EMD – Tnorm 得分规整策略在说话人确认中的应用

李燕萍[1,2]，丁　辉[2,3]，唐振民[2]

（1.南京邮电大学通信与信息工程学院，南京 210003；2.南京理工大学模式识别与智能系统实验室，南京 210094；
3.嘉兴学院数学与信息工程学院，浙江嘉兴 314001）

［摘要］ 从两个方面对确认系统进行了改进，在模型方面，扩展了 MixMax 模型，对复杂的背景噪声等干扰因素在训练说话人模型的同时也进行了建模，最大程度上消除噪声的影响，对说话人的特征分布进行了更真实的表征；在得分方面，提出了一种改进的得分规整策略，基于 EMD 距离从所有背景说话人集合中自适应选择最接近的一定数量的模型构成说话人特定的背景集合，从而进行得分归一化。实验结果表明，该方法能够同时针对说话人和测试环境的不同进行补偿，进一步降低了误识率和漏警率，获得了很好的确认性能。

［关键词］ 说话人确认；鲁棒性；EMD 距离；MixMax 模型

［中图分类号］ TN912.34 ［文献标识码］ A ［文章编号］ 1009－1742(2010)02－0095－06

1 前言

说话人确认是一个二元判决问题，给定说话人的确认语句及其所声称的身份，系统做出拒绝或接受的判断。现有的说话人确认算法在实验室环境（安静的录音环境、高质量的录音设备、训练和测试数据的采集环境相匹配）已经可以取得很好的效果，但在实际的语音交互中由于复杂的声学环境和个人语音的变化使得系统的性能急剧下降，成为说话人识别系统实用化的一个主要障碍[1]。

在说话人确认系统中，各种不同的可变因素大致可以分为两类：一类是说话人相关因素，由于说话人的个性特征具有长时变动性，会受到健康和情感等因素的影响，而且其发音时间长度，讲话风格等都会带来不同程度的影响；另一类是测试相关因素，在实际的声学环境中由于文本内容的不同，采集设备包括麦克风质量甚至摆放位置的不同和不同的信道传输，以及各种噪声的存在等都会带来不可避免的影响。研究表明，这两类可变因素都会严重影响系统的性能，因此需要采取不同的补偿方法进行处理。

目前的补偿方法主要集中在 3 个层次，特征级、模型级和得分级[1]。得分级是指在得分层进行得分补偿，也称为得分归一化，是针对由于不同说话人和不同测试环境引起的输出评分分布变化的不同因素加以补偿，将不同话者模型下的输出评分规整到同一分布范围内，然后进行确认阈值的合理选取，使得失配条件下与说话人无关的决策门限更加鲁棒。

说话人确认系统中关键的问题在于最佳阈值的选取。目前已有很多不同的得分规整方法[1,2]，例如，零规整（zero normalization，Znorm）方法主要消除不同说话人之间的差异对冒充得分分布的影响；话机规整（handset normalization，Hnorm）则是消除同一说话人在不同麦克风和传输信道环境下的语音对得分分布的影响。在这两种方法中，得分归一化参数都是通过对冒充人集合语音得分分布的估计获得。测试规整（test normalization，Tnorm）选择固定的冒充者模型来补偿由于测试文本的多变性引起的不匹配，在获得低的错误接受率性能方面有显著的改进[3]。

笔者从两个方面对确认系统进行了改进，在模型方面，扩展了 MixMax 模型，对复杂的背景噪声等干扰因素在训练说话人模型的同时也进行建模，很

［收稿日期］ 2008－04－18

［基金项目］ 浙江省自然科学基金资助项目（Y1090649）；浙江省教育厅科研资助项目（Y200805349）

［作者简介］ 李燕萍（1983－），女，陕西合阳县人，博士，研究方向为语音信号处理；E－mail：njustjsjlyp@163.com

大程度上消除噪声的影响,使得后续的地面移动距离(earth mover's distance,EMD)可以在该模型中应用;在得分补偿方面,提出了一种改进的得分规整策略,基于EMD距离从冒充者集合中自适应选择一定数量的冒充者模型构成说话人特定的冒充者集合(speaker specific cohort,SPC),同时针对说话人和测试环境的不同进行了补偿,进一步降低了误识率和漏警率,获得很好的确认性能。

2 基于EMD-Tnorm的得分归一化算法在鲁棒说话人确认系统中的应用

2.1 测试规整算法

Aucken于2000年提出了测试规整理论,原理为:设从测试语音中提取得到特征矢量序列 $\boldsymbol{O}=\{O_1,O_2,\cdots,O_N\}$,训练得到的说话人语音模型为 λ_t,计算测试语音在目标说话人模型下的似然得分为 $s(O,\lambda_t)$[4]。Tnorm首先计算测试语音在冒充者模型集合 $\Lambda_{\mathrm{I}}=\{\lambda_{\mathrm{I},1},\cdots,\lambda_{\mathrm{I},N}\}$ 下的得分 $S_{\mathrm{I}}=\{s(O,\lambda_{\mathrm{I},1}),\cdots,s(O,\lambda_{\mathrm{I},N})\}$,然后进行得分规整:

$$s_{\mathrm{Tnorm}}(O,\lambda_t)=\frac{s(O,\lambda_t)-\mu_{\mathrm{Tnorm}}}{\sigma_{\mathrm{Tnorm}}} \tag{1}$$

式(1)中,μ_{Tnorm} 和 σ_{Tnorm} 分别是假设冒充者集合得分在服从高斯分布下的均值和标准方差。其原理如图1所示,Tnorm方法在基于GMM-UBM(Gaussian mixture model-universal background model)的识别系统中得到广泛的应用,Reynolds研究表明,在固定冒充者集合中计算似然比时,如果能在考虑测试相关因素的同时考虑说话人相关因素,建立说话人特有的冒充者集合,例如通过说话人特征参数的选取(基音周期,性别等)或数据驱动的启发式策略(例如模型间距离的计算),就能够进一步改进系统的性能[5]。

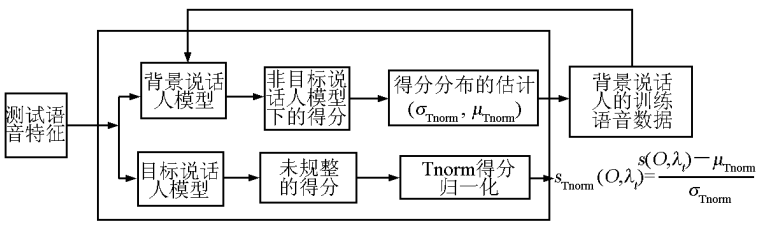

图1 Tnorm测试规整策略示意图

Fig. 1 System for test-normalization technique

Sturim提出说话人特定背景模型的测试规整方法,称为自适应测试规整(adaptive tnorm,ATnorm),通过City-Block矢量距离计算冒充者语音在目标模型的得分序列和在冒充者模型集合中的得分序列的距离,从而选择与目标模型最接近的 K 个模型[2]。笔者提出的基于EMD-Tnorm的归一化算法与ATnorm方法相比,不是基于得分序列的距离计算,而是利用模型参数直接对模型之间的相似性进行度量,不需要额外的冒充语音,算法简单,易于实现。

2.2 EMD理论

地面移动距离(EMD)定义为将"货物"从"供给者"运输给"消费者"所需的最小成本,可以用来衡量两个特征分布之间的相似性[6,7]。EMD距离作为一种典型的相似性度量广泛应用于计算机视觉中的图像检索,模式匹配和视频说话人聚类,均取得了良好效果。

该模型的描述如下:令 $P=\{(p_1,w_{p_1}),\cdots,(p_m,w_{p_m})\}$ 和 $Q=\{(q_1,w_{q_1}),\cdots,(q_n,w_{q_n})\}$ 分别表示供给者和消费者的离散分布函数,其中 p_i 和 q_j 是每一个聚类的质心;w_{p_i} 表示 p_i 可以运输的货物总数;w_{q_j} 表示 q_j 需求的货物总数,称为质心频率;$\boldsymbol{D}=[d_{ij}]$ 是"地面距离"矩阵,矩阵中每个元素 d_{ij} 表示质心 p_i 和 q_j 之间的"地面距离",可以采用不同的距离度量;f_{ij} 是从 p_i 到 q_j 的流量,即运输货物的数量,流量矩阵 $\boldsymbol{F}=[f_{ij}]$。总的运输成本为:

$$\mathrm{Cost}(P,Q,F)=\sum_{i=1}^{m}\sum_{j=1}^{n}d_{ij}f_{ij} \tag{2}$$

式(2)中隐含的约束条件包括:

$$f_{ij}\geqslant 0(1\leqslant i\leqslant m,1\leqslant j\leqslant n) \tag{a}$$

$$\sum_{j=1}^{n}f_{ij}\leqslant w_{p_i}(1\leqslant i\leqslant m) \tag{b}$$

$$\sum_{i=1}^{m}f_{ij}\leqslant w_{q_j}(1\leqslant j\leqslant n) \tag{c}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{q_j}\right) \quad \text{(d)}$$

式(a)规定是单向运输,“货物”只能从 p_i 运输到 q_j;式(b)保证了 p_i 提供给各个需求者 q_j 的货物总和不大于其拥有的货物总数;式(c)说明 q_j 接收各个供给者的货物总和不大于需求总数;归一化因子在式(d)中表示当供需不平衡时双方之间能运输的总流量是它们两者之中的最小值,表示 EMD 距离可以用于规模大小不同的模型之间计算,因此可以用来进行局部匹配。EMD 距离定义为归一化后的运输成本,如式(3)所示:

$$d_{\mathrm{EMD}}(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \quad (3)$$

2.3 MixMax 模型

在实际的说话人确认系统中,确认性能和鲁棒性是两个关键要求。由于各种背景噪声及其不同信道的影响,会使说话人的模型参数发生不同程度的改变,各种模型补偿方法都是着眼于对这些可变因素进行不同程度的抑制,而没有进行具体的建模。因此,笔者选用 MixMax 模型并且对其进行了扩展,可以应用 EMD 距离计算模型间的相似性,对复杂的背景噪声等干扰因素在训练说话人模型的同时也进行建模,很大程度上消除噪声的影响,提高了系统的鲁棒性。

高斯混合模型(Gaussian mixture model, GMM)本质上是一种多维概率密度函数,它假设说话人语音特征可以用一系列高斯函数的叠加来逼近,即用 M 个单高斯分布的线性组合来描述对应说话人的帧特征在特征空间中的概率密度分布,设 D 维特征矢量序列 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_T\}$ 其数学表达式如下:

$$p(X \mid \lambda) = \prod_{t=1}^{T}\sum_{i=1}^{M} \omega_i \prod_{d=1}^{D} g(x_{t_d}, \mu_{i_d}, \sigma_{i_d}) \quad (4)$$

式(4)中,M 是模型混合数;$\boldsymbol{x}_t$ 是特征矢量;w_i 为混合权值,且 $\sum_{i=1}^{M} w_i = 1$;μ_i 为均值矢量;σ_i 为对角化协方差矩阵,σ_{i_d} 为第 d 维对应的标准方差;模型 λ 表示说话人的特征分布服从的概率密度函数,用参数集表示:$\lambda = \{w_i, \mu_i, \sigma_i\}$,$i = 1,2,\cdots,M$,模型参数由期望最大值 EM(expectation maximization)算法训练得到。

MixMax 模型是由说话人的 GMM 模型 λ^{s} 和背景噪声 GMM 模型 λ^{b} 组成[6]。这个模型的优势在于不需要预先估计干净语音模型,在说话人模型估计阶段,含噪语音的各个成分受到背景噪声成分的不同程度掩蔽。在似然值计算过程中,特征矢量的得分通过对组合模型的计算。说话人模型的各个混合成分对最终似然得分的贡献与被噪声掩蔽的程度直接相关,掩蔽越严重,则这个成分对最终似然得分的贡献越小,具体计算公式为:

$$\lambda_{\mathrm{MixMax}} = \{\lambda^{s}_{\mathrm{GMM}}, \lambda^{b}_{\mathrm{GMM}}\} \quad (5)$$

$$p(x_{t_d} \mid i,j,\lambda) = g(x_{t_d}, \mu^{b}_{j_d}, \sigma^{b}_{j_d}) \cdot G\left(\frac{x_{t_d} - \mu^{s}_{i_d}}{\sigma^{s}_{i_d}}\right) + g(x_{t_d}, \mu^{s}_{j_d}, \sigma^{s}_{j_d}) \cdot G\left(\frac{x_{t_d} - \mu^{b}_{i_d}}{\sigma^{b}_{i_d}}\right) \quad (6)$$

$$p(X \mid \lambda) = \prod_{t=1}^{T}\sum_{i=1}^{M}\sum_{j=1}^{N} \omega^{s}_{i} \cdot \omega^{b}_{j} \cdot \prod_{d=1}^{D} p(x_{t_d} \mid i,j,\lambda) \quad (7)$$

式(6)中,$G\left(\frac{x_{t_d} - \mu^{b}_{i_d}}{\sigma^{b}_{i_d}}\right)$ 是一维标准正态分布函数。

2.4 基于 EMD – Tnorm 的得分规整策略

当说话人模型采用 GMM 时,将每个高斯混合成分当作聚类中心,对应的混合权值作为聚类权值,可以直接使用 EMD 距离进行两个模型之间的度量。但是在文章中是采用 MixMax 模型对说话人进行鲁棒建模,此时问题出现在如何将 EMD 距离应用在该模型中,MixMax 模型中噪声等干扰的掩蔽作用并不是完全体现在模型参数中,更多的是通过公式(6)和(7)作用在似然值计算过程中。笔者对 MixMax 模型进行了扩展,引入掩蔽概率的计算,在 EMD 计算过程中给每个混合成分进行掩蔽加权,模拟噪声的掩蔽过程。

公式(6)表示第 t 个特征矢量的第 d 维 x_{t_d} 由说话人模型的混合成分 i 和背景噪声模型的混合成分 j 建模的概率。式(8)给出在 $\{i,j\}$ 状态下假设现有观察特征矢量是干净语音 s_{t_d},即没有受到噪声影响的概率:

$$p(x_{t_d} = s_{t_d} \mid i,j,\lambda) = \frac{g(x_{t_d}, \mu^{s}_{i_d}, \sigma^{s}_{i_d}) \cdot G\left(\frac{x_{t_d} - \mu^{b}_{j_d}}{\sigma^{b}_{j_d}}\right)}{p(x_{t_d} \mid i,j,\lambda)} \quad (8)$$

因此说话人的 GMM 模型参数得到扩展,增加一个矢量 $\boldsymbol{m} = (m_1, m_2 \cdots m_M)$ 作为每个成分的掩蔽系数:

$$m_i = \frac{\sum_{t=1}^{T}\sum_{j=1}^{N}\sum_{d=1}^{D}1-p(x_{t_d}=s_{t_d}\mid i,j,\lambda)}{T\cdot N\cdot D} \quad (9)$$

在模型估计时即可进行掩蔽因子的计算。当混合成分 i 的掩蔽因子为 0 时意味着该混合成分未受到噪声的干扰，即为干净的语音特征分布；当为 1 时，则认为该混合成分被噪声完全破坏。在计算 EMD 距离前，在说话人模型的每个成分权值前乘以掩蔽因子 $1-m_i$，即混合成分受掩蔽作用越严重，对最后距离计算的贡献度越小。原理示意图见图 2。

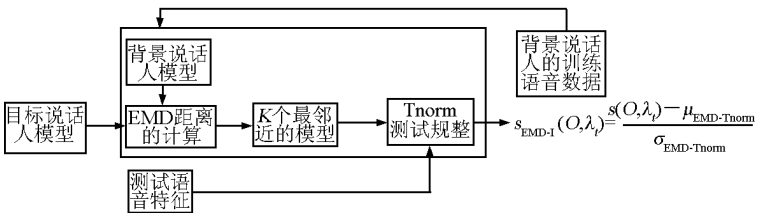

图 2　EMD – Tnorm 归一化算法示意图

Fig. 2　System for EMD – Tnorm method

算法的具体步骤如下：

1）基于 EMD 距离计算目标说话人和其他说话人的模型之间的距离，对于每一个目标说话人 λ_t，计算与其他说话人的模型 $\Lambda_I=\{\lambda_{I,1},\cdots,\lambda_{I,N}\}$ 之间的距离得到一个距离集合：

$$D_{t,I}=\{D_a(\lambda_t\mid\lambda_{I,1}),\cdots,D_a(\lambda_t\mid\lambda_{I,N})\} \quad (10)$$

2）选择 K 个最相似的模型。从距离集合中选择距离最小的 $K(K<N)$ 个模型作为目标说话人的自适应冒充者模型集：

$$\Lambda_{EMD-I}=\{\lambda_{EMD-I,1},\cdots,\lambda_{EMD-I,K}\},\ \Lambda_{EMD-I}\subset\Lambda_I \quad (11)$$

3）计算 EMD – Tnorm 得分。计算测试语音在冒充者模型集合中的得分：

$$S_{EMD-I}=\{s(O,\lambda_{EMD-I,1}),\cdots,s(O,\lambda_{EMD-I,K})\} \quad (12)$$

4）得分归一化。对测试语音在目标模型的得分 $s(O,\lambda_t)$ 进行归一化变换：

$$S_{EMD-I}(O,\lambda_t)=\frac{s(O,\lambda_t)-\mu_{EMD-Tnorm}}{\sigma_{EMD-Tnorm}} \quad (13)$$

式(13)中，$\mu_{EMD-Tnorm}$ 和 $\sigma_{EMD-Tnorm}$ 分别是假设 S_{EMD-I} 服从高斯分布下的均值和标准方差。

3　实验与结果分析

3.1　实验语料库

实验数据来自 C603 语音库，该语音库是在安静的实验室环境下录制的纯净语音。语音信号采样频率为 22.05 kHz，单声道录音，16 Bit 量化。实验中使用的语音数据包括 182 个说话人，82 个女性，100 个男性。其中所有说话人发音都是汉语普通话，每个说话人录音三部分，分别为数字串、固定文章和自由发言，分 3 个文件保存。3 次录音得到的语音长度长短不一，但同一种文件的长度基本相等。数字串以 4 个数字序列为一组，共大约 40 s；文章是伊索寓言《北风与太阳》，时间约 60 s；自由发言部分鼓励谈论生活学习天气等限定在 2 min 之内。噪声数据来自 NOISEX – 92 噪声数据库，这些噪声按不同的信噪比分别添加到干净语音中形成含噪语音。不包括交叉性别测试。

3.2　预处理和特征提取

实验中对输入系统的语音信号进行预加重，预加重系数为 0.99；按帧长 512 个采样点进行分帧，帧交叠为 50 %；之后使用汉明窗进行加窗处理。说话人的特征参数选取 14 阶 Mel 倒谱参数（mel frequency cepstrum coefficient，MFCC）参数及一阶差分 ΔMFCC 共 28 维，在 GMM 模型中，通常阶数越高，系统的识别率就越高，但计算量和存储空间的开销也随之增加，文章折中考虑，取 $M=64$。

3.3　性能评估指标

实验中采用的性能评估标准是等误识率（equal error rate，EER），定义为 DET（detection error trade – off）曲线上错误接受率（FA）和错误拒绝率（FR）充分接近基础上的算术平均值。

在 NIST 说话人识别评测中[8,9]，采用最小检测代价函数（detection cost function，DCF）来代表系统性能，它是系统对检测代价函数取最小值的工作点。DCF 函数定义为：

$$DCF = C_{FR} \times FRR \times P_{tar} + C_{FA} \times FAR \times P_{imp},$$
$$P_{imp} = 1 - P_{tar} \quad (14)$$

式(14)中，C_{FR}和C_{FA}分别是错误拒绝 FR 和错误接受 FA 的代价，P_{tar}和P_{imp}分别是真实说话人和冒充说话人的先验概率。实际测试中，给定一个阈值就会得到对应阈值下的检测代价，检测代价越小的系统性能越好。NIST 评测中定义如下：$C_{FR}=10$，$C_{FA}=1$，$P_{tar}=0.01$，$P_{imp}=0.99$。

3.4 实验结果与分析

实验分别在男女数据库中进行，每个说话人的语音从 3 种文件中分别随机选择连续 20 s 组成训练语音(共 60 s)，在剩余的语音里随机选择 10 s 语音用于自身登录，共进行 20 次；从其他说话人的语音中随机选择 10 s 进行冒认登录，分别进行 3 次；男性数据库自身登录 100×20 次，冒认登录 100×99×3 次，比例约为 1 ∶ 14.9；女性数据库自身登录 82×20 次，冒认登录 82×81×3 次，比例约为 1 ∶ 12.2，总共重复进行 5 次验证，最后取其平均值。

在选择目标说话人特有的背景模型的过程中，K 的取值会影响到最终的性能，在干净语音条件下，对 K 的不同取值进行了多次实验比较，实验结果如图 3 和图 4 所示。

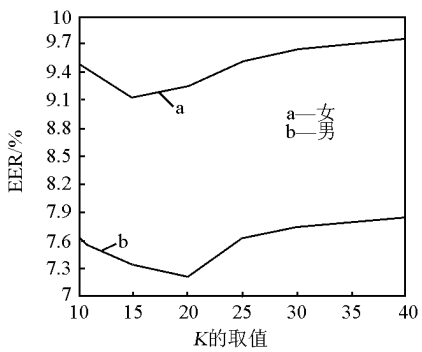

图 3 不同的 K 取值下的等误识率 EER

Fig. 3 The equal error rate during different values of K

从图 3 和图 4 可知，随着 K 取值的不同，分别在男性和女性的语料库下的等错误率和最小检测代价值都会发生改变，男性语音库的实验在 $K=20$ 时取得最小值，女性语音库的实验在 $K=15$ 时取得最小值，该 K 值将作为下一步实验的取值。同时，女性语音库下的实验数据普遍高于男性语音库，经过分析认为女声中高频成分比较丰富，而笔者所采用的

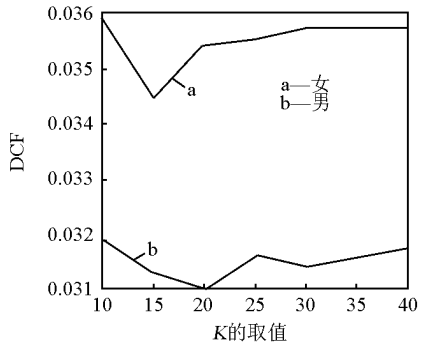

图 4 不同的 K 取值下的最小检测代价 DCF

Fig. 4 Detection cost function during different values of K

传统 ***MFCC*** 参数没有充分考虑到这一点，说明在识别系统中寻找更稳健，更适合说话人的特征也是一个重要问题。

为了验证文章提出的方法在抗噪性，以及降低错误率和最小检测代价方面的性能，笔者在混合噪声存在的环境下，与 Tnorm 方法和不做归一化变换的方法进行比较，在 15 dB 的混合噪声条件下训练模型，在不同的信噪比下进行测试，采用 White 平稳噪声与 Factory，Babble 和 F16 3 种非平稳噪声组成混合噪声。其中，男生组(M)的 K 取 20，女生组(F)取 15，Tnorm 方法和不做归一化变换(简称 None)的方法都是基于传统的高斯混合模型，笔者提出的 EMD－Tnorm 方法基于 MixMax 模型(简称 EMT)，EN Improved 表示 EMT 相对于 None 方法的相对改进率，ET Improved 表示 EMT 相对于 Tnorm 方法的相对改进率，相对改进率 =(对比方法 EER－文章方法 EER)/对比方法 EER。实验结果如表 1 所示。

表 1 不同信噪比(SNR)下各种实验方法的性能对比

Table 1 Experimental results of three methods in different SNR

信噪比	15 dB		10 dB		5 dB	
实验方法	M	F	M	F	M	F
EER None	8.05	9.98	8.94	10.67	10.23	13.08
EER Tnorm	7.74	9.86	8.37	10.31	9.95	11.27
EER EMT	7.32	9.17	7.45	9.28	8.40	10.53
EN Improved/%	9.1	8.1	16.6	13.0	17.8	19.5
ET Improved/%	5.4	6.9	10.9	9.9	15.6	6.6
DCF None	0.041	0.051	0.047	0.054	0.052	0.058
DCF Tnorm	0.038	0.049	0.043	0.052	0.050	0.055
DCF EMT	0.034	0.034	0.036	0.041	0.047	0.053

分析表1可知,一方面3种方法中EMT的等错误率EER和DCF一直是最低的,在15 dB的测试环境下取得最好的性能,EER为7.32,DCF为0.034,比不做归一化变换和Tnorm变换分别降低了9.1%和5.4%,表明了EMT方法在降低错误率和检测代价方面的有效性;另一方面,随着信噪比的降低,系统性能都随之下降,但EMT在低信噪比下仍然保持了较好的性能,从而证明了基于MixMax模型的EMT确实能够提高系统的鲁棒性。

4 结语

说话人确认作为一种典型的二元判决问题,确认性能和鲁棒性是两个关键要求,笔者提出的方法从模型和得分两个方面对确认系统进行了改进,理论分析和实验表明,该方法不仅继承Tnorm方法对测试环境和文本变化等的差异性对输出评分的影响进行了很好补偿,而且对说话人相关因素带来的自身的差异性也进行了很好的补偿。采用的MixMax模型很大程度上提高了系统的鲁棒性,从而进一步降低了误识率和漏警率,获得很好的确认性能。下一步的工作将研究如何采取有效的分级聚类策略在进一步提高系统性能的基础上降低计算复杂度,寻找更加鲁棒的参数来表征噪声干扰环境下的说话人特征。

参考文献

[1] Dijana P D, Asmaa E H, Gerard Chollet. Text - independent speaker verification state of the art and challenges [J]. LNCS, 2007, 135 - 169

[2] Sturim D E, Reynolds D A. Speaker adaptive cohort selective for Tnorm in text - independent speaker verification [J]. ICASSP, 2005, 1: 741 - 744

[3] Daniel R C, Julian F A, Joaquin G R. Speaker verification using speaker - and test - dependent fast score normalization [J]. Pattern Recognition Letters, 2007, 28: 90 - 98

[4] Auckenthaler R, Carey M, Lloyd - Tomas H. Score normalization for text - independent speaker verification systems [J]. Digital Signal Process, 2000, 10:42 - 54

[5] Reynolds D A, Quatieri T F. Speaker verification using adapted Gaussian Mixture Models [J]. Digital Signal Process, 2000, 10: 19 - 41

[6] Thilo Stadelmann, Bernd Freisleben. Fast and robust speaker clustering using the earth mover's distance and mixmax models [J]. ICASSP, 2006, 1: 989 - 992

[7] Rubner Y, Tomasi C, Guibas L J. The earth mover's distance as a metric for image retrieval [J]. International Journal of Computer Vision, 2000, 40: 99 - 121

[8] 郑 榕,张树武,徐 波. 基于特征规整和评分规整的说话人确认研究[J]. 中文信息学报, 2006, 20(6): 75 - 82

[9] 刘明辉, 陈继旭, 李 辉, 等. 基于TZNormalization规整的话者确认阈值选取[J]. 数据采集与处理, 2005, 20(3): 311 - 317

A new score normalizaion algorithm based on EMD - Tnorm for speaker verification

Li Yanping[1, 2], Ding Hui[2, 3], Tang Zhenmin[2]

(1. College of Telecommunications & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. School of Computer Science, Nanjing University of Science & Technology, Nanjing 210094, China;
3. School of Mathematics & Information Engineering, Jiaxing University, Jiaxing, Zhejiang 314001, China)

[**Abstract**] In this paper, the verification system from two aspects was improved. On one hand, we extended MixMax model that the EMD (earth mover's distance) can be applied, which can remove the disturbance of noise; on the other hand, we improved the Tnorm score normalization method based on the EMD. Experimental results show that this method can compensate the speaker - dependent and test - dependent variability, also show a stable performance improvement by decreasing the FA and FR.

[**Key words**] speaker verification; robustness; earth mover's distance; MixMax model