一种新的说话人确认方法*

张怡颖 朱小燕 张 钹

(清华大学计算机科学与技术系 北京 100084)
(清华大学智能技术与系统国家重点实验室 北京 100084)
E-mail: zxy-dcs@mail.tsinghua.edu.cn

摘要 文章在对说话人确认和说话人辨认进行比较研究的基础上,提出一种新的说话人确认方法.同传统方法相比,该方法通过建立非特定说话人模型综合多个说话人的语音特性,使其能够对于不同的待确认语音给出不同的判决阈值,从而解决了说话人确认在判决阈值设置上存在的困难.实验结果表明,该方法能够显著降低说话人确认系统的错误接受率和错误拒绝率,为说话人确认应用于保密性要求较高的环境提供了一条有效的途径.

关键词 说话人确认,说话人辨认,判决阈值,错误拒绝率,错误接受率.

中图法分类号 TP391

说话人识别分为两个范畴,即说话人确认(speaker verification,简称SV)和说话人辨认(speaker identification,简称SI).前者根据说话人的语句确定是否与所声称的参考说话人相符,这种确认只有接受和拒绝两种可能.后者则是把未标记的语句确定为属于多个参考说话人之中的某一个所说,是一个多者择一的问题.

目前,SV方法的研究主要集中在使用Bayesian规则决定判决规则[1~3],如使用分辨因子判决[2],建立反说话人模型[3,4]等.这些方法都需要为确认阶段设置一个不变的阈值,其差别在于,它们获得阈值和对似然得分进行标准化的方法不同,其局限性是,为了获得合适的阈值,需要进行大量的实验工作,并且对不同输入语音采用相同判决阈值,因而降低了系统的适应性.

本文提出使用动态阈值的说话人确认方法DTSV(speaker verification with dynamic threshold).通过在确认阶段的判别策略中使用非特定说话人模型,使判决阈值随待确认语音的不同而发生变化,提高了系统的自适应性.

1 说话人模型

自SV研究开始以来,已提出多种说话人模型[1,5].本文采用高斯混合模型,即每个参考说话人的训练语音在声学空间中的分布用高斯混合密度函数表示,说话人的模型参数为:$\lambda=((c_1,\mu_1,\Sigma_1),\ldots,(c_m,\mu_m,\Sigma_m),\ldots,(c_M,\mu_M,\Sigma_M))$,其中$\mu_m,\Sigma_m$分别为第$m$个高斯函数的均值矢量和协方差矩阵,$c_m$是第$m$个高斯函数的权重,$M$是高斯函数的个数.

令输入语音的特征矢量序列为$O=\{o_1,\ldots,o_k,\ldots,o_T\}$,则$\lambda$产生特征矢量$o_k$和特征矢量序列$O$的似然得分值分别为

$$P(o_k|\lambda)=\sum_{m=1}^{M}c_m\cdot\frac{1}{(2\pi)^{D/2}\cdot\left(\left|\sum_m\right|\right)^{1/2}}\cdot\exp\left(-\frac{1}{2}(o_k-\mu_m)^T\sum_m^{-1}(o_k-\mu_m)\right),\tag{1}$$

* 本文研究得到国家自然科学基金资助.作者**张怡颖**,女,1971年生,博士生,主要研究领域为语音识别,说话人识别.**朱小燕**,女,1957年生,博士,副教授,主要研究领域为模式识别,神经元网络,语音处理.**张钹**,1935年生,教授,博士生导师,中国科学院院士,主要研究领域为人工智能,计算机应用.

本文通讯联系人:朱小燕,北京 100084,清华大学计算机科学与技术系

本文1998-01-13收到原稿,1998-04-03收到修改稿

$$P(O|\lambda)=\prod_{t=1}^{T}P(o_t|\lambda), \tag{2}$$

其中 D 为特征矢量的维数，T 为特征矢量的帧数.

2　新的说话人确认方法

2.1　传统说话人确认方法

假设共有 N 个参考说话人，它们的说话人模型分别为 $\lambda_1,\lambda_2,\ldots,\lambda_N$，其中 λ_i 由最大化似然得分 $P(O_i|\lambda_i)$ 获得，O_i 是第 i 个参考说话人的训练数据. 当判断一段语音 X 是否为声称的参考说话人 i 所说的时候，确认策略如下：

$$\text{若 } P(X|\lambda_i)\begin{cases}>\eta, & \text{接受参考说话人 } i \text{ 的身份}\\ \leqslant\eta, & \text{拒绝参考说话人 } i \text{ 的身份}\end{cases}, \tag{3}$$

其中 η 是判决阈值.

2.2　问题的提出

SI 和 SV 的训练过程相同，只是在识别时采用不同的策略，使得二者在性能上有很大差异. 目前，SI 系统可达到的差错率性能指标远低于相应 SV 系统的等差错率. 表 1 是对 SI 和 SV 系统性能的比较实验，测试人为 20 位女性，训练数据为 10 个数字{0,1,...,9}和 20 个数字串，测试数据为每人 30 个数字串.

图 1 是说话人辨认系统的性能随人数增加发生的变化，数据内容同表 1. 随着人数的增加，SI 系统的性能降低得比较缓慢. 当系统中有 100 人时，差错率为 2.0%(平均每个测试语音长为 1.1s). Douglas 在文献[6]中的实验表明，当 SI 系统的人数达到 630 时，系统的差错率仍能达到 1.0%(平均每个测试语音长为 3s). 这些实验结果引导我们思考，既然造成 SV 和 SI 系统差异的主要原因是二者的决策策略不同，那么我们是否可以利用 SI 识别方式提高 SV 系统的性能呢？于是，我们提出 DTSV 方法.

表 1　SI和SV的实验比较

	差错率(%)	等差错率(%)
SI	0.00	—
SV	—	14.37

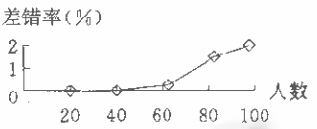

图1　说话人辨认系统性能与人数的关系

2.3　DTSV 方法

DTSV 方法在判别策略中融入说话人辨认的判别方式. 对于 SI 而言，其辨认策略为

$$\text{设 } i=\operatorname{Arg}\max_{1\leqslant k\leqslant N}\{P(X|\lambda_k)\},\text{于是 } X \text{ 是由第 } i \text{ 个参考说话人所说.} \tag{4}$$

那么，DTSV 方法为确认 X 是否为参考说话人 i 所说，先用公式(4)判断 X 由谁发出. 若 X 是由参考说话人 i 所说，则接受参考说话人 i 的身份，否则拒绝. 于是确认阶段的判别策略为

$$\text{设 } j=\operatorname{Arg}\max_{1\leqslant k\leqslant N}\{P(X|\lambda_k)\},$$

$$i\begin{cases}=j, & \text{接受参考说话人 } i \text{ 的身份}\\ \neq j, & \text{拒绝参考说话人 } i \text{ 的身份}\end{cases}. \tag{5}$$

由此判别策略可见，采用这种方案的 SV 系统的性能取决于相应的 SI 系统的性能. 在第 2.2 节中的论述已经充分说明了这种新方法的可行性和合理性.

式(5)表示的策略只有在待确认语音是某参考说话人所说的时候才是有效的，这是因为首先要判断说话人是谁. 但在实际应用中，不可避免地会发生某伪装者非法进入系统的可能性，这时，待确认语音不是参考说话人之一所说. 此时，如果使用式(5)，结果很难预知. 因此，我们对判别策略式(5)进行改进.

为了使新的判别策略适用于实际开放的应用环境，我们在现有的 N 个说话人模型之外附加一个非特定说话人模型 λ_{N+1}，这个模型的参数是通过最大化似然得分的乘积 $\prod_{i=1}^{N}P(O_i|\lambda_{N+1})$ 获得的，即训练 λ_{N+1} 所使用的数据

是所有参考说话人的训练数据.于是,在改进的判别策略中一共有 $N+1$ 个说话人模型,其中 λ_{N+1} 代表了多个说话人的共同特征.

在一般情况下,若待确认语音是某个参考说话人 i 所说,则应有

$$P(X|\lambda_i) > P(X|\lambda_{N+1}). \tag{6}$$

因为 λ_i 是由说话人 i 自己的数据训练得到的,它比 λ_{N+1} 更能精确地描述说话人 i 的语音在声学空间中的分布.同时,对于理想情况应该满足:

$$P(X|\lambda_{N+1}) > P(X|\lambda_j),\ j=1,2,\ldots,N,\text{且 } j\neq i. \tag{7}$$

这是因为 λ_{N+1} 中具有多个说话人的特征.它所包含的关于第 i 个参考说话人的信息量多于其他参考说话人模型中所包含的信息量.若待确认的语音不是某参考说话人所说,即它是外来的冒名顶替者的语音,那么,由于 λ_{N+1} 具有一定的普遍性,使其在理想情况下满足:

$$P(X|\lambda_i) < P(X|\lambda_{N+1}),\quad i=1,2,\ldots,N. \tag{8}$$

综合式(5)~式(8),得到 DTSV 方法的判别策略:

$$P(X|\lambda_i)\begin{cases} > P(X|\lambda_{N+1}), & \text{接受说话人的身份} \\ \leqslant P(X|\lambda_{N+1}), & \text{拒绝说话人的身份} \end{cases} \tag{9}$$

如果式(6)~(8)在任何情况下都成立,那么系统的错误拒绝率(false rejection rate,简称 FR)和错误接受率(false acceptance rate,简称 FA)都为零,这时系统的性能就达到理想状态.但是,我们不能保证这 3 个公式在任何情况下都成立.例如,若冒名顶替者的语音与某个参考说话人的语音十分接近,则可能违背式(7)或式(8),导致错误接受;又因为人的发音千变万化,同一个人在不同时期说同一个音时,由于诸多因素的影响,语音分布会有很大的差异,则可能违背式(6),导致错误拒绝.

通过增加非特定说话人模型,使这种新的确认策略不仅适用于确认真实身份说话人的语音,也适用于拒绝冒名顶替者的语音.若 X 不是任何参考说话人所说,则由式(8),系统拒绝此说话人的身份;若 X 由参考说话人 i 所说,则由式(6),系统接受此次确认;若 X 由某参考说话人 j 所说,则由式(7),系统将拒绝此说话人的身份.改进的判别策略式(9)同式(5)相比,不仅使确认系统适合于开放的应用环境以拒绝非法进入者,而且能够完成对参考说话人语音的正确接受和拒绝;此外,判别策略式(9)大大降低了确认所需的计算量,它只需计算两次似然得分.

若将式(9)的 $P(X|\lambda_{N+1})$ 看作传统方法中的阈值,则 DTSV 是一种可变阈值的 SV 方法.随着待确认语音的不同,DTSV 的阈值 $P(X|\lambda_{N+1})$ 不断变化,它更适于确认千变万化的输入语音,因此,DTSV 具有更强的适应性.表 2 列出 DTSV 和传统方法之间的主要区别(CSV 表示传统方法).

表 2 DTSV 和传统方法之间的区别

方法	训练过程	阈值可变	确认时计算距离的次数
CSV	训练 N 个模型	否	1 次
DTSV	训练 $N+1$ 个模型	是	2 次

3 实验和讨论

3.1 数据库

实验数据来自清华大学智能技术与系统国家重点实验室收集和整理的汉语普通话语音数据库 CIDS (Chinese isolated word, digit and syllable).我们使用 70 个人(40 位女性 1~40 和 30 位男性 41~70)的数据,每个人都要录制 10 个孤立数字{0,1,...,9}(标志为 1~10)和 50 个数字串(标志为 11~60).这些语音数据被划分为 4 个子库,如表 3 所示.每个人的训练数据的平均时间为 21s,每个测试数字串的平均时间为 1.1s.

表 3　实验所用的子数据库

	数据库 A	数据库 B	数据库 C	数据库 D
录音人	1～15,41～55	16～30,56～70	1～20	21～40
训练数据	1～30	—	1～30	—
测试数据	31～60	31～60	31～60	31～60

数据库 A 和 B 用于实验 Ⅰ，数据库 C 和 D 用于实验 Ⅱ. 数据库 A 和 C 中的训练数据分别用来训练参考说话人模型，而对 A 和 C 的测试构成了闭合集测试(closed set test)，对数据库 B 和 D 的测试构成了开放集测试(open set test)，错误率取为所有参考说话人的错误率的平均值.

特征矢量为 16 阶倒谱＋16 阶差分倒谱＋差分能量. 说话人模型采用式(1)和式(2).

3.2　实验 Ⅰ

这个实验包括 30 个参考说话人(15 位女性和 15 位男性). 传统方法和 DTSV 方法所得到的测试结果见表 4(在如下实验中，我们仍然用 CSV 表示传统方法). 由表 4 可以看出，DTSV 方法最突出的特点是无论对于闭合集测试，还是对于开放集测试，它的 FA 都远远低于传统方法的 FA，这表明 DTSV 方法具有极强的判别冒名顶替者的能力. 对于闭合集测试，它的 FR 也远低于传统方法在等差错阈值点的 FR，并且 FA 达到 0.0%. 这说明 DTSV 方法具有很强的区别说话人的能力.

为了进一步比较 DTSV 方法和传统方法，我们进行如下实验. 修改传统方法的阈值，使其 FR 和 FA 分别达到 DTSV 方法的相应值，比较这时传统方法对闭合集的 FA 和 FR，以及对开放集的 FA 与 DTSV 方法的相应值. 实验结果见表 5. 由进一步的实验看到，要想使传统方法的 FR 和 FA 分别达到 DTSV 方法的相应值，它的 FA 和 FR 将受到极大损失，其值远远高于 DTSV 方法的相应值.

表 4　实验 Ⅰ 的测试结果

方法	CSV		DTSV	
测试集	FR(%)	FA(%)	FR(%)	FA(%)
闭合集	12.25	12.25	5.89	0.0
开放集	—	19.59	—	0.015

表 5　实验 Ⅰ 的进一步比较实验

方法	CSV		CSV		DTSV	
测试集	FR(%)	FA(%)	FR(%)	FA(%)	FR(%)	FA(%)
闭合集	5.89	20.11	61.44	1.10	5.89	0.0
开放集	—	27.86	—	4.88	—	0.015

3.3　实验 Ⅱ

实验 Ⅱ 中参考说话人仅由女性组成. 实验表明，对于说话人确认系统，女性的等差错率高于男性的等差错率. 通过实验 Ⅱ，DTSV 方法的有效性将得到进一步证实.

表 6 给出了实验 Ⅱ 的测试结果，表 7 是与表 5 相似的进一步比较实验. 对于参考说话人全为女性这种较难的情形，DTSV 方法相对于传统方法仍然显示出显著的优越性. 此时，虽然 DTSV 方法的 FR 比传统方法在等差错率阈值点的 FR 高 2.88%，但其所能达到的 FA 是传统方法所不能比拟的.

表 6　实验 Ⅱ 的测试结果

方法	CSV		DTSV	
测试集	FR(%)	FA(%)	FR(%)	FA(%)
闭合集	14.87	14.87	17.75	0.00
开放集	—	5.90	—	0.013

表 7　实验 Ⅱ 的进一步比较实验

方法	CSV		CSV		DTSV	
测试集	FR(%)	FA(%)	FR(%)	FA(%)	FR(%)	FA(%)
闭合集	17.75	12.10	78.5	0.88	17.75	0.00
开放集	—	4.98	—	0.038	—	0.013

4　结论和未来的工作

本文提出了一种新的说话人确认方法——DTSV 方法. 这种方法与传统方法相比，具有如下特点：(1) 增加非特定说话人模型；(2) 融入说话人辨认策略；(3) 在确认阶段使用可变阈值，提高了系统的适应性；(4) 错误接受率极小；(5) 错误拒绝率较低. 传统方法如果要达到 DTSV 方法的错误接受率，其错误拒绝率将受到严重影响，使系统达到了几乎不可用的程度.

在本文所提出的方法中，非特定说话人模型是影响性能的一个关键因素，因此，我们在未来的工作中将进

步研究如何获得合适的非特定说话人模型以应用于实际系统.

参考文献

1 Rosenberg A E, Lee C-H, Soong F K. Sub-word unit talker verification using hidden Markov models. In: Lonnie Ludeman ed. Proceedings of the International Conference of Acoustic, Speech and Signal Processing. New Mexico: New Mexico State University, 1990. 269～272

2 Higgins A, Bahler L. Text-independent speaker verification by discriminant counting. In: John Litva ed. Proceedings of the International Conference of Acoustic, Speech and Signal Processing. Toronto, Ontario: Piscataway, 1991. 405～408.

3 Rosenberg A E, Delong J, Lee C H *et al*. The use of cohort normalized scores for speaker recognition. In: Reg Stanton ed. Proceedings of the International Conference on Spoken Language Processing. Banff, Alberta: University of Alberta, 1992. 599～602

4 Liu Chi-shi, Wang Hsiao-chuan, Lee Chin-hui. Speaker verification using normalized log-likelihood score. IEEE Transactions on Speech and Audio Processing, 1996,4(1):57～60

5 Tishby N. On the application of mixture AR hidden Markov models to text independent speaker recognition. IEEE Transactions on Acoustic, Speech, Signal Processing, 1991,39(3):563～570

6 Reynolds Douglas A. Large population speaker identification using clean and telephone speech. IEEE Signal Processing Letters, 1995,2(3):46～48

A Novel Speaker Verification Method

ZHANG Yi-ying ZHU Xiao-yan ZHANG Bo

(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)

(*State Key Laboratory of Intelligent Technology and Systems Tsinghua University Beijing* 100084)

Abstract In this paper, on the basis of comparison of speaker verification and speaker identification, a novel speaker verification method is proposed. Compared to the conventional method, the speaker-independent speaker model is established to represent the universal feature of speech, and is used during the verification phase so that the decision threshold is adapted to different input speech and decided during verification phase. The difficulties of setting the decision threshold for the conventional method are alleviated by using the proposed method. Experiments show that this novel method can significantly decrease the false acceptance rate and the false rejection rate of a speaker verification system, providing a promising way for realizing practical speaker verification systems.

Key words Speaker verification, speaker identification, decision threshold, false rejection rate, false acceptance rate.