：A 中图分类号：TP311

改进的语音特征提取方法及其应用

王安娜，王勤万，刘俊芳，袁文静
(东北大学信息科学与工程学院，沈阳 110004)

摘 要：噪音是降低语音识别系统精度的关键因素，因此，如何从带噪语音信号中提取出有效的语音特征是提高语音识别系统识别率的重要途径。该文在分析语音特征提取方法的基础上提出改进算法。实验表明，采用 LDA+MLLT+CMS 算法组合提取出的语音特征具有较好的鲁棒性，在噪音环境下的平均音节识别率为 43.79%。该组合在中文大词汇量连续语音识别系统中也有较好的性能，音节识别率达到 83.56%。

关键词：特征提取；主分量分析(PCA)；线性区分分析(LDA)；语音识别

Improved Speech Feature Extraction and Its Application

WANG An-na, WANG Qin-wan, LIU Jun-fang, YUAN Wen-jing
(School of Information Science & Engineering, Northeastern University, Shenyang 110004)

【Abstract】 Noise is a pivotal factor that reduces recognition rate of a speech recognition system. So how to extract effective speech characteristics becomes an important path for a speech recognition system to increase accuracy. This paper analyses speech feature extraction and makes improvement of it. Experimental results indicate that the algorithm combined with LDA+MLLT+CMS has better robustness than other combinations. Average syllable recognition rate reaches 43.79% by using it under conditions of noises. The algorithm combination has also a good performance in Mandarin Large Vocabulary Continuous Speech Recognition (LVCSR). Syllable recognition accuracy achieves 83.56%.

【Key words】 feature extraction; Principal Component Analysis(PCA); Linear Discriminant Analysis(LDA); speech recognition

目前纯净语音识别已达到相当成熟的阶段，语音识别系统在实验室环境下识别率很高，但在有噪声存在的开放式环境中，识别率却大幅度下降。因此，提取具有鲁棒性和较强区分能力的特征向量对语音识别系统具有重要的意义。本文分析了语音特征提取方法，提出 LDA+MLLT+CMS 算法。

1 语音特征提取方法概述

当前常用的语音特征提取方法是提取语音特征参数作为特征，例如 Mel 频率倒谱系数(MFCC)、线性预测倒谱系数(LPCC)和感知线性预测倒谱系数(PLPCC)等。该语音特征提取方法在实验室环境下识别率很高，但是在开放式环境下，效果并不理想。

为适应噪声环境下语音特征的鲁棒性要求，将语音特征的提取分为以下 3 个步骤：

(1)语音特征参数提取。相比其他的语音特征系数，MFCC 充分考虑了人耳的听觉特性，具有很好的识别性能和抗噪能力。本文采用 MFCC 作为语音特征参数。

(2)对步骤(1)中提取出的语音特征做线性变换降维，只保留具有区分力的特征成分。常用的方法是数据驱动线性特征转换(data-driven linear feature transformation)，它包括主分量分析(Principal Component Analysis, PCA)、线性区分分析(Linear Discriminant Analysis, LDA)等。

(3)使用语音增强方法来增强语音特性并减少噪音的干扰。本文采用倒频谱均值减法(Cepstral Mean Subtraction, CMS)增强语音，该算法复杂度小，实用性强[1]。

2 基于数据驱动线性特征转换的语音特征变换

在模式识别中，如何有效地选取特征参数是一个重要问题。目前使用的解决方法之一是通过数据驱动线性特征变换方法将原始的特征参数变换到一个维数更低的矢量空间，进一步降维而保留重要的具有区分力的成分。

2.1 主分量分析

主分量分析就是要寻找、保留数据中最有效、最重要的“成分”，舍去一些冗余的、包含信息量很少的“成分”。

主分量分析的目的是要找出数据中最重要“成分”所在的基底向量，且基底向量各自单位正交[2]。进行主分量分析时使用的训练语料表示为 $\boldsymbol{X}=[x_1,x_2,\cdots,x_N]$，$N$ 为语料的总数，每个 x_i 为 n 维，则模式总体的自相关矩阵 $\boldsymbol{T}$ 为

$$\boldsymbol{T}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{\boldsymbol{X}})(x_i-\bar{\boldsymbol{X}})^{\mathrm{T}} \tag{1}$$

其中，$\boldsymbol{T}$ 为 $n\times n$ 维矩阵；$\bar{\boldsymbol{X}}$ 为模式总体的均值向量：

$$\bar{\boldsymbol{X}}=\frac{1}{N}\sum_{i=1}^{N}x_i \tag{2}$$

对矩阵 $\boldsymbol{T}$ 求取特征值及其对应的特征向量，将特征值从大到小排序，取前 p 个大的特征值所对应的特征向量构成基底矩阵(也为变换矩阵) $\boldsymbol{\theta}_p$，最后所有的语料可以利用所求得的变换矩阵投射到新的特征空间 $\boldsymbol{Y}=[y_1,y_2,\cdots y_N]$，其中

$$y_i=\boldsymbol{\theta}_p^{\mathrm{T}}x_i,i=1,2,\cdots,N \tag{3}$$

$\boldsymbol{Y}$ 的第一维分量称为原始向量 $\boldsymbol{X}$ 的第一主分量，它包含了原始向量中最多的信息，$\boldsymbol{Y}$ 的第二维分量称为第二主分量，依次类推。

基金项目：教育部重点实验室基金资助项目(PAL200508)；辽宁省自然科学基金资助项目(20062033)

作者简介：王安娜(1956－)，女，教授、博士，主研方向：智能信号，信息处理，模式识别，信息融合；王勤万、刘俊芳、袁文静，硕士研究生

收稿日期：2007-03-28 **E-mail：**wangqinwan@163.com

PCA 也有较大的缺点：变换后的模式总体的协方差矩阵可实现对角化，但是各样本的协方差矩阵却不能对角化。

2.2 线性区分分析

和PCA类似，LDA也是通过求取一个变换矩阵再做线性转换来达到降维的目的。与PCA不同的是，LDA使模式样本内的分布凝聚，而使样本间的分布疏远[3]。

LDA 有如下假设：(1)特征向量投射后不是所有的维都包含具有区分力的信息，它们都被包含在前 p 维子空间，而后(n-p)维子空间因不包含有用信息而被忽略。(2)每个样本内都是高斯分布。

当语音以向量来表示时，LDA 希望模式样本间协方差矩阵 $\boldsymbol{B}$ 转换后的行列式值越大越好，且模式样本协方差矩阵 $\boldsymbol{W}$ 转换后的行列式值越小越好，即要求取一个变换矩阵 $\boldsymbol{\theta_p}$ 使两者的比值最大：

$$\hat{\boldsymbol{\theta}}_p = \arg\max_{\theta_p} \frac{\left|\boldsymbol{\theta_p}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{\theta_p}\right|}{\left|\boldsymbol{\theta_p}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{\theta_p}\right|} \tag{4}$$

要使式(4)有最大值，相当于对矩阵 $\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{B}}$ 求取特征值及其对应的特征向量，将特征值从大到小排序，取前 p 个大的特征值所对应的特征向量构成变换矩阵。

语料信息共分为 J 个样本，N_j 为第 j 个样本的语料个数，N 为所有训练语料的个数，$\boldsymbol{W}_j$ 为第 j 个样本的协方差矩阵。x_i 为训练语料的集合，$\bar{\boldsymbol{X}}_j$ 为训练语料第 j 个样本的均值向量，$\bar{\boldsymbol{X}}$ 为所有训练语料(模式总体)的均值向量，$g(i)$ 表示 x_i 所属的样本，$\boldsymbol{W}$ 与 $\boldsymbol{B}$ 的定义如下：

$$\boldsymbol{W} = \frac{1}{N}\sum_{j=1}^{J} N_j \boldsymbol{W}_j \text{，} \boldsymbol{B} = \frac{1}{N}\sum_{j=1}^{J} N_j (\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})(\bar{\boldsymbol{X}}_j - \bar{\boldsymbol{X}})^{\mathrm{T}} \tag{5}$$

其中，$$\boldsymbol{W}_j = \frac{1}{N_j}\sum_{g(i)=j} (x_i - \bar{\boldsymbol{X}}_j)(x_i - \bar{\boldsymbol{X}}_j)^{\mathrm{T}}, j = 1,2,\cdots,J \tag{6}$$

$$\bar{\boldsymbol{X}}_j = \frac{1}{N_j}\sum_{g(i)=j} x_i, j = 1,2,\cdots,J \text{，} \bar{\boldsymbol{X}} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{7}$$

同 PCA 一样，LDA 也存在较大的缺陷：经过 LDA 变换后的协方差矩阵 $\boldsymbol{W}_j$ 不能对角化。

3 改进的算法及其实现

3.1 对 PCA 和 LDA 的改进

目前的语音识别系统大多都采用隐马尔可夫模型(Hidden Markov Model, HMM)，但在实际应用中为了减少存储空间和降低计算量，通常会假设输入 HMM 的协方差矩阵仅为对角线上有值(其他元素均为 0)。这样，通过 PCA 和 LDA 得到的协方差矩阵不符合应用 HMM 的假设，造成失真从而影响识别率。本文引进最大似然线性转换(Maximum Likelihood Linear Transformation, MLLT)改进 PCA 和 LDA。与 PCA 和 LDA 相似，MLLT 也是通过求取一个变换矩阵来变换矢量空间， MLLT 不会对数据进行降维，但可使变换后模式样本的协方差矩阵对角化。这样，通过 MLLT 后得到的协方差矩阵就可以满足应用 HMM 的假设了。

基于对 PCA 和 LDA 改进的目的，通过式(8)来求取经 MLLT 后的协方差矩阵，即

$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta \in R^{n\times n}} N \ln|\boldsymbol{\theta}| - \sum_{j=1}^{J} \frac{N_j}{2} \ln\left|\mathrm{diag}\left(\boldsymbol{\theta}^{\mathrm{T}} W_j \boldsymbol{\theta}\right)\right| \tag{8}$$

其中，$\hat{\boldsymbol{\theta}}$ 为需要求取的协方差矩阵；$\boldsymbol{\theta}$ 为经过PCA或LDA转换后所得到的矩阵；J为语料信息样本的个数；N_j为第j个样本的语料个数；N为所有训练语料的个数；W_j为第j个样本的协方差矩阵；diag 表示提取矩阵对角元素。

3.2 改进后方法的实现

根据本文提出的改进的算法，语音特征提取的步骤如下：

step1 提取语音特征参数。本文采用的语音特征参数为MFCC(包括其对数能量、一阶和二阶差分系数，共 39 维)。

step2 做线性变换并降维。以 LDA 为例，对矩阵 $\bar{\boldsymbol{W}}^{-1}\bar{\boldsymbol{B}}$ 求取特征值及其对应的特征向量，选取前 p 个大的特征值对应的特征向量构成变换矩阵 $\boldsymbol{\theta}$。

step3 运用 MLLT 再作一次线性变换。通过 step2 得到的 θ 来求取符合 HMM 假设的矩阵 。

step4 使用 CMS 来对语音进行增强。完成前由 step1～step3 得到的表示语音的特征向量，设为

$$\boldsymbol{C} = \{C_1, C_2, \cdots, C_t, \cdots, C_N\}, t = 1,2,\cdots,N$$

运用 CMS 可以求得增强后的特征向量为

$$\tilde{\boldsymbol{C}}_t = C_t - M_x, t = 1,2,\cdots,N$$

其中，$M_x = \frac{1}{N}\sum_{t=1}^{N} C_t$。

4 实验结果

实验系统采用的训练语音库有两套：(1)863 大词汇量连续语音数据库(863CSL)，录音时长共计 46 h，语音采样率为 8 kHz，16 bit 线性量化；(2)Aurora 2.0，它是由欧洲电信标准协会发行的语料，提供了 8 种不同噪音，有地铁、人群、汽车、展览馆、餐厅、街道、机场、火车，包括 6 种不同的信噪比，分别为-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB。

实验 1 将 Aurora 2.0 中的不同信噪比的噪声加入到待测试的语料(取自 863CSL)中，以观测各种语音特征提取方法的鲁棒性。

求取 MFCC 时，预增强系数取值 α =0.975，语音帧长为 20 ms，帧移(相邻帧的叠加部分)10 ms，Mel 滤波器的个数为 18 个。

表 1 中“ MFCC ”表示提取MFCC作为语音特征，而不进行线性变换降维和语音增强处理，“ PCA+MLLT ”表示将提取出的MFCC进行LDA和MLLT，“ LDA+MLLT+CMS ”则表示将提取出的MFCC进行LDA, MLLT和CMS增强，以此类推。表 1 中所测数据为几种特征提取方法在各种噪声环境中的音节识别率，由于汉语和西方语言存在的较大的差异，因此中文语音识别系统的性能评估是以音节(syllable)正确率或字(character)正确率为依据，而不是以西方语言常用的词(word)正确率为依据[4]。

表 1 各种特征提取方法的鲁棒性比较 (%)

方法	-5 dB	0 dB	5 dB	10 dB	15 dB	20 dB	平均值
MFCC	-3.96	17.63	37.15	50.14	54.78	59.26	35.83
PCA	1.41	21.98	41.41	54.02	59.64	61.41	39.98
LDA	1.72	22.48	42.26	54.91	60.95	63.14	40.91
PCA+MLLT	1.58	22.14	40.96	54.24	59.25	61.86	40.00
LDA+MLLT	1.92	23.36	44.95	56.81	63.29	65.77	42.68
MFCC+CMS	-3.09	18.91	37.84	51.08	55.51	60.11	36.73
PCA+CMS	1.58	21.82	41.51	53.92	59.77	61.96	40.09
LDA+CMS	2.47	22.91	42.69	55.61	61.46	64.01	41.56
PCA+MLLT+CMS	2.83	23.48	41.85	55.21	62.26	64.32	41.66
LDA+MLLT+CMS	2.45	23.69	44.65	57.76	65.36	68.82	43.79

实验结果表明：(1)结合了 CMS 的特征提取方法在噪声环境中的识别率比没有结合它的高，证明 CMS 对提高语音特征鲁棒性有明显的作用。(2)结合了 MLLT 的特征提取方法的平均识别率比不使用它的方法识别率高。(3)LDA+MLLT+CMS 组合在噪声环境下显示出比其他组合或特征提取方法更好的性能，例如，比仅应用 LDA 的高 2.88%，比应用 PCA+MLLT+CMS 的高 2.13%。

（下转第 200 页）