基于非参数独立分量分析的说话人识别方法*)

陈 刚 陈莘萌 向广利
(武汉大学计算机学院 武汉 430072)

摘 要 首先用非参数独立分量分析方法提取表征说话人音频特性的时域基函数组,语音信号可由这些基函数线性组合而成。每个可识别的说话人对应一个不同的基函数组,对某个特定人的输入音频,只有与它对应的基函数组使其系数向量各分量之间的独立性最强(也就是互信息最小)。对待识别音频,分别用已知说话人的时域基函数组计算各自的系数向量,并计算系数向量各分量之间的互信息。互信息最小的基函数组对应的说话人即为识别结果。实验结果表明,即使用很少的测试数据,也能达到很高的识别率。

关键词 非参数独立分量分析,时域基函数组,系数向量,互信息

Speaker Recognition Based on Nonparametric Independent Component Analysis

CHEN Gang CHEN Xin-Meng XIANG Guang-Li
(School of Computer, Wuhan University, Wuhan 430072)

Abstract Time-domain basis functions are obtained through nonparametric independent component analysis first, which exhibit the main characteristics of the specific speaker. Speech signals then can be represented by the superposition of the basis functions. Every speaker candidate has his own set of basis functions, which are different from those of others. And, for a speech signal by a specific speaker, only his own set of basis functions can make the elements of the coefficient vectors most independent (namely, the mutudl information is minimal). To recognize a test speech signal, all sets of basis functions are used to produe the coefficient vectors, and the mutual information among the elements of the coefficient vectors are calculated. The speaker who has the minimum mutual information is thought of as the producer of the test speech signal. Experiments show that a high recognition rate can be achieved by a small amount of data.

Keywords Nonparametric component analysis, Time-domain basis functions, Coefficient vectors, Information

语音是一类重要的音频。同样的一段语音,可以从不同的层面传达多种信息,如语义信息、语言信息、说话人情绪、说话人性别、说话人身份等等。说话人识别就是根据人的声音特征,识别出某断语音是谁说的,而语音中包含的其他信息则被忽略。

说话人识别系统一般由两个部分组成:训练部分和测试部分。当训练与测试采用相同的语句时,称为与文本有关的说话人识别,反之则称为与文本无关的说话人识别。由于说话人识别的目的是模仿人类的学习和再认能力,因此与文本无关的说话人识别更具灵活性、更有应用价值。本文提出的方法属于与文本无关的说话人识别方法。

根据测试音频的来源范围,说话人识别又可进一步分为两类:说话人辨认(Speaker Identification)和说话人确认(Speaker Verification)。前者是指测试音频来自某个特定的说话人集合(假设有 N 个说话人),其任务是从 N 个候选人中挑出谁是真正的说话人。而后者则是判断某段音频是否为某个特定的说话人所发出,因为对测试音频的来源未作任何限定,因而又称其为一个"开集"(Open-Set)问题。利用一个合适的阈值,可将二者统一起来[1]。

说话人识别的关键是建立话者模型,也常称为声纹(Voice Print)。建立话者模型常用的方法有:矢量量化(VQ)方法[2]、隐马尔科夫模型(HMM)方法[3]、人工神经网络(ANN)方法[4]等等。要综合评价这些方法的性能,必须从算法复杂度、在不同语音信道和识别环境下的鲁棒性、识别速度等方面考虑。特别是,根据文[5]的研究结论,只要利用足够长的语音,总可以达到一个较高的识别率。因此,利用较少的测试数据而达到较高的识别率,也是衡量算法性能的重要指标。

1 方法概述

独立分量分析(Independent Component Analysis,简称ICA)是近 20 多年来发展并成熟起来的数据分析技术。最初,它用来解决所谓"盲信源分离问题"(Blind Source Separation,简称 BSS),以后它的应用范围逐渐扩大,目前已在图像处理、音频处理、脑电波分析、金融分析等各个领域得到应用。

ICA 为描述和分析多维数据提供了一种方式,其基本模型为 $x=As$,其中 x 和 s 是 N 维列向量,A 是 $N\times N$ 的满秩矩阵。在盲信源分离问题中,x 称为观测向量,A 称为混合矩阵,s 为源信号。在这种情况下,ICA 的基本假设可描述为:已知大量观测向量,求解混合矩阵 A 和源信号向量 s,使 s 的各分量 $s_i(i=1,2,\cdots,N)$ 统计独立。

如果把 A 表示成 $A=[a_1,a_2,\cdots,a_N]$,即 $a_i(i=1,2,\cdots,N)$ 是 A 的第 i 个列向量,则 ICA 基本模型可表示为 $x=\sum_{i=1}^{N}s_ia_i$。这种分解导致了 ICA 的一项重要的应用:用 ICA 方法提取多维数据特征。这时 x 称为观测向量,A 的每一列称为

*)基金项目:国家自然科学基金资助项目(10371033);国家 211 工程重大项目资助。陈 刚 博士研究生,研究方向:音频处理,模式识别。

一个特征基函数，s 为特征向量。从线性展开的角度来看，可称 s 为系数向量。

为了能够把 ICA 方法应用于提取音频特征，我们对音频采样序列作如下调整。设有个采样点，以 N 为长度，随机地从音频采样序列中抽取 N 个连续的采样点，组装成一个观测向量，一共可以得到 $T-N+1$ 个观测向量。每个观测向量代表一小段音频。接下来的任务就是求取音频的时域基函数组 A，A 的每一列代表音频的一个时域基函数。在说话人识别问题中，不同的说话人有着不同的时域基函数组，而观测向量就是用来生成时域基函数组的训练向量。

若令 $W=A^{-1}$，ICA 模型也可表示为 $s=Wx$。在把 A 叫做时域基函数组的情况下，W 称为时域基滤波器组（Time-Domain Basis Filters）。对任意观测向量 x，其对应的系数向量可由 $y=Wx$ 计算。设 $W=[w_1,w_2,\cdots,w_N]^T$（其中 $w_i(i=1,2,\cdots,N)$ 是 W 的行向量），则 w_ix 代表系数向量 y 的第 i 个分量。由 ICA 的基本假设可知，对于全体观测向量 $X=\{x^{(k)}|k=1,2,\cdots,T-N+1\}$，$W$ 使系数向量 y 的各分量之间（尽可能）统计独立。从信息论的基本理论可知，随机变量之间相互独立等价于它们之间的互信息为 0[6]。在 ICA 模型下，系数向量 y 的各分量之间尽可能统计独立，等价于 y 的各分量之间互信息最小，即 $W=\arg\min_W I(y)$，其中 $I(y)=I(y_1,y_2,\cdots,y_N)$ 是随机向量 y 的各分量之间的互信息。

对任一个随机向量 $x=[x_1,x_2,\cdots,x_N]^T$，互信息 $I(x_1,x_2,\cdots,x_N)=\sum_{i=1}^{N}H(x_i)-H(x)$，其中 $H(x_i)=-\int p_i(x_i)\log p_i(x_i)\mathrm{d}x_i$，$H(x)=-\int p(x)\log p(x)\mathrm{d}x$。而对于线性变换 $y=Wx$ 而言，$I(y_1,y_2,\cdots,y_N)=\sum_{i=1}^{N}H(y_i)-H(y)=\sum_{i=1}^{N}H(y_i)-H(x)-\log|\det W|$[7]。由于 $H(x)$ 不随 W 变化，因此可以引入另一个量 $\Omega=\sum_{i=1}^{N}H(y_i)-\log|detW|$，$\Omega$ 的最小值对应于 $I(y_1,y_2,\cdots,y_N)$ 的最小值。

基于以上分析，可以设计出如图 1 所示的采用独立分量分析技术的说话人识别方法。图中虚线框内的时域基滤波器组（与时域基函数组等价）就是表征说话人音频特性的关键数据。

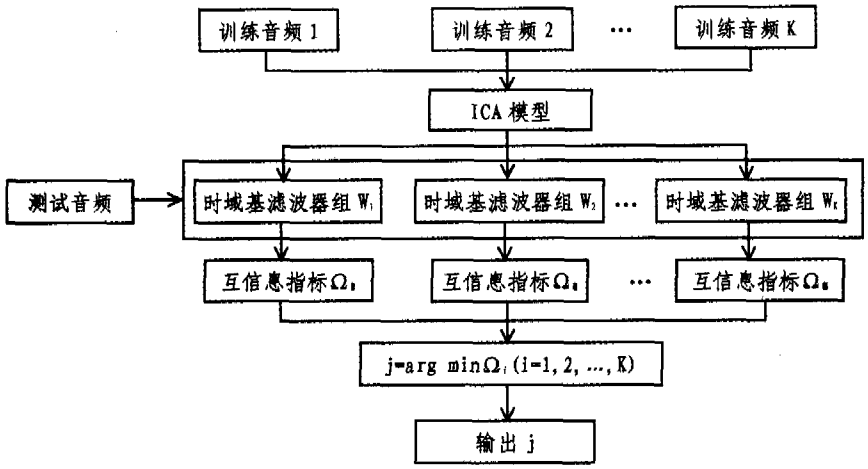

图 1　基于 ICA 的说话人识别方法流程图

2　时域基滤波器组的计算

时域基滤波器组的计算就是说话人识别的训练部分。对每一个已知的说话人，都要利用其训练数据计算出相应的时域基滤波器组。

由上节讨论可知，Ω 可以作为 $I(y_1,y_2,\cdots,y_N)$ 的替代指标。Ω 是 W 的函数（可用 $L(W)$ 表示），可得 $L(W)=\sum_{i=1}^{N}H(y_i)-\log|\det W|=-\sum_{i=1}^{N}E\{\log p_{y_i}(y_i)\}-\log|\det W|$。设训练向量集合为 $X=\{x^{(k)}|k=1,2,\cdots,M\}$，由于 $E\{\log p_{y_i}(y_i)\}\approx\frac{1}{M}\sum_{k=1}^{M}\log p_{y_i}(y_{ik})=\frac{1}{M}\sum_{k=1}^{M}\log p_{y_i}(w_ix^{(k)})$，因此 $L(W)=-\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{M}\log p_{y_i}(w_ix^{(k)})-\log|\det W|$。

在函数 $L(W)$ 中，系数向量的各分量的概率密度 $p_{y_i}(y_i)$ 的计算是一个关键。对 $p_{y_i}(y_i)$ 的不同估计形成了不同的 ICA 解决方案。总的来说，有两种方案：参数化方案和非参数化方案。参数化方案试图建立一个参数化的概率密度模型，如 Te-Won Lee 等人采用的“广义高斯模型”（Generalized Gaussian Prior）[8]，假设随机变量的概率密度服从 $p(x)\propto\exp\left(-\frac{1}{2}|x|^q\right)$。其中 q 为参数。通过改变 q 的值，可以描述一大类高斯及非高斯分布的随机变量（如 Gaussian，Platykurtic，Leptokurtic 分布等）。参数化方案的最大优点是计算效率高，且在特定情形下对随机变量的概率密度有很好的估计效果。但对现实世界的数据（Real-World Data）是否一定适用，至今没有人给出令人信服的证明。根据 Cardoso 的研究结果，对源信号的概率密度的错误估计，可能导致 ICA 方法彻底失败[9]。因此，我们采用非参数化的概率密度估计，由观测数据本身计算其概率密度。

设 y_i 是源信号 s_i 的估计值，选取 M 个观测数据（标量），重构 $p_{y_i}(y_i)$ 可采用如下的算式 $p_{y_i}(y_i)=\frac{1}{Mh}\sum_{m=1}^{M}\phi\left(\frac{y_i-Y_m}{h}\right)$，其中 h 是所谓“核带宽”（Kernel Bandwidth），ϕ 是“高斯核”（Gaussian Kernel），$\phi(u)\triangleq\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$，$Y_m$ 是“核重心”（Kernel Centroid）。这个公式就是核密度估计（Kernel Density Estimation）理论的中心结论[10]。核带宽可取 $h=1.06M^{-1/5}$，其中 M 是样本点的个数。令 $L_0(W)=\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{M}\log p_{y_i}(w_ix^{(k)})$，由 $p_{y_i}(w_ix^{(k)})=\frac{1}{Mh}\sum_{m=1}^{M}\phi\left(\frac{w_i(x^{(k)}-x^{(m)})}{h}\right)$，有：

$L_0(W)=\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{M}\log\left(\frac{1}{Mh}\sum_{m=1}^{M}\phi\left(\frac{w_i(x^{(k)}-x^{(m)})}{h}\right)\right)$，$L(W)=-L_0(W)-\log|\det W|$。

至此，ICA 的求解问题就归结为一个以 W 为决策变量的最优化问题，目标函数为 $L(W)$。可用拟牛顿(Quasi-Newton)方法对目标函数进行优化，具体步骤参见文[11]。

对于已知说话人的音频，其训练音频往往长达数十秒到数分钟，产生数十万或数百万个训练向量。为了减少计算量，我们利用向量量化的方法从所有训练向量中挑选出 M 个代表，然后用 M 个训练向量生成时域基滤波器组。这 M 个向量既在数量上远远小于原来的训练向量数，又较好地保留了原训练向量在数据空间的分布结构。具体步骤如下：

① 将所有的训练向量置于一个初始的胞腔中。

② 用 LBG 方法[12]将初始胞腔一分为二，此时总的胞腔数为 2。

③ 在所有胞腔中挑选出包含向量数最多的胞腔，用 LBG 方法将其一分为二，总胞腔数加 1。

④ if 总胞腔数达到 M 转⑤else 转③

⑤ for $i=1$ to M

在第 i 个胞腔中，挑选出离该胞腔质心最近的向量作为输出向量。

3 互信息指标的计算

在识别阶段，要计算测试音频经过线性变换 $y=Wx$ 后，其系数向量各分量之间的互信息。设 V 是测试向量 x 的白化矩阵(即 $z=Vx$ 是白向量)，则 $y=Wx=WV^{-1}z$，$I(y)=I(WV^{-1}z)=\sum_{i=1}^{N}H(y_i)-H(y)=\sum_{i=1}^{N}H(y_i)-H(z)-\log|\det WV^{-1}|$。在 V 确定的情况下，$H(z)$相对于 W 是常量，实现中用 $\Omega=\sum_{i=1}^{N}H(y_i)-\log|\det WV^{-1}|$ 作为互信息的判断指标。

4 实验结果和讨论

实验在 PⅣ-1.2G 微型计算机上完成，操作系统为 Windows 2000 Professional 版，执行程序在 Matlab 6.0 和 Microsoft Visual C++6.0 环境下生成。实验中挑选了 15 个说话人，其中女性 8 个，男性 7 个。每个说话人有 30 秒训练音频，均为单声道、11025Hz、16 位的 WAV 文件。每个时域基函数对应于一小段音频，实现中取 $N=24$，每个音频段持续时间为 $1/11025\times24\times1000=2.18$ms。

图 2 显示的是第一个女性说话人的时域基函数，图 3 显示的是第一个男性说话人的时域基函数。每个说话人对应于 24 个时域基函数，图中只挑选了其中 5 个进行显示。

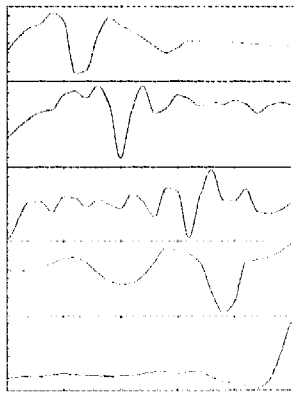

图 2 女性说话人时域基函数示例

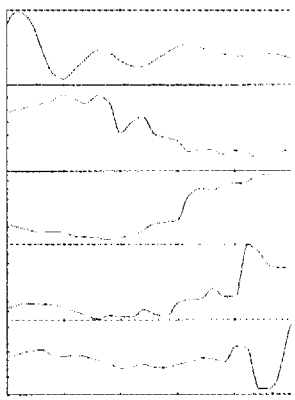

图 3 男性说话人时域基函数示例

实验中一共应用了 135 个文件进行测试，即每个说话人挑选 9 段音频进行测试(每段音频持续 9 秒钟)。图 4 显示的是第一个女性说话人的测试音频计算示例，Ω_1 明显小于其他 Ω 值，表明测试音频由话者 1 发出。图 5 显示的是第一个男性说话人的测试音频计算示例，Ω_9 明显小于其他 Ω 值，表明测试音频由话者 9 发出。由于实验中第一个男性说话人排在前 8 个女性说话人之后，因此话者 9 即为第一个男性说话人。

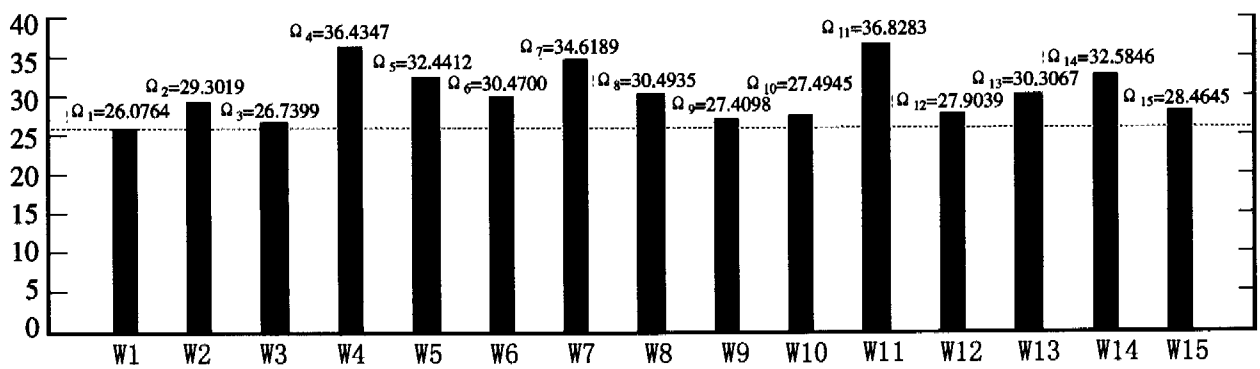

图 4 女性说话人的测试音频计算示例

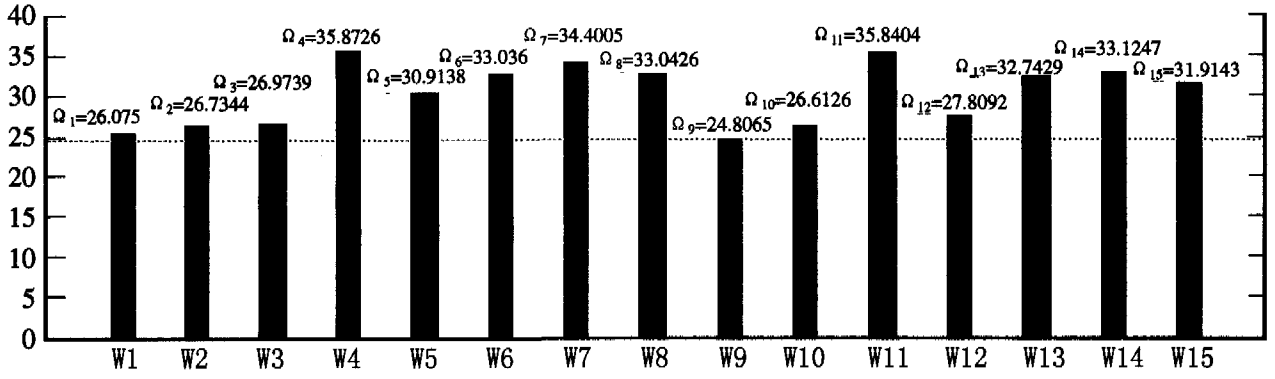

图 5　男性说话人的测试音频计算示例

实验结果显示：在测试音频长度大于 3 秒时，就能达到很高的识别率(>80%)；而当测试音频长度大于 6 秒以上时，识别率超过 85%。

小结　本文讨论了一种基于非参数独立分量分析的说话人识别方法。该方法着眼于捕捉构成特定说话人语音的基本元素——时域基函数组，说话人的任何一段语音都由这些基函数叠加而成，而系数向量各分量之间统计独立(由互信息指标度量)。不同说话人的时域基函数组是不同的，这种差别源于说话人不同的声道特性、口腔特性、发音习惯等等诸多因素。对任一测试音频，哪一组时域基函数与它"匹配"得最好(互信息最小)，该段音频即为哪一个说话人所发出。

与已有的说话人识别方法相比，本文提出得方法不用建立模拟发音器官动作的物理模型，对随机变量概率密度的估计没有使用任何"先验"模型，而是直接通过训练数据本身计算，因而具有很高的可靠性和很强的适应性。它的主要计算量集中在时域基函数组的训练阶段。在识别阶段，互信息指标的计算相对简单，用很少的测试数据就能达到很高的识别率。

参 考 文 献

1　Reynolds D A. An Overview of Automatic Speaker Recognition Technology. In: Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, Florida, 2002. 300～304

2　孙圣和，陆哲明. 矢量量化技术及应用. 北京：科学出版社，2002

3　Che C W, Yuk Q G. An HMM Approach to Text-Prompted Speaker Verification. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Atlanta, Georgia, 1996, 7(10): 673～676

4　Fakotakis N, Sirigos J. A High-Performance Text-Independent Speaker Recognition System Based on Vowel Spotting and Neural Nets. In: Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, Atlanta, Georgia, 1996, 2: 661～664

5　包威权，陈坷，迟惠生. 基于 HMM/MLFNN 混合结构的说话人辨认研究. 见：第四届全国人机语音通讯会议论文集，1995. 185～189

6　Barlow H B. Unsupervised Learning. Neural Computation, 1989, 1: 295～311

7　Cover T M, Thomas J A. Elements of Information Theory. New York: Wiley, 1991

8　Lee T W, Lewicki M S. The Generalized Gaussian Mixture Model Using ICA. In: Proc. of the Intl. Workshop on Independent Component Analysis(ICA'00), Helsinki, 2000. 239～244

9　Cardoso J F. Blind Signal Separation: Statistical Principles. Proc. of the IEEE Special Issue on Blind Identification and Estimation, 1998, 9: 2009～2025

10　Silverman B W. Density Estimation for Statistics and Data Analysis. New York: Chapman and Hall, 1985

11　Boscolo R, Pan H. Independent Component Analysis Based on Nonparametric Density Estimation. IEEE Transactions on Neural Networks, 2004, 15(1): 55～64

12　Linde Y, Buzo A, Gray R M. An Algorithm for Vector Quantizer Design. IEEE Transactions on Communications, 1980, 28(1): 84～95

(上接第 146 页)

系性导致的常见现象。

参 考 文 献

1　Pawlak Z. Rough Sets. International J of Information and Computer Sciences, 1982, 11(5): 341～356

2　Pawlak Z. Rough Sets-Theoretical Aspects of Reasoning about Data. Dordrecht: Kluwer Academic, 1991

3　曾黄麟. 粗集理论及其应用—关于数据推理的新方法. 修订版. 重庆：重庆大学出版社，1998

4　Skowron A, Rauszer C. The discernibility matrices and functions in information systems. In: Slowinski R, ed. Intelligent Decision Support-Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht, Kluwer: Academic Publishers, 1992. 331～362

5　李小霞，陈绵云. 知识表达系统的简化与集族的极小子集(I). 计算机科学，2004, 31(1): 95～97

6　李小霞，陈绵云. 知识表达系统的简化与集族的极小子集(II). 计算机科学，2004, 31(2): 9～10

7　洪帆，付小青. 离散数学基础. 第二版. 武汉：华中理工大学出版社，1983

8　李小霞，陈绵云. 知识表达系统的简化与集族的极小子集(I). 计算机科学，2003, 30(3): 54～56

9　李小霞，陈绵云. 决策表条件属性的简化. 华中科技大学学报(自然科学版)，2003, 31(8): 85～87