文章编号: 1001-0920(2007)05-0581-04

在噪声环境下的分级说话人辨识

邱政权, 尹俊勋, 杨 俊
(华南理工大学 电子与信息学院, 广州 510640)

摘 要: 将小波变换与维纳滤波结合起来对语音进行去噪. 为了提高系统的鲁棒性和辨识率, 在采用分级说话人辨识的基础上, 将基音周期的高斯概率密度对 GMM 分类器的似然度进行加权, 形成新的似然度进行说话人辨识. 实验结果显示, 所提出系统的鲁棒性和辨识率都有所提高.

关键词: 小波变换; 维纳滤波; 分级说话人辨识; 基音周期; 高斯概率密度
中图分类号: TN912.34 **文献标识码**: A

Hierarchical speaker identification under noisy environments

QIU Zheng-quan, YIN Jun-xun, YANG Jun
(School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China. Correspondent: QIU Zheng-quan, E-mail: qiuzhengquan168@sina.com)

Abstract: Wavelet transform and Wiener filtering are combined to denoise speech. In order to improve the robustness and identification rate of the system, hierarchical speaker identification is proposed. Then the likeliness of GMM classifier is weighted by using the Gauss probability density of the pitch and novel likeliness is proposed for speaker identification. The experiment result shows that the robustness and the identification rate of the system proposed are both improved.

Key words: Wavelet transforms; Wiener filter; Hierarchical speaker identification; Pitch period; Gaussian probability density

1 引 言

在说话人识别系统中, 训练和测试环境的不匹配会造成识别性能的下降. 鲁棒性噪声补偿技术主要集中在两个方面: 语音信号的前端处理和识别过程的自适应. 传统的前端处理方法有滤波处理、谱减和谱映射等. RASTA 滤波和倒谱均值减(CMS)是说话人识别较为常用的抗噪方法[1,2]. CMS 在消除慢变的信道影响方面作用比较明显, 在电话语音说话人识别中常常用到. 然而, CMS 在去除慢变的信道特性的同时, 也去除了一部分特定说话人相对稳定的特点, 损失了说话人身份相关的信息. RASTA 滤波是针对倒谱轨迹的一种高通滤波方法, 认为噪声在倒谱的高频部分影响较大, 对倒谱轨迹进行高通滤波能消除信道的影响、提高识别效果. 实际的噪声成分比较复杂, RASTA 滤波在某些环境下效果不够明显. 这两种噪声补偿方法主要解决信道的线性失真问题, 对于非线性失真作用则不明显. MAP 和 MLLP 是语音识别和说话人识别常用的两种自适应方法[3]. 以上这些补偿方法都不是从噪声本身的特征出发, 不能对噪声频谱进行准确估计. 维纳滤波在估计原始噪声语音的能量谱时, 能够动态地搜索每帧语音, 实时估计语音倒谱和噪声倒谱, 在充分保留语音信息的同时去除噪声[4]. 为此, 本文将它用于说话人特征提取之前的前端信号处理. 为了进一步增强去噪效果, 将小波变换和维纳滤波结合起来对语音进行去噪.

作者在说话人辨识实验中发现, 随着注册说话人数的增加, 一次辨识所花费的时间随之直线上升. 很明显, 每次辨识需要用测试语音去匹配所有说话人的模型, 然后找出最相近的模型对应的说话人作为辨识结果. 这样必然导致模型数(注册人数)越多, 所花费时间越长, 当注册人数达到一定数量后, 系统就很难进行实时响应. 这样的系统即使辨识率再高, 也不能满足实用的要求. 基于分类技术的分级说话

收稿日期: 2006-04-19; **修回日期**: 2006-08-26.
基金项目: 国家自然科学基金项目(60275005); 广东省自然科学基金重点项目(04105938).
作者简介: 邱政权(1972—), 男, 湖南邵阳人, 博士生, 从事语音信号处理的研究; 尹俊勋(1942—), 男, 广东东莞人, 教授, 博士生导师, 从事语音视频信号处理与通信等研究.

人辨识方法正是基于这种考虑. 该方法在基本不降低辨识率的同时,可以大大降低辨识系统的时间复杂度,提高识别的响应速度[5]. 近年来,越来越多的研究指向更为鲁棒性的高层次特征. 在众多的高层次特征中,基因周期的提取对语音数据的要求较少,受噪声和信道的影响也较少;同时,作为声带振动的频率,基音周期也是刻画说话人的一个重要特征[6,7]. 因此本文在采用分级说话人辨识的基础上,将基音周期的高斯概率密度对 GMM 分类器的似然度进行加权,形成新的似然度进行说话人辨识. 由于使用了基音周期的概率密度函数,能更全面、真实地描述说话人的声带特性,大大提高了辨识率. 实验结果显示,本文设计的系统取得了较好的抗噪效果,系统的辨识率和鲁棒性都有所提高.

2 小波变换与维纳滤波的联合去噪

2.1 维纳滤波

维纳滤波的设计准则是使实际输出与理论输出的均方误差最小,并应用随机分析原理对语音能量和噪声能量进行设计. 其频域表达式为

$$H(k) = \frac{P_s(k)}{P_s(k) + \lambda_n(k)}, \qquad (1)$$

其中 $P_s(k)$ 和 $\lambda_n(k)$ 分别为语音和噪声的功率谱密度. 语音信号只是短时平稳的,且语音功率谱密度无法得到. 为此将式(1)改写为

$$H(k) = \frac{E[|S_k|^2]}{E[|S_k|^2] + \lambda_n(k)} = \frac{E[|S_k|^2]/\lambda_n(k)}{1 + E[|S_k|^2]/\lambda_n(k)} = \frac{\xi_k}{1 + \xi_k}, \qquad (2)$$

其中 ξ_k 的物理意义是先验信噪比. 在实际处理中,通过静音段(或低音能量段)估计噪声功率谱密度 $\lambda_n(k)$,结合 $\lambda_n(k)$ 和噪声语音段的能量估计原始语音能量的期望 $E[|S_k|^2]$.

维纳滤波可看作一种改进的谱减算法. 在说话人识别中,维纳滤波的首要作用是消除背景噪声的影响,其次是提高对各说话人的区分度.

2.2 联合去噪方案

首先通过三尺度的 Daubechies 小波把输入含噪信号在不同频段分解;然后在各个频段分别通过维纳滤波去噪,并把各个频段的输出通过小波重构恢复信号;最后通过 Mel 滤波器组把小波系数转换成 MFCC. 去噪方案如图 1 所示.

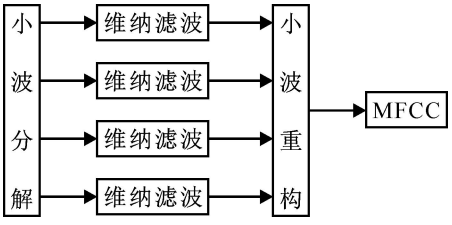

图 1 联合去噪方案

3 分级的说话人辨识系统

首先利用测试语音对训练中生成的类模型进行辨识,即在 K 个类模型中找到与测试语音最接近的类模型,这一过程称为类辨识. 假定判定语音属于类别 k,在所属该类别的说话人模型 $\{\lambda_{k1}, \lambda_{k2}, \cdots, \lambda_{kn_k}\}$ 中,找到与测试语音最接近的说话人所对应的模型,这一过程称为类内说话人辨识. 本文将该识别结果认定为系统的辨识结果.

3.1 初始分类

在运行算法之前,需先进行初始分类. 设其分为 K 类,方法如下:

1) 从 N 个注册说话人的模型中任选一个模型 λ,并设类模型 $\lambda_1^M = \lambda$;

2) 分别计算其余 $N-1$ 个说话人模型到类模型 λ_1^M 的距离,并按升序的排列方式对这些距离排序;

3) 取排在第 $[\frac{N}{K}k]$ 位处对应说话人的模型 λ',并令 $\lambda_1^M = \lambda'$,其中 $[\cdot]$ 表示取整函数,$k = 1, 2, \cdots, K$.

模型间的距离采用加权均方误差来测度.

3.2 采用基音周期高斯概率密度加权的高斯混合模型

本文利用三尺度的 Daubechie 小波系数的加权和来检测基音周期[7]. 设每个说话人的基音周期概率密度函数为 $p_i(x)$,其中 i 表示第 i 个说话人. 为了计算简单和应用方便,采用一维高斯函数来描述基音周期的概率密度,通过最大似然准则估计的均值 μ 和方差 δ,估计基音周期概率密度函数 $p_i(x)$.

设 $\theta = (\mu, \delta)$,则说话人 i 的基音周期概率密度函数的形式为

$$p_i(x/\theta) = \frac{1}{\sqrt{2\pi}\delta}\exp\left[-\frac{1}{2\delta}(x-\mu)^2\right]. \qquad (3)$$

需要求出 $\theta = (\mu, \delta)$ 的最大似然估计值 μ 和 δ^2. 设从说话人 i 的语音文件中按分析帧提取的浊音的基音周期样本值为 $X = \{x_1, x_2, \cdots, x_N\}$,相对于样本集 X 的 θ 的似然函数为

$$p(X/\theta) = p(x_1, x_2, \cdots, x_N/\theta) = \left[\frac{1}{\sqrt{2\pi}\delta}\right]^N \exp\left[-\frac{1}{2\delta}\sum_{k=1}^{N}(x_k-\mu)^2\right]. \qquad (4)$$

设 $H(\theta) = \ln p(X/\theta)$,则由最大似然估计值就是

$$\partial H(\theta)/\partial\mu = 0,\ \partial H(\theta)/\partial\delta = 0 \qquad (5)$$

的解. 求解此偏微分方程组,得到

$$\mu = \frac{1}{N_k}\sum_{1}^{N} x_k,\ \delta^2 = \frac{1}{N_k}\sum_{1}^{N}(x_k - \mu)^2. \quad (6)$$

因此说话人 i 的基音周期概率密度函数的高斯函数估计为

$$p_i(x) = \frac{1}{\sqrt{2\pi}\delta}\exp\left[-\frac{1}{2\delta^2}(x-\mu)^2\right]. \quad (7)$$

对于样本集中的每个说话人,训练阶段都可按上述方法为每个说话人估计一个基音周期概率函数. 需要强调的是,因为清音的基音周期为 0,所以最大似然估计的基音周期概率密度只是针对浊音而言.

设求得各帧的特征矢量为 $X_1, X_2, \cdots, X_N$,各帧的基音周期分别为 $x_1, x_2, \cdots, x_N$;设训练集中共有 K 个说话人,说话人 i 的基音周期概率密度函数为 $p_i(x)$. 则 GMM 分类器中基音周期概率密度函数加权的似然测度为

$$D_i = \sum_{n=1}^{N}\log(P_i(X_n)\ P_i(x_n)). \quad (8)$$

其中 $P_i(X_n)$ 表示特征矢量 X_n 在说话人 i 的高斯混合模型中得到的概率密度. 遍历训练集中的所有说话人,按照最大似然准则对应的说话人作为最终辨识结果.

4　实验结果

与文本无关的说话人辨识,是通过鉴定一个说话人发出任何测试文本来鉴别或确认说话人的身份. 实验由 30 个说话人组成,每个说话人说出 6 个句子,其中每个说话人随机选取 3 个句子组成训练集,剩下的 3 个句子组成测试集. 采用的语音是在 3 个月内分 3 次录制的. 采样率为 11 025 Hz,帧长为 30 ms,帧移为 15 ms,进行 $1 - 0.95z^{-1}$ 预加重. 说话人辨识系统框图如图 2 所示.

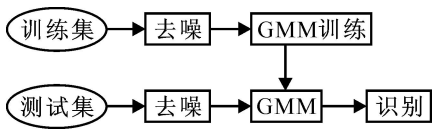

图 2　说话人辨识系统框图

为了验证所提出系统的有效性,作者进行以下实验:

实验 1　对特征的前端处理进行比较. 分别采用频谱减、维纳滤波、小波变换与维纳滤波的联合去噪(简称联合去噪) 作为前端去噪处理. 辨识统一采用 GMM,混合数统一为 64. 首先对训练集进行分帧、预加重和加窗,然后进行去噪,提取 12 阶 MFCC,建立一个 GMM,把所得的数据存储下来. 同样,对测试集进行分帧、预加重和加窗,然后提取 12 阶 MFCC,求出该 MFCC 与训练集中每个说话人的 GMM 似然得分,其中获得似然得分最大的说话人作为正确的辨识人. 实验结果见表 1.

表 1　采用不同前端处理时的辨识率

前端去噪处理	SNR			
	0	5	10	20
频谱减/ %	62.3	71.6	79.9	86.7
维纳滤波/ %	69.5	76.2	83.7	91.2
联合去噪/ %	75.6	84.8	91.4	95.1

从表 1 的辨识率看出,维纳滤波与频谱减相比,系统的辨识率和鲁棒性都有所提高,而联合去噪的辨识率和鲁棒性进一步增强.

实验 2　对辨识阶段进行比较. 前端处理采用小波变换与维纳滤波的联合去噪,在分级说话人辨识的基础上,将基音周期的高斯概率密度对 GMM 分类器的似然度进行加权,形成新的似然度进行说话人辨识,并与 GMM 辨识进行比较. 混合数统一为 64,实验结果见表 2.

表 2　采用不同似然度的辨识率

辨识方法	SNR			
	0	5	10	20
GMM/ %	75.6	82.8	91.4	95.1
基音 + GMM/ %	81.3	89.5	94.7	97.3

采用基音 + GMM 的联合辨识与采用 GMM 辨识相比,系统的辨识率和鲁棒性都有所提高,表明所提出的方法确实能提高系统的性能.

实验 3　验证分级辨识的效率. 对采用 GMM 辨识和分级辨识的测试时间进行比较,实验结果见表 3.

表 3　辨识的测试时间比较

测试时间	GMM	分级辨识
t/ s	12.36	5.28

从表 3 可以看出,采用分级辨识的测试时间明显缩短,可见分级辨识能提高系统的响应速度.

5　结　　语

维纳滤波因其对噪声的有效估计,在不同信噪比下有利于说话人识别性能的提高,但在非平稳情况下效果却不够理想. 采用小波变换与维纳滤波的联合去噪作为前端处理,使这一问题得到了较好的解决.

本文在采用分级说话人辨识的基础上,将基音周期的高斯概率密度对 GMM 分类器的似然度进行加权,形成新的似然度进行说话人辨识. 由于使用了基音周期的概率密度函数,能更全面、真实地描述说话人的声带特性,大大提高了辨识率. 实验结果显

示,本文提出的辨识方法取得了较好的抗噪效果,系统的辨识率和鲁棒性都有所提高.

参考文献(References)

[1] Hermansky H, Morgan N. RASTA processing of speech [J]. IEEE Trans on Speech and Audio Processing, 1994, 2(4): 578-589.

[2] de Lima C B, da Silva D G, Alcaim A, et al. AR-vector using CMS for robust text independent speaker verification [C]. 14th Int Conf on Digital Signal Processing. Greece, 2002, 2: 1073-1076.

[3] Kosaka T, Yamamoto H, Yamada M, et al. Instantaneous environment adaptation techniques based on fast PMC and MAP-CMS methods [C]. Proc of the 1998 IEEE Int Conf on Acoustics, Speech and Signal Processing. Seattle, 1998, 2: 789-792.

[4] 白俊梅,张世磊,张树武, 等. 噪声环境下的鲁棒性说话人识别[J]. 中文信息学报,2006, 1: 91-97.
(Bai J M, Zhang S L, Zhang S W, et al. Robust speaker recognition in noisy environment [J]. J of Chinese Information Processing, 2006, 1: 91-97.)

[5] 刘文举,孙兵,钟秋海. 基于说话人分类技术的分级说话人识别研究[J]. 电子学报,2005, 7: 1230-1233.
(Liu W J, Sun B, Zhong Q H. Research on hierarchical speaker recognition based on speaker clustering technology[J]. Acta Electronical Sinica, 2005, 7:1230-1233.)

[6] 段新,黄新宇,吴淑珍. 与文本无关的说话人辨认系统中一种新的使用基音周期方法研究[J]. 北京大学学报, 2003, 5: 690-696.
(Duan X, Huang X Y, Wu S Z. A new method of using pitch period in text-independent speaker identification system[J]. J of Beijing University, 2003, 5: 690-696.)

[7] 李香春, 杜利民. 一种基于多尺度边缘特征提取的基音检测算法[J]. 电子学报, 2003,(10) :1500-1502.
(Li X C, Du L M. A pitch detection algorithm using multiscale edges feature extraction[J]. Acta Electronical Sinica, 2003, (10) :1500-1502.)

[8] 曲天书,戴逸松. 基于离散小波变换的自适应语音消噪方法[J]. 电工技术学报, 2001, 4(2): 75-78.
(Qu T S, Dai Y S. A new method for adaptive speech noise canceling based on wavelet transform [J]. J of China Electrotechnical Society, 2001, 4(2) :75-78.)

(上接第 580 页)

[2] Zhou Y, Leung H, Yip P. An exact maximum likelihood registration for data fusion[J]. IEEE Trans on Signal Processing, 1997, 45(6): 1560-1573.

[3] Leung H, Blanchette M, Harrison C. A least squares fusion of multiple radar data[C]. Proc Int Conf Radar'94. Paris, 1994: 364-369.

[4] Halm Karniely, Hava T Siegelmann. Sensor registration using neural networks[J]. Trans on AES, 2000, 36(1): 85-101.

[5] Zhou Y. A Kalman filter based registration approach for multiple asynchronous sensors [R]. Ottawa: Defence R&D, 2003.

[6] 陈非, 敬忠良, 姚晓东. 空基多平台多传感器时间空间数据配准与目标跟踪[J]. 控制与决策, 2001, 16(11): 808-811.
(Chen F, Jing Z L, Yao X D. Time and spatial registration and target tracking for multiple airbone mobile platforms and sensors[J]. Control and Decision, 2001, 16(11): 808-811.)

[7] Friedland B. Treatment of bias in recursive filtering[J]. IEEE Trans on Automatic Control, 1969, 14(4): 359-367.

[8] Ignagni M B. An alternative derivation and extension of Friedland's two-stage Kalman estimator [J]. IEEE Trans on Automatic Control, 1981, 26(3): 746-750.

[9] 何友, 熊伟. 带反馈信息的多传感器分层估计算法[J]. 电子学报, 2000, 28(12): 85-89.
(He Y, Xiong W. Multisensor hierarchical estimation algorithm with feedback information [J]. Acata Electronic Sinica, 2000, 28(12): 85-89.)

[10] 邓自立. 最优滤波理论及其应用[M]. 哈尔滨: 哈尔滨工业大学出版社, 2000.
(Deng Z L. Optimal filtering and application [M]. Harbin: Publication of Harbin Industry University, 2000.)

[11] Petr Tichavsk, Carlos H Muravchik. Posterior Cramer-Rao bounds for discrete-time nonlinear filtering [J]. IEEE Trans on SP, 1998, 46(5): 1386-1396.