文章编号:1001 - 9081(2005)10 - 2401 - 03

基于独立分量分析和矢量量化的说话人识别

屈 微,刘贺平
(北京科技大学 信息工程学院,北京 100083)
(wei747734@sina.com)

摘 要:使用独立分量分析(ICA)来提取说话人特征并与矢量量化(VQ)判决方法相结合,实现了一个高性能的基于 ICA 特征的 VQ (ICA-VQ)说话人识别系统。通过 ICA 变换得到说话人语音特征基函数系数用于生成 VQ 码书,并导出包含能量失真的 ICA-VQ 码书失真测度和质心确定条件,生成最终的判决。仿真实验中 ICA 提取的特征分别用于不同系统实现说话人确认任务,各系统的 DET 曲线对比验证了 VQ 方法用于 ICA 特征分类判决的优势,同时不同码书尺寸下的等差率(EER)对比证明了 VQ 码书设计的有效性。

关键词:独立分量分析(ICA);矢量量化(VQ);说话人识别;失真测度
中图分类号: TP18 **文献标识码**:A

Speaker recognition system based on ICA and VQ

QU Wei, LIU He-ping
(*Information Engineering School, University of Science and Technology Beijing, Beijing* 100083, *China*)

Abstract: The paper combined the speaker feature extracted by ICA with VQ technique to the ICA-VQ speaker recognition system with high performance. A speaker speech ICA synthesis model was presented to get the speaker speech feature bases with ICA algorithm, and the coefficients of the bases were used in designing codebooks. A novel distortion measurement including energy and a new centroid condition were given. In the simulation experiment of speaker verification, the EER contrast results of VQ with different sizes prove that VQ codebooks are efficient and the DET cures of various methods show that VQ is a more suitable method to speaker recognition with the coefficients of ICA feature bases.

Key words: Independent Component Analysis (ICA); Vector Quantization (VQ); speaker recognition; distortion measurement

0 引言

说话人识别中提取说话人语音特征是一个基本环节。特征提取得好,不同说话人就能表现出很大的差别,可以很容易地设计出高性能的分类器,得到准确的识别结果。语音的独立分量(ICs)基函数可以作为语音的一种有效特征,具有紧致,精确和高分辨的特点并且在说话人识别中具有很好分类性能[1],提取独立分量特征基函数通常使用独立分量分析(ICA)方法[2]。但基本 ICA 模型有一个重要的假设条件,即最多只有一个源是高斯分布的,如果具有高斯分布的源信号个数超过一个,则各源信号是不可分的,只有满足这一假设 ICA 算法才能成立。Darmois 定理严格证明了这一结论[3]。

根据特征识别说话人可以采用参数化(统计量)和非参数化(模板)方法[4]。参数化方法对特征分布有严格的约束,而非参数方法对特征分布只有很少的假设。典型的参数化方法有高斯混合模型(GMM)和隐马尔可夫(HMM)模型。它们都是基于将信号重构为多个高斯分布的子随机变量的组合的假设[5]。因此使用 GMM 和 HMM 系统[6]分类 ICA 提取的说话人特征,在假设条件上是不一致的,影响 ICA 特征分类性能的发挥,导致判决时出现偏差。VQ 是一种非参数方法,与 GMM 和 HMM 不同,VQ 使用码字组成的码书作为说话人模型,码字的分布遵从于训练数据的基本分布[7],而没有其他的约束条件。另外,当可用于训练的数据量较少时,基于 VQ 的方法比 GMM 和 HMM 具有更好的鲁棒性。而且,VQ 方法比较简单,实施性也较好。

本文将 ICA 特征提取与 VQ 判决结合设计出基于 ICA 特征的 VQ (ICA-VQ)说话人识别系统。

1 ICA 说话人特征提取

稀疏编码和 ICA 概念已经成功地应用于图像编码和自然信号表示。与基本视觉皮层的单细胞接收场相似,自然图像的稀疏编码体现的是局部方向基滤波器功能[8]。一个图像块可以由一组基图像块乘以尽可能稀疏的系数形成线性组合来表示。ICA 用来阐释自然图像和声音信号的基函数,信号具有稀疏性是基本假设条件,超高斯分布的语音信号满足这一条件。

连续语音可以表示为一组语音特征基的加权组合。独立分量分析(ICA)是一种利用线性非正交变换来获得数据有效表示的统计方法,能用于提取语音组合的特征基。借鉴自然图像的处理模型[9],将语音信号表示成为独立系数作为基函数加权值的线性组合,建立一个 ICA 合成模型表示语音信号生成:

$$X = AS \tag{1}$$

$$WX = U \tag{2}$$

收稿日期:2005 - 04 - 21;**修订日期**:2005 - 07 - 10 **基金项目**:国家十五科技攻关课题(2004BA616A1103)

作者简介:屈微(1974 -),女,辽宁鞍山人,博士研究生,主要研究方向:语音信号处理、盲信号处理、神经网络在数字信号处理中的应用;刘贺平(1951 -),男,辽宁沈阳人,教授,博士生导师,主要研究方向:神经网络、数字信号处理、自适应控制理论.

式中:A 的列向量为语音特征基函数,$W = A^{-1}$,U 的列为具有独立性的系数向量。X 可以是由时域的语音帧组成,也可以是帧的平滑谱域变换。S 是未知的生成语音的基本成分,作为语音生成的激励信号。A 的列向量是说话人语音的特征。U 和 W^{-1} 分别为 S 和 A 的估计,给定的语音可以重构为 $X = W^{-1}U$,表示为 X 的每列即一帧,是 W^{-1} 的列的线性组合,线性组合的系数为 U 的对应列,是用于语音编码的统计独立变量。模型的因子化的表示为:

$$x_k = \sum_{m=1}^{M} u_{mk}(W^{-1})_m \tag{3}$$

基于语音 ICA 合成模型的特征提取过程如图 1 所示。先将语音训练数据经过预处理,包括 DFT 变换及 Mel 域滤波变换到频谱域,为下一步处理提供频域高分辨率的语音数据。归一化处理包括均值提取,方差归一以及数据的去相关处理。均值为零可以消除通道卷积效应的影响。PCA 变换或特征值分解可以实现数据维数在最小方差意义下的消减,使其在新的空间的去相关和方差为 1。这两步得到的变换矩阵的列向量虽然也可以作为特征基,但其成分之间只是不相关的,还不是具有强分辨性能的特征。所以使用 ICA 变换来获得独立的特征基函数。最后分析和选取具有代表性的 ICA 基函数,作为指定说话人的特征基函数。

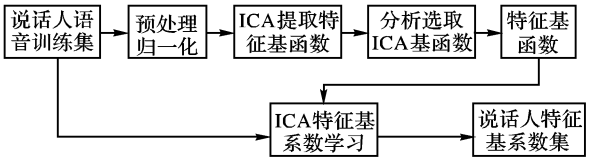

图 1 ICA 提取说话人语音特征

2 ICA-VQ 说话人识别系统

如图 2 所示,ICA-VQ 系统设计的基本思想是:将每个待识别的说话人看作是一个信源,用一个码书来表征,码书是由该说话人的 ICA 特征基函数系数向量聚类生成的,只要训练的数据量足够,就可以认为这个码书有效地包含了说话人的个人特征,而与说话的内容无关。识别时,首先将待识别的语音段以 ICA 特征基函数做线性重构(可称之为 ICA 滤波过程)得到的 ICA 特征基系数作为特征向量序列,然后用系统已有的每个码书依次对其进行量化,计算各自的平均量化失真 $D_i, i = 1,\cdots,N$。最后根据识别任务做出相应的判决。

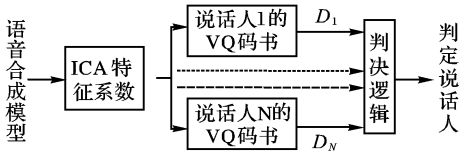

图 2 ICA-VQ 说话人识别系统

在 ICA-VQ 系统的设计中,使用 LBG(Linde-Buzo-Gray)算法生成码书。码字搜索采用一种简单有效的最近临码字搜索算法 - 部分失真搜索算法(PDS)。本文导出了适合 ICA 特征的失真测度和该测度下生成码书的质心条件用于生成 ICA 特征下的 VQ 码书。

2.1 ICA-VQ 码书的失真测度

研究表明,ICA 特征基系数具有与 LPCC 和 MFCC 等倒谱系数相同的谱失真测度。

证明: 如果 X 上时域语音数据的对数谱,使用 Fourier 序列给定的滤波器 W,那么(3)式的因子化表示就是对应 LPCC:

$$[\log X(\omega)] = \sum_{-\infty}^{+\infty} u_m \cdot [e^{-jm\omega}] \tag{4}$$

假设语音是一个最小相位的全极点模型 $\frac{\sigma}{A(e^{j\omega})}$,其频谱为 $S(\omega) = \frac{\sigma^2}{|A(e^{j\omega})|^2}$。如果 $A(e^{j\omega})$ 在单位园内是解析的,$\log S(\omega)$ 的泰勒展开就是一个稳定的全极点模型。LPC 倒谱展开为:

$$[\log \frac{\sigma^2}{|A(e^{j\omega})|^2}] = \sum_{-\infty}^{+\infty} c_m \cdot [e^{-jm\omega}] \tag{5}$$

$c_0 = \log(\sigma^2); c_{-m} = c_m$

因此,ICA 的系数 U 具有谱失真测度的性质:

$$\sum (u_m - u'_m)^2 \sim \frac{1}{2\pi}\int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 d\omega \tag{6}$$

证毕。

使用对数似然比[10] 作为 ICA 系数失真测度:

$$d_{ICA} = d_{LLR} = \log(\frac{a^T R a}{a'^T R a'}) \tag{7}$$

式中,$a^T = (1, u_{1k}, u_{2k}, \cdots, u_{Mk})$,$a'^T = (1, u'_{1k}, u'_{2k}, \cdots, u'_{Mk})$。

$\boldsymbol{R}$ 是 $(N+1)\times(N+1)$ 阶自相关矩阵。

$$a^T R a = r(0) r_a(0) + 2\sum_{i=1}^{M} r(i) r_a(i) \tag{8}$$

式中,$r(i) = \sum_{j=0}^{L-1-|i|} x(j)x(j+|i|)$,$r_a(i) = \sum_{j=1}^{M-i} a_j a_{j+1}$,$(i = 0,\cdots,M)$,$L$ 为信号 $x(n)$ 的长度,$r(i)$ 为信号的自相关函数,$r_a(i)$ 为 ICA 系数的自相关函数。

频谱与能量都携带有语音信号的信息,因此考虑能量的 ICA 系数失真测度为:

$$d_{ICA-E} = d_{ICA} + \alpha \cdot g(|E - E'|) \tag{9}$$

式中,E 和 E' 分别是输入的 ICA 系数向量和码书重构向量的归一化能量,$g(x)$ 可取为:

$$g(x) = \begin{cases} 0 & (x \leqslant x_d) \\ x & (x_F \geqslant x > x_d) \\ x_F & (x \geqslant x_F) \end{cases} \tag{10}$$

式中,α 为加权因子。x_F、x_d 和 α 经实验确定。

2.2 ICA-VQ 码书的质心确定

ICA 特征系数的失真测度不是简单的欧氏距离,其胞腔形状也是不规则的,因此使用一种等价于蒙特卡洛方法的质心确定方法,即通过估计多重积分值来确定质心。

对于离散输入向量,定义选择函数:

$$G_j(v) = \begin{cases} 1, v \in R_j \\ 0, v \notin R_j \end{cases} \tag{11}$$

则,胞腔 R_j 的质心条件描述为:

$$y_j = E[v | v \in R_j] = \frac{E[vG_j(v)]}{E[G_j(v)]} \tag{12}$$

并得到:

$$y_j = \frac{\frac{1}{K}\sum_{i=0}^{K-1} x_i G_j(x_i)}{\frac{1}{K}\sum_{i=0}^{K-1} G_j(x_i)} \tag{13}$$

式中,$j = 0,1,\cdots,N-1$。其中 K 为训练集的大小,N 为码书大小。

对于离散输入情况,给定胞腔划分的最近邻条件描述为:

$$R_i = \{v \in X : d(v,y_i) < d(v,y_j); \forall j \neq i\} \tag{14}$$

相应的平均失真描述为:

$$D = \frac{1}{K}\sum_{i=0}^{K-1}\sum_{j=0}^{N-1} d(x_i, y_j) G_j(x_i) \tag{15}$$

式中,$d = d_{ICA-E}$,训练集 $X = \{x_0, x_1, \cdots, x_{K-1}\}$。

3　仿真实验

说话人确认是说话人识别的一个范畴,它是根据说话人的语音来确定是否与其所声明人相符。确认的结果有两种:肯定(即接受)或是否定(即拒绝)。一个说话人确认系统存在着两种类型的错误,即冒认者被接受的错误(False Acceptance, FA)和真实说话人被拒绝的错误(False Rejection, FR),根据这两种错误率相等而得到的等差率(Equal Error Rate, EER),常被作为一种直观的确认系统评估标准,其值越小,系统的性能越好。另一种常用的评估标准是 DET(Detection Error Tradeoff)曲线,其横坐标和纵坐标分别是对数刻度下的 FA 和 FR,曲线越靠近坐标轴则系统的性能越好[5]。本文针对 100 个说话人进行文本无关确认仿真实验。

实验数据来自英语声调变化数据集[11](Intonational Variation in English,IViE)。语音信号的采样频率为 8kHz,训练和测试数据都来自一段 22.9s 的朗读材料,其中 17.9s 用于训练,5.0s 用于测试。用汉明窗进行分帧处理,取窗口尺度为 16ms,每帧包含 256 个采样值,帧重叠为 32 个采样值。

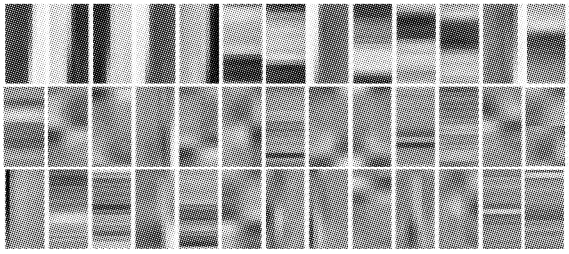

图 3　ICA 提取的特征基函数

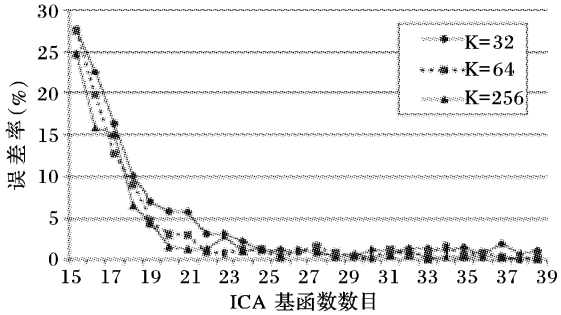

图 4　不同码书尺寸 ICA 系数性能

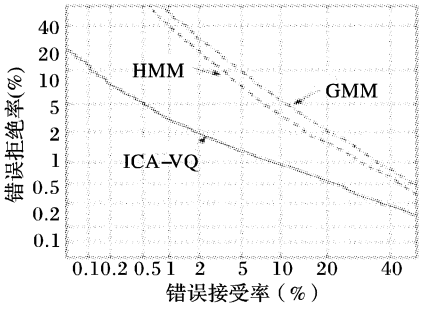

图 5　不同系统的 DET 曲线

以图像块形式展示的空间—时间域的 ICA 基函数如图 3 所示,体现出基函数的空间和时间结构。纵坐标是时间,横坐标 MEL 对数倒谱系数。图像块的灰度等级代表基函数的值,从最小值到最大值被归一化为从 0(黑)到 255(白)。图像块的高度由 MEL 带宽的数目确定。

实验中对比了不同码书尺寸下,ICA 提取特征的性能,图 4 中曲线表明 ICA 基函数数目达到 27 时,ICA-VQ 系统识别误差已接近零。

每个说话人对应的基函数形成 ICA 滤波,说话人的语音训练信号经过滤波得到的 ICA 系数作为该说话人的特征参数,分别作为 ICA-VQ 的码书训练集及 GMM 和 HMM 的状态输入序列,得到实验中各点的 DET 曲线见图 5。ICA-VQ 系统的 DET 曲线最靠近坐标轴,其性能明显优于 GMM 和 HMM 系统。以等差率(EER)作为整个系统的性能指标 GMM 和 HMM 系统 EER 分别为 16.8% 和 14.9%,码书大小为 256 的 ICA-VQ 系统比两者分别下降 85% 和 83%,不同码书大小的 ICA-VQ 系统的 EER 见表 1。可以看出 ICA-VQ 系统性能在最小码书(大小为 32)时也大大好于 GMM 和 HMM。并且随着码书大小的增加等差率下降很快,系统性能明显提高,证明了码书设计的有效性。

表 1　不同码书大小的 ICA-VQ 系统的等差率

码书大小	32	64	128	256
系统 EER	8.6%	5.3%	3.5%	2.5%

4　结语

ICA 语音合成模型,用于提取 ICA 特征基函数。说话人语音信号经 ICA 基函数的滤波作用生成的基系数作为说话人的语音的特征参数,并将 ICA 特征参数与 VQ 方法有效结合,设计出了 ICA-VQ 说话人识别系统。说话人确认的仿真实验表明,该系统性能明显好于 ICA 特征应用于 GMM 及 HMM 系统。可以说,ICA 特征与 VQ 结合建立的说话人识别系统的是具有理论依据,同时又是高效、实用的。

参考文献:

[1] JANG GJ, YUN SJ, OH YH. Feature vector transformation using ICA and its application to speaker verification [J]. Eurospeech, 1999:767 – 770.

[2] JANG GJ, LEE TW, OH YH. Learning statistically efficient features for speaker recognition [A]. In proceedings ICASSP[C], 2001.

[3] CAO XR, LIU RW. General approach to blind source separation [J]. IEEE Trans. Signal Processing 1996, 78(4): 753 – 766.

[4] DUDA R, HART P, STORK D. Pattern Classification [M]. Seconded. Wiley Interscience, New York, 2000.

[5] REYNOLDS D, QUATIERI T, DUNN R. Speaker verification using adapted gaussian mixture models [J]. Digital Signal Processing 2000, 10 (1):19 – 41.

[6] SONMEZ, M, HECK L, WEINTRAUB M, *et al.* A lognormal tied mixture model of pitch for prosody-based speaker recognition [A]. In Proc. 5th European Conference on Speech Communication and Technology, Eurospeech (Rhodos, Greece) [C], 1997. 1391 – 1394.

[7] GERSHO A, GRAY R. Vector quantization and signal compression [M]. Kluwer Academic Publishers, Boston, 1991.

[8] OLSHAUSEN BA, FIELD DJ. Emergence of simple – cell receptive Geld properties by learning a sparse code for natural images [J]. Nature, 1996(381): 607 – 609.

[9] BELL AJ, SEJNOWSKI TJ. The independent components of natural scenes are edge filters[J]. Vision Research, 3 1997, 7(23): 3327 – 3338.

[10] ROSCA J, KOFMEHL A. Cepstrum-like ICA representations for text independent speaker recognition [A]. In proceedings ICASSP[C], 2003.

[11] IViE (Intonational Variation in English, UK ESRC award R000237145) [EB/OL]. http://www.phon.ox.ac.uk/~esther/ivyweb.