文章编号:1671-6833(2009)03-0141-04

基于归一化自相关语音筛选的鲁棒性说话人确认研究

王 珊,谷 城

(河南大学 物理与电子学院,河南 开封 475000)

摘 要:语音在受到加性噪声污染时,不同部分受噪声污染的程度存在差异,保留受污染严重的语音将会对说话人确认系统产生负面影响,笔者提出了一种基于归一化自相关的鲁棒性语音筛选方法.通过归一化自相关舍弃受污染严重的语音.实验表明,通过归一化自相关舍掉受污染严重的语音,能提高噪声环境下的说话人确认性能.

关键词:归一化自相关;语音筛选;鲁棒性;说话人确认

中图分类号:TP 206 **文献标识码**:A

0 引言

噪声一直是制约说话人确认性能的重要因素.目前,与文本无关的说话人确认系统大多采用基于短时频谱特征(如MFCC,LPCC参数)的GMM-UBM系统.文献[1]指出,基于短时频谱特征的系统性能随着噪声的污染程度或环境的失配都将迅速下降.因此,提高短时频谱特征系统的鲁棒性一直是一个重要的问题.近年来,国内外很多研究机构也提出了很多的方法,如CMS,RASTA,特征补偿等,减少了参数受噪声污染和环境失配的影响,不同程度地提高了短时频谱参数系统的鲁棒性.然而,这些做法基本都是针对所有的特征,而忽略这些特征参数受噪声影响的大小.而对于受噪声污染严重的参数即使经过各种处理,提供的信息也很难有效利用;而对于与文本无关的说话人确认来说,应舍弃这些语音.笔者根据语音信号的时间相关性,和这种相关性同时随噪声变化的规律,提出了一种基于归一化自相关的语音筛选算法.在MSRA数据库与NIST2006年8side-1side任务上的实验表明,该方法能筛选出受噪声污染小的语音,相对于使用所有语音的基准系统,提高了系统的鲁棒性.

1 归一化自相关

语音是激励源通过声道产生的有序序列,语音采样点前后有一定的相关性,用归一化自相关R来表征语音的相关程度.记$y(n)(n=1,2\cdots N$,N为帧长)是一帧语音,设$y(n)=x(n)+w(n)$,$x(n)$为干净语音,$w(n)$为加性噪声,语音与噪声不相关.

$$r_y(k)=\frac{\left|\frac{1}{N-k}\sum_{n=1}^{N-k}y(n)y(n+k)\right|}{\frac{1}{N-k}\cdot\sqrt{\sum_{n=1}^{N-k}y(n)^2\sum_{n=1}^{N-k}y(n+k)^2}}$$

$$k=20\cdots160 \tag{1}$$

$$R=\sqrt{\max(r_y(k))}=\sqrt{r_y(K)} \tag{2}$$

而$\left|\frac{1}{N-k}\sum_{n=1}^{N-k}y(n)y(n+k)\right|=$

$$\left|\frac{1}{N-k}\sum_{n=1}^{N-k}[x(n)x(n+k)+w(n)x(n+k)+x(n)w(n+k)+w(n)w(n+k)]\right|$$

$$=R_x(k)+R_w(k) \tag{3}$$

$$R_x(k)=\frac{1}{N-k}\sum_{n=1}^{N-k}x(n)x(n+k) \tag{4}$$

$$R_w(k)=\frac{1}{N-k}\sum_{n=1}^{N-k}w(n)w(n+k) \tag{5}$$

由(3-5)式得

$$r_y(k)=\frac{|R_x(k)+R_w(k)|}{R_x(0)+R_w(0)} \tag{6}$$

对于加性白噪声,

$$R=\sqrt{r_y(K)}=\sqrt{\frac{|R_x(K)|}{R_x(0)(1+R_w(0)/R_x(0))}} \tag{7}$$

收稿日期:2008-01-14;修订日期:2008-03-13

作者简介:王 珊(1981-),女,河南开封人,河南大学助教,硕士,主要从事电子电路系统及计算机网络方面的研究.

从式7可以看出,对于加性白噪声,随着信噪比降低,R将变小.

图1为归一化自相关高的一帧语音(采样率16 k,每帧20 ms)与归一化自相关低的一帧语音(选自微软语音库)的归一化自相关值随高斯白噪声的变化情况.从图中可以看出,不管是归一化自相关高的语音还是归一化自相关低的语音,随着信噪比的降低(语音受污染程度越来越严重),其归一化自相关值都呈现下降趋势.图2是该条语音加高斯白噪声后的的实际信噪比与归一化自相关分布图,表1为相应的各归一化自相关范围内的信噪比统计表,从表1可看出,随着归一化自相关的提高,相应区间的信噪比也明显提高,即归一化自相关与信噪比有很大的相关性.故可以采取依据归一化自相关值来筛选出信噪比较高的语音,舍弃受污染较严重的语音,从而提高说话人确认的性能.

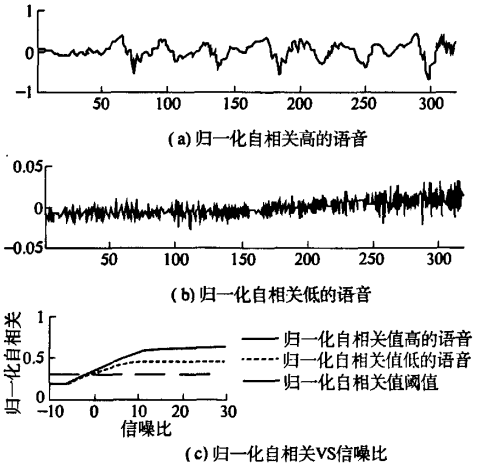

图1 高度自相关语音序列

Fig.1 Highly autocorrelation voice series

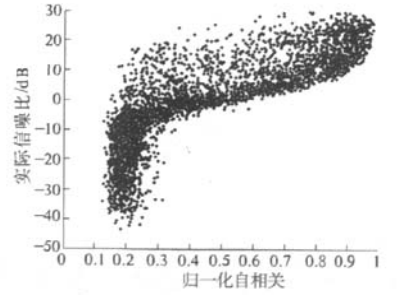

图2 归一化-信噪比分布图

Fig.2 Normalization - SNR diagram

2.2 基于归一化自相关的语音筛选

图3是微软数据库的一条语音(加10 dB高斯白噪声)采用归一化自相关的语音筛选结果.

表1 归一化-信噪比表

Tab.1 Normalization - SNR diagram

归一化自相关范围	0.1-0.2	0.2-0.3	0.3-04	0.4-0.5
平均信噪比/dB	-17.22	-10.46	0.01	2.76
0.5-0.6	0.6-0.7	0.7-0.8	0.8-0.9	0.9-1
4.90	8.12	11.77	15.26	19.87

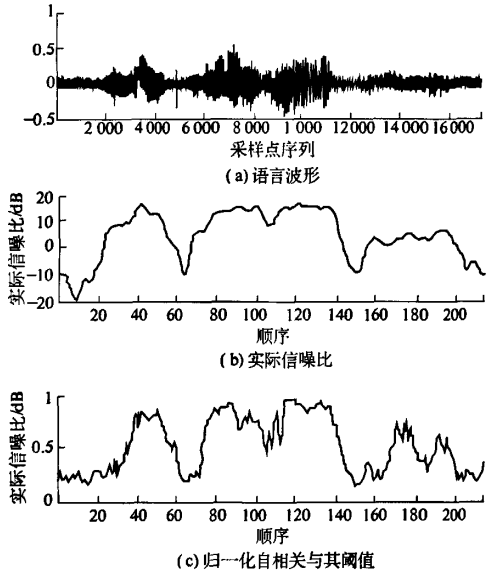

图3 采用归一化后的图像

Fig.3 figures after normalized

从图3可以看出,本算法筛选出的各帧语音的信噪比较高,舍弃了一部分受污染较严重的语音,这有可能有利于对说话人的确认;如果阈值选择过高,将会舍弃部分信噪比较高的语音,导致可供利用的数据不足,对说话人确认带来不利影响.因此要在信噪比高低与数据量之间做一个折中,定一个合适的归一化自相关筛选阈值.

2 GMM-UBM话者确认系统

全局背景模型UBM(Universal Background Model)是一个大型的GMM.在话者确认系统中,用不同说话人在各种环境下的语音信号数据集训练一个高混合度的GMM来获得,这类数据集越大越好.而各个说话人的GMM模型则是由UBM自适应而来.

如图1所示:训练时,训练语音先经过特征提取,然后利用已经训练好的UBM,通过最大后验概率算法(Maximum a Posteriori)自适应得到该说话人的高斯混合模型;识别时,测试语音同样经

过特征提取,然后分别与说话人模型和 UBM 模型匹配,设 $\boldsymbol{X}=\{\vec{x}_1,\vec{x}_2,\cdots,\vec{x}_T,\}$ 为某条测试语音所提取出的特征向量序列,该条语音被分成了 T 帧,对应着 T 个特征向量,X 相对于说话人模型 λ_i 和 UBM 模型 λ_o 的对数似然度 P_i 和 P_o 的计算方法分别如公式(2)、(3) 所示,两者相减得到最后的输出评分,将评分跟阈值相比较得到确认结果.

$$p_i = \log p(X \mid \lambda_i) = \log \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_i) \tag{8}$$

$$p_o = \log p(X \mid \lambda_o) = \log \prod_{t=1}^{T} p(\vec{x}_t \mid \lambda_o) \tag{9}$$

$$p = p_i - p_o = \log \frac{p(X \mid \lambda_i)}{p(X \mid \lambda_o)} \tag{10}$$

UBM 代表的是各种说话人的平均信息,自适应的方法突出了特定人与公共背景的差别,尤其是对于电话语音,噪声比较复杂,通过评分相减后可以部分抵消噪声的干扰,提高系统的鲁棒性;另外一个优点就是可以用比较少的语音数据训练出混合度较大的模型. 目前最常用的是最大后验概率算法.

3　实验和结果分析

作者采用微软亚洲研究院提供的 100 个男性话者的麦克风干净语音数据库(简称微软数据库)和 NIST2006 年 8side-1side 的电话语音数据库,对笔者提出的基于归一化自相关的语音筛选策略进行验证.

系统性能的评估采用等误识率 EER(Eaqual Error Rate)来衡量. 对于说话人确认,需要定一个阈值 T,对于每一个特定阈值,都有 2 错误率,系统的错误接受率(系统判断为目标话者的冒认者语音数量占总的冒认者语音数量的比率)与错误拒绝率(系统判断为冒认者的目标话者语音数量占总的目标话者测试语音数量的比率). 两类错误率相等的错误率就是 EER.

3.1　微软数据库

3.1.1　数据库描述

数据库共包括 100 个男性说话人,每个人约有 200 条语音,每条语音的长度约为 6 ~ 7 s,采样频率为 16 kHz,16 bit 量化. 本实验中,挑选了每个人的 15 条语音用来训练 UBM,15 条用来做训练 GMM 模型,6 条进行测试,共计 100 × 100 × 6 次测试.

3.1.2　前端处理

语音首先加入一定信噪比的高斯白噪声,然后预加重,再通过 Hamming 窗分帧处理,帧长 20 ms,帧移 10 ms,每帧提取 16 阶 MFCC 参数(不包括第 0 阶),并在此基础上计算其一阶动态参数,计 32 维.

4.1.3 实验结果

实验中 UBM 混合度采用 256,实验结果如表 2 和图 5 所示. 可以看出,在语音信噪比较高时(20 dB 下),随着归一化自相关阈值的提高,经归一化自相关筛选后 EER 会下降,这主要是因为语音比较干净,随着阈值的提高,越来越多的归一化自相关低的较干净语音被筛去,保留的语音越来越少,数据的减少使系统的性能下降. 而随着信噪比的降低(10 dB 与 0 dB),系统性能(EER)都会在合适的归一化自相关阈值处有个最优值,说明虽然语音减少了,但由于筛去的多是归一化自相关较低的语音,而这部分又比较容易受噪声污染,系统性能反而提高;随着语音数量的进一步减少,数据不足的缺陷将进一步凸显,系统性能将会下降.

表 2　不同信噪比和归一化自相关阈值下的 EER

Tab. 2　The EER of different SNR and normalization

	阈值	所有语音	0.3	0.4
EER/%	20 dB	1.00	1.09	1.17
	10 dB	2.50	2.17	2.30
	0 dB	10.83	5.02	4.76
	0.5	0.6	0.7	0.8
EER/%	1.12	1.29	1.33	2.00
	2.50	2.47	2.99	4.53
	6.77	10.16	17.62	30.16

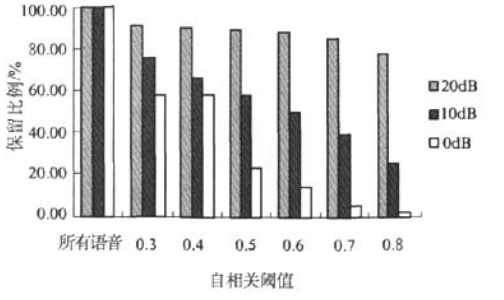

图 5　不同信噪比与自相关阈值下保留语音比例

Fig. 5　Remain voice proportion of different SNR and autocorrelation

3.2　NIST'06 数据库上实验

3.2.1　数据库说明

选取 8side - 1side 任务中的一部分数据. 依据 NIST 说话人评测任务 8side - 1side 男性的测试列表,将包含了 126 个人的训练语音及相关的 808 条语音分布作为我们的训练和测试语音. 每个人有 8 条 5 min 的训练语音(包含静音),测试

语音也是5 min(含静音),测试为开集测试,同一条语音可能供多个说话人测试,计5 000次测试. UBM数据为NIST'05 8side-1side任务中的240条男性话者的语音.

3.2.2 参数处理

语音先经过预加重,通过归一化自相关切分出归一化自相关高的语音,对这些语音通过Hamming窗分帧处理,帧长和帧移分别是20 ms和10 ms.对于每一帧,提取16阶的MFCC参数(不包括第0阶)及其一阶动态衍生参数共32维特征参数,并针对NIST电话语音的特点对MFCC应用RASTA、CMS进行电话通道影响的补偿.

3.2.3 实验结果

实验中UBM与GMM的混合度是1 024.从图6与表3可以看出,由于NIST电话语音并非干净语音,语音受到不同程度的污染,通过设定合适的归一化自相关阈值,可以切掉受污染比较严重的语音;而同时,若阈值过大,则会导致语音的过分减少,对于实验采用的统计模型GMM来说是不利的.因此,系统性能随着归一化自相关阈值的增大,先提升而后随着数据量的减少又下降.

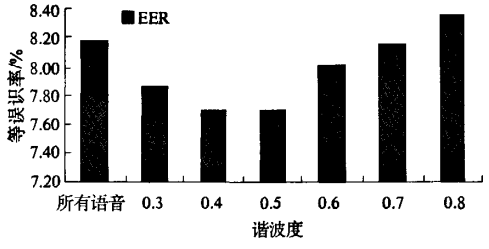

图6 等误识率图

Fig.6 EER diagram

表3 不同归一化自相关阈值下筛选出的语音比例

Tab.3 Selected voice proportion of different normalization

归一化自相关阈值	0.3	0.4	0.5	0.6	0.7	0.8
保留语音比例/%	99.31	93.45	84.83	76.90	68.72	57.24

4 结论

根据语音的归一化自相关随信噪比的变化特性,提出了一种基于归一化自相关的语音筛选策略,并在微软数据库和NIST'2006数据库上做了验证实验,实验表明,选择合适的归一化自相关阈值筛选语音可以提高系统性能.而目前归一化自相关阈值都是人工设定的,下一步将研究在未知数据库上如何设定合适的归一化自相关阈值以提高系统性能.

参考文献:

[1] REYNOLDS, D. A. QUATIERI, T. F DUNN. R. B. Speaker verification using adapted gaussian mixture models[J]. Digital Signal Processing, 2000, 10: 19-41.

[2] Back, G. Hsieh, W. C. J. Lepreau. Processes in KaffeOS: isolation, resource management, and sharing in Java. In Proceedings of the Fourth Symposium on Operating Systems Design and Implementation [C]. USENIX Association, San Diego, CA, 2000, 333 - 346.

[3] ATANAS O. Robust features and neural network for Noisy Speech Detection[J]. Cybernetics and Information Technologies, 2006, 6(3): 2-10.

Confirmation Researching on Robustness Speaker based-on Normalized Autocorrelation Voice Selection

WANG Shan, GU Cheng

(School of Physics and Electronics, Henan University, Kaifeng 475000 China)

Abstract: Aiming at the situation that the voice was vitiated by positive noise and the difference of pollution extent between different parts of the voice, preserving the polluted parts of voice will bring in negative affection to the confirmation system of speaker. This paper gives a method that based on normalized autocorrelation robustness selection. Normalized autocorrelation not only shows the correlation of short - time voice frame, but also shows the polluted extent of voice. As a result, we can wipe off the deeply polluted voice through the normalized autocorrelation. Experiments showing, through the normalized autocorrelation wiping off the polluted voice, can improve confirmation ability of speaker in high noise environment.

Key words: normalized autocorrelation; voice selection; robustness; speaker confirmation