文章编号：2095-2163(2020)10-0019-05 中图分类号：TN912.34 文献标志码：A

基于得分规整的说话人确认

曹 伟，梁春燕
（山东理工大学 计算机科学与技术学院，山东 淄博 255049）

摘 要： 对于说话人确认系统，由于不同识别对得分分布的差异性，如果对原始得分直接使用统一门限判决，会严重影响系统的性能，因此需要对得分进行规整。针对现有得分规整方法的不足，本文提出一种基于对数似然值归一化（Log-likelihood Normalization，LLN）得分规整的说话人确认系统。LLN 在不需要先验知识的情况下，通过扩大同一测试语音在目标说话人模型与非目标说话人模型上的得分差距，使同一测试语音对应的两类识别对得分混叠现象得到有效改善，进而利于系统的区分判决。实验证明，LLN 是一种有效的得分规整方法，且与已有的零规整和测试规整形成互补，可进一步提高说话人确认系统的性能。

关键词： 说话人确认；得分规整；对数似然值归一化；零规整；测试规整

Speaker Verification Based on Score Normalization

CAO Wei, LIANG Chunyan
（College of Computer Science and Technology, Shandong University of Technology, Zibo Shandong 255049, China）

【Abstract】 For the speaker verification system, due to the difference in the score distribution of different trials, if the original score is directly judged by the unified threshold, the performance of the system will be seriously affected, so it is necessary to regulate the score. In view of the shortcomings of the existing score normalization methods, this paper proposes a speaker verification system based on log-likelihood normalization (LLN). LLN does not require prior knowledge, by expanding the gap between the scores of the same test speech on the target speaker model and the non-target speaker model, the overlap of two kinds of scores corresponding to the same test speech is reduced, which is conducive to the discrimination of the system. Experiments show that LLN is an effective score normalization method, and complements the existing zero normalization and test normalization, which can further improve the performance of speaker verification system.

【Key words】 Speaker verification; Score normalization; Log-likelihood normalization; Zero normalization; Test normalization

0 引 言

说话人识别，也称为声纹识别，是指利用语音波纹中所包含的信息自动识别说话人身份的技术[1]。由于语音获取方便，采集设备简单，并能通过网络远程识别，说话人识别正成为一种主要的生物特征识别手段[2]。

根据识别的目的不同，说话人识别可以分为说话人辨认（Speaker Identification）和说话人确认（Speaker Verification）两种方式[3]。说话人辨认是从给定说话人集合中找到与测试语音匹配的说话人；说话人确认是判断测试语音是否属于某个预先声明的说话人，即需要将测试识别对（由测试语音和其声明的说话人身份构成）作出“True”或“False”的二类判决。根据是否依赖于语音的内容，说话人识别可以分为与文本有关和与文本无关两种类型[4]。本文主要基于文本无关的说话人确认展开研究。

在说话人确认的测试阶段，不同识别对的得分分布存在着很大的差异性[5]，差异性主要来自以下方面：

（1）相同说话人的不一致性。由于受时间、健康状况、心理状态、录音条件等因素的影响，同一说话人的不同测试语音在目标说话人模型上的得分并不是一个恒定值，而是呈现某种概率分布。

（2）不同说话人之间的不一致性。由于受说话习惯、嗓音、语言等因素的影响，不同说话人模型对应的识别对得分表现出不一致性。有的说话人模型对应的识别对得分普遍偏高，有的说话人模型对应的识别对得分则相对偏低。

基金项目： 国家自然科学基金（11704229，61701286，61562068）；山东省自然科学基金（ZR2017LA011，ZR2015FL003，ZR2017MF047）；山东省高等学校科技计划项目（J17KA078）。

作者简介： 曹 伟（1993-），男，硕士研究生，主要研究方向：说话人识别；梁春燕（1986-），女，博士，讲师，主要研究方向：说话人识别、语种识别、说话人分段聚类。

通讯作者： 梁春燕 Email：liangchunyan_sdut@163.com

收稿日期： 2020-05-08

(3)不同测试语音间的不一致性。在时长、环境噪声、信道情况等影响下,不同测试语音对应的识别对得分也会表现出不一致性,比如有的测试语音对应的识别对得分普遍偏高,有的测试语音对应的识别对得分则偏低,而有的测试语音在目标说话人模型和非目标说话人模型上的得分比较接近,不容易区分。

综合以上方面的原因,如果将所有识别对的得分汇集在一起,"True"和"False"两类识别对的得分会出现严重的交叉和混叠;在这种情况下使用统一的门限对每一个识别对作"True"或"False"的判决,会严重影响说话人确认系统的性能[6]。因此,需要在识别对原始得分的基础上进行得分规整[7]。

目前最常用也是最典型的得分规整方法有零规整(Zero normalization, Znorm)、测试规整(Test normalization, Tnorm)以及二者的结合算法 ZTnorm 等,通过估计"False"识别对的得分分布,对测试识别对的得分进行规整,将"False"识别对的得分规整为均值为 0、方差为 1 的分布,从而消除不同说话人模型间的差异或不同测试语音之间的差异,有效减小两类识别对得分汇集后的混叠部分,从而提高说话人确认的系统性能。一般来说,得分规整不受限于系统所使用的说话人模型建立方法,无论是简单基础的高斯混合模型-通用背景模型(Gaussian Mixture Model-Universal Background Model, GMM-UBM),还是目前比较主流的联合因子分析(Joint Factor Analysis, JFA)、总变化因子分析(Total Variability Factor Analysis)技术等,原始测试得分均需要进行得分规整,而现有的得分规整方法也都适用于基于以上不同说话人模型的确认系统。

现有的得分规整方法中,大多数都是通过规整"False"识别对得分分布的方式,以减小两类识别对得分汇集后的重叠部分,却没有有效扩大同一说话人模型或同一测试语音对应的两类识别对得分之间的差距;在这些得分规整方法中,都需要预先收集和选取大量的非目标说话人语音数据来估计"False"识别对得分的均值和方差,而非目标说话人语音数据选取的好坏会影响最终得分规整的效果。

针对现有得分规整方法的不足,本文提出一种对数似然值归一化得分规整算法(Log-likelihood Normalization, LLN),通过扩大同一测试语音在目标说话人模型与非目标说话人模型上的得分差距,使同一测试语音对应的两类识别对得分混叠现象得到有效改善;与 Znorm、Tnorm 和 ZTnorm 等方法相结合,可同时从不同角度解决两类识别对得分汇集后的混叠问题,从而进一步提高系统识别性能。

1 说话人确认系统

1.1 说话人确认系统的基本框架

说话人确认系统如图 1 所示,主要分为三部分:提取特征、建立模型和打分判决[8]。

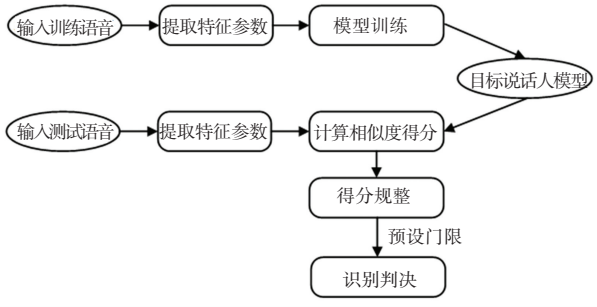


图 1 说话人确认系统

Fig. 1 Speaker verification system

1.2 说话人确认系统的评价指标

在说话人确认系统中,每一次测试,就是将一组识别对进行"True"和"False"判决的过程。当本是"False"的识别对判决为"True"(非目标说话人被接受)时,称之为"虚警"(False Alarm);当本是"True"的识别对判决为"False"(目标说话人被拒绝)时,称之为"漏检"(Miss),这两种错判出现的概率分别称为虚警率和漏检率。

(1)等错率(Equal Error Rate, EER)。实际应用中,应同时降低虚警率和漏检率,然而这二种错误概率相互约束,随着判决门限设定的不同,二者呈相反趋势变化,只有当虚警率和漏检率大致相等的时候,系统的性能被认为达到了最大发挥,此时的错误率称为等错率(EER)。

(2)最小检测代价(Minimum Value of Detection Cost Function, minDCF)。不同的应用场景对虚警率和漏检率要求不同,系统门限的设定会按需调整,为了对不同情况下系统性能进行更加贴切地描述,引入了检测代价函数(Detection Cost Function, DCF)的概念,其数学表达式(1)为:

$$C_{\text{det}} = C_{\text{M}} \times P_{\text{M|T}} \times P_{\text{T}} + C_{\text{FA}} \times P_{\text{FA|NT}} \times (1 - P_T). \tag{1}$$

其中,C_{M} 和 C_{FA} 分别是漏检率 $P_{\text{M|T}}$ 和虚警率 $P_{\text{FA|NT}}$ 对应的代价,P_{T} 是测试中应该判决为"True"的识别对出现的概率,$(1 - P_{\text{T}})$ 是应该判决为"False"的识别对出现的概率。检测代价函数是描述识别错误发生后损失大小的一个函数,可以很好地表示系统的性能。设定门限可以得到该门限对应

的 DCF 值，遍历判决门限，获得最小检测代价（minDCF），这是目前美国国家标准技术研究院说话人识别评测（NIST SRE）中最重要的指标。

1.3　零规整（Znorm）和测试规整（Tnorm）

Znorm 方法是用大量非目标说话人语音对目标说话人模型打分，计算出对应于目标说话人模型 λ 的辅助参数均值 μ_λ 和方差 σ_λ，用来规整得分分布的差异，其得分规整公式（2）如下：

$$S_\lambda^* = \frac{S_\lambda - \mu_\lambda}{\sigma_\lambda}. \tag{2}$$

其中，S_λ 是测试语音对模型 λ 的原始得分，S_λ^* 为规整后的得分。

Tnorm 是用测试语音对大量非目标说话人模型计算得分，得到对应于测试语音的辅助参数，同样是均值和方差，用来减少测试语音环境不同对得分分布的影响，最终得分公式同（2）。

对于说话人确认系统，Znorm 参数计算在模型训练阶段完成，Tnorm 参数计算在测试阶段完成。ZTnorm 是在得分域将训练模型和测试语音的信息结合起来，即将 Znorm 和 Tnorm 相结合的得分规整方法。上述 3 种得分规整方法的不足之处是没有有效扩大同一说话人模型或同一测试语音对应两类识别对得分之间的差距；并且必须引入先验知识，需要将训练数据中的一小部分预留出来作为开发集，用来估计得分规整时需要的参数，而开发集选取的好坏会影响最终得分规整的效果。

2　对数似然值归一化（Log-likelihood Normalization，LLN）

本文提出一种基于 LLN 的得分规整方法，该方法相对于 Znorm、Tnorm 和 ZTnorm 的优势在于扩大了同一测试语音在目标说话人模型与非目标说话人模型上的得分差距，使同一测试语音对应的两类识别对得分混叠现象得到有效改善；并且可以直接对测试得分进行规整，不需要引入先验知识，因此不需要预留训练数据。

假设 $\hat{S} = [S_1 \quad S_2 \quad \cdots, S_L]^t$ 是某测试语音在所有 L 个说话人模型上的得分。设 S_t 为测试语音与其目标说话人模型的得分，即该测试语音对应的“*True*”识别对得分；其余 $L-1$ 个得分 $S_n(n \neq t)$ 为测试语音与非目标说话人模型的得分，即该测试语音对应的“*False*”识别对得分。通常情况下，测试语音在目标说话人模型上的得分会高于其在非目标说话人模型上的得分，即 $S_t > S_n(n \neq t)$。用对数似然值归一化公式（3）来规整每个得分：

$$S_i' = S_i - \ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right). \tag{3}$$

其中，S_i 表示该测试语音在第 i 个说话人模型上的原始得分，S'_i 是经过规整后的得分，$\ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right)$ 表示对得分 S_i 的规整量，令 $N_i = \ln\left(\frac{1}{L-1}\sum_{j\neq i} e^{s_j}\right)$，则 N_i 由除 S_i 之外的其余 $L-1$ 个得分计算得到。根据上面的分析可推出：

（1）如果 $i = t$，则 S_i 较大，规整量 N_i 因不包含 S_t，故数值较小；

（2）如果 $i \neq t$，则 S_i 较小，规整量 N_i 因包含 S_t，故数值较大。

公式（3）中每个得分 S_i 作为 e 的指数是考虑目标说话人模型得分的独特性（较大且数目少），充分扩大其得分的影响，求和是利用非目标说话人模型得分的共同特点（较小且数目多），减少单个得分的影响，取对数可避免非目标说话人模型得分的规整量差距过大。经过（3）式规整，测试语音对目标说话人模型和非目标说话人模型得分差距会进一步拉大，即可以使识别对中“True”识别对和“False”识别对的得分具有更好的区分性，从而更容易设定门限区分“True”识别对和“False”识别对，提升了系统确认性能。

3　实　验

3.1　实验配置

本文实验在 NIST SRE 2008 核心测试集（short2-short3）的电话训练、电话测试（tel-tel）情况下开展。实验主要针对女声测试集，该测试情况下共23 385个测试对，涉及1 674个测试语音和1 140个目标说话人模型，在 LLN 得分规整阶段，每个识别对得分都是基于测试语音数据与全部 1140 个说话人模型的匹配得分经公式（3）得到。

本实验中所使用的特征为 36 维的梅尔频率倒谱系数（Mel Frequency Cepstral Coefficents，MFCC）特征，其每帧特征由 18 维的基本倒谱系数及其一次差分（delta）构成。首先用音素解码器来对语音数据进行语音活动性检测（Voice Activity Detection，VAD），以去除数据中的静音部分；然后根据 25ms 的窗长和 10 ms 的窗移提取 36 维的 MFCC 特征。由于得分规整方法具有普适性，不受限于系统所使用的说话人建模方法，且目前主流的说话人建模技术大多基于 GMM-UBM 模型，因此本实验的说话人建模方法选用简单基础的 GMM-UBM。使用 NIST

SRE 2004 1side 的目标说话人训练数据训练与性别相关的 UBM，UBM 高斯数为 1023[9]。并利用本征信道(Eigenchannel)技术在模型域做了信道补偿，训练 Eigenchannel 信道空间的数据，选择的是 NIST SRE 2004、2005 以及 2006 的电话语音数据，包含 755 个说话人的数据，共9 855个语音文件。另外，从 NIST SRE2006 的数据中挑选了 340 条数据，用于 Tnorm 得分规整和 340 条数据用于 Znorm 得分规整，基本上保证这两个小数据集每个说话人只有一条语音数据。

3.2 实验结果

表 1 比较了 Znorm、Tnorm、ZTnorm 和 LLN 不同得分规整方法的实验结果。从表 1 可以看出，LLN 在不需要开发集的条件下，具有良好的规整性能，相比无得分规整的情况，EER 相对提升 9.7%，minDCF 相对提升 4.57%，本身的规整性能可以和 Znorm、Tnorm 相当。

表 1 NIST SRE 2008 测试集上 Znorm、Tnorm 和 LLN 性能比较

Tab. 1 Performance comparison of Znorm, Tnorm and LLN on NIST SRE 2008 test set

	无得分规整	Znorm	Tnorm	ZTnorm	LLN
EER/%	10.82	9.73	9.66	8.63	9.77
minDCF	4.6	4.46	4.11	3.96	4.39

表 2 是在 Znorm、Tnorm 和 ZTnorm 基础上做 LLN 规整的实验结果。结合表 1 和表 2 中的实验结果可以看出，LLN 可以大幅度提升原有说话人确认系统的性能。在 Znorm 基础上做 LLN 和不做 LLN 相比，系统的 EER 和 minDCF 分别有 20.45% 和 24.44%的性能提升；在 Tnorm 基础上做 LLN 和不做 LLN 相比，系统的 EER 和 minDCF 分别有 5.59% 和 9.98%的性能提升；在 ZTnorm 基础上做 LLN 和不做 LLN 相比，系统的 EER 和 minDCF 分别有 11.7% 和 18.69%的性能提升。

表 2 NIST SRE 2008 测试集上做 LLN 的性能

Tab. 2 Performance of LLN on NIST SRE 2008 test set

	Znorm+LLN	Tnorm+LLN	ZTnorm+LLN
EER/%	7.74	9.12	7.62
minDCF	3.37	3.7	3.22

对比 LLN 规整前后某测试语音在 15 个说话人模型上的得分变化，如图 2 所示。其中，spk13 为该测试语音的目标说话人，其余为非目标说话人。从图 2 可以看出经 LLN 规整后，测试语音对目标说话人模型和非目标说话人模型得分差距会进一步拉大。如果门限保持不变，相比 LLN 规整前，系统的虚警率会明显降低。

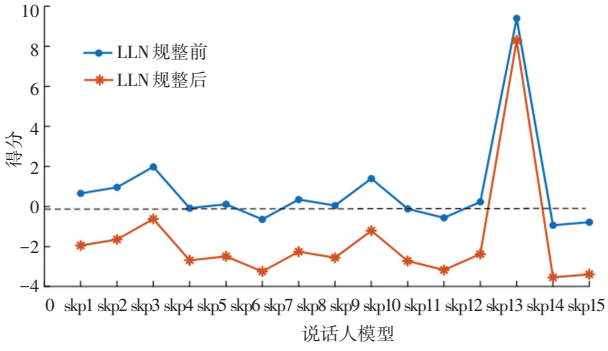
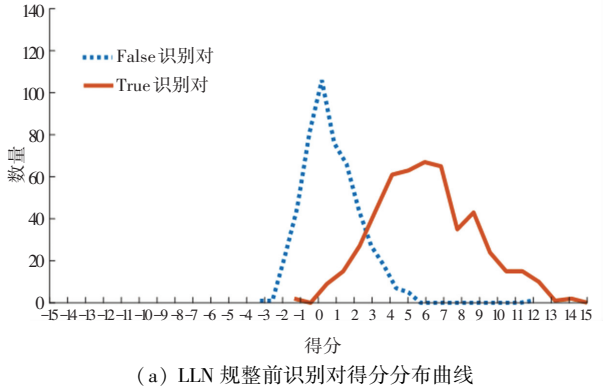


图 2 某测试语音在不同说话人模型上得分

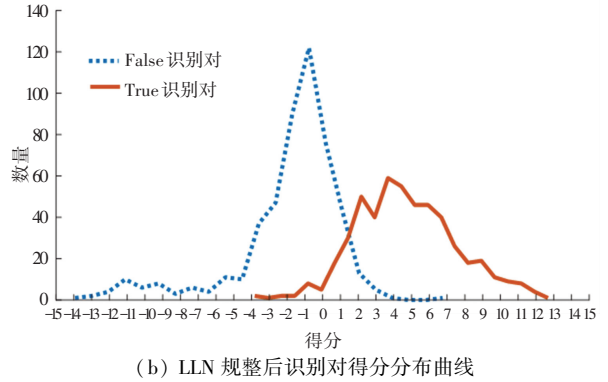
Fig. 2 Score of a test voice in different speaker models

随机选取 500 个“True”识别对和 500 个“False”识别对，比较 LLN 规整前后的得分分布，如图 3 所示。从图 3 可以看出经 LLN 规整后，“False”识别对的得分分布明显向左偏移，而“True”识别对的得分分布变化不明显，“True”识别对和“False”识别对的得分差距拉大，区分性增强，有效降低了虚警率。因此，用统一的门限进行判决时会更有优势。LLN 虽然不会改变同一测试语音在每个目标说话人上得分的排序，但可以有效降低 EER 和 minDCF。



(a) LLN 规整前识别对得分分布曲线

(a) Score distribution curve of verification pairs before LLN normalization



(b) LLN 规整后识别对得分分布曲线

(b) Score distribution curve of verification pairs after LLN normalization

图 3 LLN 得分规整后的得分分布变化

Fig. 3 Changes in score distribution after LLN normalization

(下转第 26 页)