基于自适应直方图均衡化的鲁棒性说话人辨认研究

徐利敏 [1,2] 唐振民 [2] 何可可 [2] 钱 博 [2]

摘 要 在噪声环境下, 为提高说话人识别系统的鲁棒性, 需要对系统进行各种抗噪声处理. 本文基于说话人特征的统计特性和直方图均衡化在说话人识别中的应用特点, 提出了直方图均衡化的自适应方法. 实验结果表明, 与普通直方图均衡化变换方法相比, 自适应直方图均衡化能进一步提高辨认系统的辨认率; 并且无论在平稳噪声还是非平稳噪声环境下, 该算法都能取得较好辨认率, 进一步增强系统的鲁棒性.

关键词 说话人识别, 直方图均衡化, 高斯混合模型, 鲁棒性说话人辨认
中图分类号 TP391.42

Research on Robust Speaker Identification Based on Adaptive Histogram Equalization

XU Li-Min[1,2] TANG Zhen-Min[2] HE Ke-Ke[2] QIAN Bo[2]

Abstract Diversified methods of decreasing the influence of noise have appeared to improve the performance of speaker recognition system in noise. In the paper, based on the statistical characteristics of speaker feature and the particularity of histogram equalization applying to speaker recognition, an adaptive histogram equalization method for speaker recognition is presented. The experiments showed improvements in performance with the proposed method in comparison with ordinary histogram equalization, and that the robustness of the system could be improved with the method under different noise environments.

Key words Speaker recognition, histogram equalization, Gaussian mixed model, robust speaker identification

说话人识别是指通过对说话人语音信号的分析处理, 自动确认说话人是否在所记录的话者集合中, 以及进一步确认说话人是谁. 在理想条件下, 比如安静的录音环境、高质量的录音设备以及训练和测试环境相匹配, 说话人识别已经可以达到令人满意的识别结果. 然而由于说话人的个性特征具有长时变动性, 而且其发音常常与环境背景噪声等干扰、说话人情绪、说话人健康状况等有密切关系[1−2], 这些都使得说话人识别的识别率大幅度下降. 这主要是因为训练和测试语音之间的声学失配造成的, 从统计的观点上看就是训练和测试语音特征所服从的概率分布不一致造成的.

通常的抗噪声方法主要可以分为三种: 前端处理、特征值处理以及模型补偿. 前端处理的目标是消除测试语音中噪声的影响, 所有操作基本上都是针对原始语音波形进行的, 和以后的特征提取及模型匹配没有直接联系. 针对白噪声, Pandey[3] 采用谱减法降低噪声的影响, 但对于有些噪声干扰, 谱减法不仅不能降低或消除其影响, 而且还带来了严重的 MUSIC 噪声[4]. Tadj[5] 等提出了自适应噪声抵消技术来降低噪声的影响. Soon[6] 采用二维傅氏变换和 wiener 滤波对含噪语音进行降噪处理.

特征值的抗噪声处理主要集中在寻找稳健性的特征和对含噪语音产生的特征进行处理. 特征加权算法[7] 就是通过把由噪声引起的使含噪语音信号特征值与纯净语音特征值的偏差部分去除, 从而使进入识别器的特征值接近纯净语音的特征值. 虽然特征加权算法在平稳噪声等情况下取得良好的结果, 但在非平稳噪声环境下有其一定的局限性[7−8].

模型补偿的方法属于后端处理, 当说话人的个性特征不断变化、语音与噪声不能很好地分离或者降噪算法对语音有损伤、模型不能很好地匹配时, 需要对似然概率进行补偿, 例如归一化补偿变换[9] 和基于最小风险的得分判决方法[10].

直方图均衡化 (Histogram equalization, HEQ)[11−12] 属于特征值处理抗噪声方法的类型, 该方法最初是数字图像处理中增强图像整体对比度的一种技术[13]. 近几年来不少学者将其成功地应用到语音处理上[14−17], 比如, Torre 等[18] 将其应用到语音识别上以提高系统鲁棒性, Skosan 等[19] 提出了修正的分段直方图均衡化方法来提高说话人确认

收稿日期 2007-01-17 收修改稿日期 2007-08-18
Received January 17, 2007; in revised form August 18, 2007
1. 南京财经大学电子商务重点实验室 南京 210003 2. 南京理工大学计算机科学与技术学院 南京 210094
1. Key Laboratory of Electronic Business, Nanjing University of Finance and Economics, Nanjing 210003 2. School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094
DOI: 10.3724/SP.J.1004.2008.00752

系统在电话环境中的鲁棒性, 都取得较好的识别效果.

虽然直方图均衡化方法近年来已被广泛地应用, 但仍然有许多需要改善的地方, 例如查表式直方图均衡化 (Table look-up based histogram equalization, THEQ)[20] 需要将庞大的表格信息加载到内存中才能进行转换匹配动作, 而且若要有良好的补偿效果, 表格所记录的点数不能太少, 但当表格记录点数增加时, 需耗费更大量的内存空间与进行查表转换的处理器运算时间; 又如分位差统计图等化法 (Quantile-based histogram equalization, QHEQ)[21−22], 虽然转换过程不需通过查表动作, 只需使用少量的参数即可进行等化动作, 但是对每一句待转换的语句在进行转换动作前, 必须利用格式搜寻以在线实时运算求取参数, 因此所需的处理器运算时间也是相当可观的.

本文基于说话人特征的统计特性和直方图均衡化在说话人识别中的应用特点, 提出了应用于说话人辨认中的自适应直方图均衡化方法, 该方法使得变换后的特征更符合实际特征的分布, 进一步提高了噪声环境下说话人识别系统的识别率和鲁棒性.

1　直方图均衡化

直方图均衡化技术起初是在数字图像处理中提出的, 是一种采用压缩原始图像中像素数较少的部分, 拉伸像素数较多的部分, 从而使整个图像的对比度增强、图像变清晰的方法. 实际上, 直方图均衡化就是一个样本的非线性变换, 目的是使得变换后的样本服从我们所需要的参考分布[18]. 假设原样本矢量为 $\boldsymbol{x}_0$, 其样本的概率密度函数为 $p_0(\boldsymbol{x}_0)$, 参考概率密度函数为 $p_{ref}(\boldsymbol{x}_0)$; 变换后的矢量为 $\boldsymbol{x}_1$, 其概率密度函数为 $p_1(\boldsymbol{x}_1) = p_{ref}(\boldsymbol{x}_1)$, 其变换记为 $\boldsymbol{x}_1 = F(\boldsymbol{x}_0)$. 因此直方图变换可以看成将原矢量的直方图变换到参考的直方图, 以达到将原矢量 $\boldsymbol{x}_0$ 变换到目标矢量 $\boldsymbol{x}_1$ 的过程.

根据直方图的定义, 经过变换后的小面积元应相等, 即

$$p_{ref}(\boldsymbol{x}_1)\mathrm{d}\boldsymbol{x}_1 = p_0(\boldsymbol{x}_0)\mathrm{d}\boldsymbol{x}_0 \tag{1}$$

假定 $G(\boldsymbol{x}_1) = \boldsymbol{x}_0$ 是 $\boldsymbol{x}_1 = F(\boldsymbol{x}_0)$ 的反函数, 那么参考概率密度函数可以写成

$$p_{ref}(\boldsymbol{x}_1) = p_0(\boldsymbol{x}_0)\frac{\mathrm{d}\boldsymbol{x}_0}{\mathrm{d}\boldsymbol{x}_1} = p_0(G(\boldsymbol{x}_1))\frac{\mathrm{d}G(\boldsymbol{x}_1)}{\mathrm{d}\boldsymbol{x}_1} \tag{2}$$

又根据概率统计的定义, 样本的概率密度函数为 $p_0(\boldsymbol{x}_0)$ 和参考概率密度函数为 $p_{ref}(\boldsymbol{x}_1)$ 的分布函数可以分别写为

$$C_0(\boldsymbol{x}_0) = \int_{-\infty}^{\boldsymbol{x}_0} p_0(\boldsymbol{x}_0')\mathrm{d}\boldsymbol{x}_0' \tag{3}$$

$$C_{ref}(\boldsymbol{x}_1) = \int_{-\infty}^{\boldsymbol{x}_1} p_{ref}(\boldsymbol{x}_1')\mathrm{d}\boldsymbol{x}_1' \tag{4}$$

因此根据式 (2) ∼ (4), 可以得到原分布函数和参考分布函数之间的关系为

$$\begin{aligned} C_0(\boldsymbol{x}_0) = &\int_{-\infty}^{\boldsymbol{x}_0} p_0(\boldsymbol{x}_0')\mathrm{d}\boldsymbol{x}_0' = \\ &\int_{-\infty}^{\boldsymbol{x}_1} p_0(G(\boldsymbol{x}_1)')\frac{\mathrm{d}(G(\boldsymbol{x}_1)')}{\mathrm{d}\boldsymbol{x}_1'}\mathrm{d}\boldsymbol{x}_1' = \\ &\int_{-\infty}^{\boldsymbol{x}_1} p_{ref}(\boldsymbol{x}_1')\mathrm{d}\boldsymbol{x}_1' = \\ &C_{ref}(\boldsymbol{x}_1) = C_{ref}(F(\boldsymbol{x}_0)) \end{aligned} \tag{5}$$

从式 (5) 可以得到将原样本空间变换到参考分布空间的变换函数为

$$F(\boldsymbol{x}_0) = C_{ref}^{-1}(C_0(\boldsymbol{x}_0)) \tag{6}$$

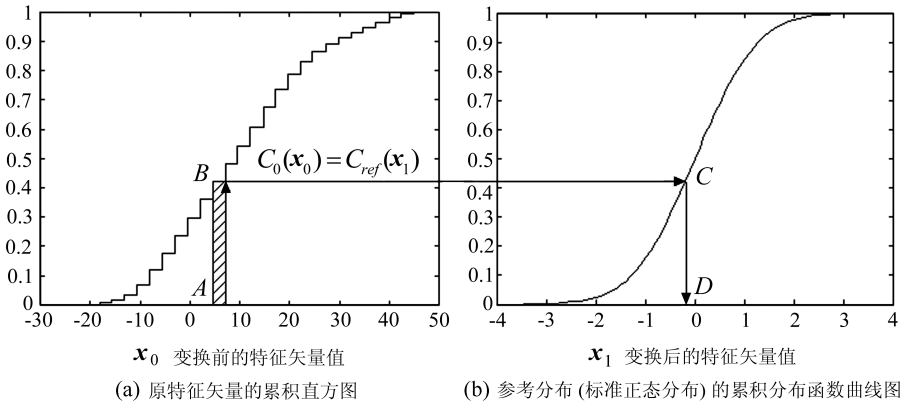
其中, C_{ref}^{-1} 是参考概率分布函数的反函数.

值得注意的是, 在实际应用中观察值的个数是有限的, 因此在上文中所提到的概率密度函数其实是样本的直方图, 相应的分布函数为直方图的累积函数.

事实上, 直方图均衡化应用到语音及说话人识别技术中可以看成是倒谱均值归一化 (Cepstral mean normalization, CMN)[23] 和均值方差归一化 (Mean and variance normalization, MVN)[24] 的延伸和扩展[19]. 均值归一化是对特征分布的一阶矩 (均值) 进行标准化; 均值方差归一化通过对分布的一阶矩 (均值) 和二阶矩 (方差) 的归一化来提高特征的抗噪能力. 这两种方法都是基于线性变换的技术, 而实际的噪声大多引起特征的非线性失真, 因此限制了其提高系统鲁棒性的能力. 直方图均衡化方法是一种非线性的补偿变换, 它不仅仅对特征分布的一阶和二阶矩进行归一化, 而是对所有阶矩都进行归一化, 使得训练和测试的语音特征之间的失配程度降低, 从而提高系统的识别性能.

2　改进的自适应直方图均衡化

直方图均衡化方法用于图像处理中, 由于灰度级是一定的, 因此计算直方图非常简便; 然而说话人识别中的特征矢量是多维的, 而且其值也是随机的, 所以不能直接应用. 为了简化模型, 通常假定说话人特征矢量各维分量相互独立, 由此我们可以在特征的每一维分量上独立进行直方图的非线性变换[19].



(a) 原特征矢量的累积直方图 (b) 参考分布 (标准正态分布) 的累积分布函数曲线图

图 1 基于直方图均衡化的累积分布曲线变换图

Fig. 1 Transformation chart of cumulative distributing curve based on histogram equalization

在每一维特征分量上构造直方图时, 通常的做法是将特征值的值域分成 M 个大小相等且不交叠的区间, 然后计算特征值落在各区间的频率, 这样生成的直方图累积函数就是一个区间大小相等的分段单调递增函数, 如图 1 中左边曲线所示. 参考分布在本文中取标准正态分布, 其累积分布函数曲线图如图 1 中右图所示. 那么其特征矢量值的变换过程如图中 A、B、C 和 D 所示: 对于原特征值 (A 点), 在原累积直方图上找到其对应的频率值 (B 点); 然后根据式 (5) 在参考标准正态分布的累积分布函数上找到相同的概率点 (C 点); 最后该点对应的横坐标值 (D 点) 即为变换后的特征矢量值.

从原特征矢量的累积直方图曲线中我们可以看出, 变换前的特征值值域被分割成大小相等的区间, 其对应的频率增量在曲线的中部普遍比两侧大些. 换言之, 特征值在越靠近其均值处, 其密集程度越高; 离均值处越远特征值越稀疏. 那么就出现这样的问题: 在靠近均值处的特征值区间 (如图 1 中阴影部分) 内聚集了大量的特征值样本 (对于 2 000 个特征矢量样本集大约有 150 个样本落在该区间); 而在远离特征值均值的两端却只有少量甚至没有特征值落在大小相等的区间上. 对于后者, 少量的特征值通过直方图均衡化变换到同一值上是合理的; 而对于前者, 大量属于同一区间的特征值变换到同一值使生成的特征集在一定程度上背离了实际分布. 解决此问题最简单的方法是增加区间个数. 然而这种方法虽然减少了靠近均值处区间内的特征值样本个数, 但是也在远离均值处划分了无需划分的区间, 当区间个数达到一定程度时甚至出现了很多零频率增量的区间, 这不仅增加了变换的计算量, 而且浪费了资源.

为此, 本文提出了一种自适应确定区间大小的直方图均衡化方法. 该方法首先用较大的区间长度来构造直方图的累积函数, 然后根据各区间内特征值频率增量的大小来自适应确定该区间是否需要再划分以及划分的程度. 并且在同一区间内并不把区间内所有的原特征值变换到参考分布的同一值上, 而是计算当前区间对应的变换特征值和前一区间对应的值, 然后根据原区间内特征值出现的个数对变换特征值区间进行线性划分, 变换时将划分后大小依次排序的各值代替相应排序位置的原特征值. 采用这种方法不仅使计算量降低, 而且得到的变换特征值的分布更符合实际特征空间. 将其应用到每个说话人的训练和测试特征矢量分布变换时, 步骤如下:

1) 选定合适的参考分布 (文中为标准正态分布), 其概率密度函数为 $P_{ref}(\boldsymbol{x}_1)$, 累积分布函数为 $C_{ref}(\boldsymbol{x}_1)$.

2) 对特定的每一维特征值序列找到其最大值和最小值, 分别为 $x_{0\max}$ 和 $x_{0\min}$.

3) 将特征值值域预划分成 M 个大小相等且无交叠的区间 $[x_{0\min}, x_{0\max}]$, 那么 $x_{0\min}=b_1<\cdots<b_{M+1}=x_{0\max}$. 假定特征值序列长度为 N_f, 通常 M 可以取为 $N_f/50$.

4) 在第 3) 步中设置的每一个区间上计算序列的特征值落入该区间的个数 $n_i\,(i=1,\cdots,M)$.

5) 设置区间内特征值个数最大域值 n_{th}. 当区间内特征值个数小于等于该域值时, 该区间不作处理; 反之对该区间再划分处理. 若当前区间内特征值个数为 $n_i(>n_{th})$, 那么对该区间作 $\lfloor n_i/n_{th}\rfloor$ 等分. 其中, 符号 $\lfloor\cdot\rfloor$ 表示下取整运算. 由此将生成新的区间序列 $B_j\,(j=1,\cdots,N)$ 和对应的特征值个数序列 $n_j\,(j=1,\cdots,N)$, 其中 $N\geq M$.

6) 根据新的区间序列和相应的特征值个数序列, 计算当前维特征值分布的累积直方图, 计算公式如下

$$C_0(\boldsymbol{x}_0 : \boldsymbol{x}_0 \in B_j) = \sum_{k=1}^{j} \frac{n_k}{N_f} \tag{7}$$

7) 根据式 (5) 对每一区间点 $b_j(j = 1, \cdots, N+1)$ 找到其变换后在参考分布上的新特征值 $c_j(j = 1, \cdots, N+1)$, 生成了变换后的区间序列 $C_j = [c_j, c_{j+1})$; 然后对 C_j 进行 n_j 次线性划分, 得到 $c_j = c_j^1 < \cdots < c_j^k < \cdots < c_j^{n_j+1} = c_{j+1}$. 同时对相应原区间内的 n_j 个特征值进行排序, 得到: $b_j \le x_{0,j}^1 \le \cdots \le x_{0,j}^k \cdots \le x_{0,j}^{n_j} < b_{j+1}$, 最后分别将 $x_{0,j}^k$ 变换到参考空间中的 c_i^k, 实现整个特征值分布到参考分布的变换.

自适应直方图均衡化 (Adaptive histogram equalization, AHEQ) 主要改变了传统方法构造累积直方图时区间大小完全相等的缺点. 采用这种自适应的方法可以使分割的区间大小随着特征值样本聚集程度的变化而变化, 并且对同一区间内的特征值按照其大小线性地映射到参考空间内对应的区间上, 因此在得到相同性能的条件下, 使变换过程的计算量降低, 得到的变换特征值的分布更符合实际, 即在相同区间个数条件下, AHEQ 比 HEQ 对因噪声引起的失真的补偿性能更加有效.

3　鲁棒说话人辨认实验的建立

自适应直方图均衡化方法将在 CST603 语音库上进行验证, 该语音库录制的是纯净的语音数据. 语音信号采样频率为 22 050 Hz, 单声道录音. 实验中使用的语音数据包括 60 个说话人 (28 个女性, 32 个男性), 其中所有说话人发音都是汉语普通话, 每个说话人录音 3 次, 得到 3 个文件, 分别为数字串文件、固定文章文件和自由发言文件. 3 次录音得到的文件中的语音长度长短不一, 但同一种文件的长度基本相等. 数字串文件为 30 s 左右, 文章文件为 60 s 左右, 而自由发言部分也限定在 1 分钟以内. 说话人训练时使用数字串文件和自由发言文件中的各 30 s 的录音语料. 测试时 3 个文件都被采用, 测试语音长度取为 4 s.

同时, 本文也在 Childers[25] 提供的公开 Normal 语音测试集上进行了实验. 该测试集采样频率为 10 kHz, 包括 52 个说话人 (25 个女性, 27 个男性), 年龄分布在 20 ∼ 80 岁之间. 训练时采用其中前 20 个语音文件, 而所有语音都参与测试.

3.1　预处理和特征提取

实验中需要对输入的语音信号进行预加重、分帧、加窗和语音/非语音检测的处理. 首先采用一阶高通滤波器对语音信号进行预加重, 按帧长 12 ms 分帧, 帧交叠 6 ms 取帧, 之后使用汉明窗进行加窗处理, 最后采用基于短时能量和短时过零率的语音/非语音检测器把每一帧信号分成语音帧或非语音帧.

说话人的特征参数采用能够反映人对语音感知特性的美尔倒谱系数 (Mel frequency cepstrum coefficient, MFCC). 实验中, 求出 12 维 MFCC 系数和 12 维一阶差分倒谱动态系数作为说话人辨认的特征参数.

3.2　高斯混合模型

M 阶高斯混合模型 (Gaussian mixture model, GMM) 用 M 个单高斯分布的线性组合来描述帧特征在特征空间中的分布, 如下式所示

$$p(\boldsymbol{x}) = \sum_{i=1}^{M} P_i b_i(\boldsymbol{x}) \tag{8}$$

其中

$$b_i(\boldsymbol{x}) = N(\boldsymbol{x}, \mu_i, R_i) = \frac{1}{(2\pi)^{\frac{p}{2}} |R_i|^{\frac{1}{2}}} \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathrm{T}} R_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i)) \tag{9}$$

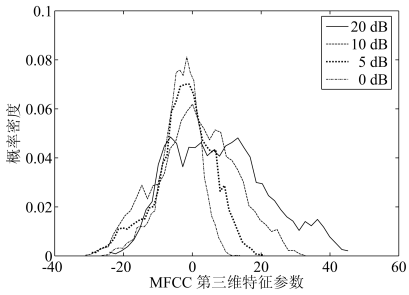
式中, p 为特征的维数; $b_i(\boldsymbol{x})$ 称为核函数, 是均值矢量为 $\boldsymbol{\mu}_i$、协方差矩阵为 R_i 的高斯分布函数; M 称为 GMM 模型的阶数, 设为一确定整数; $\lambda = P_i, \boldsymbol{\mu}_i, R_i|_{i=1,2,\cdots,M}$ 为说话人特征分布 GMM 模型中的参数.

通常 GMM 的阶数越高, 系统识别率越高, 但同时计算量和存储空间的开销也增加, 因此模型阶数应恰当选取, 本文折中考虑, 取 $M = 16$.

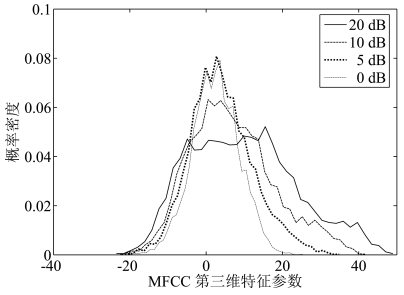
4　实验与结果分析

为了验证自适应直方图均衡化方法的优越性, 进行了 3 个实验. 3 个实验分别采用纯净语音附加不同信噪比的平稳 White 噪声和非平稳 Babble 噪声进行实验. 噪声数据取自 NoiseEx-92 噪声数据库[26].

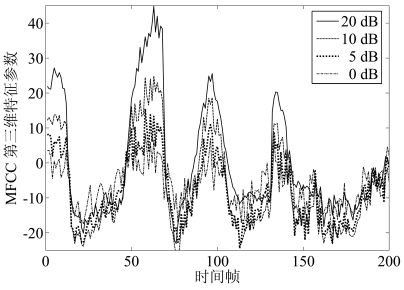
实验 1 测试了 AHEQ 在平稳和非平稳噪声环境下的变换性能, 并且在相同的条件下测试了 AHEQ 和 HEQ 方法的变换效果. 实验中将某一特定说话人的纯净语音 (取自 CST603 语音库) 按照信噪比 (SNR) 为 0, 5, 10, 20 dB 分别添加 White 噪声和 Babble 噪声形成含噪语音信号. 然后对其 MFCC 特征参数作 AHEQ 变换和 HEQ 变换, 其第三维 MFCC 特征参数的实验结果如图 2 所示, 图中变换前后的直方图显示仍然用大小一致的区间, 而用于 AHEQ 变换的累积直方图采用本文方法. 为了比较 HEQ 和 AHEQ 两种方法, 对比实验中 HEQ 的区间个数与 AHEQ 最终划分的区间个数相等.



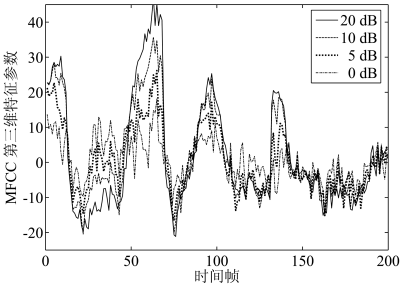
White 噪声下 AHEQ 变换前的直方图



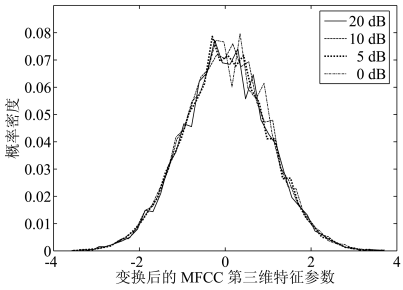
Babble 噪声下 AHEQ 变换前的直方图



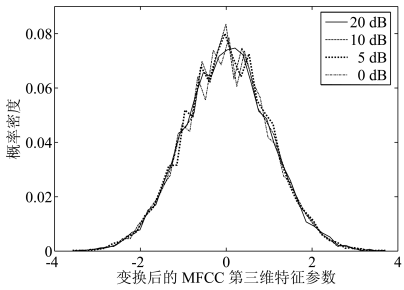
White 噪声下 AHEQ 变换前特征值曲线图



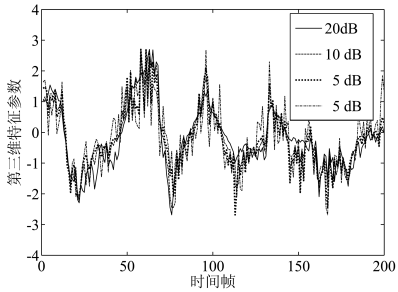
Babble 噪声下 AHEQ 变换前特征值曲线图



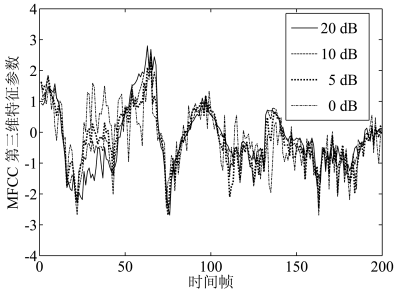
White 噪声下 AHEQ 变换后的直方图



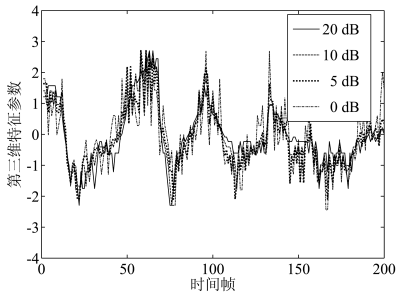
Babble 噪声下 AHEQ 变换后的直方图



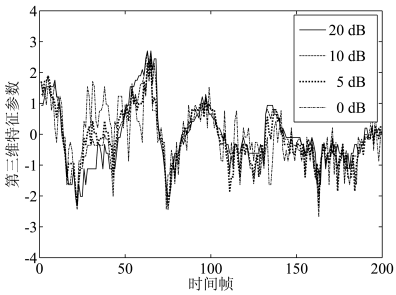
White 噪声下 AHEQ 变换后特征值曲线图



Babble 噪声下 AHEQ 变换后特征值曲线图



White 噪声下 HEQ 变换后特征值曲线图



Babble 噪声下 HEQ 变换后特征值曲线图

图 2 White 和 Babble 噪声下 AHEQ 对第三维 MFCC 特征参数的影响
以及相同区间个数下 AHEQ 与 HEQ 特征值曲线对比图

Fig. 2 Effect of AHEQ over the speech representation for the third MFCC coefficient in the White and Babble noise

图 2 中第一行的直方图显示了在不同信噪比的噪声影响下第三维特征值分布的变化. 显然, 不同信噪比的噪声在不同程度上影响了特征值的分布, 信噪比越大, 影响程度越严重. 特征值随时间变化以及受不同信噪比的噪声影响下的曲线图如图 2 中第二行所示. 同样, 从这两幅图中可以明显地看到不同信噪比噪声对特征产生不同程度的失真. 而且 Babble 非平稳噪声和 White 平稳噪声在同等信噪比下对特征的影响程度相似. 第三行中的两个图显示的是经 AHEQ 变换后的特征值直方图, 从图中可以看出它们的分布图与参考标准正态分布基本一致. 第四行是经 AHEQ 变换后的特征值曲线图, 从中可以看出由噪声引起的特征值失真明显地降低了. 第五行是经区间个数与 AHEQ 相同的 HEQ 变换后的特征值曲线图. 从图中可以看到有一些细小的水平线段, 这是由于 HEQ 某些区间内落入的特征值样本过多, 而且共同映射到参考空间的同一值上造成的, 也就是说在相同条件下, AHEQ 比 HEQ 变换后得到的特征值分布更符合实际, 更好地描述了说话人的个性特征.

实验 1 中 HEQ 的区间个数与 AHEQ 最终划分的区间个数是相等的; 在实验 2 中, 我们将比较 HEQ 的区间个数为 AHEQ 最终划分个数的 4 倍、7 倍和 10 倍的情况下的识别效果和处理速度. 实验中将取自 NoiseEx-92 噪声数据库[26] 的 White 噪声 (平稳噪声) 和 Babble 噪声 (非平稳噪声) 按一定的信噪比 (SNR= 10 dB) 添加到纯净语音 (取自 CST603 语音库) 中形成带噪语音. 对带噪语音信号采用高斯混合模型进行训练和识别, 其说话人辨认率和处理时间如表 1 和表 2 所示, 其中处理时间为 Matlab 7.0 环境下处理一个测试语音的平均时间.

从表 1 和表 2 中可以看出无论在 White 噪声还是在 Babble 噪声环境下, 当区间个数比值为 4 时, HEQ 较 AHEQ 变换方法的辨认率变化不大, 在 White 噪声环境下, 辨认率竟然下降了 0.1%; 而 HEQ 的处理时间比 AHEQ 变换方法增加了 4.1 s. 这主要是因为当 HEQ 的区间个数为 AHEQ 最终划分个数的 4 倍时, 在特征值样本密集的地方两种方法划分的疏密程度相似, 增加的区间个数主要集中在特征值样本稀少的地方, 而这些地方区间划分并无多大意义. 当区间个数比值为 7 和 10 时, HEQ 方法的辨认率平均增加了 0.9% 和 1.05%, 增值变化不大, 而处理时间增加了 12.9 s 和 26.2 s. 这主要因为当区间个数比值增大一定程度时, HEQ 在整个特征值区域的区间划分比 AHEQ 变换方法更加细致, 所以识别率有所提高; 而同时付出的代价是由于区间个数增加带来的处理时间的增加. 总之, 通过实验 2 可以看出采用 AHEQ 变换方法可以达到识别率和处理时间较优的权衡, 尤其当测试时间变长时, 特征值个数变多, 而每个特征值有 24 维, 当区间个数成倍增加时, 即使采用一些优化方法, 其处理时间也将大幅度增加. 而采用 AHEQ 变换方法可以避免这样的问题.

表 1　不同区间个数比值下的说话人辨认率 (%)
Table 1　Speaker identification rates under different ratios of interval number (%)

变换方法	区间个数比值					
	White 噪声			Babble 噪声		
	4	7	10	4	7	10
HEQ	82.4	83.4	83.5	80.2	81.0	81.2
AHEQ	82.5			80.1		

表 2　White 噪声下不同区间个数比值的处理时间 (s)
Table 2　The computation time under different ratios of interval number in White noise (s)

变换方法	区间个数比值		
	4	7	10
HEQ	15.9	24.7	38.0
AHEQ		11.8	

为了测试本文提出的 AHEQ 方法在不同信噪比下平稳噪声和非平稳噪声两种类型的噪声环境下的识别性能, 我们进行了实验 3. 该实验分别针对 CST603 语音和 Childers 测试集语音, 将噪声按一定的信噪比 (SNR = 0, 5, 10, 20 dB) 添加到纯净语音中形成带噪语音. 实验中分别采用不作变换、HEQ 变换和 AHEQ 变换 (HEQ 和 AHEQ 的区间个数相同) 对两种语音数据库中的说话人进行辨认, 采用高斯混合模型进行训练和识别, CST603 语音库上的识别结果如表 3 和表 4 所示; Childers 测试集的识别结果如表 5 和表 6 所示.

从 4 个表中可以得出以下几个结论:

1) 无论在 White 还是 Babble 噪声环境下, 两种测试集下的 HEQ 和 AHEQ 变换方法的辨认率普

表 3　CST603 库 White 噪声环境下的说话人辨认率 (%)
Table 3　Speaker identification rates in White noise with CST603 database (%)

变换方法	SNR (dB)				
	0	5	10	20	∞
None	35.1	49.5	68.7	92.9	98.5
HEQ	56.2	63.5	79.5	93.4	98.2
AHEQ	62.4	68.3	82.5	94.2	98.4

表 4 CST603 库 Babble 噪声环境下的说话人辨认率 (%)
Table 4 Speaker identification rates in Babble noise with CST603 database (%)

变换方法	SNR (dB)				
	0	5	10	20	∞
None	34.7	47.8	67.5	93.2	98.5
HEQ	54.9	62.9	78.4	93.8	98.2
AHEQ	61.3	67.8	80.1	95.1	98.4

表 5 Childers 测试集 White 噪声环境下的说话人辨认率 (%)
Table 5 Speaker identification rates in White noise with Childers database (%)

变换方法	SNR (dB)				
	0	5	10	20	∞
None	34.7	49.2	67.9	92.1	99.0
HEQ	55.9	62.9	79.0	93.5	98.5
AHEQ	61.9	68.2	82.6	93.9	98.4

表 6 Childers 测试集 Babble 噪声环境下的说话人辨认率 (%)
Table 6 Speaker identification rates in Babble noise with Childers database (%)

变换方法	SNR (dB)				
	0	5	10	20	∞
None	33.7	47.1	67.2	93.3	99.0
HEQ	54.1	62.5	77.9	94.0	98.5
AHEQ	60.9	67.9	79.4	94.8	98.4

遍比无变换的情况下提高很多, 而且信噪比越低时辨认率提高得越大. 当信噪比为零时, 提高的辨认率均大于 20 %; 而当信号为纯净语音 (SNR = ∞) 时, HEQ 和 AHEQ 的辨认率反而比无变换时平均降低了约 0.4%. 这主要是因为当信噪比越低时, HEQ 和 AHEQ 对失真的特征值补偿的越多, 当为纯净语音时, HEQ 和 AHEQ 的变换反而引起了特征信息的小部分丢失.

2) 在各种信噪比下的两种噪声环境中, AHEQ 的辨认率比 HEQ 的辨认率要高一些; 同样随着信噪比的增加辨认率的提高幅度逐渐下降. 当 SNR = 0 时, 在两种噪声环境 AHEQ 的辨认率比 HEQ 在 CST603 和 Childers 测试集下分别平均增加 6.3% 和 6.4%, 而 SNR = ∞ 时, 辨认率分别平均提高了 0.2% 和 0.55%. 这说明本文中提出的 AHEQ 方法在区间个数相等的条件下确实提高了说话人辨认的性能.

3) 同一变换方法、同一信噪比下, Babble 噪声环境下的辨认率大多数都比 White 噪声环境下要小一些, 但差别并不大, 而且随着变换方法和信噪比的变化, 辨认率的差值变化相对稳定. 这可以表明两种环境下导致辨认率的差异主要是因为噪声本身的性质引起的, 也就是说 AHEQ 方法在两种不同类型的噪声环境下都具有较强的鲁棒性.

4) 无论是 CST603 语音库还是 Childers 测试集, 对于相同的噪声环境和信噪比下, 三种方法的辨别率变化比较相似. 这主要因为两种测试集中的语音都是纯净语音, 加噪后在相同的噪声环境和信噪比下其识别率也就差别不大.

总之, 通过实验我们发现自适应直方图均衡化确实能够提高系统识别率, 降低误识别率. 而且该算法在两种不同的测试集和不同类型的噪声环境下都具有一定的鲁棒性.

5 结论

在说话人识别研究中, 噪声和干扰一直是影响系统识别率提高的主要原因. 本文基于说话人特征的统计特性和直方图均衡化在说话人识别中的应用特点, 提出了直方图均衡化的自适应方法. 实验结果表明, 采用 AHEQ 变换方法可以达到识别率和处理时间较优的权衡. 当区间个数一致时, 与普通的直方图均衡化变换方法相比, 自适应直方图均衡化能进一步提高辨认系统的辨认率; 并且在不同的测试集以及无论在平稳噪声还是非平稳噪声环境下, 该算法都能取得较好的辨认率, 进一步增强系统的鲁棒性.

References

1 Zhao Li. *Digital Processing of Speech Signals*. Beijing: China Machine Press, 2003
(赵力. 语音信号处理. 北京: 机械工业出版社, 2003)

2 Ma Da-You. *Foundation of Modern Acoustics Theory*. Beijing: Science Press, 2004
(马大猷. 现代声学理论基础. 北京: 科学出版社, 2004)

3 Pandey P C, Bhandorkar S M, Bachher G K, Lehana P K. Enhancement of alaryngeal speech using spectral subtraction. In: Proceedings of the 14th International Conference on Digital Signal Processing. New York, USA: IEEE, 2002. 591−594

4 Zhong L, Goubran R. Musical noise reduction in speech using two-dimensional spectrogram enhancement. In: Proceedings of the 2nd International Workshop on Haptic, Audio and Visual Environments and Their Applications. Ottawa, Canada: IEEE, 2003. 61−64

5 Tadj C, Gabrea M, Gargour C, Ramachandran V. Towards robustness in speaker verification: enhancement and adaptation. In: Proceedings of the 45th Midwest Symposium on Circuits and Systems. New York, USA: IEEE, 2002. 320−323

6 Soon I Y, Koh S N. Speech enhancement using 2-D Fourier transform. *IEEE Transactions on Speech and Audio Processing*, 2003, **11**(6): 717−724

7 Zhen Y X, Zheng T F, Wu W H. Weighting observation vectors for robust speech recognition in noisy environment. In: Proceedings of International Conference on Spoken Language Processing. Jeju Island, Korean: ISCA, 2004. 819−822

8 Yu Peng, Xu Yi-Fang, Cao Zhi-Gang. Robust speaker identification based on weighted feature compensation. *Signal Processing*, 2002, **18**(6): 513−517
(于鹏, 徐义芳, 曹志刚. 基于加权特征值补偿的说话人识别. 信号处理, 2002, **18**(6): 513−517)

9 Bao Yong-Qiang, Zhao Li, Zou Cai-Rong. Text-independent speaker recognition using normalization compensation transformation. *Acta Acustica*, 2006, **31**(1): 55−60
(包永强, 赵力, 邹采荣. 采用归一化补偿变换的与文本无关的说话人识别. 声学学报, 2006, **31**(1): 55−60)

10 Chen K. Towards better making a decision in speaker verification. *Pattern Recognition*, 2003, **36**(2): 329−346

11 Matthews J. Histogram equalization [Online], available: http://www.generation5.org/content/2004/histogramEqualization.asp, September, 2004

12 Skosan M. Histogram Equalization for Robust Text-independent Speaker Verification in Telephone Environments [Master dissertation], University of Cape Town, 2005

13 Skosan M, Mashao D J. Matching feature distributions for robust speaker verification. In: Proceedings of Annual Symposium of the Pattern Recognition Association of South Africa. Grabouw, South Africa: IEEE, 2004. 93−97

14 De la Torre A, Segura J C, Benftez C, Peinado A M, Rubio A J. Non-linear transformations of the feature space for robust speech recognition. In: Proceedings of the 27th International Conference on Acoustics, Speech, and Signal Processing. Orlando, USA: IEEE, 2002. 401−404

15 Molau S, Keysers D, Ney H. Matching training and test data distributions for robust speech recognition. *Speech Communication*, 2003, **41**(4): 579−601

16 Wan C Y, Lee L S. Joint uncertainty decoding (JUD) with histogram-based quantization (HQ) for robust and/or distributed speech recognition. In: Proceedings of the 31st International Conference on Acoustics, Speech, and Signal Processing. Toulouse, France: IEEE, 2006. 125−128

17 Wan C Y, Lee L S. Histogram-based quantization (HQ) for robust and scalable distributed speech recognition. In: Proceedings of the 9th European Conference on Speech Communication and Technology. Lisbon, Portugal: IEEE, 2005. 957−960

18 de la Torre A, Peinado A M, Segura J C, Perez-Cordoba J L, Benitez M C, Rubio A J. Histogram equalization of speech representation for robust speech recognition. *IEEE Transactions on Speech Audio Processing*, 2005, **13**(3): 355−366

19 Skosan M, Mashao D J. Modified segmental histogram equalization for robust speaker verification. *Pattern Recognition Letters*, 2006, **27**(5): 479−486

20 Dharanipargda S, Padmanabhan M. A nonlinear unsupervised adaptation technique for speech recognition. In: Proceedings of the 6th International Conference on Spoken Language Processing. Beijing, China: ISCA, 2000. 556−559

21 Hilger F, Ney H. Quantile based histogram equalization for noise robust speech recognition. In: Proceedings of the 7th European Conference on Speech Communication and Technology. Aalborg, Denmark: ISCA, 2001. 1135−1138

22 Hilger F, Ney H. Quantile based histogram equalization for noise robust large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(3): 845−854

23 Jankowski C R, Vo H D, Lippmann R P. A comparison of signal processing front ends for automatic word recognition. *IEEE Transactions on Speech and Audio Processing*, 1995, **3**(4): 286−293

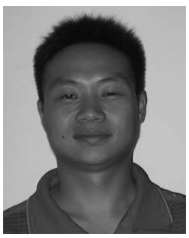
24 Jain P, Hermansky H. Improved mean and variance normalization for robust speech recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, USA: IEEE, 2001

25 Childers D G. *Speech Processing and Synthesis Toolboxes (English Version of Photocopying)*. Beijing: Tsinghua University Press, 2004
(Childers D G. Matlab 之语音处理与合成工具箱 (影印版). 北京: 清华大学出版社, 2004)

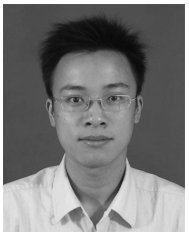
26 Varga A P, Steeneken H J M, Tomlinson M, Jones D. The Noisex-92 Study on the Effect of Additive Noise on Automatic Speech Recognition. Technical Report, Speech Research Unit, Defense Research Agency, Malvern, UK, 1992

徐利敏 南京财经大学电子商务重点实验室讲师, 南京理工大学计算机科学与技术学院博士研究生. 主要研究方向为模式识别, 数据挖掘, 语音识别, 说话人识别. 本文通信作者.
E-mail: meiwen_xu@yahoo.com.cn
(**XU Li-Min** Lecturer at the Key Laboratory of Electronic Business, Nanjing University of Finance and Economics and Ph. D. candidate at the School of Computer Science and Technology, Nanjing University of Science and Technology. Her research interest covers pattern recognition, speech recognition, and speaker recognition. Corresponding author of this paper.)

唐振民 南京理工大学计算机科学与技术学院教授. 主要研究方向为模式识别, 图像与语音处理.
E-mail: tang_zm@mail.njust.edu.cn
(**TANG Zhen-Min** Professor at the School of Computer Science and Technology, Nanjing University of Science and Technology. His research interest covers pattern recognition, image and speech processing.)



何可可 南京理工大学计算机科学与技术学院博士研究生. 主要研究方向为模式识别, 语音识别, 说话人识别.
E-mail: he_keke@126.com
(**HE Ke-Ke** Ph. D. candidate at the School of Computer Science and Technology, Nanjing University of Science and Technology. His research interest covers pattern recognition, speech recognition, and speaker recognition.)



钱 博 南京理工大学计算机科学与技术学院博士研究生. 主要研究方向为模式识别, 语音识别, 说话人识别.
E-mail: sandson6@163.com
(**QIAN Bo** Ph. D. candidate at the School of Computer Science and Technology, Nanjing University of Science and Technology. His research interest covers pattern recognition, speech recognition, and speaker recognition.)