基于双高斯 GMM 的特征参数规整及其在语音识别中的应用[1)]

刘 波 戴礼荣 王仁华 杜 俊 李锦宇

(中国科学技术大学电子工程与信息科学系 合肥 230027)
(E-mail: lrdai@ustc.edu.cn)

摘 要 对特征参数概率分布的实验分析表明, 在有噪声影响的情况下, 特征参数通常呈现双峰分布. 据此, 本文提出了一种新的, 基于双高斯的高斯混合模型 (Gaussian mixture model, GMM) 的特征参数归一化方法, 以提高语音识别系统的鲁棒性. 该方法采用更为细致的双高斯模型来表达特征参数的累积分布函数 (CDF), 并依据估计得到的 CDF 进行参数变换将训练和识别时的特征参数的分布都规整为标准高斯分布, 从而提高识别正确率. 在 Aurora 2 和 Aurora 3 数据库上的实验结果表明, 本文提出的方法的性能明显好于传统的倒谱均值规整 (Cepstral mean normalization, CMN) 和倒谱均值方差规整 (Cepstral mean and variance normalization, CMVN) 方法, 而与非参数化方法 — 直方图均衡特征规整方法的性能基本相当.

关键词 语音识别, 前端, 噪声鲁棒性, 语音特征参数规整, 直方图均衡
中图分类号 TN912

Double Gaussian GMM Based Feature Normalization and Its Application in Speech Recognition

LIU Bo DAI Li-Rong WANG Ren-Hua DU Jun LI Jin-Yu

(*Department of Electronic Engineering & Information Science, University of Science & Technology of China, Hefei* 230027)
(E-mail: lrdai@ustc.edu.cn)

Abstract In this paper, a new feature normalization approach based on double Gaussian mixture model is proposed. Since speech features in noisy environments usually follow bimodal distributions, to fully utilize this characteristic we represent the cumulative density function (CDF) of the features with a more delicate Gaussian mixture model. Finally, feature normalization process is performed according to the estimated CDF to improve speech recognition performance. Experimental results on Aurora 2 and Aurora 3 tasks show that the performance of our method is much better than those of the conventional cepstral mean normalization (CMN) and cepstral mean and variance normalization (CMVN) methods, and is comparable to that of the histogram equalization method, which is a non-parametric method.

Key words Speech recognition, front-end, noise-robustness, speech feature normalization, histogram equalization

1) 国家自然科学基金项目 (60275038), 国家高技术研究发展计划 (863 计划)(2004AA114030) 资助
Supported by National Natural Science Foundation of P. R. China (60275038), National 863 Program (2004AA114030)
收稿日期 2005-2-26 收修改稿日期 2006-1-26
Received February 26, 2005; in revised form January 26, 2006

1 引言

尽管目前在实验室环境中, 许多语音识别系统的性能都非常出色, 但在实际环境中, 它们的表现往往很难令人满意, 这严重影响了语音识别的实用化进程. 导致这种现象的原因非常复杂, 包括了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 (说话风格, 口音, 以及环境影响引起的说话风格的变化等). 其中加性噪声和信道畸变又是最常见的两个原因.

鲁棒性语音识别研究并试图解决的就是如何在实际环境下提升语音识别系统性能的问题. 在理论上, 噪声鲁棒性所面临的问题其实就是训练和识别环境之间的不匹配. 而这种不匹配通常都会体现在特征参数概率分布的差异上, 特征参数规整 (归一化, Normalization) 可以在一定程度上减小这种不匹配的程度, 进而提升系统性能.

本文所要讨论的累积分布函数匹配 (Cumulative distribution function matching, CDF-Matching) 原理, 就是一个表述这种特征规整思想的很好理论框架. 有两大类特征参数规整方法可以统一到 CDF-Matching 框架之下.

首先是以直方图均衡方法[1] 为代表的非参数化方法. 它是 CDF-Matching 原理的直接应用, 其基本思想是设法直接估计出随机变量的累积分布函数, 然后再依据累积分布函数对特征参数进行规整. 直方图均衡方法有一系列变形, 例如基于分位数的直方图均衡[2~4] 和基于滑动窗口的直方图均衡方法, 它们适用于不同的场合, 前者用较少量的数据便可以获得特征分布的累积直方图, 后者则可以及时的适应和跟上分布的变化. 直方图均衡方法还可以与一些降噪方法 (如谱相减[5], 矢量泰勒级数 (Vector taylor series, VTS)[6]) 一起使用, 以取得进一步的性能提升. 还有对语音段和噪声段分别计算累积直方图的均衡方法[7], 但这种方法需要一个 VAD 算法来很好的区分语音段和噪声段, 有时这并不容易做到; 将直方图均衡作为特征矢量自适应变换方法的实验也取得了相当好的效果[8,9]. 直方图均衡方法还可以用于非线性无监督自适应方法[9].

除此之外, 在一定的假定下, 许多参数化的特征参数规整方法也都可以用 CDF-Matching 原理来解释, 例如倒谱均值规整 (Cepstral mean normalization, CMN), 倒谱均值方差规整 (Cepstral mean and variance normalization, CMVN) 等. 这类方法往往不是直接地以非参数的方法估计累积分布函数, 而往往是先假设特征参数的概率分布的形式 (对 CMN 是常均值偏移模型, 对 CMVN 是单高斯分布模型), 然后用数据来估计分布的参数, 最后进行规整, 这也可以达到缩小概率密度函数的差别的目的.

但是, 有分析表明, 噪声环境下的语音特征参数通常呈现双峰分布[7], 而 CMVN 等简单的参数化方法无法很好的表达这种双峰结构, 直方图均衡则是非参数化的方法. 而我们则期望能找到一种能够更细致的表达特征分布 (特别是双峰结构) 的参数化方法. 为此, 我们提出了一种基于双高斯 GMM 的特征参数归一化方法.

本文的内容是这样安排的. 在第 2 节中, 将阐述特征参数规整的基本原理 — 累积分布函数匹配原理, 并简单介绍 CMN、CMVN 和直方图均衡方法. 第 3 节中, 将具体说明我们提出的双高斯归一化方法的原理及其实现过程. 在第 4 节中, 将比较这三种方法在 Aurora2 以及 Aurora3 数据集上的结果. 最后, 我们将在第 5 节中给出本文结论.

2 累积分布函数匹配原理

语音识别所用参数, 比如 MFCC, 本身都是随机矢量, 因而具有相应的概率分布, 训练和识别的不匹配也就体现在概率分布的差别上. 受实际环境的影响, 特征参数的概率分

布往往发生改变. 这时，一个很自然的想法就是对特征参数进行规整，使得训练和识别时候的特征参数的概率分布比较接近，这样两者之间不匹配的问题就应该能得到改善[1,7].

虽然特征参数的概率密度函数匹配应该是我们最直接的目标，但由于对它的估计既不方便也很难准确，所以我们一般还是通过概率密度函数的积分 — 累积分布函数 (Cumulative distribution function, CDF) 来表述概率分布匹配原理.

根据这个原理，特征参数变换函数可以由数据的累积分布函数获得，如下

设特征参数变换函数为 $x = T[y]$, y 是规整前的特征参数，x 是规整变换后的特征参数;

再设 x 的累积分布函数为 $C_X(x)$, y 的累积分布函数是 $C_Y(y)$, 则特征参数变换函数应该使得

$$C_Y(y) = C_X(x) \tag{1}$$

由此可以得到

$$x = T[y] = C_X^{-1}(C_Y(y)) \tag{2}$$

上述方法也被称为参数补偿，实际应用当中，为了简化算法实现过程，经常把训练和测试的数据概率分布都变到同一个事先给定的标准分布 (通常是标准高斯分布), 这称作参数规整.

2.1　直方图均衡方法 (Histogram equalization)

直方图均衡是 CDF Matching 原理的一种最直接的实现. 做法如下：

首先是利用各维独立假设，这样各维可分别处理. 然后将整个训练集参数的可能取值范围等分为 M 个不相重叠的小区间 (bins), M 的数值一般在几十到几百左右. 之后统计落在每个小区间中的参数个数，由此可以得到特征参数的累积直方图，并用于逼近真实的累积分布函数.

2.2　参数化特征参数规整方法

前面已经提到，还有许多参数化的特征参数规整方法，尽管他们最初的出发点也许并不是 CDF-matching 原理，但在满足一些假定的条件下，它们也可以用 CDF-matching 原理来表达. 常用的 CMN 和 CMVN 方法都可以归到这一类中.

2.2.1　倒谱均值相减 (CMN)

倒谱均值归一化 (Cepstral mean normalization, CMN) 方法，有时也称为 CMS, 是一种非常常用的特征参数规整方法. 当假设训练和识别的参数概率密度分布相同，只是相差一个常数时，可以通过减去均值来去掉这个常数偏移的影响.

从 CDF-matching 的观点来看, CMN 是将变换函数 $x = T[y]$ 简化为 $x = y - C$, C 在这里是一个常数. 这时候，两者的 CDF 也是在横轴上做一个平移. 所以，CMN 在这种条件下就完全等效于 CDF-matching.

2.2.2　倒谱均值方差规整 (CMVN)

倒谱均值方差规整 (Cepstral mean and variance normalization, CMVN) 是一种对倒谱特征参数的均值和方差都做规整的方法. 我们也可以把 CMVN 看作是累积分布函数匹配原理应用的一个特例：当特征参数符合高斯分布时，累积分布函数原理等价于 CMVN 方法. CMVN 方法中仅仅用到均值和方差两个参数，只能较好地表示单峰分布. 但是，如果特征参数分布比较复杂的话，仅仅用两个参数来描述是不够充分的.

有分析表明，噪声环境下的语音特征参数通常呈现双峰分布[7]. 而无论是 CMN 和 CMVN 方法无法很好的表达这种双峰结构，直方图均衡方法则通常是应用非参数方法来表达特征参数的概率分布. 为此，我们提出了一种基于双高斯 GMM 的特征参数归一化方法，以达到很好地用参数化方法表达特征参数的双峰结构.

3 双高斯特征参数规整方法

为了更精细的表达特征的概率分布，一个常用的方法是用混合高斯模型 (GMM) 来逼近复杂的概率密度分布. 一般来说，所使用的混合高斯数越多，对真实概率密度分布的逼近程度越好，但是估计参数时所需数据量也越大，运算量也越大. 本文中采用的混合高斯数为 2.

3.1 方法原理

如果 $\boldsymbol{y}$ 的概率密度分布可以用下面的 GMM 表示

$$p(\boldsymbol{y}) = \sum_{k=1}^{K} c_k N(\boldsymbol{y}; \boldsymbol{\mu}_k, \Psi_k) \tag{3}$$

那么 $\boldsymbol{y}$ 的累积分布函数为

$$C(\boldsymbol{y}) = \sum_{k=1}^{K} c_k F[(\boldsymbol{y} - \boldsymbol{\mu}_k)\Psi_k^{-1}] \tag{4}$$

上式中，$\boldsymbol{y}, \boldsymbol{\mu}_k, \Psi_k$ 分别是特征参数行向量，均值行向量，对角协方差矩阵；函数 $F(t)$ 是标准高斯分布的累积分布函数，但是这个函数并没有解析表达式，所以我们用查表法来近似实现 $F(t)$ 及其反函数.

3.2 模型参数估计

在获得语音特征参数后，需要估计模型的参数. 我们采用了 EM 算法[10], 该算法适用于含有隐含变量的模型参数估计问题. 而 GMM 的参数估计就属于这类问题.

目前我们的算法是建立在特征参数各维独立假设之上的，在计算 GMM 模型参数和进行规整的时候，认为各维相互独立，互不相关. 这样做虽然简单，但是也存在一些问题，因为实际上特征参数都是多维的，做参数映射的时候各维之间也有影响. 为把这种影响考虑进来，我们使用了多维的 GMM, 通过 EM 算法在多维空间中聚成两个 (或者多个) 高斯分布.

如果一个双高斯的 GMM 模型如下所示

$$p(\boldsymbol{y}|\phi) = \sum_{k=1}^{2} c_k p_k(\boldsymbol{y}|\phi_k) = \sum_{k=1}^{2} c_k N_k(\boldsymbol{y}|\boldsymbol{\mu}_k, \Psi_k) \tag{5}$$

给定数据序列 $Y = (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_N)$ 后，得到

$$\gamma_k^i = [c_k p_k(\boldsymbol{y}_i|\phi_k)]/[p(\boldsymbol{y}_i|\phi)] \tag{6}$$

$$\gamma_k = \sum_{i=1}^{N} \gamma_k^i \tag{7}$$

那么经过一次迭代后的 GMM 的各个高斯的权重，均值，方差分别如下所示：

$$c_k' = \gamma_k / N \tag{8}$$

$$\mu_k' = [\sum_{i=1}^{N} \gamma_k^i \boldsymbol{y}_i]/\gamma_k \tag{9}$$

$$\boldsymbol{\psi}_k = [\sum_{i=1}^{N} \gamma_k^i (\boldsymbol{y}_i - \boldsymbol{\mu}_i)(\boldsymbol{y}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}]/\gamma_k \tag{10}$$

因为采用了多维 GMM, 各维特征参数将共享同一个权重. 另外, 为了实现上的简单, 我们的规整算法是基于每一句话的, 因此可用于估计模型参数的特征参数帧数并不太多. 为求得可靠的参数, 通常需要进行几次迭代.

4　语音识别实验

4.1　实验数据和设置

我们的比较实验是在 Aurora2 数据库和 Aurora3 数据库上进行的.

Aurora2 是人工加入噪声和信道影响的 TI 数字串数据库. 它给定了一个在规定的训练和测试数据集上, 以及识别后端的具有互比性的数字串识别系统, 它的目的就是在完全相同的训练和测试条件下, 比较不同的前端鲁棒性参数方法的效果.

Aurora3 数据库是欧洲语言车载语音数据库 SpeechDataCar 的一个子集, 共包含四种欧洲语言的实际车载环境数字串语音数据文件. 这个数据库主要用于车载环境下噪声鲁棒性方法的评测.

4.2　实验结果与讨论

我们做了以下比较实验.

首先, 我们实现了 CMN 和 CMVN 方法. 实验所采用的特征参数为 MFCC 和 logE (对数能量). 对每一句话, 分别计算出特征参数每一维的均值和方差, 然后将均值减去 (CMN 方法) 或将其归一化到标准高斯分布 (CMVN 方法).

之后, 我们实现了双高斯的特征参数规整方法. 实验所采用的特征参数是 MFCC 和 logE. 对 Aurora2 训练集和测试集中的每一句话, 在提取出 MFCC 参数后, 用 EM 算法迭代计算出双高斯模型的参数, 实验中采用的 EM 迭代次数为 3 次 (实验表明, 3 次迭代较为合适, 更多次迭代没有明显改善); 然后根据这个模型将特征参数的分布归一化到标准高斯分布; 在用归一化后的训练集参数训练得到声学模型后, 对归一化后的测试集数据进行测试, 得到测试结果.

直方图均衡方法虽然属于非参数化特征规整方法, 但与前两种方法也具有可比性, 因而我们也实现了直方图均衡方法以便比较.

表 1 和表 2 分别是在 Aurora2 和 Aurora3 任务上三种方法及基线系统 (基于 MFCC) 的实验结果. 从表中可以得出如下一些结论: 首先, 三种方法均较基线系统有较为明显的改善; 其次, 同为参数化方法的 CMVN 方法和我们所提出的双高斯特征参数规整方法相比, 后者要明显优于前者; 最后, 我们的方法的性能与直方图均衡方法基本相当, 在 Aurora2 的 Clean-Condition 和 Aurora3 的 Well-Matched 及 Medium-Mismatch 情况下还要稍好一些.

表 1　各种方法在 Aurora2 任务上的性能比较

Table 1　Comparison of performance on Aurora2 between different methods

方法	Clean-condition	Multi-condition	总计 (Overall)	
	词错误率 (WER)	词错误率 (WER)	词错误率 (WER)	相对错误率下降 (ERR)
基线 (Baseline)	13.61%	39.94%	26.77%	0%
CMN	11.94%	32.23%	22.08%	15.78%
CMVN	11.68%	30.35%	21.02%	19.08%
直方图均衡	10.35%	19.08%	14.72%	38.08%
双高斯特征规整	10.07%	21.34%	15.70%	36.31%

表 2 各种方法在 Aurora3 任务上的性能比较
Table 2 Comparison of performance on Aurora3 between different methods

方法	Well matched WER(%)	Medium mismatch WER(%)	High mismatch WER(%)	总计 (Overall)	
				WER(%)	ERR(%)
基线 (Baseline)	8.96	21.96	48.85	23.48	0
CMN	7.76	17.01	32.73	17.24	20.92
CMVN	7.22	15.24	35.76	17.16	25.71
直方图均衡	6.72	14.41	23.03	13.49	35.81
双高斯特征规整	6.57	14.21	26.68	14.27	34.99

另外，我们也对采用更多高斯数 (三高斯) 的特征参数规整方法进行了初步的尝试，在 Aurora2 上仅获得 34.72% 的相对错误率下降．相对于双高斯情况，并无性能提升．

5 结论

鲁棒性语音识别面对的主要就是训练和识别环境失配的问题．而语音参数本质上都是随机变量 (矢量), 具有某种概率分布，失配的问题也就会体现在概率分布的差异上．一个很自然的想法是对这些随机变量进行某种变换，以使得训练和识别所用的参数的概率分布尽可能接近，这样失配的影响将减小，系统的鲁棒性将会得到提高．这就是基于 CDF-matching 的方法的基本思想．

在 CDF-matching 的框架下，由于具体实现方法的差异，这一类方法又可以分为两大类：作为非参数方法的直方图均衡方法，作为参数化方法的多高斯拟合的方法 (CMN 和 CMVN 等方法其实可以看做这一方法在一定条件下的特例)．

我们提出了一种基于双高斯 GMM 模型，并使用 EM 算法来估计模型参数的语音特征参数归一化方法．与传统的 CMN 和 CMVN 等参数方法相比，我们的方法能够更加精细的表现带噪语音特征参数的特点．在 Aurora2 及 Aurora3 任务上的实验表明，我们的方法在总体性能与直方图均衡方法 (非参数方法) 持平，而要明显优于 CMN 以及 CMVN 方法．我们认为这是因为双高斯的 GMM 模型可以更好的表达带噪语音特征参数的双峰结构，从而能够更好的对特征参数归一化，使得训练和识别的不匹配能够进一步减小．

与双高斯相比，更多高斯并未带来性能提升．这可能有两个原因，一是我们的规整是基于句子的，因此数据可能不足以用来充分估计 GMM 参数；另外，特征参数映射方法需要获取的是不同环境间特征参数的关系，而应该排除掉语音信息等非环境因素对特征参数带来的影响，因此太过精细的映射模型也不一定就能带来性能增益．解决这些问题将是后继研究的方向之一．

References

1 de la Torre A, Segura J C, Benitez C. Non-linear transformations of the feature space for robust speech recognition. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2002. Piscataway, USA: IEEE Press, 2002. 401∼404

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

3 Hilger F, Molau S, Ney H. Quantile based histogram equalization for online application. In: Proceedings of International Conference of Spoken Language Processing 2002. Rundle Mall, Australia: Causal Productions, 2002. 237∼240

4 Hilger F, Molau S, Ney H. Evaluation of quantile based histogram equalization with filter combination on the Aurora 3 and 4 Databases. In: Proceedings of European Conference on Speech Communication and Technology. Grenoble, France: ISCA, 2003. 341∼344

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. In: 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. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2002. Piscataway, USA: IEEE Press, 2002. 409∼412
7 Molau S, Hilger F, Keysers D, Ney H. Enhanced histogram normalization in the acoustic feature space. In: Proceedings of International Conference of Spoken Language Processing 2002. Rundle Mall, Australia: Causal Productions, 2002. 1421∼1424
8 Molau S, Hilger F, Ney H. Feature space normalization in adverse acoustic conditions. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing 2003. Piscataway, USA: IEEE Press, 2003. 656∼659
9 Dharanipragada S, Padmanabhan M. A nonlinear unsupervised adaptation technique for speech recognition. In: Proceedings of International Conference of Spoken Language Processing 2000. Beijing, China: Institute of Acoustics, 2000. 556∼559
10 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data *via* the EM algorithm. *Journal of the Royal Statistical Society*, 1977, **39**(1): 1∼38
11 Chen C P, Bilmes J, Kirchhoff K. Low-resource noise-robust feature post-processing on AURORA2.0. In: Proceedings of International Conference of Spoken Language Processing 2002. Rundle Mall, Australia: Causal Productions, 2002. 2445∼2448

刘　波　硕士研究生，研究方向为鲁棒性语音识别.
(**LIU Bo**　Master student. His research interest includes robust speech recognition.)
戴礼荣　博士，副教授，研究方向：语音信号处理、语音通信、及 DSP 技术应用.
(**DAI Li-Rong**　Ph. D., associate professor. His research interests include speech signal processing, speech communication, and the application of DSP technology.)
王仁华　教授，博士生导师，从事数字信号处理、语音通信、多媒体通信等方面的研究.
(**WANG Ren-Hua**　Professor. His research interests include digital signal processing, speech communication, and multimedia communication.)
杜　俊　硕士研究生，研究方向为鲁棒性语音识别.
(**DU Jun**　Master student. His research interest includes robust speech recognition.)
李锦宇　硕士，研究领域为低比特率语音编码，语音识别.
(**LI Jin-Yu**　Master. His research interests include low-bit rate speech coding and speech recognition.)