

基于高斯语音滤波的稳健文本无关说话人识别

周静芳, 陈一宁, 李科, 刘加
(清华大学电子工程系, 北京 100084)

摘要: 基于高斯模型, 提出一种去除实际电话录音中噪音、静音等非语音信号的新方法。与传统的语音检测器方法相比, 基于高斯语音滤波的方法在不同信道条件下都可以自动进行, 更好地保留了与说话人身份有关的信息。实验结果表明, 采用该方法的系统的等错误率比传统方法最多下降了21.2%。

关键词: 说话人识别; 高斯语音滤波; 高斯混合模型

Robust Text-independent Speaker Identification Based on Gaussian Speech-filter

ZHOU Jingfang, CHEN Yining, LI Ke, LIU Jia

(Department of Electronic Engineering, Tsinghua University, Beijing 100084)

【Abstract】 In this paper, a novel approach based on Gaussian model is proposed to remove silence and noise in real-life telephone recordings. Compared to the commonly used energy-based speech detector approach, the painful procedure to choose good thresholds in different channel conditions are avoided, and the useful frames containing information about the speaker identities are better retained. Experiment results show that this approach gives reasonably good performance.

【Key words】 Speaker identification; Gaussian speech-filter; GMM

说话人辨识的任务是, 给定一段语音, 在一些候选说话人中确定说话人的身份^[1]。说话人辨识可以分为开集识别和闭集识别。在后者情况下, 所有真实说话人都是候选人之一, 而前者则可能有候选人集合之外的说话人参与辨识。根据说话人是否需要根据系统提示发音, 说话人辨识又可以分为文本相关的和文本无关的识别。本文重点关注已知说话人集内人数和集外人数都相当大的开集、文本无关说话人辨识在实际电话录音环境中的应用。

基于高斯混合模型(Gaussian Mixture Model, GMM)^[2]的方法近年来在文本无关说话人识别中取得了成功的应用。然而和隐含马尔科夫模型(Hidden Markov Model, HMM)不同, 静音在GMM模型中并没有得到单独的建模, 因此会被看作是说话人声学特征的一部分记入最终的声学模型, 降低说话人身份识别的准确性。此外, 实际电话录音中常常出现的噪音也会大大影响识别性能。在许多情况下, 如采用一种基于能量的语音检测器^[3], 去除输入信号的20~30%以获得可用的语音。在实际应用中, 噪声情况往往是未定的, 因此准确的可用语音比例也很难事先确定。本文提出一种基于高斯模型的, 去除实际电话录音中噪音、静音等非语音信号的新方法, 称之为高斯语音滤波器。首先, 建立两个高斯模型代表语音和非语音信号, 用录音中最高能量的帧和最低能量的帧分别对其进行初始化, 然后用EM算法迭代训练直至模型收敛。语音帧到语音高斯模型的距离小于到非语音模型的距离, 因而可以从录音中与非语音自动分离, 实现稳健的说话人识别。

1 高斯语音滤波器

静音在GMM模型中并没有得到单独的建模, 会被看作是说话人声学特征的一部分记入最终的声学模型, 从而降低说话人身份识别的准确性。在文献[4]中指出, 爆破音、清音等低能量音的说话人区分能力比元音、鼻音等高能量音

弱。此外, 实际电话录音中常常出现的噪音也会大大影响识别性能。因此, 在说话人识别系统的前端处理中, 往往需要进行静音、低能量音和噪音的消除。

在许多情况下, 如采用一种基于能量的语音检测器^[3], 去除输入信号的20%~30%以获得可用的语音。然而在实际应用中, 噪声情况往往是未定的, 因此准确的可用语音比例也很难事先确定。而且单独采用能量这一个特征也不足以进行语音和非语音信号的区分。

高斯语音滤波器是本文提出的一种基于高斯模型的, 去除实际电话录音中噪音、静音等非语音信号的新方法。基于高斯语音滤波器的框图如图1所示。

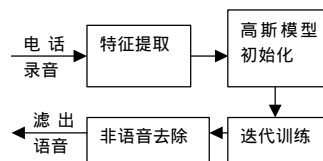


图1 基于高斯语音滤波器框图

首先, 从电话录音中提取每一帧信号的三维特征, 包括能量、过零率和FFT谱方差。

对一个D维特征矢量 X_i , 计算高斯似然度的公式定义如下:

$$p(X_i/I_i) = \frac{a_i(2p)^{-D/2}}{|\Sigma_i|^{1/2}} \exp\{-\frac{1}{2}(X_i - \mu_i)^T (\Sigma_i)^{-1} (X_i - \mu_i)\} \quad (1)$$

其中: μ_i 是高斯模型的均值矢量; Σ_i 是协方差阵; i

基金项目: 国家自然科学基金资助项目(60272016)

作者简介: 周静芳(1980-), 女, 硕士生, 主研方向: 说话人识别; 陈一宁, 博士生; 李科, 硕士生; 刘加, 教授

定稿日期: 2004-01-05 E-mail: zhoujif01@mails.tsinghua.edu.cn

是比例因子 λ_i ，用于平衡两个高斯模型之间的帧分布。一个完整的高斯模型记为

$$I_i = \{m_i, \Sigma_i, a_i\} \quad (2)$$

其次，用录音中能量最高的和最低的帧，分别对语音高斯模型 λ_s 和非语音高斯模型 λ_n 初始化：均值分别等于对应帧的特征向量，协方差阵置单位阵，比例因子置1/2。

接下来按照EM算法，用整段录音特征迭代训练 λ_s 和 λ_n ，直至收敛。比例因子的迭代公式：

$$\hat{a}_i = T_i / T \quad (3)$$

其中：T是录音特征的总帧数， T_i 是本次迭代中判决为归属模型 λ_i 的帧数。判决基于距离公式：

$$D_i = p(X_t / \lambda_s) - p(X_t / \lambda_n) \quad (4)$$

若 $D_i > 0$ ，判决该帧属于 λ_s ，否则属于 λ_n 。

最后，依据收敛的语音和非语音模型对每一帧录音进行判决，滤出纯净的语音帧。

这种方法和传统的语音检测器的方法的不同之处在于：一是利用信号与高斯模型之间的似然度来去除非语音信号，不再依赖于固定的语音-非语音比例值；二是在能量之外增加过零率和FFT谱方差两个区分性的特征，同时考虑到静音和不利于说话人身份识别的其他信号。在实际电话应用环境中，语音和非语音信号在录音中的比例值往往很难确定：某些电话录音的质量可能相当好，信噪比(SNR)超过30dB；而有些录音的信噪比却可能低于15dB，包含显著的噪声。这使得依赖于固定语音-非语音比例的传统语音检测器方法不能在实际环境中正常应用。

2 其他改进

由于本文关注的是说话人辨识系统在实际电话录音环境中的应用，下面讨论几点其他具体改进。

首先，在说话人辨识系统中引入双音多频信号(DTMF)的检测和去除。尽管电话录音中经常出现拨号音、忙音、回铃音等双音多频信号和电话信号音，并且明显对系统识别性能造成影响，它们却很少在说话人识别文献讨论中涉及。其原因可能在于这些信号通常出现在电话录音的前几秒，可以简单地通过抛弃前10s左右录音来处理。而这种方法并不总是有效，比如录音记录本身相当短，或者这些信号超过10s，或者10s后也可能出现这些信号。因此，在说话人识别系统中引入双音多频信号的自动检测处理是非常必要的。

根据国际电信联盟(ITU)的公用信道信令系统7号信令，一个DTMF信号是由一个低频率信号和一个高频率信号组合而成。具体频率设定如表1所示。

表1 DTMF信号频率组成

	1209Hz	1336Hz	1477Hz	1633Hz
697Hz	1	2	3	A
770Hz	4	5	6	B
852Hz	7	8	9	C
941Hz	*	0	#	D

对待处理电话信号 $x(n)$ ，先计算每一帧的FFT $\hat{X}(k)$ ，

然后得到幅度谱 $\left| \hat{X}(k) \right|$ 。如果该帧在DTMF各频率点或任一

电话信号音频率点上的能量超过该帧总能量的某一比例，则判决该帧是DTMF信号或者电话信号音，并从录音中去除。为了保证这一检测算法的稳定性，还必须在相邻帧之间做一定平滑处理。

另一点是关于本文采用的高斯混合模型——统一背景模型(GMM-UBM)^[3]中的背景模型UBM的改进。

在单人说话人识别检测系统中，通常假设实际说话人和目标说话人的性别一致，因此自然可以采用性别相关的UBM进行识别。然而对于开集说话人辨识系统而言，集内目标说话人和集外冒充说话人的性别往往不是单一确定的。为了能够在开集说话人辨识中采用识别性能更高的性别相关UBM，本文提出一种预分类的解决方法。首先，用男声和女声语料训练两个性别相关的UBM。这两个模型本身具有性别区分性，可以用来对待处理的语音进行预分类为男性或者女性语音。然后采用性别相关的UBM，在该性别的集内目标说话人中进行开集说话人辨识。

3 实验

3.1 实验数据库

本文实验采用的数据库是在实际电话对话中收集的，包含625个说话人(365名男性，260名女性)。已知说话人集合包括100个说话人(55男性，45女性)，每个说话人都有一段不少于2分钟的训练语音记录。集外冒充说话人数目为525(310个男性，215个女性)。大部分的话音记录来自于长途电话，少部分是手机语音。语音的信噪比分布在14dB到25dB之间，平均信噪比为19.6dB。

由于集内和集外的说话人数目都非常大，而且噪音水平也比较高，本文在实验中设置了一组不同的训练、识别段长条件。对每个集内说话人，采用2~4min录音训练其GMM模型。识别录音长度从20s到1min，并且不与训练语音重复。在所有的实验中，集外录音段总数都超过6 000次，以保证据识精度。

3.2 开集说话人辨识系统

本文所采用的识别系统在NIST Speaker Recognition Evaluation 1999评测库^[5]上取得了和文献[3]同样好的性能。并集说话人辨识系统框图如图2所示。

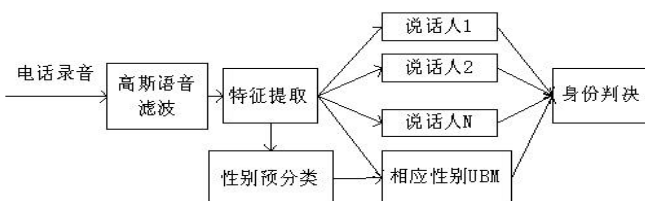


图2 开集说话人辨识系统框图

在DTMF信号、静音和噪音等非语音信号从电话录音中滤除之后，按照下列步骤进行特征提取：

- 分帧(10ms帧移，20ms帧长)；
- 加汉明窗；
- 在电话频带(300~3 400Hz)上提取19维Mel频标倒谱系数及其一阶差分。

根据性别预分类的结果，对应于相应性别的UBM模型，进行集内说话人GMM模型训练或者识别。识别时，计算录音段特征序列对相应UBM和集内说话人的似然度，做出拒绝或者接受的判决；如果判决为集内说话人，还需要提供具体的身份识别结果。

3.3 实验结果

3.3.1 采用语音检测器传统方法的系统

这个系统采用一种基于能量的语音检测器，去除输入信号的20%~30%以获得可用的语音。实验结果如表2所示。

表2 采用传统方法系统的辨识结果

训练/识别段长	去除比例	EER (%)	集内辨识正确率 (%)
2min/20s	20%	23.2	88.3
2min/40s	20%	17.6	91.7
4min/40s	25%	13.6	96.6
4min/1min	30%	13.2	96.4

3.3.2 采用高斯语音滤波器的系统

采用高斯语音滤波器系统的辨识结果如表3所示。

表3 采用高斯语音滤波器系统的辨识结果

训练/识别段长	EER (%)	集内辨识正确率 (%)
2min/20s	21.5	86.9
2min/40s	17.1	89.5
4min/40s	11.8	95.3
4min/1min	10.4	96.1

等错误率(Equal Error Rate, EER)是集外虚警错误概率和集内漏报错误概率相等时的系统错误率。集内辨识正确率指判决为集内说话人，并且正确辨识其身份的段占所有正确接受为集内说话人的段的比例。

从上面的实验结果可以看到，高斯语音滤波的方法要优于仅基于固定比例的传统语音检测器方法。在训练和识别段

长充分的情况下，系统的等错误率最多下降了21.2%。可能的解释是，在更好地保留与说话人身份有关的信息的同时，高斯语音滤波器趋向于去除更多一些录音，因此在段长充分的情况下性能改善更明显。

4 结论

本文提出了一种去除实际电话录音中噪音、静音等非语音信号的新方法，在不同信道条件下都可以自动进行，并且更好地保留了与说话人身份有关的信息。实验结果表明，采用这种方法的系统的等错误率比传统方法最多下降了21.2%。

参考文献

- 1 Doddington G R, Przybocki M A, Martin A V, et al. The NIST Speaker Recognition Evaluation — Overview, Methodology, Systems, Results, Perspective. *Speech Communication*, 2000, 31(2): 225-254
- 2 Reynolds D A, Rose R C. Robust Text-independent Speaker Identification Using Gaussian Mixture Models. *IEEE Trans. SAP*, 1995, 3(1): 72-83
- 3 Reynolds D A, Quatieri T F, Dunn R B. Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 2000, 10(1): 19-41
- 4 Petrovska-Delacretaz D, Cernocky J, Hennebert J, et al. Segmental Approaches for Automatic Speaker Verification. *Digital Signal Processing*, 2000, 10(1): 198-212
- 5 Martin A, Przybocki M. The NIST 1999 Speaker Recognition Evaluation — An overview. *Digital Signal Processing*, 2000, 10(1): 1-18

(上接第124页)

(3)后台真实节点的负载状况(截取一部分),见表1。在测试这一组数据时，主要访问的文件是静态文件，主要的负载变化是网络负载的变化，而对于网络负载，可以看出利用本文实现的方法负载均衡的比较好，基本没有出现负载很不均衡的情况，出现的差异基本上最大为单个请求引起的负载。

表1 本文实现的后台负载数据

		eth	(kbps)	
bt1	158.33	164.67	0.00	225.00
bt2	144.00	190.33	128.33	187.67
bt3	140.00	0.00	169.00	134.67
bt4	146.00	0.00	158.00	137.00
Bt5	172.67	0.67	152.67	179.67
Bt6	0.33	170.00	129.00	130.67
Bt7	127.00	179.00	204.67	126.33

4 小结

测试表明本文提出的算法是可行的，但是，算法中有的参数(如权值调整时间、内容分类方法等)需要优化才能得到更好的结果，本文今后的工作是对算法的相关参数进行优化，得到一个优化的参数配置，并且希望今后将使用遗传算法的方式来进行控制，希望得到更好的性能。

参考文献

- 1 Colajanni M, Yu P S. A Performance Study of Robust Load Sharing Strategies for Distributed Heterogeneous Web Server Systems. *IEEE Transactions on Knowledge and Data Engineering*, 2002-03/04, 14: 398

- 2 Cardellini V, Colajanni M, Yu P S. Dynamic Load Balancing on Web-server Systems. *Internet Computing*, IEEE, 1999-05/06, 3: 28-39
- 3 Colajanni M, Yu P S, Dias D M. Analysis of Task Assignment Policies in Scalable Distributed Web-server Systems. *IEEE Transactions on Parallel and Distributed Systems*, 1998-06, 9: 585-600
- 4 Rumsewicz M, Dwyer M. Preferential Load Balancing for Distributed Internet Servers. *First IEEE/ACM International Symposium on Cluster Computing and the Grid*, 2001 Proceedings, Brisbane, Qld., Australia, 2001: 363-370
- 5 Zhang Jian, Hamalainen T, Joutsensalo J. QoS-aware Load Balancing Algorithm for Globally Distributed Web Systems. *2001 International Conferences on Info-tech and Info-net*, 2001 Proceedings, ICII 2001 - Beijing, 2001, 2: 60-65.
- 6 Zomaya A Y, Yee-Hwei Teh. Observations on Using Genetic Algorithms for Dynamic Load-balancing. *IEEE Transactions on Parallel and Distributed Systems*, 2001-09, 12: 899-911
- 7 Kostin A E, Aybay I, Oz G. A Randomized Contention-based Load-balancing Protocol for a Distributed Multiserver Queuing System. *IEEE Transactions on Parallel and Distributed Systems*, 2000-12, 11: 1252
- 8 Tan Ling, Zahir Tari. Dynamic Task Assignment in Server Farms: Better Performance by Task grouping. *2002 Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications*, 2002: 175-180
- 9 Buttazzo G C, Lipari G, Caccamo, et al. Elastic Scheduling for Flexible workload Management. *IEEE Transactions on Computers*, 2002-03, 51: 289-302
- 10 章文嵩. Linux服务器集群系统(四). <http://www-900.ibm.com/developerWorks/cn/linux/cluster/lvs/part4/index.shtml>