采用特征分类直方图均衡化的鲁棒性语音识别

姜 莹 俞一彪
(苏州大学电子信息学院,江苏 苏州 215006)

摘 要:大部分噪声会引起语音倒谱域特征参数的非线性失真,导致识别系统性能下降。直方图均衡化方法是一种非线性补偿变换技术,较传统的基于线性变换技术的抗噪声方法进一步提高了系统的鲁棒性。但实际识别系统中,除了噪声引起语音特征的非线性失真外,还存在训练和测试数据的语音特征类分布不一致问题,从而难以保证传统的直方图均衡化方法发挥其优势。本文提出一种基于特征分类的直方图均衡化方法,首先对初步均衡化后的含噪语音特征矢量进行 K 均值分类,然后对各类别下的特征矢量再进行直方图均衡变换。实验结果表明,低信噪比时无论在平稳噪声还是非平稳噪声环境下,与传统的直方图均衡化方法相比都进一步增强了识别系统的鲁棒性。

关键词:语音识别;直方图均衡化;特征分类;鲁棒性
中图分类号:TN912 **文献标识码**:A **文章编号**:1003-0530(2011)06-0896-05

Robust Speech Recognition Using Histogram Equalization of Classified Features

JIANG Ying YU Yi-biao
(Soochow University, Suzhou Jiangsu 215006)

Abstract: Noises cause feature distortion of speech and make the performance of speech recognition system seriously poor. Comparing with classical methods, the histogram equalization can reduce non-linear distortion and improve the robustness of speech recognition system quite well. However in many applications, the feature distribution between training and test speech is usually not identical because of their difference in phonetics or acoustics, then the validity of HEQ can be weaken. The proposed algorithm in this paper utilizes K-means clustering to classify the pre-equalized noisy features into several classes, then further equalizes the features belong to the same class. The experiments show the proposed method improves the performance of system with comparison of usual histogram equalization.

Key words: Speech recognition; Histogram equalization; Feature classification; Robustness

1 引言

目前在实验室环境中,语音识别可以达到令人满意的识别结果。但在实际应用中,由于语音采集环境的影响(加性噪声、信道畸变、录音设备等)和说话人的影响(说话口音、风格、情绪以及健康状况等),识别系统性能严重下降。

鲁棒性语音识别研究试图解决如何在实际环境下提升语音识别系统性能的问题[1]。通常的抗噪声方法主要可以分为三种,即前端处理、特征值处理以及模型补偿。特征值的抗噪声处理直方图均衡化(Histogram equalization,HEQ)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[2],主要对含噪语音的特征参量进行处理,寻找稳健性的特征参量。近几年来研究人员开始尝试将其应用在语音识别上以提高系统的鲁棒性。

实际环境下大部分噪声会引起语音倒谱域特征参数的非线性失真,与 CMN 等去噪补偿方法不同,直方图均衡化是一种非线性的补偿变换方法[3],并且只对语音的特征参数进行处理,不需要在识别时预先知道含噪语音的信噪比和噪声类型,计算量也并不复杂。但是传统的直方图均衡化方法也存在一些缺陷:(1)对

收稿日期:2010 年 10 月 29 日;修回日期:2011 年 3 月 16 日
基金项目:北京市“现代信息科学与网络技术”重点实验室暨铁道部“铁路信息科学与工程”开放实验室(编号:XDXX1006)的资助

语音特征参数各维矢量分别进行处理是基于语音特征参数各维矢量相互独立的假设,但实际上各维矢量之间存在相关性;(2)准确计算累积分布函数需要充分多的特征样本数据,但是有些识别系统(如孤立词识别)中测试语音数据较少,不可能达到这一要求;(3)不能保证参考语音和测试语音的特征分布的一致性,两者包含的语音成分往往存在差异。

本文提出了一种基于语音特征分类的直方图均衡化方法应用于语音识别系统。由于测试语音数据较短,首先采用统计顺序累积分布函数估计方法对含噪语音各维特征参数进行初步直方图均衡化处理,补偿训练和测试环境的声学失配,然后将均衡化后的特征矢量用 K 均值聚类算法分类,并对属于每个类的特征矢量再进一步进行均衡化处理。实验证明该方法在低信噪比环境下能有效提高语音识别系统的性能,相对传统的直方图均衡化方法鲁棒性更好。

2　直方图均衡化原理

直方图变换的原理是将原矢量的直方图变换到参考的直方图,以达到将原矢量 x 变换到目标矢量 y 的过程。假设原样本矢量为 x,其概率密度函数为 $P_x(x)$,累积分布函数为 $C_x(x)$。变换后的矢量为 y,其参考概率密度函数为 $P_{ref}(y)$,累积分布函数为 $C_{ref}(y)$,且有 $y=T(x)$。特征参数的变换函数应使得

$$C_x(x)=C_{ref}(y)=C_{ref}(T(x)) \tag{1}$$

由此可得

$$y=T(x)=C_{ref}^{-1}(C_x(x)) \tag{2}$$

在实际实现中,通常把训练和测试数据概率分布都变换到标准高斯分布,实现参数规整。

直方图均衡化方法又有一系列的变形,例如:基于分位数的直方图均衡[4]、直方图均衡与一些降噪方法(如谱减法[5],适量泰勒级数(Vector taylor series, VTS)[6])联合使用的方法以及对语音段和噪声段分别计算累积直方图的均衡方法等。由(2)式可知,直方图均衡方法中的一个关键问题是得到合理的累积分布函数 $C_x(x)$ 的估计,而数据量越少越难逼近其真实的累积分布。有实验结果表明,测试语音数据比较短时,由于样本数据的减少,以上方法明显逊色于基于统计顺序的直方图估计方法 OS-HEQ(Order-Statistic-based HEQ)[7],以下进行简单的描述:

假定一 $N\times K$(帧数×维数)的特征矢量,其中第 k $(1\le k\le K)$ 维的特征矢量 V_k 表示为:

$$V_k=\{x_k(1),x_k(2),\cdots,x_k(N)\} \tag{3}$$

其中 $x_k(n)$ 是第 n 帧特征矢量第 k 维的参数。将 V_k 中元素按升序排列有:

$$V_k'=\{x_k([1]),x_k([2]),\cdots,x_k([r_k])x_k([N])\} \tag{4}$$

其中

$$x_k([1])\le x_k([2])\le\cdots\le x_k([r_k])\cdots\le x_k([N]) \tag{5}$$

$[r_k]$ 表示在重排后的特征矢量 V_k' 中处于第 r_k 位置处的元素对应于原 V_k 中所处于的原始帧数。

由式(4),(5)对处在第 r_k 处的元素进行处理,其基于统计顺序估计的累积分布函数表示为:

$$C_{X(k)}(x_k([r_k]))=\frac{r_k-0.5}{N},1\le r_k\le N,1\le k\le K \tag{6}$$

那么原始特征矢量 $x_k(n)$ 经 OS-HEQ 变换后的矢量 $y_k(n)$ 为:

$$y_k(n)=C_{ref}^{-1}(C_{X(k)}(x_k(n)))=C_{ref}^{-1}(\frac{r_k(x_k(n))-0.5}{N}) \tag{7}$$

3　特征分类直方图均衡化方法(FC-HEQ)

当测试数据时长较短时,由于样本数据减少,基于统计的累积分布方式估计难以逼近特征参数真实的累积分布。基于统计顺序的直方图均衡化方法 OS-HEQ,采用了更为合理有效的累积分布函数估计方法,把各维分量在所有时间帧上的特征值作为一个整体,得到其统计排列顺序估计累积分布函数,但每一维特征分量必定包含不同语音成分的特征分布信息,对所有语音成分采用统一的基于统计顺序的累积分布函数估计方法并不合理。此外基于统计顺序的直方图均衡化方法 OS-HEQ 仍然是对各维分量独立处理,假设各维分量相互独立,而实际上各维分量之间存在相关性。

本文提出基于语音特征分类的直方图均衡化方法 FC-HEQ(Feature Classification HEQ)。首先将特征矢量进行直方图均衡化变换,补偿声学失配引起的失真,然后将均衡化后的矢量进行 K 均值聚类,再对属于每个子类的特征参数计算累积分布函数,实现参数规整。这里,首先对特征矢量进行直方图均衡化变换,可以补偿由于训练和测试环境不匹配引起的声学失真,以便于后续特征分类的正确性。在此基础上,对初步均衡化后的特征矢量进行特征分类,把相似的语音特征矢量划分到同一类别,以增加对应类别中训练和测试语音的特征分布一致性。最后,对属于每个类的特征矢量进行均衡化处理,实现特征矢量的进一步均衡化处理。这一方法不同于传统的直方图均衡化方法,无需假设特征矢量各维独立的情况下计算各维参数集的累积分布函数进行参数规整,同时通过特征分类使得训练和测试语音在对应子类上特征分布具有更好的一致性。

图 1 为具体的实现方法。M_n 表示为含噪语音某一帧(含 K 维)的特征矢量,$M_n=\{x_1(n),x_2(n),\cdots,x_K(n)\}$。

首先假定特征矢量各维参数相互独立,进行 OS-HEQ 处理后得到均衡化后的矢量 $M_n^{'}$

$$M_n^{'}=\{y_1(n),y_2(n),\cdots,y_K(n)\}$$
$$=\{C_{Y(1)}^{-1}(C_{X(1)}(x_1(n))),\cdots,C_{Y(K)}^{-1}(C_{X(K)}(x_K(n)))\}$$

对 $M_n^{'}$ 进行 K 均值聚类,聚类的数目可根据语音所含的语音成分而定。数据帧较短的时候往往意味着语音成分较少,因此适当减少聚类数目将使得划分到每一类的数据帧保持一定数量,有利于计算直方图累积函数。假定属于第 i 类的数据个数为 N_i,那么整个均衡化后的特征分量表示为:

$$y_k^{'}(n)=C_{Y(i,k)}^{-1}(C_{X(i,k)}(x_k(n)))=C_{Y(i,k)}^{-1}\left(\frac{r_{ik}(x_k(n))-0.5}{N_i}\right) \tag{8}$$

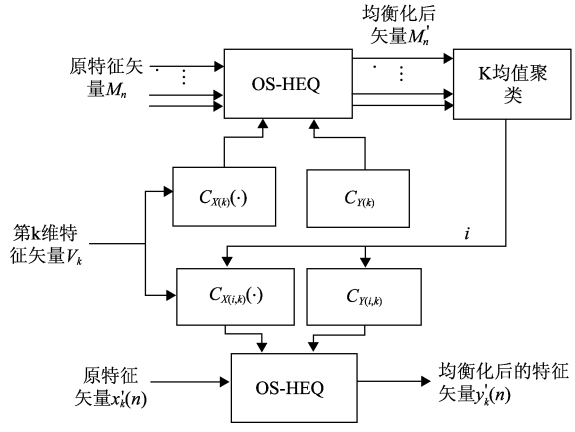


图 1 FC-HEQ 处理流程

Fig. 1 Block diagram of FC-HEQ method

4 实验分析

实验采用 SUDA2008 语音库,其中包括 30 个人(15 男 15 女)的地名词汇语音。词汇包含国内 26 个主要城市地名,其中每个人每个词汇发音十遍。添加的噪声信号取自 NoiseX-92 噪声数据库。实验中采用纯净语音训练,测试语音为纯净语音按照信噪比 0,5,10,15,20db 添加 White 噪声(平稳噪声)和 Babble 噪声(非平稳噪声)形成含噪语音。库中 22 个人(11 男 11 女)的每个城市地名 6 遍发音组成训练数据。识别测试由两部分组成:InSideTest 测试集由参与训练的 22 个人(11 男 11 女)构成,每人的另外 4 遍各城市地名发音组成 2288 个测试数据;另一个是 OutSideTest 测试集,由其余 8 个人(4 男 4 女)的各城市地名的 10 遍发音组成,共 2080 个测试数据。语音信号分帧处理,帧长 25ms,帧移 10ms,加汉明窗,预加重系数取 0.97。参数采用 39 维分量,其中包括 12 阶 MFCC,1 阶归一化能量加及其一阶、二阶差分分量[8],[9]。语音模型采用 6 状态、3 个分量的高斯混合分布 HMM,在 HTK 平台下进行模型训练和识别工作。

实验对原始语音叠加 White 噪声(平稳噪声)和 Babble 噪声(非平稳噪声),在不同信噪比和不同语音特征分类数目下的识别性能进行了比较分析。表 1 和表 2 分别是 InsideTest 和 OutSideTest 测试集在 White 噪声下的识别率,表 3 和表 4 是 Babble 噪声下的识别率。均衡化方法中 None 表示对噪声语音的特征参数 MFCC 不进行均衡化处理,表中的 OS-HEQ 方法,实际上相当于 FC-HEQ 方法中,分类数目 M 取 1。

表 1 InSideTest 测试集 White 噪声下的识别率(%)

Tab. 1 Recognition rates for InSideTest in White noise

均衡方法		SNR(db)				
		0	5	10	15	20
None (MFCC)		23.94	47.70	64.47	81.39	88.85
OS-HEQ		60.26	81.20	90.85	95.24	96.78
FC-HEQ No. of classes	$M=2$	64.88	83.38	92.62	96.75	97.98
	$M=3$	62.18	82.56	92.15	96.42	97.28
	$M=4$	61.62	81.59	90.92	96.04	96.99
	$M=5$	60.48	81.42	90.56	95.41	95.98

表 2 OutSideTest 测试集 White 噪声下词的识别率(%)

Tab. 2 Recognition rates for OutSideTest in White noise

均衡方法		SNR(db)				
		0	5	10	15	20
None (MFCC)		22.19	46.78	62.84	79.72	88.77
OS-HEQ		58.64	78.52	89.76	94.48	95.35
FC-HEQ No. of classes	$M=2$	63.57	81.37	91.78	95.76	96.45
	$M=3$	61.92	80.52	91.24	95.08	95.72
	$M=4$	60.56	79.41	90.78	94.52	95.45
	$M=5$	59.74	78.81	89.87	94.48	95.39

表 3 InSideTest 测试集 Babble 噪声下词的识别率(%)

Tab. 3 Recognition rates for InSideTest in Babble noise

均衡方法		SNR(db)				
		0	5	10	15	20
None (MFCC)		24.32	48.63	65.52	83.16	88.92
OS-HEQ		60.93	82.20	91.16	95.48	97.12
FC-HEQ No. of classes	$M=2$	65.12	84.54	92.45	96.65	97.88
	$M=3$	63.18	83.88	91.57	96.08	97.49
	$M=4$	62.37	82.52	91.31	95.69	97.24
	$M=5$	61.53	82.40	91.18	95.48	97.15

表 4　OutSideTest 测试集 Babble 噪声下词的识别率(%)

Tab. 4　Recognition rates for OutSideTest in Babble noise

均衡方法		SNR(db)				
		0	5	10	15	20
None (MFCC)		22.78	47.72	64.35	82.71	87.98
OS-HEQ		59.75	81.06	90.42	93.82	96.48
FC-HEQ No. of classes	$M=2$	64.22	83.45	91.68	95.03	97.12
	$M=3$	63.02	82.92	90.89	94.72	97.02
	$M=4$	61.49	81.39	90.57	94.13	96.58
	$M=5$	61.11	81.22	90.47	93.91	95.52

从表中可以得出以下结论:

1)无论在 White 还是 Babble 噪声环境下,采用直方图均衡化方法的识别性能比对特征参数不进行任何处理时的识别性能提高很多,而且信噪比越低,直方图均衡化方法对失真的特征值补偿效果越明显,识别率提高得越多,说明直方图均衡方法有效地提高了识别系统的鲁棒性。

2)在各等级信噪比下,FC-HEQ 方法的识别率在不同分类数目下都比 OS-HEQ 方法的要高。尤其在分类数目 $M=2$ 时,在各等级信噪比 0、5、10、15、20db 下,FC-HEQ 比 OS-HEQ 方法的识别率提高最多,在 White 噪声下,InSideTest 测试分别提高了 4.62%、2.18%、1.77%、1.51%、1.20%;OutSideTest 测试分别提高了 4.93%、2.85%、2.02%、1.28%、1.10%。在 Babble 噪声下,InSideTest 测试分别提高了 4.19%、2.34%、1.29%、1.17%、0.76%;OutSideTest 测试分别提高了 4.47%、2.39%、1.26%、1.21%、0.64%。可见,随着信噪比增加,FC-HEQ 方法的识别率提高的幅度降低,在其它分类数目下,同样如此。尤其在 SNR=0 时,识别率提高最多。说明本文提出的 FC-HEQ 方法较传统的 OS-HEQ 方法能有效提高系统的识别性能,尤其在低信噪比时其优越性更加明显。

3)在各等级信噪比下,FC-HEQ 方法的分类数目为 2 时,识别率达到最大,随着分类数目的增加识别率反而降低,增加分类数目已没有意义。由于测试语音数据为地名词汇,平均时长 60 帧左右,因此,当类别数进一步增加时,划分到每个类的数据逐步减少,反而因为没有充分的数据估计类分布而导致识别率的降低。实际上,这里 M 的取值不一定由语音中的语音单元数目决定,而由识别语音数据长短经验估计。

4)同一变换方法、同一信噪比下,Babble 噪声环境下的识别率大多数都比 White 噪声环境下要高,但差别并不大,而且不同变换方法和信噪比下这种识别率的差值变化相对稳定。这表明两种环境下导致识别率的差异主要是因为噪声本身的性质引起的,也就是说基于特征分类均衡化方法在两种不同类型的噪声环境下都具有较强的鲁棒性。

当测试语音较短时,FC-HEQ 方法和 OS-HEQ 方法都采用了合理有效的分布函数估计方法,与 OS-HEQ 方法相比,FC-HEQ 方法考虑了各语音成分的分布信息和特征矢量之间的相关性,减少了语音特征分布不一致的影响,从而进一步提高了识别率。尤其在低信噪比下,体现了更为明显的优越性。图 2 是三种方法在不同信噪比下的平均识别率,数据基于 OutSideTest 测试集在两种噪声下的实验结果计算平均值得到,其中 FC-HEQ 方法采用 $M=2$ 的结果数据计算。由图可见,直方图均衡方法有效地提高了噪声环境下的识别率,而 FC-HEQ 更进一步提高了识别系统的性能。

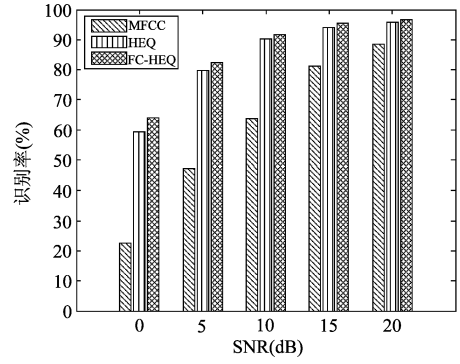


图 2　OutSideTest 测试集在 White 和 Babble 噪声下词的平均识别率(FC-HEQ 中 $M=2$)

Fig. 2　Average recognition rates for OutSideTest in White and Babble noise($M=2$, FC-HEQ)

5　结论

传统的直方图均衡化方法忽略各维参数之间的相关性,当测试语音帧很少的时候,采用统计顺序方法估计累积分布函数模糊了各语音成分的分布信息,从而影响直方图均衡化方法发挥其优势。本文提出的基于 K 均值聚类的直方图均衡化方法考虑了特征矢量之间的相关性,既补偿了训练环境和测试环境的声学特征不匹配,也减少了语音特征分布不一致的影响。实验结果表明,无论在平稳噪声还是在非平稳噪声下都比传统的直方图均衡化方法具有更高的识别性能,尤其在低信噪比时其优越性更加明显。

参考文献

[1] 刘波,戴礼荣,王仁华,杜俊,李锦宇. 基于双高斯GMM的特征参数规整及其在语音识别中的应用[J]. 自动化学报,2006,4(32):519-525.
Liu Bo, Dai Lirong, Wang renhua, Du Jun, Li Jingyu. Double Gaussian GMM Based Feature Normalization and Its Application in Speech Recognition[J]. ACTA AUTOMATICA SINICA, 2006,4(32):519-525. (in Chinese)

[2] R. C. Gonzalez, R. E. Woods. Digital Image Processing [M], New Jersey, Prentice-Hall, 2002.

[3] O. Viikki, K. Laurila. Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition[J]. Speech Communication, 1998, 1(25): 133-147.

[4] Hilger F, Molau S, Ney H. Quantile based histogram equalization for online application. Proceedings of International Conference of Spoken Language Proceessing, Rundle Mall, Australia, Causal Productions, 2002, 237-240.

[5] Segura J C, Benitez M C, de la Torre A, Rubio A J. Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR[J]. Proceedings of International Conference of Spoken Language Processing 2002, Rundle Mall, Australia, Causal Productions, 2002, 225-228.

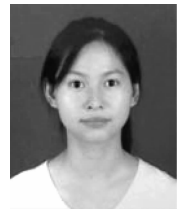
[6] Segura J C, Benitez M C, de la Torre A. VTS residual noise compensation [J]. Proceedings of International Conference on Acoustics and Signal Processing 2002. Piscataway, USA, IEEE Press, 2002, 209-212.

[7] J. C. Segura, C. Benítez, Á. de la Torre, A. J. Rubio, J. Ramírez. Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speec Recognition[J]. IEEE Signal Processing Letters, 2004, 5(11): 517-520.

[8] Young S, Evermann G, Hain T et al. The HTK Book (for HTK Version 3.2.1). 2002, http://htk.eng.cam.ac.uk.

[9] H. Y. Jun. Filtering of Filter-Bank Energies for Robust Speech Recognition[J]. ETRI, 3(26), 2004, 273-276.

作者简介



姜 莹(1986-),女,苏州大学电子信息学院硕士研究生,研究方向为语音识别。E-mail:jiangyingawu@163.com

俞一彪(1962-),教授,男。1985年7月毕业于北方交通大学信息处理技术专业,获工学硕士学位,骨干教师,现为苏州大学通信与信息系通专业硕士生导师,主要研究方向为:数字信号处理。曾在国家核心刊物上发表了《基于关键词的句法分析及在连续语音识别中的应用》、《ActiveX及在Intranet应用程序设计中的应用》、《计算机应用基础课程教学方案设计》等论文。从事的主要科研项目有:汉语文语转换中语音韵律的控制,省重点学科资助项目等。
E-mail:yuyb@suda.edu.cn