基于听觉模型的说话人语音特征提取*

何朝霞，潘 平

（贵州大学 计算机科学与信息学院，贵州 贵阳 550025）

摘 要： 基于听觉模型的特性，仿照 MFCC 参数提取过程，提出了一种基于 Gammatone 滤波器组的说话人语音特征提取方法。该方法用 Gammatone 滤波器组代替三角滤波器组求得倒谱系数，并且可以调整 Gammatone 滤波器组的通道数和带宽。将该方法所求得的特征在高斯混合模型识别系统中进行仿真实验，实验结果表明，该特征在一定情况下优于 MFCC 特征在系统的识别率，同时在 Gammatone 滤波器组通道数较高或滤波器带宽较小的情况下，系统具有较高的识别率。

关键词： 听觉模型；Gammatone 滤波器组；MFCC；特征；识别率

中图分类号： TN912.3 **文献标识码：** A **文章编号：** 1674-7720(2012)01-0037-03

Feature extraction for speaker recognition based on auditory model

He Zhaoxia, Pan Ping

(College of Computer Science & Information, Guizhou University, Guiyang 550025, China)

Abstract: In this paper, a novel feature based on an auditory model and Gammatone filter band is proposed for speaker recognition, which imitates the parameters extraction process of MFCC. The frequency cepstrum coefficient features are calculated using a Gammatone filter band instead of commonly used triangle filter band. Moreover, the dimension and the equivalent rectangular bandwidth of Gammatone filter band could be adjusted. Simulation results with Gaussian mixture model indicate that the recognition rate is significantly improved compared with MFCC in some condition, and the correct recognition rate is higher by more dimensions or smaller equivalent rectangular bandwidth.

Key words: auditory model; Gammatone filter band; MFCC; feature; recognition rate

声音的感受细胞在内耳的耳蜗部分，而基底膜是耳蜗接收声音最重要的组织。声波在外耳腔引起空气振动，从而引起行波沿基底膜的传播[1]。基底膜内有许多平行走向的胶原样纤维，称为听弦。听弦长短不同，靠近蜗底较窄，靠近蜗顶较宽。基底膜约有 24 000 条听弦，能够对不同频率的声音产生共鸣，分别反映不同频率的声音[2]。不同频率的声音产生不同的行波，其峰值出现在基底膜的不同位置上，研究发现，不同的声音频率沿着基底膜的分布是对数型的[3]。

早在 1992 年，PATTERSON R 就提出了耳蜗模型，该模型是基于一系列带通滤波器——Gammatone 滤波器组[4]实现的，该滤波器组能够很好地模拟基底膜的分频特性。本文提出了一种基于 Gammatone 滤波器组的特征提取方法，该方法能够很好地提取说话人语音信号的特征，并且具有很高的识别率。

* 基金项目：国家科技计划基金资助项目（2008RR0003）；贵州省国际科技合作计划基金资助项目（[2009]700109，[2009]700125）

1 Gammatone 滤波器

Gammatone 滤波器的时域表达形式[5]为：

$$g(t)=\frac{at^{n-2}\cos(2\pi ft+\phi)}{\exp(2\pi ERB(f)t)},t\geqslant 0 \quad (1)$$

其中，a 为滤波器增益，f 为中心频率，ϕ 为相位，n 为滤波器阶数。各种研究表明，$n=4$ 时，Gammatone 滤波器就有很好的模拟特性。$ERB(f)$ 为 Gammatone 滤波器的等效矩形带宽[6]，它与中心频率 f 的关系[7]为：

$$ERB(f)=24.7(4.37f/1\ 000+1) \quad (2)$$

式(2)还可以写成如下形式：

$$ERB(f)=\frac{f}{EarQ}\cdot order+\min BW \quad (3)$$

其中，$EarQ=9.264\ 49$，$\min BW=24.7$，$order=1$。

由于在实际应用中，增益 a 和初始相位 ϕ 不会影响滤波器的性能，因此可以忽略，所以只要确定 Gammatone 滤波器的中心频率，其性能也就确定了。中心频率 f 的计算公式[8]为：

$$f=(f_H+228.7)\exp\left(-\frac{v}{9.264\ 49}\right)-228.7 \quad (4)$$

其中，f_H 为滤波器的截止频率，v 为滤波器的重叠因子。

Gammatone 滤波器的时域表达式为冲击响应函数，将其进行傅里叶变换就可以得到其频率响应特性。不同中心频率的 Gammatone 滤波器的幅频响应曲线如图 1 所示。

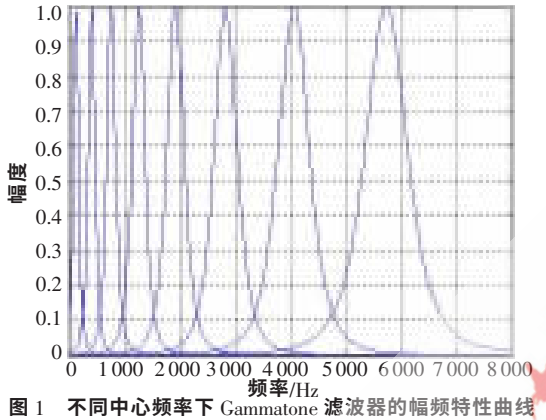


图 1 不同中心频率下 Gammatone 滤波器的幅频特性曲线

取 $n=4$，将式(1)进行拉普拉斯变换得到：

$$G(S)=6a\frac{s^4+4bs^2+6(b^2-c^2)s^2+(4b^2-12bc^2)s+b^4-6b^2c^2+c^4}{[(s+b)^2+c^2]^4} \quad (5)$$

其中，$b=2\pi ERB(f)$，$\omega=2\pi f$。

将 $G(s)$ 转换为 z 变换 $G(z)$，再反变换得到：

$$g(n)=\frac{1}{2\pi j}\int G(z)z^{n-1}dz \quad (6)$$

将语音信号与 $g(n)$ 卷积就可以得到滤波器的输出。

2 特征提取过程

从上述 Gammatone 滤波器的介绍，仿照 MFCC 参数提取过程，考虑将 Gammatone 滤波器组运用到说话人识别中参数的提取过程，这样就更加符合人耳的听觉特性。该提取过程如图 2 所示，具体步骤如下。

图 2 基于 Gammatone 滤波器组的参数提取流程

(1)为了提升高频部分，使信号的频谱变得平坦，将语音信号经过预加重数字滤波器 $H(z)=1-0.937\ 5z^{-1}$。

(2)将预加重后的信号进行分帧，帧长 256 点，帧移 100 点，加汉明窗；再经过离散傅里叶变换(DFT)得到频谱特性，求出频谱平方，即能量谱。

(3)设计 Gammatone 滤波器组。Gammatone 滤波器组的中心频率在 50 Hz~3 000 Hz 之间。这里采用的是 4 阶 Gammatone 滤波器，其通道数 N 和带宽可以调节，根据式(3)，取 $0<order\leqslant 1$。图 3 为同一中心频率下不同 $order$ 值的 Gammatone 滤波器的幅频曲线。

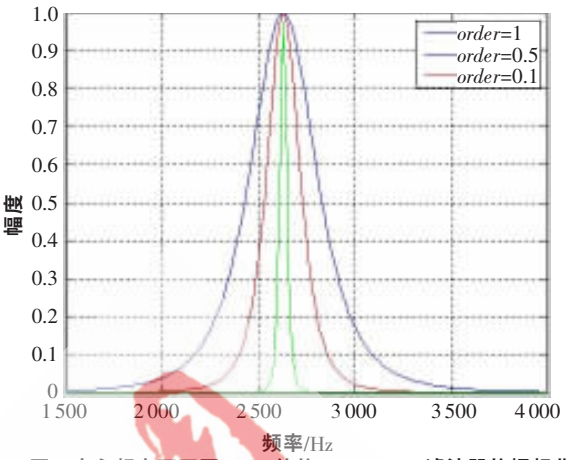


图 3 同一中心频率下不同 $order$ 值的 Gammatone 滤波器的幅频曲线

(4)经过 Gammatone 滤波器组后得到相应频带的能量，再进行对数运算和反离散余弦 IDCT 变换，就可以得到静态特征参数了。

3 仿真实验

仿真实验的语料库来源于贵州省公安厅提供的语料及学校部分学生随机利用 MP4 所得的录音，语料时间各不相同。采用高斯混合模型(GMM)进行与文本相关的说话人确认实验。

图 4 为 48 通道 Gammatone 滤波器组($order=1$ 时)的幅频曲线。

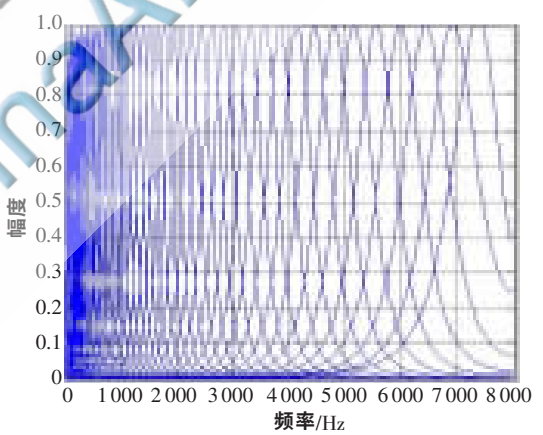
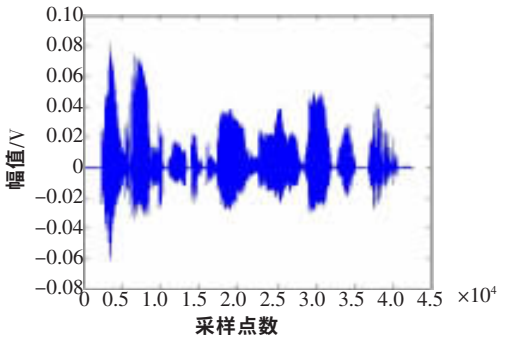


图 4 48 通道 Gammatone 滤波器组的幅频曲线

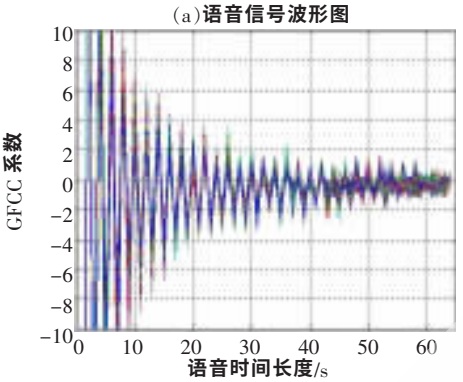
图 5 为某一说话人语音的波形及其经过特征提取系统(系统中 Gammatone 滤波器组为 48 通道)所得到的 GFCC 系数输出曲线。

从图 5(b)可以可以看出，该系数含有丰富的内容，对识别率的提高有很大的帮助。下面将该特征运用到 GMM 识别系统中，具体结果如下。

首先是不同时间长度的语音信号，时间长度分别为 5 s、20 s、50 s，将其在 64 通道 Gammatone 滤波器组所得到的静态特征参数(简称 GFCC)与 MFCC(Mel 滤波器组维数为 24)静态参数在识别系统中进行了识别率的对比，其结果如图 6 所示。



(a)语音信号波形图



(b)语音信号 GFCC 系数输出波形图

图 5 某说话人语音信号及 GFCC 波形曲线

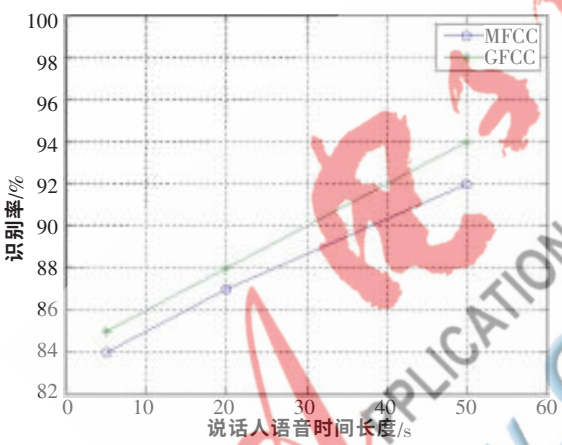


图 6 不同时间长度信号在系统中识别率

从图 6 可以看出，64 通道的 GFCC 静态特征参数比 MFCC 静态特征具有更好的识别率。

同时，将不同 *order* 值、不同滤波器组通道数所得的 GFCC 参数在识别系统中进行了识别率比较，如图 7 所示。其中，*order* 值分别为 0.1、0.5、1，滤波器组通道数分别为 48、64、128。从图 7 可以看出，滤波器组通道数越高，识别率越高；*order* 值越小，识别率越高。

本文介绍了基于人耳听觉特性的 Gammatone 滤波器组的特征提取方法，并通过实验验证了该特征在滤波器通道数较多或 $ERB(f)$ 较小时具有较高的识别率。但是同时也得出只有在滤波器组通道数较高时才有较高的识别率，增加了数据的复杂度。在以后的研究中需要考虑通过降低滤波器组的通道数提高识别率的方法。

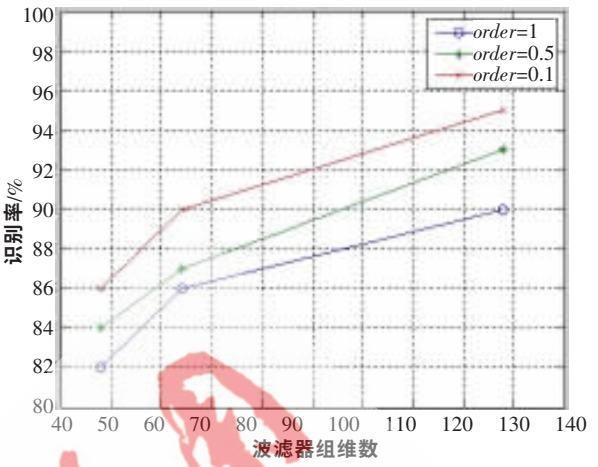


图 7 不同 *order* 值，不同滤波器组维数情况下系统识别率

参考文献

[1] JOHANNESMA P I M. The pre-response stimulus ensemble of neurons in the cochlear nucleus [C]. Proceedings of the Symposium on Hearing Theory, 1972:58-69.

[2] COOKE M P. Modeling auditory proeessing and organization[M]. Cambridge, U.K: Cambridge University Press, 1993.

[3] 韩纪庆，张磊，郑浩然.语音信号处理[M].北京：清华大学出版社，2008.

[4] SLANEY M. An efficient implementation of the patterson-holdswort auditory filter bank. Apple ComPuter Teehnieal RePort#35 Pereeption GrouP-Advaneed Technology GrouP [R]. ComPuter, Inc:Apple, 1993.

[5] Shao Yang, Wang Deliang. Robust speaker identification using auditory features and computational auditory scene analysis [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008, 5:1589.

[6] SRINIVASAN S, Wang Deliang. Transforming Binary uncertainties for robust speech recognition [C]. IEEE Transactions on Audio, Speech and Language Processing, 2007, 15(7):2130-2140.

[7] Wang Deliang, BROWN G J. Computational auditory scene analysis: principles, algorithms, and applications[M]. Hoboken, NJ: Wiley-IEEE Press, 2006.

[8] 王男，钱志鸿，王雪，等.基于伽马通滤波器组的听觉特征提取算法研究[J].电子学报，2010，38(3).

（收稿日期：2011-10-19）

作者简介：

何朝霞，女，1984 年生，硕士研究生，主要研究方向：语音信号处理。

潘平，男，1962 年生，副教授，主要研究方向：信息处理。