

基于 LPC 和 MFCC 得分融合的说话人辨认

单燕燕

(南京邮电大学 通信与信息工程学院, 江苏 南京 210003)

摘要:实验室环境下,说话人识别研究已经取得很大进展,但是在实际生活中,说话人识别系统的性能受到环境噪声、健康状况等因素的影响很大。日常生活中,感冒是不可避免的。而感冒往往会诱发鼻腔的炎症,改变鼻腔的容积和形状,引起说话人声音的改变,导致说话人识别性能下降。文中研究测试者感冒时说话人识别系统的性能。为了有效利用不同特征参数得分的互补性,针对基于 GMM 模型的说话人辨认系统,提出了将特征 LPC 和 MFCC 分别应用于该系统,并将二者的得分归一化后进行融合计算。实验结果表明,对正常语音来说,与 LPC 特征系统相比,该方法能够有效提升辨认性能;对感冒语音来说,当高斯成分为 16 时,较之 LPC 特征系统,该方法提升辨认性能 12.5% 左右,较之 MFCC 特征系统,该方法也能提升 8.5% 左右的辨认性能。

关键词:感冒语音;说话人辨认;得分融合;得分归一化

中图分类号: TN912.3

文献标识码: A

文章编号: 1673-629X(2016)01-0039-04

doi: 10.3969/j.issn.1673-629X.2016.01.008

Speaker Identification Based on Score Combination of LPC and MFCC

SHAN Yan-yan

(College of Communication and Information Engineering, Nanjing University of Posts and
Telecommunications, Nanjing 210003, China)

Abstract: At present, speaker recognition technology has made great progress in clean voice. But in daily life, there are various factors, such as environmental noise and healthy condition, impacting recognition rate of speaker recognition system. The cold tends to induce the nasal cavity's inflammation, and changes the volume and shape of the nasal cavity and then changes the vocal characteristics of the speaker. In order to effectively use the complementarity of scores from different feature parameter, the performance's change of speaker identification system was studied when the speaker gets the cold. So the method was proposed using linear prediction coefficient and MEL cepstrum coefficient to train the speaker model respectively, and then score normalization method is used to process scores from two feature systems. Finally, two outputs were weighted. The experimental results show that for normal speech, this method can improve the identification performance; for cold speech, the method improves the identification performance by 12.5% when the number of Gaussian components equals to sixteen compared with the system taking MFCC as feature, by 8.5% to the LPC system.

Key words: cold speech; speaker identification; score combination; score normalization

0 引言

语音信号不仅传递了所要表达的语义信息,还传递了说话人的健康状况以及情绪等信息。例如,当说话人身体不舒服或生病时,他说出的语音往往比身体健康时的低沉,给人一种有气无力的感觉,有时甚至是声音沙哑的,所以说说话人生病时产生的语音波形也随着而改变,从而降低了说话人识别系统的性能。实际生活中,说话人识别系统的性能受到两方面因素的影响:外部因素和内部因素^[1]。外部因素主要指的是环

境噪音、编码方式不同以及通道变化。说话人识别的研究目前主要集中在环境噪音和通道失配等外部因素的影响。在这方面已取得了非常大的进展^[2]。内部因素,也称自身因素,主要是指说话人的声道特征或者独特的行为特征发生变化,按照时间长短可分为短时和长时两大类。长时变化^[3]通常指的是随着说话人年龄的增大发声器官产生的缓慢变化,包括疾病、物理损伤或者发育期变化等带来发声器官的长久性变化。与长时变化不同,短时变化^[1]则是由于发声方式的变化、说

收稿日期:2015-01-07

修回日期:2015-05-08

网络出版时间:2016-01-04

基金项目:国家自然科学基金资助项目(61271335);国家重点基础研究发展计划(2011CB302303)

作者简介:单燕燕(1988-),女,硕士研究生,研究方向为说话人识别、语音信号处理。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20160104.1607.064.html>

话人伪装、情绪变化以及短时疾病(如感冒)等因素使得说话人的声音发生暂时性的变化。长时变化通常可利用自适应的方式得到很好地解决,然而短时变化则因其具备复杂性和突变性等特点而成为当前说话人识别中的一个难题。短时疾病一般指感冒、咳嗽、扁桃体炎等造成发声器官变化的短暂的可康复的疾病。P. Rose^[4]指出感冒往往会伴随着鼻腔(nasal cavities)中的炎症和肿大,这会改变鼻腔的容积和形状,从而改变鼻腔对声源激励信号的调制作用,引起说话人声音的改变,从而导致说话人识别性能急剧下降。R. G. Tull 等^[5]也发现用正常语音训练的说话人识别系统,在说话人感冒时的识别率明显下降。但是对于说话人感冒时引起语音的具体变化以及如何减小感冒时语音的短时变化对说话人识别系统的影响缺乏进一步的研究。

日常生活中,人们总是难以避免地感冒,感冒使得说话人语音发生变化进而影响说话人识别系统的性能,所以,研究测试者感冒时的说话人识别系统具有很大的现实意义。

文中定义说话人未感冒时录制的中性语音作为正常语音,而患有感冒时录制的语音为感冒语音。用正常语音训练说话人 GMM 模型,而待识别语音分别为正常语音和感冒语音,文中提出了将线性预测系数和梅尔倒谱系数分别应用于基于 GMM 模型的说话人辨认系统,归一化处理匹配得分,然后进行融合计算,融合得分最高者即为最终的匹配结果。实验结果表明,该方法能显著提高系统的性能。

1 基于 GMM 模型的说话人辨认系统

说话人辨认分为两个阶段:训练阶段和辨认阶段。在训练阶段,对说话人的语音信号进行一系列的处理,提取其特征参数之后,然后对这些特征参数进行聚类以表征这个特定说话人,即建立说话人模型。而辨认阶段则提取出待辨认说话人的测试语音特征参数,并将其与已建立的说话人模型进行相似性比较,根据相应的评估准则判定说话人身份。

基于 GMM^[6-8]模型的说话人辨认系统框图如图 1 所示。

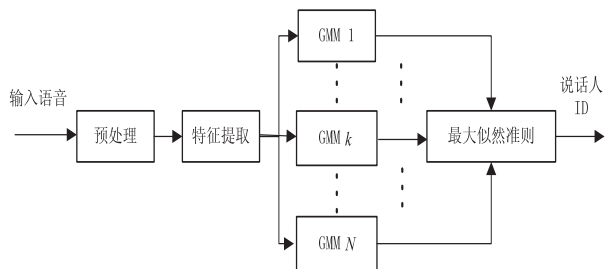


图 1 基于 GMM 模型的说话人辨认系统框图

2 特征提取

2.1 线性预测系数(LPC)

语音信号的线性预测的基本思想是:在特定时间内,语音信号采样点之间具有一定的相关性,因而可以用过去的样点值的线性组合来表示现在或未来的样点值。通过使语音的预测样点值在最小均方误差准则下逼近实际样点值,可以求得唯一的一组预测系数。这组预测系数表征了语音信号的特性。语音信号下一时刻的样点值可以利用该语音信号过去 p 个时刻的样点值的线性组合来逼近,用过去 p 个时刻语音采样值的线性组合以最小预测误差预测语音信号下一时刻的采样值,称为对语音信号的 p 阶线性预测^[9]。设语音的采样序列为 $\{x(n) | n = 0, 1, \dots, N-1\}$, 则 $x(n)$ 的 p 阶线性预测值为:

$$\tilde{x}(n) = - \sum_{i=1}^p a_i x(n-i) \quad (1)$$

式中, p 为预测系数; $a_i (i = 1, 2, \dots, p)$ 称为线性预测系数。

每一帧语音求解的线性预测系数构成一个 p 维矢量。线性预测分析就是用这 p 维矢量来表示每帧语音。若用 $e(n)$ 来表示预测误差,则

$$e(n) = x(n) - \tilde{x}(n) = x(n) + \sum_{i=1}^p a_i x(n-i) = \sum_{i=0}^p a_i x(n-i) \quad (2)$$

式中, $a_0 = 1$ 。由均方误差最小准则可令 $\frac{\partial E[e^2(n)]}{\partial a_i} = 0, i = 1, 2, \dots, p$, 推得

$$\sum_{k=1}^p a_k R(i-k) = -R(i), i = 1, 2, \dots, p \quad (3)$$

由上式可得 p 个方程,其矩阵表示为:

$$\begin{pmatrix} R(0) & R(1) & \cdots & R(p-1) \\ R(1) & R(0) & \cdots & R(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & \cdots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} -R(1) \\ -R(2) \\ \vdots \\ -R(p) \end{pmatrix} \quad (4)$$

上述方程的解就是 LPC 系数。经典的求解方法有两种:自相关法和协方差法^[10]。其中 Durbin 递推算法^[11]是目前广泛采用的一种自相关方法。

相比于它的预测功能,线性预测能够提供非常好的声道模型和模型参数估计方法,因而被应用于语音信号处理。

线性预测是一种分析语音信号频谱的谱估计方法,之所以是目前最重要的语音特征参数之一,有以下

几个原因:

(1)它提供了很好的短时语音信号的声道模型以及求解模型的方法。

(2)基音、共振峰等模型参数数据量小,容易计算,便于实时处理。

(3)LPC 参数训练得到的模型参数可以存储起来,在语音识别等应用中减少识别时间。

(4)参数数据量小,传输速率低。

LPC 模型阶数 p 的选择主要从两方面考虑:共振峰个数和对口唇辐射影响的补偿。通常 p 的取值范围为 8 至 12 之间。12 阶的 LPC 模型可以以非常小的误差逼近几乎所有的语音信号产生的声道模型。

2.2 梅尔倒谱系数(MFCC)

噪声环境下及其他变异情况下,人耳仍能分辨出语音内容甚至说话人身份,这是因为耳蜗对输入信号的调制作用。对于不同频率的信号,耳蜗基础膜的振动位置也是不同的。实际的声音频率与人耳听到的声音高低不是线性关系,它经过了一个非线性的频率变换。即不同的频率信号,人的听觉系统有不同的响应灵敏度。在 1 000 Hz 以下,实际声音频率与人耳感知到的声音高低成线性关系,而 1 000 Hz 以上为对数关系。Mel 频率反映了实际频率与感知频率的转换关系。其表达式为:

$$f_{\text{mel}} = 2595 \log_{10}(1 + f/700) \quad (5)$$

式中, f 的单位是 Hz。

当两个频率成分的差值超出某个特定值时,这时人耳才能够区分它们。这个特定值被称为临界带宽。根据以上特性,人耳的听觉特性可以用临界频带滤波器来模拟。

一般,采用三角滤波器组^[12]来逼近临界频带滤波器组。

MFCC 就是基于人耳听觉系统的临界效应提出出来的一种倒谱参数。图 2 即为 MFCC 参数提取框图。



图2 MFCC 特征参数提取框图

具体流程为:

(1)将采样后的语音信号进行归一化、端点检测、预加重和分帧加窗预处理后,得到语音信号的矩阵形式,其中每个行向量表示一帧语音。

(2)将预处理后的矩阵形式的语音信号进行离散傅里叶变换,并对语音频谱取模的平方得到能量谱。

(3)通过三角滤波器组对语音信号进行滤波处理。计算出语音信号通过第 m ($1 \leq m \leq M$) 个滤波器后的能量和,其中 M 为滤波器个数。

(4)对每个三角滤波器输出的能量求对数,将得

到 M 个系数。

(5)对这 M 个系数进行离散余弦变换,即得到 MFCC 参数。

文中使用的 MFCC 取其前 1 ~ 12 个,共 12 阶。

3 基于得分融合的说话人辨认系统

语音信号不仅包含说话人特有的个性信息,还蕴含了语义信息,是二者的综合体。迄今为止学者们仍未找出一个能够将二者分离的语音特征参数。现有的语音特征参数都只是表示了语音信号的某些信息。为了比较充分地表征语音信号,提高说话人识别系统的识别率和鲁棒性,特征参数的融合、不同说话人识别系统的融合以及得分融合已经成为了许多学者考虑的一个重要的研究方向。文中提出用 LPC 和 MFCC 分别训练得到的 GMM 的说话人识别系统,并将这两种特征的测试得分进行融合的说话人识别方法。其系统框图见图 3。

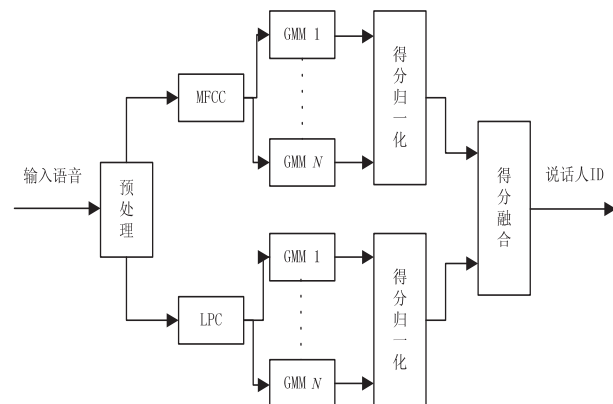


图3 基于LPC和MFCC得分融合的

GMM 模型说话人辨认系统框图

预处理主要包含预加重、分帧、加窗等操作,预加重技术^[13]可以消除口鼻辐射,文中采用预加重滤波器的系数为 0.91,其传递函数为:

$$H(z) = 1 - \alpha z^{-1} \quad (6)$$

建立的说话人模型,考虑了高斯混合成分分别为 16、32、64 三种情况。

3.1 得分归一化

由于测试语音的特征不同,提取得到的数据之间有很大差异,通过系统的匹配计算部分获得的得分变化幅度较大。如果将不同特征获得的得分直接进行融合,很难获得得分融合分布规律,故而文中采用数据归一化方法来处理得分。系统获得的得分是一个向量 $\mathbf{x} = (\alpha_1, \alpha_2, \dots, \alpha_N)$,归一化处理如下:

$$\hat{\mathbf{x}} = (\alpha_1, \alpha_2, \dots, \alpha_N) / \sum_{i=1}^N \alpha_i \quad (7)$$

其中, α_i 表示该测试语音与第 i 个说话人模型的匹配得分; N 为模型库中说话人总数。

3.2 得分融合

得分融合的方法通常有算术平均值^[14],或者是线性加权法^[15],文中采用的是线性加权法。对于不同特征的归一化得分,文中采用将二者进行线性组合的方式进行融合计算。测试语音在提取特征为 MFCC 时,其归一化得分为 $x_1 = (\alpha_1, \alpha_2, \dots, \alpha_N) / \sum_{i=1}^N \alpha_i$; 提取特征为 LPC 时,其归一化得分为 $x_2 = (\beta_1, \beta_2, \dots, \beta_N) / \sum_{i=1}^N \beta_i$ 。将二者进行加权融合,其表达式如下:

$$s = w_1 x_1 + w_2 x_2 \tag{8}$$

文中使用的加权系数为 $w_1 = 0.7, w_2 = 0.3$ 。

4 实验结果和分析

4.1 实验设置

文中采用的语音数据库有 10 名 20 ~ 30 岁说话人,其中 7 名男性,3 名女性。采样频率 f_s 为 8 kHz, 语音文本包含一段约 1 分半钟的长语句、10 个短句子。每个说话人分别在身体正常和感冒的情况下,朗读以上的长语音和 10 句短句子。正常情况下朗读的长语音用来训练说话人模型,说话人身体正常和感冒情况下说的 10 个短句子作为测试语音。说话人身体正常时朗读的长语句作为训练集,共 10 句,短句子 80 句作为测试集 1;说话人感冒时朗读的 80 句短句子作为测试集 2。

4.2 LPC 和 MFCC 系统性能

表 1、表 2 分别给出了特征参数为 LPC 的 GMM 模型说话人辨认系统和特征参数为 MFCC 的 GMM 模型说话人辨认系统的识别率,文中采用的是正确识别率作为系统的评价标准。分别是对测试集 1 和测试集 2 的识别结果。

表 1 基于 LPC 参数的 GMM 说话人辨认

| 高斯混合 成分数 | 正常语音 识别率/% | 感冒语音 识别率/% |
|-------------|---------------|---------------|
| 16 | 93.25 | 86.25 |
| 32 | 96.25 | 90 |
| 64 | 97.5 | 91.25 |

表 2 基于 MFCC 参数的 GMM 说话人辨认

| 高斯混合 成分数 | 正常语音 识别率/% | 感冒语音 识别率/% |
|-------------|---------------|---------------|
| 16 | 98.75 | 90 |
| 32 | 98.75 | 91.25 |
| 64 | 98.75 | 93.75 |

从表 1、表 2 可知,对于同一特征构建的说话人识别系统,感冒语音的识别率比正常语音的识别率低,这

说明说话人感冒时发出的语音的个性特征发生了变化,使得说话人识别系统的性能下降。

4.3 得分归一化和融合后性能分析

将测试语音分别提取特征参数 LPC 和 MFCC,并输入相应的系统模型,经匹配计算求得两个得分,将两者线性加权。其中,LPC 特征参数获得的得分权重为 0.3,MFCC 特征参数的得分权重为 0.7,其实验结果见表 3。

表 3 基于 LPC 和 MFCC 得分融合的 GMM 模型说话人辨认

| 高斯混合 成分数 | 正常语音 识别率/% | 感冒语音 识别率/% |
|-------------|---------------|---------------|
| 16 | 98.75 | 98.75 |
| 32 | 98.75 | 98.75 |
| 64 | 98.75 | 97.5 |

由表 3 可得,对于正常语音,与 LPC 特征系统相比,得分融合系统性能提高的比较显著;对于感冒语音,高斯成分为 16、32、64,得分归一化和决策融合系统性能明显优于单一特征系统,其中在 16 个高斯成分时,与 LPC 特征系统相比,其性能提高 12.5% 左右,比 MFCC 特征系统性能提高 8.5% 左右。

5 结束语

文中提出将 LPC 和 MFCC 两特征分别应用于 GMM 模型的说话人辨认系统,提取测试语音的 LPC 和 MFCC 特征参数,输入相应的说话人识别模型库中进行匹配计算,将这两个特征的得分进行归一化处理后进行线性加权融合,并得出最终的判决结果。实验结果表明,对于正常语音集,文中提出的系统能够较显著地提高说话人辨认系统的性能;对于感冒语音集,说话人辨认系统的性能得到了显著提高,在高斯混合成分 $M = 16$ 时,相对于 LPC 单特征系统提高了 12.5 个百分点,相对于 MFCC 特征系统也提高了 8.75 个百分点。

参考文献:

[1] Furui S. Recent advances in speaker recognition[M]//Audio – and video – based biometric person authentication. Berlin: Springer-Verlag,1997:237–252.

[2] Kinnunen T, Li H. An overview of text – independent speaker recognition;from features to supervectors[J]. Speech Communication,2010,52(1):12–40.

[3] Pawlewski M, Jones J. URU plus – a scalable component – based speaker – verification system for BT’s 21st century network [J]. BT Technology Journal,2007,25(3):170–178.

$$\lambda_k \|A^*\| \sum_{j=1}^r \beta_j d_{\bar{\rho}}(Q_{j,k}, Q_j) \quad (18)$$

其中, $\bar{\rho} \geq \max\{\|\bar{x}^k\|, \|A\bar{x}^k\|, \|y\|\}$ 。

因此,由定理 3 和式 (12) 得到 $\{x^k\}$ 收敛到 H 的一个不动点,即 MSSFP 的解,证毕。

4 结束语

多集合分裂可行问题在现实中的许多领域有着广泛的应用。到目前为止,多集合分裂可行问题的许多算法的求解都是在 R^n 空间中完成的,而在 Hilbert 空间中的推广应用还有待完善。文中基于 R^n 空间中求解多集合分裂可行问题的 KM 迭代算法,给出了 Hilbert 空间中的一种自适应不精确算法,以及算法的收敛性证明。

参考文献:

- [1] Censor Y, Elfving T. A multi-projection algorithm using Bregman projections in a product space [J]. Number Algorithms, 1994, 8: 221–239.
- [2] Xu Hong kun. A variable Krasnosel'ski-Mann algorithm and the multiple-set split feasibility problem [J]. Inverse Problems, 2006, 22: 2021–2034.
- [3] Censor Y. Row-action methods for huge and sparse systems and their applications [J]. SIAM Review, 1981, 23(4): 444–466.
- [4] Censor Y, Bortfeld T, Martin B, et al. Unified approach for inversion problems in intensity-modulated radiation therapy [J]. Physics in Medicine and Biology, 2006, 51: 2353–

2365.

- [5] 杨庆之,赵金玲. 分裂可行问题 (SFP) 的投影算法 [J]. 计算数学, 2006, 28(2): 121–132.
- [6] 王新艳,屈 彪. 求解分裂可行问题逆问题的算法推广 [J]. 泰山学院学报, 2010(6): 10–14.
- [7] Zhao Jinling, Yang Qingzhi. A note on the Krasnoselski-Mann theorem and its generalizations [J]. Inverse Problems, 2007, 23: 1011–1016.
- [8] Censor Y, Motova A, Segal A. Perturbed projections and sub-gradient projections for the multiple-sets split feasibility problem [J]. Journal of Mathematical Analysis and Applications, 2007, 327(2): 1244–1256.
- [9] Bauschke H H, Borwein J M. On projection algorithms for solving convex feasibility problems [J]. SIAM Review, 1996, 38: 367–426.
- [10] Eicke B. Iteration methods for convexly constrained ill-posed problems in Hilbert space [J]. Numerical Functional Analysis and Optimization, 1992, 13(5–6): 413–429.
- [11] Wang C, Xiu N. Convergence of the gradient projection method for generalized convex minimization [J]. Computational Optimization and Applications, 2000, 16(2): 111–120.
- [12] Zarantonello E H. Projections on convex sets in Hilbert space and spectral theory [D]. Wisconsin: University of Wisconsin, 1971.
- [13] 何炳生. 论求解单调变分不等式的一些投影收缩算法 [J]. 计算数学, 1996(1): 97–103.
- [14] 徐成贤,陈志平,李乃成. 近代优化方法 [M]. 北京: 科学出版社, 2002: 18–35.

(上接第 42 页)

- [4] Rose P. Forensic speaker identification [M]. London: Taylor & Francis, 2002.
- [5] Tull R G, Rutledge J C, Larson C R. Cepstral analysis of “cold-speech” for speaker recognition: a second look [J]. Journal of Acoustical Society of America, 1996, 100(4): 2760–2760.
- [6] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Trans on Speech and Audio Processing, 1995, 3(1): 72–83.
- [7] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models [J]. Digital Signal Processing, 2000, 10(1): 19–41.
- [8] Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models [J]. Speech Communication, 1995, 17(1–2): 91–108.
- [9] Akhoul M. Linear prediction of speakers from their voice

[J]. Proc of IEEE, 1976, 64: 460–475.

- [10] 张军英. 说话人识别的现代方法与技术 [M]. 西安: 西北大学出版社, 1994: 14–16.
- [11] 张玲华,郑宝玉. 随机信号处理 [M]. 北京: 清华大学出版社, 2003.
- [12] Zhu Weizhong, O'Shaughnessy D. Incorporating frequency masking filtering in a standard MFCC feature extraction algorithm [C]//Proc of 7th international conference on signal processing. [s.l.]: IEEE, 2004: 617–620.
- [13] 王 青. 基于神经网络的汉语语音情感识别的研究 [D]. 杭州: 浙江大学, 2008.
- [14] Yu P, Seide F T B. A hybrid word/phoneme-based approach for improved vocabulary-independent search in spontaneous speech [C]//Proc of INTERSPEECH 2004. Jeju Island, Korea: [s.n.], 2004: 293–296.
- [15] Chen B. Voice retrieval of Mandarin broadcast news speech [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2006, 20(1): 91–109.

基于 LPC 和 MFCC 得分融合的说话人辨认

作者：[单燕燕](#)，[SHAN Yan-yan](#)
作者单位：[南京邮电大学 通信与信息工程学院, 江苏 南京, 210003](#)
刊名：[计算机技术与发展](#)[ISTIC](#)
英文刊名：
年，卷(期)：2016(1)

引用本文格式：[单燕燕](#),[SHAN Yan-yan](#) [基于 LPC 和 MFCC 得分融合的说话人辨认](#)[期刊论文]-[计算机技术与发展](#)
2016(1)