文章编号:1004-9037(2014)02-0238-05

小训练语料下基于均值超矢量聚类的说话人确认方法

花 城 李 辉

(中国科学技术大学电子科学与技术系,合肥,230027)

摘要:讨论了一种新的基于均值超矢量聚类的说话人确认方法,在确保性能的情况下放宽对语料的要求,聚类训练语料是每个说话人只有一段语音的小语料。以女性统一背景模型(Universal background model,UBM)为基准,对所有女性训练语音均值超矢量相对该 UBM 的偏移聚类,判别待映射男性语音所属类别后进行特征映射,在特征参数域同时削减掉匹配到的通道信息和一部分女性说话人信息。实验表明,不论从性能还是语料角度,采用本文方法相对其他方法均具有一定优势。

关键词:说话人确认;特征映射;语料;超矢量

中图分类号:TN912.34 **文献标志码**:A

Speaker Verification Based on Supervector Clustering With Poor Corpus

Hua Cheng, *Li Hui*

(Department of Electronic Science and Technology, University of Science and Technology of China, Hefei, 230027, China)

Abstract: A new speaker verification method based on supervector clustering is discussed, in order to ensure the performance and reduce the data requirements. An approach based on supervector clustering under poor training corpus using the inter-speaker variability between male and female is presented. Mixed effects of speaker and channel information are clustered, then after the decision on categories of unprocessed speech feature mapping is conducted. Experiments show its advantage compared with other methods under poor corpus from corpus and performance perspective.

Key words: speaker verification; feature-mapping; poor corpus; supervector

引 言

在文本无关说话人确认中[1],主流的方法是林肯实验室 Reynolds 提出的 GMM-UBM-MAP 结构[2],该方法在统一背景模型(Universal background model,UBM)基础上根据最大后验概率(Maximum a posterior,MAP)得到目标话者的模型。当前影响说话人确认系统性能的一个重大且待解决的问题是通道失配问题[3]。通道失配是指训练和测试语音的失配,是由传输信道、话筒及环境等差异造成。特征映射方法是解决通道失配问题的主要方法之一。传统的特征映射方法要对每个通道分别建立高斯混合模型(Gaussian mixture model,GMM)[4-5],该方法需要训练语音带有通道先验信息。

文献[6]提出了一种基于均值超矢量聚类的特征映射方法,该方法不要求语音明确标注通道信息,但是对语料库要求比较苛刻:每个话者必须具有多段语音。然而在很多实际情况下,每个人的语音只能在同一通道下录制,并且只有一段语音。如何利用这些较小的语料制定均值超矢量聚类策略是本文研究的目标。

本文参考美国国家标准技术署(National institute of standard and technology,NIST)评测[7]的部分要求,对语料库中的男性话者进行测试。算法需要事先知道语音的性别标记,将语料库中的一部分女性语音作为聚类的训练语音,每个女性话者只要一段语音即可。用 K-means 和期望最大化(Expectation maximization,EM)算法对这些女性

收稿日期:2012-12-12;修订日期:2013-03-09

语音均值超矢量相对女性 UBM 的偏移量进行聚类。对需要进行特征映射的男性语音先判别属于哪一类,然后通过映射关系在特征域上去除匹配到的通道信息和女性话者信息。由于男性和女性这两类群体本身说话人信息差别就较大,在特征域上削减掉的这部分女性话者信息并不对男性语音中的主要说话人信息产生太大影响。

1　盲通道聚类和特征映射方法

文献[6]中盲通道聚类方法的主要步骤如下:

(1)设一个人有 K 段话,对每段话 j 都通过 MAP 形成超矢量 $\boldsymbol{\Phi}(x_{ij})$,通过下式

$$\bar{\boldsymbol{\Phi}}(x_i)=\frac{1}{K}\sum_{j=1}^{K}\boldsymbol{\Phi}(x_{ij}) \tag{1}$$

求其平均值,式(1)是与通道无关的矢量。

(2)求得与通道无关的矢量后,利用式(2)求得通道引起的说话人空间偏移

$$\Delta\boldsymbol{\Phi}(x_{ij})=\boldsymbol{\Phi}(x_{ij})-\bar{\boldsymbol{\Phi}}(x_i) \tag{2}$$

(3)每句话都形成一个超矢量 $\boldsymbol{\Phi}(x_{ij})$,将与人相关的下标 i 省去,标为 $\boldsymbol{\Phi}(x_j)$,如果得到了很多 $\boldsymbol{\Phi}(x_j)$,采用 Kmeans 和 EM 聚类方式,将这些高维矢量聚到几十个类别上去,类别中心为 λ_k,包含超矢量聚类后的权重、均值和方差。

(4)对于待映射的语音矢量 $\boldsymbol{x}$,先通过 MAP 映射到一个超矢量 $\boldsymbol{\Phi}(x)$ 上,然后获得一个相对 UBM 的偏移 $\Delta\boldsymbol{\Phi}(x)$,再通过最大后验概率求出最近似的一个类 k,将这个类的均值 $\boldsymbol{\mu}_k$ 作为通道引起的偏移。

(5)对于每一帧矢量 $\boldsymbol{x}_t$,首先计算相对于说话人模型中每一个高斯的占有率 $Pr(i\mid\boldsymbol{x}_t)$,然后利用式(3)进行特征映射

$$\boldsymbol{x}_t{}'=\boldsymbol{x}_t-\sum_{i=1}^{M}Pr(i\mid\boldsymbol{x}_t)\boldsymbol{\mu}_{ki} \tag{3}$$

由于式(1)要求一个人话者需要有 K 段语音,所以当说话人只有一段语音可供训练时,这种方法就无法使用。

2　小训练语料下聚类方法

在所给语料库中收集说话人确认任务所在性别的异性语料。本文的评测任务是针对男性话者,就先采集语料库中的异性(女性)语音 F 条。这 F 条语音来自语料库中的不同说话者。聚类方法分以下几步:

(1)将 F 条女性语音训练出一个 UBM 模型,称之为 UBM_{female},UBM_{female} 是女性话者的通用背景模型。女性语音与男性语音采集自同一语料库,通道分布类似。UBM_{female} 是与通道无关的代表女性话者背景的大 GMM 模型。

(2)将每条女性语音 f 由 UBM_{female} 自适应得到一个 GMM 模型。只对 UBM_{female} 的均值作自适应,而方差和权重保持不变。将这个高斯混合模型的均值拼接起来形成一个维数较高的均值超矢量 $\boldsymbol{V}_f$。如图 1 所示,该均值超矢量包含了语音的说话人信息和通道信息[8]。

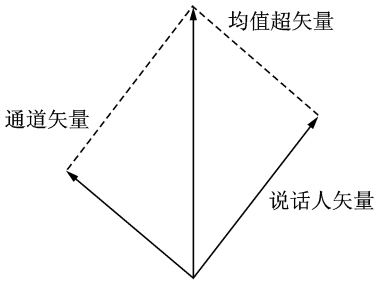

图 1　均值超矢量分解示意图

Fig. 1　GMM-supervector decomposition

求出每条女性语音相对于 UBM_{female} 的偏移量

$$\Delta\boldsymbol{V}_f=\boldsymbol{V}_f-UBM_{\text{female}} \tag{4}$$

这样求得的偏移量同时包括通道空间的偏移和说话人空间的偏移。

(3)将 F 条女性语音相对 UBM_{female} 的偏移量 $\Delta\boldsymbol{V}_f$ 聚成 N 类,聚类方法和训练 GMM 模型的 EM 方法类似,每一类的中心为 $\lambda_k=\{\omega_k,\boldsymbol{\mu}_k,\Sigma_k\}$,其中 $\omega_k,\boldsymbol{\mu}_k,\Sigma_k$ 依次为第 k 类的权重、均值和方差。由于整个聚类过程中并没有单独提取出通道信息,因此这种方法是依据通道信息和说话人信息的综合影响进行聚类的。

3　类别判定与特征映射

对于待特征映射的男性语音 m,首先在男性通用背景模型 UBM_{male} 的基础上用 MAP 方法自适应出均值超矢量 $\boldsymbol{V}_m$,$\boldsymbol{V}_m$ 中包括男性语音 m 的说话人信息和通道信息。

求出每条男性语音相对于女性通用背景模型 UBM_{female} 的偏移

$$\Delta\boldsymbol{V}_m=\boldsymbol{V}_m-UBM_{\text{female}} \tag{5}$$

式(4-5)是相对于 UBM_{female} 的偏移。将式(4)展开

$$\Delta\boldsymbol{V}_f=\Delta\boldsymbol{V}_{fs\text{-}female}+\Delta\boldsymbol{V}_{fc} \tag{6}$$

$\Delta\boldsymbol{V}_{\text{fs-female}}$ 表示该条女性语音相对于女性通用模型说话人信息的偏移量,$\Delta\boldsymbol{V}_{fc}$ 表示该女性语音相

对女性通用背景模型通道信息的偏移量。

将式(5)展开

$$\Delta \boldsymbol{V}_m = \Delta \boldsymbol{V}_{\text{ms-female}} + \Delta \boldsymbol{V}_{mc} \tag{7}$$

$\Delta \boldsymbol{V}_{\text{ms-female}}$表示该条男性语音相对于女性通用背景模型说话人信息的偏移量，$\Delta \boldsymbol{V}_{mc}$表示该条男性语音相对于女性通用背景模型通道信息的偏移量。

将每条待映射的男性语音 m 形成的均值超矢量 $\Delta \boldsymbol{V}_m$ 通过最大后验概率来判断类别

$$k = \arg\max_k \{P(\Delta \boldsymbol{V}_m \mid \lambda_k)\} \tag{8}$$

式(6-7)是相对于女性通用背景模型 UBM_{female} 的偏移量，同时包含了通道信息和说话人信息，而式(8)是依据通道信息和说话人信息的综合影响进行判断。如果男性语音 m 被判定属于类别 k，说明 m 中的通道和说话人综合影响最为接近第 k 类。

由于 UBM_{female} 与通道无关，因此 $\Delta \boldsymbol{V}_{fc}$ 和 $\Delta \boldsymbol{V}_{mc}$ 均为语音中的通道信息。本文的关键是异性语料聚类，女性群体和男性群体的说话人信息本身差别就较大，尽管 m 被判断出属于第 k 类，但第 k 类中的女性说话人信息和 m 语音中男性说话人信息相距较远，男性语音 m 中表征说话人信息的主要成分并不在由女性语料聚类而成的第 k 类中。因此如果将第 k 类中的信息(包括说话人和通道)从男性语音 m 的特征域中去除，意味着去除匹配到的通道信息与次要说话人信息。

对于需要特征映射的语音，计算每一帧对于 UBM_{female} 中每个高斯的占有率

$$Pr(i \mid \boldsymbol{x}_t) = \frac{\omega_i N_i(\boldsymbol{x}_t)}{\sum_{j=1}^{M} \omega_j N_j(\boldsymbol{x}_t)} \tag{9}$$

若该段语音与第 k 类匹配，利用式(9)对第 k 类聚类中心均值加权求和，得到特征映射式

$$\boldsymbol{x}_t' = \boldsymbol{x}_t - \sum_{i=1}^{M} Pr(i \mid \boldsymbol{x}_t)\boldsymbol{\mu}_{ki} \tag{10}$$

式中：M 是 UBM_{female} 的高斯混合度。

为了验证男性语音经过上述特征映射后仍保留主要的说话人信息成分，选取大量经过特征映射后的测试语音观察在原话者模型下的输出评分，结果如表 1 所示。

表 1 测试语音特征映射前后性能对比 %

Table 1 Performance before and after feature mapping

性能指标	特征映射前	特征映射后
EER	5.12	5.23

由表 1 可以发现，经过上述特征映射处理后的测试语音基本不影响原模型的判决结果，说明测试语音经过特征映射去除匹配到的通道信息的同时仍然保留了主要的说话人信息成分。

4 实验和结果

4.1 实验数据库

对 NIST 中的男性说话人进行话者确认实验。实验任务中共有 40 个目标说话人，每个人有大约 5 min 训练语音(Voice activity detection，VAD 后大约 3 min)，20 条目标语音测试(每条 10 s)，195 条冒认测试(每条 10 s)。

用于聚类的语料是采集自 NIST 的女性语料 212 条，每条在 1 ～ 3 min，根据 NIST 计划书[7,9,10]，NIST 语料通道差异按照传输信道影响来分，有以下 3 类：Celluar(蜂窝电话)，Cordless(无线)，Regular(普通，比如陆上线路)；按照话筒设备影响分为 4 类：Speakerphone(免提电话)，Headmounted(头戴式)，Ear-bud(入耳式)，Regular(普通，比如手持设备)。

对原始语音预加重后加窗分帧，窗长 20 ms，窗移 10 ms，采用 Hamming 窗。提取 16 维静态梅尔倒谱参数(Mel frequency cepstral coefficients，MFCC)和 16 维一阶动态差分参数，联合组成 32 维的 MFCC 参数。随后在特征参数层面上切除静音，RASTA 滤波[11]和倒谱均值相减(Ceptrum mean substraction，CMS)[12]。

4.2 实验任务描述

对所有 212 条女性语料训练出一个女性通用背景模型 UBM_{female}，模型混合度为 256，然后将这所有的 212 条语料分别由 $\text{UBM}_{\text{female}}$ 自适应到 GMM 空间，获得 8 192 维的均值超矢量。按照第 2 节中的式(4)求出超矢量偏移后聚类，并对所有的男性 MFCC 参数按照式(10)做特征映射。

说话人确认系统的高斯混合度为 1 024 个，用特征映射后的男性语料训练得到一个 1 024 混合度的 UBM 模型，然后通过 MAP 算法得到各个说话人模型。FA 是所有测试语音中对于冒认者语音接收的次数比率，FR 是所有测试语音中对于目标说话人语音拒绝的次数比率。FA 和 FR 均随阈值的变化而变化，确认阈值升高，错误接收率(False acceptance rate，FA)变小，错误拒绝率(False rejection rate，FR)变大；确认阈值降低，FA 变大，FR 变小。当 FA 与 FR 相等时所对应的错误率就是等误识率(Equal error rate，EER)，本文用 EER 来衡量系统性能优劣。

4.3　不同聚类类别下系统性能

小训练语料情况下,如果每个话者只有一段训练语音,就无法实现文献[6]的聚类方法。实验中聚类所用训练语音只有 212 条,为保证每个类别都有一定量的训练语音,聚类个数不超过 10。

文献[6]中实验结果表明,盲通道聚类类别数与实际通道类别数并无直接联系,较高的聚类类别数(32 类)能更细致地描绘通道情况,从而达到更好的效果。在本文的小训练语料情况下,聚类类别受到限制,且由于针对通道和说话人信息的综合影响进行聚类,因此聚类类别的确定更加复杂,图 2 绘出了不同聚类类别下系统性能曲线。

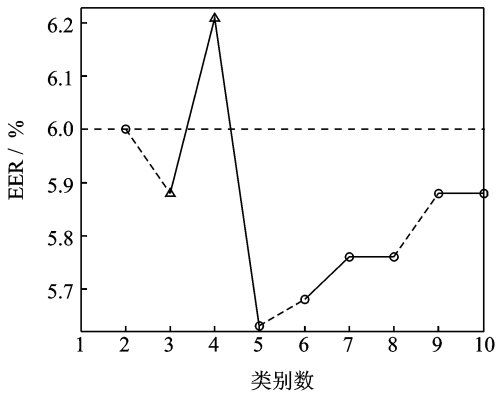

图 2　不同聚类类别对系统性能的影响

Fig. 2　Performance of different clustering categories

图 2 中虚线表示基线系统性能(EER 为 6.00%),三角形坐标点表示实际通道类别数,其中 3 是按传输信道划分的类别数,4 是按话筒设备划分的类别数。从图中可以发现,在聚类类别数等于实际通道数 3 或 4 时,系统性能较差,在类别数是 4 时系统性能甚至低于基线系统,这一情况验证了本文中的聚类方法并不是单纯按照通道进行聚类。聚类类别数为 5 时系统性能达到最优,EER 为 5.63%,相对基线系统下降了 6.17%。聚类类别数较高时,每个类别得不到足够量的训练语音,从而降低了对最终系统性能的提升效果。

4.4　不同方法之间性能比较

为了比较本文方法与文献[6]中方法在聚类训练语料数量相当情况下的性能差异,用 55 个说话人共 220 条语音作为文献[6]中的聚类训练语料,具体结果如表 2 所示。

从表 2 可以发现,在训练语料较少的情况下,本文方法在聚类个数为 5 时可以将系统 EER 相对基线系统降低 6.17%,采用文献[6]中的方法系统性能则明显不及本文方法,甚至劣于基线系统,而且本文方法中训练聚类并不要求一个说话人有多条语音,因此,无论是从结果还是语料角度,在小训练语料下本文方法均优于文献[6]方法。

表 2　不同方法之间性能差别

Table 2　Performance among different methods

序号	实验方法	EER/%
1	基线系统	6.00
2	文献[6](聚类数为 5)	9.76
3	文献[6](聚类数为 7)	9.13
4	本文方法(聚类数为 5)	5.63
5	本文方法(聚类数为 7)	5.76

5　结束语

本文给出了一种在小训练语料下基于均值超矢量聚类的特征映射方法。文献[6]中的算法需要一个比较大的数据库,每个人必须要有多段语音,并且这些语音是从不同通道或录音话筒采集而成。但在大多实际情况下,每个人往往只有一种通道下的训练语音,且有可能只有一段语音。本文所采用的聚类训练语料更加容易实现足够训练 256 个混合度 UBM 的女性语料。由于男性群体和女性群体说话人之间的较大差异,本文在聚类时并没有考虑说话人信息和通道信息的分离,而是对其混合聚类。实验表明,在小训练语料情况下,运用本文方法可以使 EER 相对基线系统下降 6.17%,不论从结果还是语料角度均相对文献[6]方法具有一定优势。

参考文献:

[1] Bimbot F, Bonastre J, Fredouille C, et al. A tutorial on text-independent speaker verification[J]. EURASIP Journal on Applied Signal Processing, 2004, 4: 430-451.

[2] Reynolds D A, Quatieri T, Dunn R. Speaker verification using adapted gaussian mixture models[J]. Digital Signal Processing, 2000, 10(1-3): 19-41.

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

[4] Reynolds D. Channel robust speaker verification via feature mapping[C]//ICASSP. Hong Kong, China: IEEE, 2003(II): 53-56.

[5] 王敏,赵鹤鸣,张庆芳. 基于瞬时频率估计和特征映射的汉语耳语音话者识别[J]. 数据采集与处理,2011,

26(6):686-690.
Wang Min, Zhao Heming, Zhang Qingfang. Speaker identification with Chinese whispered speech based on instantaneous frequency estimation and feature mapping[J]. Journal of Data Acquisition and Processing, 2011, 26(6):689-690.

[6] 郭武,戴礼荣,王仁华. 基于均值超矢量聚类和特征映射的说话人确认[J]. 数据采集与处理, 2009, 24(1): 19-22.
Guo Wu, Dai Lirong, Wang Renhua. Speake verification based on supervector clustering and feature mapping[J]. Journal of Data Acquisition and Processing, 2009, 24(1):19-22.

[7] NIST. 2006 NIST speaker recognition evaluation(EB/OL). http://www.itl.nist.gov/iad/mig/tests/sre/2006/index.html, 2006.3/2012.11.

[8] Kenny P, Ouellet P, Dehak N, et al. A study of interspeaker variability in speaker verification[J]. IEEE Transactions on Audio Speech and Language Processing, 2008, 16(5):980-988.

[9] NIST. 2004 NIST speaker recognition evaluation(EB/OL). http://www.itl.nist.gov/iad/mig/tests/sre/2004/index.html, 2004.01/2012.11.

[10] NIST. 2005 NIST speaker recognition evaluation(EB/OL). http://www.itl.nist.gov/iad/mig/tests/sre/2005/index.html, 2005.03/2012.11.

[11] Hermansky H, Morgan N. RASTA processing of speech[J]. IEEE Transactions on Speech and Audio Processing, 1994, 2(4):578-589.

[12] Garcia A, Mammone R. Channel robust speaker identification using modified-mean cepstral mean normalization with frequency warping[C] //Proceedings of ICASSP. Phoenix. Ariz, USA:IEEE, 1999(1): 325-328.

作者简介:花城(1989-)男,硕士研究生,研究方向:说话人识别、语音信号处理,E-mail:huacheng@mail.ustc.edu.cn;李辉(1959-),男,博士,副教授,研究方向:语音信号处理、电子系统设计。