2014 年 3 月 第 3 期
Vol. 35 No. 3 2014

利用 i-vectors 构建区分性话者模型的话者确认

方 昕,李 辉,刘青松

(中国科学技术大学 电子科学与技术系,合肥 230027)
E-mail: klg@ mail. ustc. edu. cn

摘 要: 对于电话手机语音的文本无关话者确认,运用联合因子分析构建话者信息子空间与信道信息子空间来进行失配信道补偿取得了较好的效果. 然而研究表明,信道信息子空间仍然包含了可以用来区分话者的信息. 因此,本文运用一种既包含话者信息又包含信道信息的全变量信息子空间来提取 i-vectors 低维特征矢量,再运用类内协方差规整进行失配信道补偿,最后用补偿后的 i-vectors 特征矢量构建支持向量机话者模型. 在 NIST08 数据库上实验表明,本文所构建系统的性能在等误识率和最小检测代价函数上有相对近 70% 的提高.

关 键 词: 话者确认; 全变量信息子空间; 类内协方差规整; 支持向量机; i-vectors
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2014)03-0685-04

Discriminative Speaker Models Based on i-vectors for Speaker Verification

FANG Xin, LI Hui, LIU Qing-song
(Department of Electronic Science & Technology, University of Science & Technology of China, Hefei 230027, China)

Abstract: Joint Factor Analysis provides an effective means for text independent speaker verification system. It is a powerful technique for compensating the variability caused by different channels and speakers. However, studies show that, the channel information subspace also contains information that can be used to distinguish between speakers. In this study, we propose a new speaker representation called i-vectors which is a low-dimensional vector. Firstly , it is extracted from a total variability space which models both the speaker and channel variability. Then, within this total variability space, Within-Class Covariance Normalization, a common used channel compensation method, is performed to reduce the channel variability. Finally, the compensative i-vectors are used to train discriminative models based on Support Vector Machines. Experiments on NIST08 SRE database show that the proposed strategy can improve the system performance as much as 70% both in EER and MinDCF over the baseline system.

Key words: speaker verification; total-variability subspace; within-class covariance normalization; support vector machine; i-vectors

1 引 言

近年来,在与文本无关话者确认中,运用联合因子分析(JFA)[1]方法分别构建话者信息子空间与信道信息子空间来进行失配信道补偿,取得了较好的效果. 然而研究表明[2],JFA 中构建的信道信息子空间仍然包含了可以用来区分话者的信息,因此,运用 JFA 方法得到的话者均值矢量损失了一部分的话者信息. 为此,i-vectors[3]特征参数被提出. i-vectors 特征参数是通过全变量信息子空间估计而来的一种低维度的矢量,全变量信息子空间是一种既包含了话者信息又包含了信道信息的矩阵,通过它估计得到的 i-vectors 特征参数既包含了话者信息,又包含了信道信息. 为了从 i-vectors 特征参数中消弱信道信息的影响,突出话者信息,可以采用类内协方差规整(WCCN)[4]对 i-vectors 特征参数进行失配信道补偿.

支持向量机(SVM)[5]一直是话者确认中应用最广泛的区分性模型,而经过失配信道补偿后的 i-vectors 特征矢量是代表了话者信息的低维度的矢量,用它作为 SVM 话者模型的训练矢量,必将得到较好的效果.

因此,本文构建了一种将 i-vectors 特征参数结合 SVM 的话者确认系统. 将代表目标话者信道失配补偿后的 i-vectors 特征矢量作为目标 SVM 话者模型的"+1"类训练样本,挑选若干冒认话者的信道失配补偿后的 i-vectors 特征矢量作为目标 SVM 话者模型的"-1"类训练样本,采用余弦核函数训练得到目标话者的 SVM 模型. 测试时,同样对测试语音提取 i-vectors 特征矢量,代入目标话者的 SVM 模型进行判决,得到最终确认结果. 实验表明,本文构建的系统取得了较好的性能.

2 基于 i-vectors 的话者确认

2.1 基于 i-vectors 的话者确认原理

在经典的基于 GMM-UBM 的话者确认中,通用背景模型(UBM)是由大量说话人的语音通过期望最大化(EM)算法训练得到,它代表着统计平均的说话人信息和信道信息. 在 UBM 的基础上,注册语音通过最大后验概率算法(MAP)自适应得到目标话者模型. MAP 通常只自适应 UBM 的均值. 在实际环境中,语音中不仅包含说话人的信息,还包含传输信道等信息,因此,目标话者的GMM均值矢量既包含了话者信

收稿日期: 2012-10-10 收修改稿日期: 2012-11-12 作者简介: 方 昕,男,1988 年生,硕士,主要研究方向为语音信号处理; 李 辉,男,1959 年生,博士,副教授,主要研究方向为语音信号处理与电子系统设计; 刘青松,男,1984 年生,博士,主要研究方向为语音信号处理与模式识别.

息也包含了信道信息. 在经典的基于联合因子分析的话者确认中,分别构建了话者信息子空间(V 与 D) 与信道信息子空间(U) ,从而估计出话者因子(y 与 z) 与信道因子(x) 进行失配补偿,因此,目标话者的 GMM 均值矢量可以表示为(1) 式所示:

$$M_{tar} = m + Vy + Ux + Dz \tag{1}$$

然而,在[1]中证明,信道因子仍然包含了可以用来区分话者的话者信息. 因此,如果将 GMM 均值矢量表示为(2) 式所示[3],就可以解决上述存在的问题:

$$M_{tar} = m + Tw \tag{2}$$

在(2) 式中,M_{tar}代表目标话者的 GMM 均值矢量,既包含了话者信息也包含了信道信息; m 代表 UBM 的均值矢量,是一个与话者和信道都无关的矢量; T 是一个低秩的矩阵,称之为全变量信息子空间,它代表了话者信息与信道信息二者在空间中的分布; w 是一个全变量信息因子,既包含话者因子也包含了信道因子,与均值矢量相比,它一般是一个低维的矢量,简称为 i-vectors 特征矢量,将它进行信道失配补偿减弱信道因子影响后,作为本文进行话者确认的语音特征参数.

2.2 i-vectors 的提取

对于给定的语音 s,首先计算语音 s 特征参数(维数为 F) 对于 UBM(混合度为 C) 的零阶统计量 $N_s[c]$ 和一阶统计量 $f_s[c]$,如(3) 式所示:

$$\begin{aligned} N_s[c] &= \sum_t \gamma_c(o_t) \\ f_s[c] &= \sum_t \gamma_c(o_t)\, o_t \end{aligned} \tag{3}$$

其中,$\gamma_c(o_t)$ 是观察矢量 o_t 对于给定 UBM 的第 c 个混合度的后验概率输出. 对于语音 s,它的所有混合度的一阶统计量 $f_s = (f_s^{(1)}, \cdots, f_s^{(c)})$.

为了计算方便,我们对一阶统计量和 UBM 的均值进行归一化[4],如(4) 式所示:

$$\begin{aligned} f_s[c] &\leftarrow f_s[c] - N_s[c]\, m[c] \\ m[c] &\leftarrow 0 \end{aligned} \tag{4}$$

同样,我们再对一阶统计量和全变量信息子空间 T 用 UBM 的协方差矩阵进行“归整”,假定 UBM 的协方差矩阵 $\Sigma[c]$ 是对角正定矩阵,如(5) 式所示:

$$\begin{aligned} f_s[c] &\leftarrow \Sigma[c]^{-1/2} f_s[c] \\ T[c] &\leftarrow \Sigma[c]^{-1/2} T[c] \\ \Sigma[c] &\leftarrow I \end{aligned} \tag{5}$$

其中,$\Sigma[c]^{-1/2}$是协方差矩阵$\Sigma[c]$的逆的 Cholesky 分解[7],$T[c]$ 是维数为 $F \times M$ 的全变量信息子空间矩阵 T 的子矩阵,$T = (T^{(1)\prime}, \cdots, T^{(c)\prime})$.

根据估计出的一阶统计量与零阶统计量,提取 i-vectors 如(6) 式所示:

$$w_s = L_s^{-1} T' f_s \tag{6}$$

其中 L_s 是一个 $M \times M$ 的矩阵,由(7) 式计算得来:

$$L_s = I + \sum_{c=1}^{C} N_s(c)\, T(c)'\, T(c) \tag{7}$$

w_s 为就是语音 s 对应的 i-vectors 特征参数. 在上述计算过程中,在全变量信息子空间 T 矩阵已知的情况下,就可以对任意给定的语音求出其 i-vectors 特征参数,因此,如何估计出全变量信息子空间 T 是提取 i-vectors 特征参数的关键.

2.3 全变量信息子空间 T 的构建

全变量信息子空间 T 代表了话者信息与信道信息二者在空间中的统计分布,可以采用大量话者语音,采用 EM 算法训练得到.

在 E 阶段,对于所有训练语音 S 条,计算得到(8) 式和(9) 式的变量:

$$C(c) = \sum_{s=1}^{S} f_s(c)\, w_s' \tag{8}$$

$$A(c) = \sum_{s=1}^{S} N_s(c)\, (L_s^{-1} + w_s w_s') \tag{9}$$

其中,$N_s[c]$,$f_s[c]$,w_s 分别由 2.2 节中的(3) 式和(6) 式得到.

在 M 阶段,T 的更新公式如(10) 式所示:

$$T(c) = C(c)\, A(c)^{-1} \tag{10}$$

在 i-vectors 特征参数提取过程中,全变量信息子空间 T 的大小只需能够精细描述话者和信道信息在空间中的变化即可,因此,T 的大小一般只需几百维,这就直接导致了 i-vectors 特征矢量相比与话者模型的均值矢量是一个低维度的矢量,所以,T 起到了将话者信息和信道信息进行数据压缩和降维的目的.

3 类内协方差规整及 SVM 话者模型的构建

上述方法得到的 i-vectors 特征矢量,既包含话者信息又包含信道信息,因此,对 i-vectors 特征矢量进行失配信道补偿是必要的. 类内协方差规整(WCCN) [8] 运用一个 WCCN 矩阵子空间,消弱信道信息的影响. 在本文构建话者确认系统中,WCCN 矩阵可以由(11) 式估计得来:

$$S_w = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{N_s} (w_i^s - w_s)(w_i^s - w_s)' \tag{11}$$

其中 w_s 为话者的 i-vectors 特征参数的均值,S 为话者总数,n_s 为话者 s 的语音条数.

在得到 S_w 矩阵后,可以对得到的 i-vectors 进行规整,如(12) 式所示:

$$w_{(norm)} = B' w \tag{12}$$

其中,$BB' = S_w^{-1}$,$w_{(norm)}$,即为规整后的 i-vectors 特征参数.

在得到规整后的 i-vectors 特征参数后,即可将目标话者的 $w_{s(norm)}$ 作为“+1”类,若干条冒认话者的 $w_{s(norm)}$ 作为“-1”类,训练得到目标话者的 SVM 模型. 研究表明[8],信道信息主要影响 i-vectors 特征矢量的模,而话者信息主要影响 i-vectors 特征矢量的方向,而余弦核函数恰好可以消除 i-vectors 特征矢量模的影响,从而进一步消弱信道信息的影响,所以训练 SVM 模型采用余弦核函数是最合适的.

4 实验结果

本文构建的系统流程图如图 1(见下页) 所示.

实验数据都取自于 NIST[9] 电话手机语音数据库,全变量信息子空间 T 的训练数据集取自 NIST05 和 NIST06 男性语音数据库,共 7200 条 5 分钟时长训练语音; UBM 训练集取自 NIST04 和 NIST05 的 1side-1side 数据库,共 500 条 5 分钟时长训练语音,WCCN 矩阵的训练数据取自 NIST06 的 8side-1side

数据库,每人有 8 条 5 分钟时长语音, 共 295 人,2950 条语音. 因子分析方法中失配信息子空间 U 的训练集取自 NIST06 的

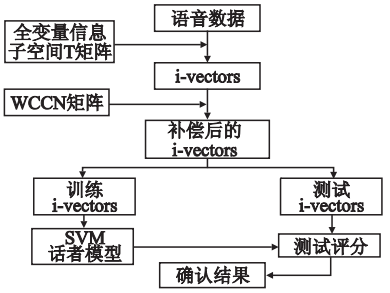

图 1　基于 i-vectors-SVM 话者确认系统框图

Fig. 1　Block diagram of speaker verification system based on i-vectors-SVM

8side-1side 数据库,每人有 8 条 5 分钟时长语音, 共 100 个人,800 条语音. 训练测试数据集取自 NIST08,如表 1 所示.

表 1　实验数据集

Table 1　Experiment database

	语音长度	训练集	测试语音	测试次数
数据集	5 分钟	40 条	280 条	5600 次

采用梅尔频率倒谱系数(MFCC)参数作为系统的前端语音特征参数,在提取 MFCC 之前,对语音进行语音活动检测(VAD),去掉其中的静音段. MFCC 参数维数为静态的 16 维加上一阶动态 16 维共 32 维,并对其进行 Rasta 滤波[10] 和均值方差规整(CMVN)[11].

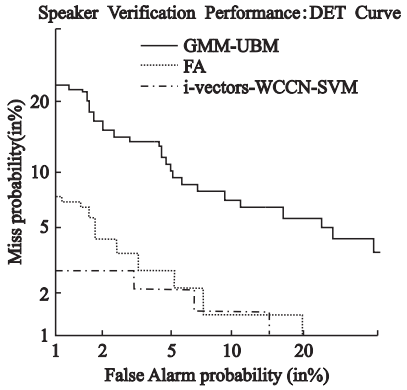

图 2　三种系统 DET 曲线

Fig. 2　DET curves of three systems

实验的评测标准采用等误识率(Equal Error Rate,EER)和 NIST 评测中的检测代价函数[11](Detection Cost Function,DCF),用公式表示为(13)式所示:

$$DCF = C_{fr} \cdot FR \cdot P_{Tar} + C_{fa} \cdot FA \cdot P_{Imp} \quad (13)$$

其中 C_{fr} 和 C_{fa} 分别是错误拒绝和错误接受的代价,P_{Tar} 和 P_{Imp} 则分别是真实说话人和冒认者出现的先验概率,在 NIST 的评测任务中定义 $C_{fr}=10$,$C_{fa}=1$,$P_{Tar}=0.01$,$P_{Imp}=0.99$. 以最小检测代价函数(MinDCF)作为系统性能的评测标准.

4.1　本文系统与基准系统,因子分析系统比较

图 2 与表 2 给出了在训练测试数据集上,本文构建系统与 GMM-UBM 基准系统和因子分析系统性能比较. 系统 UBM 混合度为 512,全变量信息子空间矩阵 T 大小为 16384 * 200,SVM 的惩罚因子为 1,因子分析信道信息子空间 U 大小为 16384* 40.

表 2　三种不同系统性能比较

Table 2　Performance of three different systems

系统	EER(%)	MinDCF
GMM-UBM 基准系统	7.86	0.0322
因子分析系统	3.50	0.0151
本文系统	2.86	0.0096

由图 2 与表 2 可知,本文构建系统的 EER 从 7.86% 下降到 2.86%,MinDCF 从 0.0322 下降到 0.0096,性能分别相对提高 64% 和 70%. 表明本文构建系统性能较基准系统有很大的提高,与因子分析系统相比,系统的 EER 和 MinDCF 也好于因子分析系统.

4.2　全变量信息子空间大小对系统性能影响

在此组实验中,我们在训练测试数据集上比较了 T 的大小对系统性能的影响. T 列数 M 分别取 100,200,400. 实验结果如图 3 和表 3 所示.

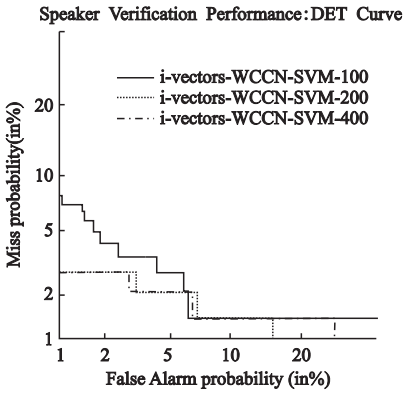

图 3　T 的大小不同时系统 DET 曲线

Fig. 3　DET curves of baseline system and different size of T

由图 3 和表 3 可知,当全变量信息子空间 T 大小逐渐增大时,系统性能也随之提高. 空间大小代表着话者信息和信道信息在空间中的分布变化情况,所以空间越大,则对话者信息和信道信息描述的越精细. 实验结果也表明,空间大小 M 为

表 3　T 的大小对系统性能影响比较

Table 3　Comparison of performance of different size of T

M	EER(%)	MinDCF
100	3.57	0.0138
200	2.86	0.0096
400	2.86	0.0070

200 和 400 的性能优于 M 为 100 的系统. 当然,M 越大,系统的计算复杂度越高,需合理选择 M 的大小.

5　结　论

本文将 i-vectors 特征矢量与 SVM 结合,与 JFA 系统相比,本文构建系统不需要独立训练话者信息子空间和信道信

息子空间,避免了话者信息的损失,且话者注册和测试都是同样的过程,降低了系统的复杂度. 在 NIST 数据库上的实验结果表明本文的系统性能有明显的提升. 和一般 SVM 系统相比,本文都建系统的训练矢量和测试矢量都是 i-vectors 特征矢量, i-vectors 特征矢量的维度 M 一般只有几百维,类比采用 GMM 均值矢量作为 SVM 的特征矢量,大大降低了特征矢量的维度,也就大大减少了系统的训练和测试时间,提高了方法的实用性.

References:

[1] Kanagasundaram A, Vogt R, Dean D, et al. i-vector based speaker recognition on short utterances [C]. Interspeech 2011, Florence, 2011: 2341-2344.

[2] Dehak N. Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification: application to speaker verification [D]. Montreal: École de Technologie Supérieure, 2009.

[3] Dehak N, Kenny P, Dehak R, et al. Front end factor analysis for speaker verification [J]. IEEE Transactions on Audio, Speech and Language Processing, 2010, 19(4): 788-798.

[4] Hatch A O, Kajarekar S, Stolcke A. Within-class covariance normalization for SVM-based speaker recognition [C]. Interspeech 2006, Pittsburgh, 2006: 53-56.

[5] Dehak N, Dehak R, Kenny P, et al. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification [C]. Interspeech 2009, Brighton UK, 2009: 53-56.

[6] Glembek O, Burget L, Matějka P, et al. Simplification and optimization of i-vector extraction [C]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, 2011: 4516-4519.

[7] Matthias Seeger. Low rank updates for the cholesky decomposition [EB/OL]. http://upseeger.epfl.ch/papers/cholupdate.pdf, 2008.

[8] Dehak N, Dehak R, Glass J, et al. Cosine similarity scoring without score normalization techniques [C]. Odyssey Speaker and Language Recognition Workshop, Brno, 2010.

[9] The NIST year 2006 speaker recognition evaluation plan [EB/OL]. http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf, 2006.

[10] Kinnunen T, Li H. An overview of text-independent speaker recognition: from features to supervectors [J]. Speech Communication, 2010, 52: 12-40.

[11] Kenny P, Boulianne G, Ouellet P, et al. Speaker and session variability in GMM-based speaker verification [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(4): 1448-1460.