似然得分归一化及其在与文本无关说话人确认中的应用

邓浩江 杜利民 万洪杰
(中国科学院声学研究所语音交互技术研究中心 北京 100080)

摘 要：该文研究了似然得分归一化方法的原理，建立了基于自适应 GMM 模型的说话人确认系统，并将非特定人的背景模型与特定人的 cohort 模型相结合，提出了混合归一化的方法。在电话语音条件下，该文比较了不同得分归一化方法对确认系统性能的影响。实验表明，在自适应 GMM 模型似然比得分的基础上，T-cohort 与通用背景模型混合归一化能获得最佳识别效果。当错误拒绝率为 5%时，该方法可以获得 0.5%的错误接受率，远远低于采用通用背景模型归一化方法的 2%。

关键词：说话人确认，高斯混合模型，得分归一化，与文本无关
中图分类号：TP391.42 **文献标识码**：A **文章编号**：1009-5896(2005)07-1025-05

Likelihood Score Normalization and Its Application in Text-Independent Speaker Verification

Deng Hao-jiang Du Li-min Wan Hong-jie
(*SITR, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

Abstract In this paper, the methodology of likelihood score normalization is studied. The text-independent speaker recognition system based on the adapted Gaussion Mixture Models(GMMs) is established, and the approach to normalize scores combining speaker-independent background model and the speaker-dependent models of cohort speaker sets are proposed. The speaker verification experiments over telephone channels show that based on the likelihood ratio of adapted GMMs system, both cohort normalization and hybrid score normalization approaches can improve the verification performance of baseline system using Universal Background Model (UBM). Specially, the hybrid approach combining UBM and cohort models selected during testing (T-cohort normalization) achieve the best performance. At a miss probability of 5%, the hybrid approach using UBM and T-cohort models reduce the false alarm rate to 0.5% compared to 2% for the baseline.

Key words Speaker verification, Gaussian mixture model, Score normalization, Text-independent

1 引言

随着数字时代的来临，为保证个人信息和私人交易的正确访问，保证计算机和通信网络的安全，个人身份认证、安全密码口令将成为现代生活不可分割的一部分。为了进一步增强安全性，提高识别精度，许多生物特征，如指纹、签名、手形、眼睛虹膜和声音等，都被用作识别的对象。在这些生物特征中，人的声音与其它主要基于图像的识别对象不同，它的产生、捕获以及通过电话在网络上传输都非常容易，对于个人身份认证是最便利的。

说话人确认(Speaker verification)是根据语音波形中所包含的说话人生理和行为特征参数，由机器自动判断说话人是否是所声言人。语音中包含了多种信息，不同信息所代表的特征是不同的。近年来，说话人确认系统所使用的特征主要还是基于倒谱系数的声学特征，在此特征基础上的主流的识别方法是基于概率统计的识别方法[1]。

在现实生活中，通过电话信道进行语音传输是非常普遍的，但电话语音的带宽比较窄，且存在信道干扰和各种背景噪声的影响；语音受时间、环境、身体状况以及说话的内容等条件的影响变化比较大。实际语音交互中的复杂声学环境和个人语音的变化，使得用于训练说话人模型的数据仅仅代表应用过程中所有声学条件下的很小一部分，训练和测试样本之间的不匹配是系统性能下降的主要原因。为了减少话筒、信道等环境因素的影响，可以采用信道、话筒的归一化技术[2,3]；为了在确认过程中采用统一阈值，可以调整不同说话人的似然值分布的参照系，二者都可以归结为似然得分的归一化。得分归一化(Score normalization)技术可以说是当前

2004-02-23 收到，2004-07-05 改回

基于统计的说话人确认系统的一个重要组成部分。本文研究的就是在电话语音条件下，基于统计模型的说话人识别系统的鲁棒性增强技术——得分归一化，本文采用的统计模型是自适应的高斯混合模型(Gaussian Mixture Models，GMM)。

2 似然比检验

说话人识别时，给定一句话 U，对于统计模式识别方法需要估计出参考说话人模型 λ，对于这句话的概率似然值(likelihood) $p(\lambda|U)$。假设每一个参考说话人都是等概率出现，即 $p(\lambda_i)$ 都相等。根据贝叶斯(Bayes)理论，说话人统计模型的一般输出可以用该语音段在统计意义上对于参考说话人模型 λ_i 的似然概率 $p(U|\lambda_i)$ 与对于全体说话人 λ 似然概率 $p(U|\lambda)$ 的比值来近似，

$$P(\lambda_i|U)=\frac{P(U|\lambda_i)P(\lambda_i)}{\sum_{i=1}^{n}P(U|\lambda_i)P(\lambda_i)}=\frac{P(U|\lambda_i)}{P(U|\lambda)} \tag{1}$$

说话人确认是一个二元检测问题，为了获得最佳的识别效果，确认过程中广泛采用了似然比检验的方法[4-6]。将式(1)中全体说话人用非说话人的统计模型 $\bar{\lambda}$ 来表示。给定待测语音段 U 的特征矢量序列 $\boldsymbol{X}$，在对数域似然比，即归一化似然值 $S^{(n)}(\boldsymbol{X})$ 可以表示为

$$S_i^{(n)}(\boldsymbol{X})=\log\left[p\left(\boldsymbol{X}|\lambda_i\right)\right]-\log\left[p\left(\boldsymbol{X}|\bar{\lambda}\right)\right] \tag{2}$$

非说话人模型 $\bar{\lambda}$ 的建模有两种主要方式。一种是使用一系列其他说话人的模型覆盖可能的冒名顶替者的特征空间，如 cohorts[7,8]，似然比集合(likelihood ratio sets)[9]以及背景说话人(background speakers)[10]等，本文统称为背景说话人。另一种建模方法是将若干说话人的语音数据组合到一起训练一个单独的背景模型，即通用模型[10,4]。

3 似然得分混合归一化

非特定人的通用模型有助于减少话筒、信道等背景因素的影响，而特定人的背景说话人集合有助于减少一些异常因素的影响，降低错误拒绝率。下面讨论基于自适应 GMM 的通用背景模型(Universal Background Model，UBM)归一化和特定人的 cohort 归一化方法，然后结合二者的优点提出混合归一化的方法。

3.1 Cohort 归一化

对于特定的背景集合，必须选择参考说话人 i 的背景说话人集合的 M 个模型，那么，待测语音段的特征矢量序列 $\boldsymbol{X}$ 不是说话人 i 的似然得分，是其对于 M 个背景说话人模型得分的函数组合，一般是求平均或求最大值。背景说话人的选择可以是随机的，也可以按照某种相似度的原则选择最相似的或最不相似的或二者的组合。Cohort 归一化方法是选择最具竞争力的说话人(cohort)，即背景说话人的模型与参考说话人模型最相近。然后用 $\boldsymbol{X}$ 对于 cohort 说话人集合的似然平均作为 $\bar{\lambda}$ 的似然得分(式(2)右边第 1 项)：

$$S_i^{(c)}(\boldsymbol{X})=\log p\left(\boldsymbol{X}|\bar{\lambda}_i\right)=\log\sum_{j=1}^{M}\exp\left\{\log\left[p\left(\boldsymbol{X}|\lambda_{ij}^{(c)}\right)\right]\right\} \tag{3}$$

其中，$\lambda_{ij}^{(c)}(j=1,2,\cdots,M)$ 是第 i 个参考说话人的 cohort 模型集合，M 是 cohort 数目。实验中，我们采用了对称的相似度[10]计算公式：

$$d(i,j)=\frac{\log\left[p\left(\boldsymbol{X}_i|\lambda_i\right)\right]}{\log\left[p\left(\boldsymbol{X}_j|\lambda_i\right)\right]}\frac{\log\left[p\left(\boldsymbol{X}_j|\lambda_j\right)\right]}{\log\left[p\left(\boldsymbol{X}_i|\lambda_j\right)\right]},\quad j\neq i \tag{4}$$

选择与参考说话人 i 最相近的 M 个人组成 cohort 集合。

在识别时，参考说话人的 cohort 集合一般是固定的，即在训练时就确定下来了。这时，当某个冒名者的测试语句与参考说话人及其 cohort 集合的模型的似然值几乎相同时，就可能出现较高的归一化得分从而导致冒名者的错误接受。在识别阶段选择 cohort 集合的方法可以解决这个问题[6]，该方法计算待测特征矢量序列与一系列候选 cohort 说话人模型的似然得分，然后从中选择与参考说话人相似的 cohort 集合。为了以示区别，我们称这种方法为测试时 cohort(T-cohort)归一化。

3.2 自适应 GMM 模型

GMM 模型[11]是用多元混合高斯概率密度描述说话人语音特征矢量在特征空间的概率分布。GMM 模型既可以用最大似然 EM 迭代算法直接估计出模型的参数[11]，也可以用最大后验贝叶斯学习算法由一个通用背景模型自适应而得到[4,10]。后者将说话人的模型和背景模型有机地结合在一起，可以在训练和识别的过程中简化计算的复杂度。

在基于自适应 GMM 模型的识别系统中，UBM 由一个大型的 GMM 模型构成，而说话人的模型是与 UBM 相同规模的 GMM 模型，它利用说话人的训练语音数据对 UBM 中的每一个混合高斯项的数学期望、一阶矩以及二阶矩等统计项进行充分的估计。然后将新估计出来的统计项与通用模型中充分估计的旧参数进行线性组合就可以得到说话人模型的自适应参数。

对于测试特征矢量序列 $\boldsymbol{X}$，一般采用一种快速的方法计算自适应 GMM 模型的似然比得分：首先计算并确定 $\boldsymbol{x}_t$ 对于通用背景模型 $\lambda^{(b)}$ 的得分贡献最大的前 N 个高斯混合项，并只用这 N 个混合项计算 $\boldsymbol{x}$ 对于 $\lambda^{(b)}$ 的似然值 $p\left(\boldsymbol{x}_t|\lambda^{(b)}\right)$；然后只用说话人模型 λ_i 中与 $\lambda^{(b)}$ 相对应的 N 个混合项计算

$p(\boldsymbol{x}_t|\lambda_i)$；最后计算对数似然比的平均。

3.3 混合归一化

在自适应 GMM 模型似然比得分的基础上，我们采用了 cohort 归一化和 T-cohort 归一化的方法，并将非特定人的背景模型与特定人的 cohort 模型相结合，提出了混合归一化的方法。Cohort 和 T-cohort 归一化时，无论是参考说话人模型的得分还是 cohort 说话人模型的得分，都只计算与通用背景模型相对应的 N 个贡献最大的混合项的输出，而混合归一化则按下面的公式计算：

$$S_i^{(n)}(\boldsymbol{X}) = \log\left[p\left(\boldsymbol{X}|\lambda_i\right)\right] - \eta\log\left[p\left(\boldsymbol{X}|\lambda^{(b)}\right)\right] - (1-\eta)\log\left[p\left(\boldsymbol{X}|\lambda_i^{(c)}\right)\right] \tag{5}$$

其中，$\eta(0\le\eta<1)$是混合因子，用于控制通用背景模型和 cohort 模型在对数归一化得分中所占的比例。公式的前两项与自适应 GMM 模型的似然比计算相似，但需要注意的是，必须分别计算特征矢量序列 $\boldsymbol{X}$ 对于说话人模型 λ_i 中与 $\lambda^{(b)}$ 对数似然值的平均，在按照式(5)计算它们的差分。公式的第 3 项按式(4)计算，cohort 说话人模型的得分，只计算与通用背景模型相对应的 N 个贡献最大的混合项的输出。混合因子 η 的设定依赖于通用背景模型和 cohort 模型对可能出现的非说话人特征空间的覆盖程度。

4 说话人实验及其结果

在电话语音条件下，针对不同的似然得分归一化方法，我们进行了与文本无关的说话人确认实验。

4.1 语音数据库和基线系统

实验中使用的电话语音数据库是中国科学院声学研究所语音交互技术研究中心制作的 SITR 固定和移动电话语音数据库。该数据库通过播放多话者、连续、纯净、汉语语音数据库语音，并使之通过多种话筒与公用电话交换网(PSTN)和构成的多个电话语音信道录音，从而获得的多话筒、非特定人、连续电话语音数据库。

在电话语音条件下的说话人确认实验中，我们采用了基于自适应 GMM 模型的识别系统。背景模型和说话人模型的训练语音数据以及测试语音数据都来自于 SITR 电话语音数据库。其中，我们从语音库中选择 80 个说话人(男 40 人，女 40 人)，其中 50 人(男 25 人，女 25 人)是固定电话录音，30 人(男 15 人，女 15 人)是移动电话录音，每人随机选取 25 句话用于 UBM 的训练，训练语音的时长(除去静音与噪音段)大约为 1.88h。UBM 是 1024 个混合高斯项的 GMM 模型，采用 LBG 矢量量化的方法初始化参数。用户的说话人模型由 UBM 自适应得到，实验中仅对高斯混合项的均值自适应。

我们选择另外 40 人(男 20 人，女 20 人)作为说话人确认系统的用户。训练语音来自于相同的 40 个人，只是所用话筒和话筒的组合不同，它们分别是

(1) 40 个用户每人 20 句话，固定电话，每个人的语音数据来自于相同话筒。每个人使用的话筒为不同品牌的驻极体式电容话筒，标记为 S1h。

(2) 40 个用户每人 20 句话，固定电话，每个人的语音数据来自于相同话筒，但与 S1h 中同一人的话筒不同。每个人使用的话筒为不同品牌的驻极体式电容话筒，标记为 S2h。

(3) 40 个用户每人 20 句话，固定电话，分别来自于两种话筒(与 S1h, S2h)相同，每种话筒语音各 10 句话，相当于(S1h +S2h)/2。每个人使用的话筒为不同品牌的驻极体式电容话筒，标记为 S3h。

测试用的语音数据分成以下几个部分，它们的不同组合用于不同目的实验。

(1) 40 个用户每人 5 句话，固定电话，话筒与训练语料 S1h 相同。标记为 T1h。

(2) 40 个用户每人 5 句话，固定电话，话筒与训练语料 S2h 相同。标记为 T2h。

(3) 60 个冒名者(男 30 人，女 30 人，不同于 T1h、T2h)每人 5 句话，来自于相同话筒。每个人使用不同的话筒，固定电话和移动电话各 30 人，标记为 T3h。

在确认实验中，测试以一句话为单位，计算每一句话的特征矢量序列 $\boldsymbol{X}$ 对于系统的归一化似然值。每一测试语句均对系统所有用户进行测试，对于闭集(T1h，T2h)测试，每一测试语句作为真实说话人测试一次，作为冒名入侵者测试 $L-1$ 次(L 为系统用户的数目，实验中 $L=40$)；对于开集(T1h，T2h，T3h)测试，冒名者的每一测试语句均对系统所有用户模型进行冒名入侵测试 L 次。系统性能的测试结果用 DET (Detection Error Tradeoff)[12]曲线表示，其横坐标代表误警(False alarm)率，即错误接受率，纵坐标代表漏警(Miss)率，即错误拒绝率。

实验中，我们首先对语音信号进行端点检测，去除信号中的静音段和噪音段，然后进行预加重，再用汉明窗提取 MFCC 特征矢量：采用 20 通道滤波器组，12 阶 mel 倒谱，12 阶 Δ-mel 倒谱，构成 24 维特征向量。其中，汉明窗帧长为 25s，帧移为 10s。倒谱分析的 mel-滤波器组的覆盖带宽限定在 300－3400Hz 之间，Δ-倒谱系数采用一阶正交多项式拟合的方法计算，拟合的长度包含当前矢量及其前后各两个矢量。最后，用倒谱均值归一化的方法消除话筒和电话传输

信道所带来的频率响应失真，即卷积噪声。

4.2 实验及结果

我们首先在不同话筒条件下，对话筒的匹配问题进行了闭集和开集实验，闭集测试的说话人确认系统性能比较如图 1 所示。

图中，似然值得分采用 T-cohorts 与背景模型混合归一化的方法，cohorts 人数为 5，混合因子 η 取值 0.5。实验表明，无论是闭集还是开集测试，话筒的不匹配都会使系统的性能下降很多。同样都是 20 句话，用来自两种话筒的语音数据训练的说话人确认系统的性能就比单话筒提高很多，如闭集实验的等差错率由大于 6%左右降到 1%左右。

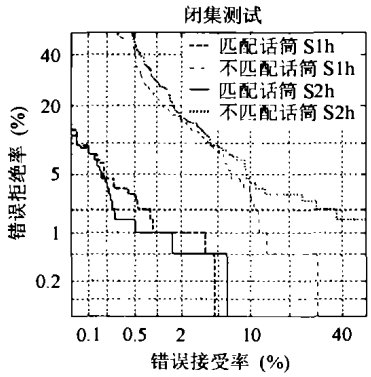

图 1 不同话筒条件下闭集测试的 DET 曲线

匹配话筒 S1h——用 S1h 训练，用 T1h 测试；

匹配话筒 S2h——用 S2h 训练，用 T2h 测试；

不匹配话筒 S1h——用 S1h 训练，用 T2h 测试；

不匹配话筒 S2h——用 S2h 训练，用 T1h 测试；

然后针对各种似然值得分归一化方法及其参数的不同设定，我们进行了实验比较研究，并得出了所能获得的最佳系统性能及其参数设置。采用最佳参数的各种归一化方法的系统性能比较如图 2 所示。其中，用户的说话人模型用 S3h 训练，用 T1h+T2h+T3h 测试，‘bkgd’表示采用非特定人的背景模型归一化；‘cohort 15’表示采用特定人的 cohorts 归一化，cohorts 人数为 15；‘bkgd +cohort 15’表示采用非特定人的背景模型和 cohorts 模型混合归一化，cohorts 人数为 15；‘T-cohort 3’表示采用 T-cohorts 归一化，cohorts 人数为 3；‘bkgd+T-cohort 1’表示采用背景模型与 T-cohorts 模型混合归一化，cohorts 人数为 1。cohort 数目都是相应的归一化方法获得最佳性能时的 cohort 人数，混合因子 η 取值 0.5。

可以看出，在自适应 GMM 模型似然比得分的基础上，无论是 cohort，T-cohort 归一化的方法，还是混合归一化的方法都能获得比通用背景模型归一化更好的确认效果。当错误拒绝率为 5%时，采用通用背景模型归一化能获得 2%的错误接受率，而 T-cohort 归一化，T-cohort 与通用背景模型混合归一化方法分别可以获得 0.65%和 0.5%的错误接受率，其中，T-cohort 与通用背景模型混合归一化的方法效果最好。当获得最佳性能时，T-cohort 归一化所需的 cohort 人数为 3，而 T-cohort 与通用背景模型混合归一化所需的 cohort 人数为 1；远远少于 cohort 归一化，cohort 与通用背景模型混合归一化方法的 15 人左右，这是因为在训练时选择的 cohort 集合仍然受到训练数据与测试数据不匹配的影响，通过增加 cohort 的人数可以减少这种影响。

最后我们针对 T-cohort 与通用背景模型混合归一化采用不同的混合因子 η 进行了实验，仍然用 S3h 训练用户的说话人模型，用 T1h+T2h+T3h 测试，结果如图 3 所示。其中，η = 1 相当于仅采用通用背景模型归一化；η = 0 相当于仅采用 T-cohort 归一化，cohort 数目为 1。可见，η 取 0.2~0.6 之间是比较适合的。

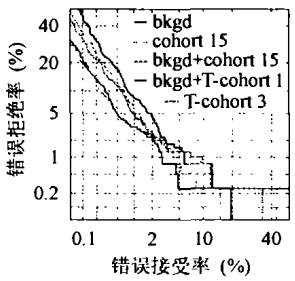

图 2 采用不同似然得分归一化方法的说话人确认系统性能比较

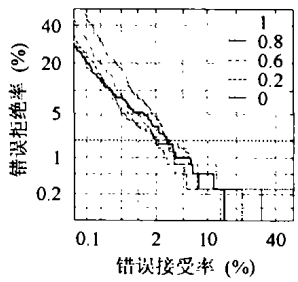

图 3 混合归一化方法混合因子 η 对说话人确认系统性能的影响

5 结束语

本文在基于自适应 GMM 模型的说话人识别系统的基础上，研究了特定人和非特定人的归一化方法，并将通用背景模型与特定人的 cohort 模型相结合，提出了混合归一化的方法。在电话语音条件下，针对不同似然得分归一化方法的与文本无关的说话人确认实验表明，在自适应 GMM 模型似然比得分的基础上，无论是 cohort，T-cohort 归一化的方法，还是混合归一化的方法都能获得比通用背景模型归一化更好的确认效果。其中，T-cohort 与通用背景模型混合归一化的方法效果最好，当错误拒绝率为 5%时， T-cohort 与通用背景模型混合归一化方法可以获得 0.5%的错误接受率，远远低于采用通用背景模型归一化方法的 2%。当获得最佳性能时，T-cohort 归一化，T-cohort 与通用背景模型混合归一化所需的 cohort 人数都远远少于 cohort 归一化，cohort 与通用背景模型混合归一化方法的人数。

参 考 文 献

[1] Doddington G R, Przybocki M A, Martin A F, Reynolds D A. The NIST speaker recognition – Overview, methodology, systems,

results, perspective. *Speech Communication*, 2000, 31(2 − 3): 225 − 254.

[2] Reynolds D A. The effects of handset variability on speaker recognition performance: experiments on the switchboard corpus. In: Proc. ICASSP-1996, Atlanta, USA, May 1996: I13 − 116.

[3] Heck L P, Weintraub M. Handset dependent background models for robust text-independent speaker recognition. In: Proc. ICASSP-1992, Munich, Germany, 1997: 1071 − 1074.

[4] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, 10(1 − 3): 19 − 40.

[5] Dunn R B, Reynolds D A, Quatieri T F. Approaches to speaker detection and tracking in conversational speech. *Digital Signal Processing*, 2000, 10(1 − 3): 93 − 112.

[6] Ariyaeeinia A M, Sivakumaran P. Analysis and comparison of score normalization methods for text-dependent speaker verification. In Proc. EUROSPEECH'97, Rhodes, Greece, 1997: 1379 − 1382.

[7] Rosenberg A E, *et al*.. The use of cohort normalized scores for speaker verification. In Proc. ICSLP-1992, Banff, Canada, Nov. 1992: 599 − 602.

[8] Colombi J, Ruck D, Rogers S, Oxley M, Anderson T. Cohort selection and word grammar effects for speaker recognition. In Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Atlanta, GA, 1996: 85 − 88.

[9] Higgins A, Bahler L, Porter J. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1991, 1(2): 89 − 106.

[10] Reynolds D A. Comparison of background normalization methods for text-independent speaker verification. In Proc. EUROSPEECH'97, Rhodes, Greece, 1997: 963 − 966.

[11] Reynolds D, Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech Audio Processing*, 1995, 3(1): 72 − 83.

[12] Martin A, *et al*.. The DET curve in assessment of detection task performance. In Proc. EUROSPEECH'97, Rhodos, Greece, 1997, 1895 − 1898.

邓浩江：　男，1971 年生，副研究员，博士，主要研究方向为语音信号处理、说话人识别以及神经网络信息处理.

杜利民：　男，1957 年生，博士，研究员，博士生导师，IEEE 高级会员，中国电子学会理事，研究领域为语音识别、自然语言理解、语音交互信息技术.

万洪杰：　男，1978 年生，博士生，从事说话人识别、语音信号处理方面的研究.