基于伽马通滤波器组的听觉特征提取算法研究

王 ▮[1,2],钱志鸿[1],王 雪[1],程光明[1]
(1.吉林大学,吉林长春 130025; 2.国家知识产权局,北京 100190)

摘 要: 本文从模拟人类听觉角度出发,给出了基于人耳耳蜗听觉模型的伽马通滤波器组模型,测试语音通过该滤波器组输出得到了高维听觉特征向量.经过主成分分析和离散余弦变换,分别得到了可用于表征说话人的伽马通系数和伽马通滤波器倒谱系数及其衍生特征.实验证明,与传统梅尔倒谱特征相比,采用本文提出特征的说话人识别系统在识别率及鲁棒性上均有明显提高.

关键词: 语音信号处理; 伽马通滤波器; 听觉特征提取; 倒谱系数
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2010) 03-0525-04

An Auditory Feature Extraction Algorithm Based on γ-Tone Filter-Banks

WANG Yue[1,2], QIAN Zhi-hong[1], WANG Xue[1], CHENG Guang-ming[1]
(1. *Jilin University*, *Changchun*, *Jilin* 130025, *China*; 2. *State Intellectual Property Office of the PRC*, *Beijing* 100190, *China*)

Abstract: By means of emulating human auditory, gamma-Tone filter-banks models based on the auditory system in human cochlea are presented. The speech to be detected goes through the gamma-Tone filter-banks, thereby multi-dimension eigenvectors are obtained. By PCA (principal component analysis) and DCT (discrete cosine transform), it is yielded to represent a speaker's gamma-Tone coefficients, gamma-Tone filter-banks cepstral coefficients respectively and their derivative features as well. Compared to the ordinary Mel-frequency cepstral coefficients, the speaker recognition system presented turns out to have better recognition rate and robustness characteristics.

Key words: speech signal processing; gammatone filter; auditory feature extraction; cepstral coefficients

1 引言

一个典型的说话人识别系统通常提取的说话人特征为时变特性参数如梅尔倒谱系数(Mel-frequency Cepstral Coefficients, MFCC)[1],感知线性预测系数(Perceptual Linear Prediction – PLP)[2]或韵律特征[3].然而,实际使用时由于受到噪音干扰,训练与识别传输通道不匹配时,说话人系统通常不能表现良好[4].为了解决这一问题,通常采用谱减法来消除噪声,但噪声为非平稳时其有效性大大降低[5].RASTA 滤波和倒谱均值归一化(CMN)[6]也被用于去除卷积噪声,消除信道失真.然而,这些去噪手段不能处理未知的干扰类型,因为它们依赖于事先预知的噪音形式.因此,提取具有鲁棒性的语音特征在说话人识别中显得尤其重要,文献[7]中提到了一种基于听觉模型的汉语声调检测算法,采用听神经平均发放率作为特征得到了较好的声调识别效果.本文中,提取了两种基于人类听觉特性的说话人识别特征:伽马通滤波器系数(γ-Tone Filter, GTF)和伽马通滤波器倒谱系数(γ-Tone Filter Cepstral Coefficients, GFCC)以及其衍生特征.实验证明,基于人耳听觉特性的这两类特征在说话人识别系统中的表现要优于 MFCC 及其差分特征.在三种不同噪声背景下进行测试比对,GTF 系数与 GFCC 倒谱系数及其衍生特征仍然获得了较高的识别率.

2 耳蜗听觉滤波器组模型

2.1 γ-Tone 滤波器组

γ-Tone 滤波器是一个标准的耳蜗听觉滤波器,该滤波器组冲击响应的典型模式为

$$g_i(t) = At^{N-1}\exp(-2\pi \mathrm{ERB}(f_i)t)\cdot\cos(2\pi f_i t + \varphi_i)U(t),\quad t\geqslant 0, 1\leqslant i\leqslant N \qquad (1)$$

其中,A 为滤波器增益,N 为滤波器阶数,f_i 是中心频率,φ_i 是相位,简化模型中取 $\varphi_i = 0$. $\mathrm{ERB}(f_i)$为等效矩形带宽(Equivalent Rectangular Bandwidth, ERB),它决定了脉冲响应的衰减速度,与滤波器带宽有关,而每个滤波器带宽与人耳听觉临界频带(Critical Band, CB)有关,听觉心理学中,$\mathrm{ERB}(f_i)$可以由式(2)得到

$$\mathrm{ERB}(f_i) = 24.7(4.37\frac{f_i}{1000}+1) \qquad (2)$$

收稿日期:2009-03-23;修回日期:2010-01-07

每个滤波器的带宽由上式决定,其中中心频率 f_i 在后面的计算中给出.本文取 $N=64$,由 64 个滤波器叠加成 64 通道 γ-Tone 滤波器组来实现耳蜗滤波器模型.各个 γ-Tone 滤波器的中心频率按照 ERB 的关系,在 30Hz 到 4000Hz 之间分布.每个中心频率 f_i 可由式(3)计算得出

$$f_i=(f_H+228.7)\exp(-\frac{v_i}{9.26})-228.7\ ,\ 1\leqslant i\leqslant N \tag{3}$$

其中 f_H 为滤波器的截止频率,v_i 是滤波器重叠因子,用来指定相邻滤波器之间重叠百分比.每个滤波器的中心频率确定后,相应的带宽可由式(2)计算得出.

2.2 γ-Tone 滤波器的实现

对式(1)进行拉普拉斯变换

$$\begin{aligned}G_i(s)&=\int_{-\infty}^{\infty}g_i(t)\mathrm{e}^{-st}\mathrm{d}t\\&=\int_{-\infty}^{\infty}At^{N-1}\exp(-2\pi\mathrm{ERB}(f_i)t)\\&\quad\cdot\cos(2\pi f_it+\phi_i)U(t)\mathrm{e}^{-st}\mathrm{d}t\\&=\frac{A}{2}\int_0^{\infty}t^{N-1}\mathrm{e}^{(-2\pi\mathrm{ERB}(f_i)t)}(\mathrm{e}^{\mathrm{j}2\pi f_it}+\mathrm{e}^{-\mathrm{j}2\pi f_it})\mathrm{e}^{-st}\mathrm{d}t\\&=\frac{A}{2}\left[\frac{(n-1)!}{(s+b-\mathrm{j}w)^n}+\frac{(n-1)!}{(s+b+\mathrm{j}w)^n}\right]\end{aligned} \tag{4}$$

其中 $b=2\pi\mathrm{ERB}(f_i)$,$w=2\pi f_i$,将 $G_i(s)$转换为 Z 变换 $G_i(z)$形式,再反变换得到

$$g_i(n)=\frac{1}{2\pi\mathrm{j}}\int G_i(z)z^{n-1}\mathrm{d}z \tag{5}$$

将语音信号与 $g_i(n)$卷积后就得到 γ-Tone 滤波器的滤波输出 $G(n)$.图 1 所示是覆盖 30 ~ 4000Hz 的 64 通道 γ-Tone 滤波器组的对数幅频响应(每四组一条曲线).为了尽可能去除所得到的 γ-Tone 特征矢量的冗余度,我们对高维数据进行主成分分析(Principal Component Analysis,PCA)处理,将样本从原始高维空间降到中间维,去除噪声和变化不大的特征,降低计算复杂度.

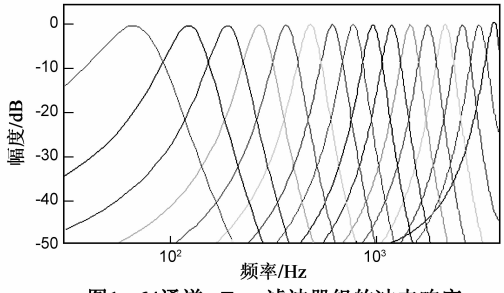


图1 64通道γ-Tone滤波器组的冲击响应

2.3 PCA 变换

首先计算 64 维 γ-Tone 滤波器系数矩阵的协方差矩阵和 PCA 转换矩阵,文献[8]证明,PCA 的主方向就是协方差阵的特征值对应的特征向量,要将原向量降到目标维数,只需取最大特征值对应的特征向量组成 PCA 转换矩阵即可,根据主成分累计贡献率计算公式 $w_k=\sum_{i=1}^{p}\lambda_i/\sum_{i=1}^{N}\lambda_i$,其中 λ_i 为 S 的第 i 个特征根,按照累积贡献率不小于 85% 的准则,将 64 维降到 25 维.

2.4 DCT 变换

另一方面,对于 64 维 γ-Tone 滤波系数,由于相邻的特征分量之间相关性很大,因此对 γ-Tone 滤波器系数做离散余弦变换(Discrete Cosine Transform,DCT).将通过 DCT 变换之后的新特征称为 γ-Tone 滤波器倒谱系数 GFCC.实际数据表明,经过 DCT 变换后的 GFCC 参数,低 22 维系数占据了全部 GFCC 参数的主要特征信息,而高于 22 维的 GFCC 值都接近于 0,提供的信息几乎可以忽略.因此本文采用 22 维 GFCC 作为特征向量.

3 实验结果

实验采用 PKU-SRSC[9]语音数据库.测试集选 80 人.混入噪声选自 NOISEX-92 库.采用 Gauss 混合模型(GMM)作为分类器[10],每个 GMM 由 16 个分量构成.采用最大似然法进行训练,用扩展 Bauman-Welch 算法迭代 8 次.基线系统采用 HTK[11]算法包得到的 24 阶的 MFCC 特征以及衍生的一阶 MFCC_D 和二阶 MFCC_D_A.

首先在纯净语音理想情况下,对所提取的 GTF、GFCC 以及其一二阶衍生特征与 MFCC 基线系统进行比较.训练语音为 2min,测试语音约 10s.实验采用检测错误代价函数 C_{DET}分别从漏检和误检两个角度进行评测.

对比采用 MFCC 的基线系统,采用 GTF 与 GFCC 特征的系统性能都有比较明显的提升,实验中采用的所有特征的 DET 性能曲线图如图 2,图 3 所示.图 2 中可以看出 GFCC 的 DET 曲线更接近图中左下方,因此要优于其他两种特征,GTF 与 MFCC 从 DET 曲线上没有明显差异.在 x 轴体现的误检率上 GTF 要优于其他两种特征.图 3 是这三种特征的一阶二阶衍生特征共 6 种特征的 DET 曲线.最接近坐标原点的三条曲线分别是 GFCC_D_A,GTF_D 和 GTF_D_A,说明二阶衍生特征的

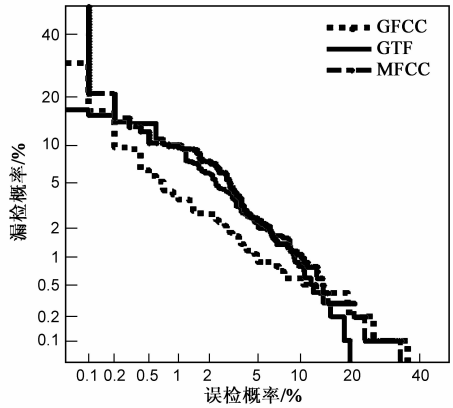


图2 采用MFCC,GTF,GFCC特征的DET性能曲线

性能要普遍要好于一阶衍生特征,其中采用 22 维 GFCC_D_A 参数的系统性能最佳,优于采用其他一阶二阶衍生特征的性能.

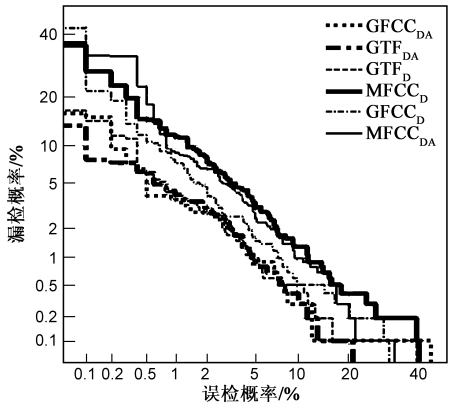


图3　采用差分系数特征的DET性能曲线

为了测试噪声环境下新特征的识别性能,选取噪声库中 3 种典型噪声作为测试系统的背景噪声.基线系统选取 MFCC,MFCC_D 和 MFCC_D_A 作为特征参数,不同于 DET 曲线,这里我们只考虑误检的情况.表 1 给出了 3 种典型噪声在信噪比为 15dB 下各种特征的测试结果.

表 1　基线系统、GTF 和 GFCC 特征系统识别率

	White	Babble	Factory
MFCC	75.0	70.7	74.3
MFCC_D	82.3	75.3	79.3
MFCC_D_A	82.7	74.7	80.0
GTF	74.3	72.0	75.7
GFCC	78.0	73.0	77.0
GTF_D	76.7	72.3	75.7
GFCC_D	81.3	78.7	82.3
GTF_D_A	85.7	78.3	84.7
GFCC_D_A	85.3	81.3	82.0

对比于基线系统,在三种不同噪声背景下,GTF 与 GFCC 在系统识别率上均有不同程度的提高.三种噪声中,白噪声平均识别率最高,Babble 噪声由于受到"鸡尾酒会"效应的干扰,系统识别率要低于其他噪声背景.同纯净语音测试结果一样,GTF 和 GFCC 以及其一阶二阶衍生特征在识别率上均要高出基线系统,说明该类特征对加性噪声具有一定的抑制作用.如 White 噪声背景下,具有最优表现的 GFCC_D_A 特征参数的识别率要高出 MFCC 特征 10.3 个百分点,验证了基于人耳耳蜗听觉特征的噪声鲁棒性.

4　结论

本文给出了基于人耳耳蜗听觉特性的 γ-Tone 滤波器组模型,通过 PCA 变换和 DCT 变换,分别得到了 GTF 和 GFCC 特征,实验仿真表明,这两类特征及其衍生特征在说话人识别任务中表现均优于传统的 MFCC 特征.在加性噪声条件下也能将噪声影响最小化,得到较好的识别效果.在说话人识别应用中,与传统的 MFCC 及其衍生特征相比,GTF 与 GFCC 及衍生特征参数均表现出了更好的鲁棒性以及更高的识别率.

参考文献:

[1] S Furui. Digital Speech Processing, Synthesis, and Recognition [M]. New York: Marcel Dekker, 2001.

[2] H Gish, M Schmidt. Text-independent speaker identification [J]. IEEE Signal Proc, 1994, 11(4): 18 – 32.

[3] D A Reynolds, et al. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing [C]. Hong Kong, China: IEEE, 2003. 4: 784 – 787.

[4] A Drygajlo, M El-Maliki. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing [C]. Seattle, USA: IEEE, 1998. 1: 121 – 124.

[5] SHAO Y, WANG D L. Robust speaker recognition using binary time-frequency masks [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing [C]. Toulouse: IEEE, 2006. 1: 645 – 648.

[6] WNG L, KITAOKA N, NAKAGAWA S. Analysis of effect of compensation parameter estimation for CMN on speech/speaker recognition [A]. 9th International Symposium on Signal Processing and Its Applications [C]. Sharjah: IEEE, 2007. 1 – 4.

[7] 陈雪勤,赵鹤鸣.基于听觉模型的汉语耳语音声调检测 [J].电子学报,2009,37(4):864 – 867.
CHEN Xue-qin, ZHAO He-ming. Perceiving of tone in whispered chinese based on auditory model [J]. Acta Electronica Sinica, 2009, 37(4): 864 – 867. (in Chinese)

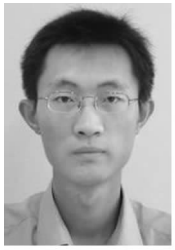
[8] Z Wanfeng, Y Yingchun, W Zhaohui, S Lifeng. Experimental evaluation of a new speaker identification framework using PCA [A]. IEEE International Conference on Systems, Man and Cybernetics [C]. Washington, DC: IEEE, 2003. 4147 – 4152.

[9] WU Xihong. A Chinese Speech Database for Speaker Recognition [EB/OL]. http://nlpr-web.ia.ac.cn/english/irds/chinese/sinobiometrics-pdf/wuxihong.pdf, 2002.

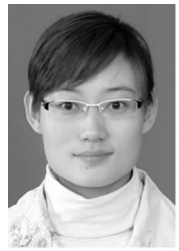
[10] D A Reynolds, R C Rose. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. Proc IEEE Trans Speech Audio Process, 1995, 3(1): 72 – 83.

[11] YOUNG S, EVERMANN G, GALES M, et al. The HTK Book [M]. Cambridge: Cambridge University, 2006.

作者简介:



王 ▮ 男,1980 年出生于黑龙江大庆.2006 与 2009 年毕业于吉林大学通信工程学院,分别获得硕士和博士学位,研究方向为语音信号处理和 DSPs 设计与应用.
E-mail:wawa543@yeah.net



王 雪 女,1984 年生,吉林大学通信工程学院博士研究生,研究方向为基于正交频分复用的超宽带同步技术.
E-mail:yeti_1019@yahoo.com.cn

钱志鸿 男,1957 年生,教授,博士生导师,主要从事领域为无线网络通信系统的信号分析和处理、通信系统微弱信号检测理论与应用等.
E-mail:dr.qzh@163.com

程光明 男,1957 年生,教授,博士生导师,主要研究方向为压电驱动与控制技术.