第 38 卷 第 21 期
Vol.38 No.21

计 算 机 工 程
Computer Engineering

2012 年 11 月
November 2012

·人工智能及识别技术· 文章编号：1000—3428(2012)21—0168—03 文献标识码：A 中图分类号：TP391

基于 Gammatone 滤波器组的听觉特征提取

胡峰松[1,2]，曹孝玉[1]

(1. 湖南大学信息科学与工程学院，长沙 410082；2. 北京师范大学管理学院，北京 100875)

摘 要：目前主流说话人特征参数在噪声环境中的鲁棒性较差。为此，提出一种可用于说话人识别的听觉倒谱特征系数。分析人耳听觉模型的工作机理，采用 Gammatone 滤波器组代替传统的三角滤波器组模拟人耳耳蜗的听觉模型，用指数压缩代替固定的对数压缩，模拟人耳听觉模型处理信号的非线性特性。在基于高斯混合模型分类器的识别算法下进行仿真实验，结果表明，该听觉特征具有比梅尔频率倒谱系数和线性预测倒谱系数更好的抗噪声能力。

关键词：说话人识别；特征提取；Gammatone 滤波器；听觉模型；倒谱系数；鲁棒性

Auditory Feature Extraction Based on Gammatone Filter Bank

HU Feng-song[1,2], CAO Xiao-yu[1]

(1. College of Information Science and Engineering, Hunan University, Changsha 410082, China;
2. School of Management, Beijing Normal University, Beijing 100875, China)

【Abstract】Aiming at the problem that speaker's feature coefficients have poor robustness in noise environment, this paper proposes an auditory cepstral coefficient for speaker recognition. It analyzes the working mechanism of the human auditory model, simulates the auditory model of human ear cochlea by Gammatone filter banks replaces the traditional triangular filter banks. Based on the nonlinear signal processing capability of human auditory model, exponential compression is used instead of the fixed logarithm compression. Simulation experiment is conducted based on Gaussian Mixed Model(GMM) recognition algorithm. Experimental results show that the auditory feature has better noise robustness than Mel Frequency Cepstral Coefficient(MFCC) and Linear Prediction Cepstral Coefficient(LPCC).

【Key words】speaker recognition; feature extraction; Gammatone filter; auditory model; cepstral coefficient; robustness

DOI: 10.3969/j.issn.1000-3428.2012.21.045

1 概述

说话人识别是指从说话人的语音中提取说话人的个性特征对说话人身份进行认证的技术，其特征参数提取即提取语音信号中表征说话人的个性特征，它是说话人识别的关键技术之一。目前，在说话人识别中常用的特征参数有梅尔频率倒谱系数(Mel Frequency Cepstral Coefficient, MFCC)、线性预测倒谱系数(Linear Prediction Cepstrum Coefficient, LPCC)、基音周期等。

众所周知，在有噪声及多个说话人的复杂环境下，人耳仍然能够辨认出说话人的身份。因此，将人耳听觉处理特性融入到说话人识别系统中可以极大地提高系统的性能。近年来的研究发现，人耳的听觉系统具有十分优异的语音识别能力及噪声鲁棒性，这种优良特性吸引了众多的研究者从事人耳听觉模型的研究。文献[1]提出了基于人类听觉特性的伽马通滤波器系数和伽马通滤波器倒谱系数。文献[2]利用动态压缩 Gammachirp 听觉滤波器组提取话者特征参数，提高了系统的识别率。文献[3]对基于听觉滤波器模型的特征参数及其历史进行了研究，分析了各模型的优缺点。文献[4]对近 30 年听觉外周计算模型的研究及其在语音识别领域的应用进行了评述。文献[5]对听觉系统的非线性压缩进行了研究，并论证了 MFCC 提取过程中对数压缩的缺点。

本文在对人耳听觉模型研究的基础上，用 Gammatone 滤波器模拟人耳耳蜗的听觉模型，用指数压缩代替固定的对数压缩来模拟人耳听觉模型处理信号的非线性特性，提出了一种基于 Gammatone 滤波器组的听觉特征提取方法。

2 基于人耳听觉模型的特征参数提取

人耳生理学研究表明，人耳听觉系统主要由外耳、中耳和内耳构成。语音信号在听觉系统中，依次通过外耳、中耳和内耳，在经过耳蜗基底膜的频带分解作用后，沿听觉通路进入听觉中枢系统[6]。在整个听觉系统中，耳蜗是非常重要的核心部件。当外界的语音信号传入到耳蜗基底膜之后，基底膜将产生以行波传递形式的振动，且基底膜振动的听觉响应与受刺激的语音信号频率有关；基底膜的这种频率分解作用是人耳听觉系统进行声音信号处理的重要环节。在语音识别中，通常采用一组相互交叠的带通滤波器组模拟实现耳蜗基底膜的频率分解作用，本文采用 Gammatone 滤波器组实现耳蜗模型。

作者简介：胡峰松(1969－)，男，副教授、博士，主研方向：语音识别，人脸识别；曹孝玉，硕士研究生

收稿日期：2012-02-14 **修回日期：**2012-03-12 **E-mail：**cxy131517@163.com

2.1　Gammatone 滤波器组

Gammatone 滤波器是一个标准的耳蜗听觉滤波器，其滤波器的时域脉冲响应为：

$$g_i(t)=At^{n-1}\exp(-2\pi b_i t)\cos(2\pi f_i+\phi_i)U(t),\ t\geqslant 0, 1\leqslant i\leqslant N \quad (1)$$

其中，A 为滤波器增益；f_i 是滤波器的中心频率；$U(t)$ 为阶跃函数；ϕ_i 是相位，为了简化模型，本文取 $\phi_i=0$；n 是滤波器的阶数，本文取 n=4；b_i 为滤波器的衰减因子，它决定了脉冲响应的衰减速度，并与相应的滤波器的带宽有关，$b_i=1.019ERB(f_i)$，$ERB(f_i)$ 为等效矩形带宽，它可以由式(2)得到：

$$ERB(f_i)=24.7\times(4.37\times\frac{f_i}{1\ 000}+1) \quad (2)$$

其中，N 为滤波器个数，本文取 $N=64$，即由 64 个 Gammatone 滤波器叠加成的带通滤波器组实现耳蜗模型。各滤波器的中心频率在 ERB 域上等间距分布，整个滤波器组的频率覆盖范围为 80 Hz~8 000 Hz。图 1 给出了其频率响应示意图。

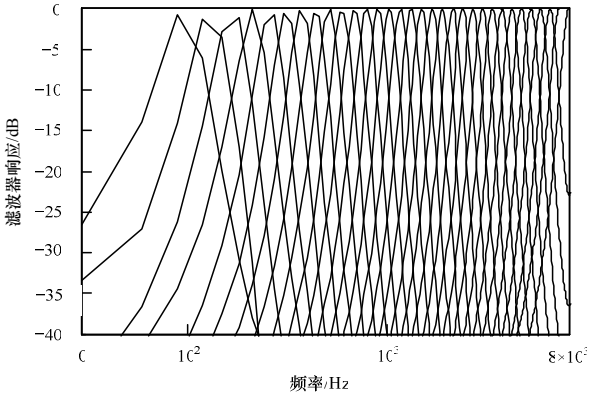

图 1　Gammatone 滤波器组的冲击响应

2.2　听觉系统的非线性压缩——指数压缩

语音信号处理及听觉研究表明，非线性特性是听觉系统具有抗干扰能力的重要原因之一[4]。实际上听觉系统的非线性特性是“指数压缩”的且由低频到高频非线性逐渐增强[7]。听觉系统的非线性估计的基本方法就是测量语音信号通过听觉系统的输出输入比，而输出输入通常采用信号的声压级来表示，如式(3)所示：

$$\frac{10\times\lg(\frac{P_o}{P_r})}{10\times\lg(\frac{P_i}{P_r})}=\lambda,\lambda\leqslant 1 \quad (3)$$

其中，P_i 表示输入信号的功率；P_o 表示输出信号的功率；P_r 表示参考声音信号的功率；λ 表示输出输入信号的比值。式(3)可以进一步转换为：

$$\frac{P_o}{P_r}=(\frac{P_i}{P_r})^{\lambda} \quad (4)$$

上式说明听觉系统的非线性是服从指数压缩的。非线性压缩具体数值的选择对倒谱系数的性能非常重要，由于具体实验方法的不同，各文献中给出的值也不同。但目前有 2 点结论在听觉系统领域得到了普遍认可：(1)频率在 1 kHz 以上的信号，非线性压缩行为应该比较强，且具体数值也比较接近；(2)频率在 1 kHz 以下的信号，非线性压缩行为随着频率的降低越来越弱。

通过对仿真实验结果的多次分析，本文对 1 kHz 以上的非线性压缩指数采用常数值 0.2；对 1 kHz 以下部分，规定 500 Hz 对应的非线性压缩值为 0.7，0 Hz 对应的非线性压缩值为 0.8，其他频率处的压缩值由线性插值的方法获得，其压缩值与频率的关系如图 2 所示。

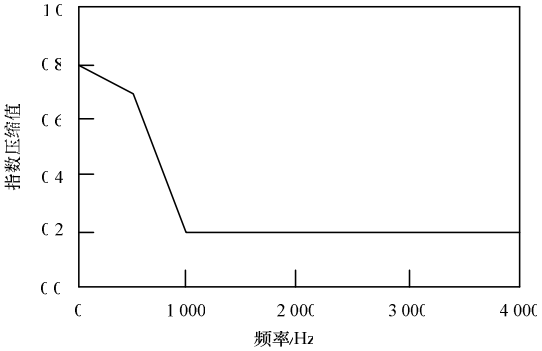

图 2　指数压缩值与频率的关系

2.3　听觉特征提取

本文采用 Gammatone 滤波器组模拟人耳耳蜗听觉模型，同时采用指数压缩来实现人耳听觉系统的非线性特性，提出了一种基于 Gammatone 滤波器的听觉模型倒谱特征参数，记为 GFCC(Gammatone Frequency Cepstrum Coefficient)。GFCC 特征提取流程如图 3 所示。

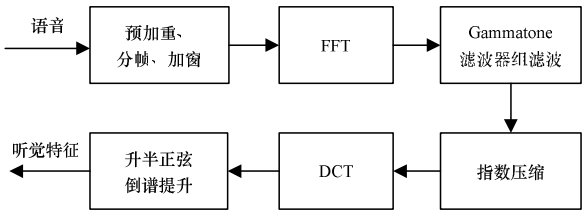

图 3　GFCC 特征提取流程

GFCC 特征参数提取算法如下：

(1)预加重、分帧和加窗。为加强高频信号，需要对语音信号进行预加重处理，预加重系数为 0.97。假设 $x(n)$ 是原始的语音信号，则其预加重之后的信号 $y(n)$ 为：

$$y(n)=x(n)-0.97\times x(n-1) \quad (5)$$

根据语音信号的短时平稳特性，把语音信号分成若干帧，每一帧的帧长为 256 采样点、帧移为 50%。

为了减少语音帧的边缘影响，对语音信号加汉明窗。汉明窗的数学公式见式(6)，加窗后的语音信号 $s_w(n)$ 见式(7)：

$$w(n)=\begin{cases}0.54-0.46\times\cos(\frac{2\pi n}{N-1}), & \text{if } n=0,1,\cdots,N-1\\ 0 & \text{otherwise}\end{cases} \quad (6)$$

$$s_w(n)=y(n)\times w(n) \quad (7)$$

(2)快速傅里叶变换(Fast Fourier Transform, FFT)。对加窗后的语音信号进行快速傅里叶变换，把语音信号由时域变到频域，得到语音信号的离散功率谱 $X(k)$。

(3)Gammatone 滤波器组滤波。对功率谱 $X(k)$ 取平方

得到能量谱，然后用 Gammatone 滤波器组进行滤波处理。

(4)指数压缩。对每个滤波器的输出进行指数压缩，得到一组对数能量谱 $m_1, m_2, \cdots, m_P$。

$$m_i = \sum_{k=1}^{N} [X(k)^2 \times H_i(k)]^{e(f)} \tag{8}$$

其中，e(f) 是 2.2 节中介绍的指数压缩值。

(5)离散余弦变换(Discrete Cosine Transform, DCT)。对经过指数压缩的能量谱进行离散余弦变换，得到 GFCC，其计算公式如下：

$$C_{\mathrm{GFCC}}(i) = \sqrt{\frac{2}{N}} \sum_{j=1}^{P} m_j \cos[\frac{\pi i}{P}(j-0.5)], i = 1,2,\cdots,M \tag{9}$$

其中，M 为 GFCC 特征的维数；P 为滤波器的个数。

(6)升半正弦倒谱提升。对经过 DCT 得到的特征进行升半正弦倒谱提升，升半正弦窗函数如式(10)所示，倒谱提升后的特征如式(11)所示：

$$w(i) = 0.5 + 0.5 \times \sin(\pi i / N), 1 \leqslant i \leqslant N \tag{10}$$

$$\hat{C}_{\mathrm{GFCC}}(i) = C_{\mathrm{GFCC}}(i)_i \times w(i) \tag{11}$$

3 仿真实验与结果分析

3.1 实验数据库简介

本文实验采用的数据库为 TIMIT 和 NOIZEUS 语音库。TIMIT 语音库是语音识别研究中最常用的纯净语音库，在 TIMIT 语音库中，每个说话人包含 10 段 3 s~6 s 的语音，每段语音的内容都不同，其采样率为 16 kHz。NOIZEUS 是一种噪声语音库，其中含有 Babble noise、Airport noise、Car noise、Restaurant noise 等，它们的信噪比分别为 0 dB、5 dB、10 dB、15 dB，有关 NOIZEUS 语音库的详细介绍见文献[8]。

3.2 实验设计

实验 1 测试 GFCC 听觉特征的有效性，采用不含噪声的 TIMIT 语音库中的 dr1 和 dr6 部分作为 2 个子数据集进行实验。其中，dr1 部分共选取 40 个说话人(男 26 个，女 14 个)；dr6 部分共有 46 个说话人(男 30 个，女 16 个)。对于每个说话人，分别从 SA、SX 和 SI 中各选取一段语音作为测试语句，剩下的 7 段语音作为训练语句。

实验 2 测试 GFCC 听觉特征的抗噪声能力，采用 NOIZEUS 语音库，分别在 Babble noise、Airport noise、Car noise、Restaurant noise 条件下进行实验。

在实验中首先对语音信号进行预处理，然后对每一帧语音分别提取 LPCC、MFCC 和 GFCC 特征参数。最后采用高斯混合模型作为分类器进行识别，混合度为 16。

3.3 实验结果与分析

实验 1 的识别结果如表 1 所示。可以看出，GFCC 与 MFCC 的识别率相当，但比 LPCC 高 2 个~3 个百分点，表明本文的听觉倒谱系数 GFCC 用于说话人识别是有效的。

表 1 TIMIT 数据库识别准确率 (%)

特征	数据集 1	数据集 2
LPCC	87.50	86.95
MFCC	90.00	89.13
GFCC	89.16	90.57

实验 2 的结果见图 4~图 7。可以看出，在各种噪声环境下，GFCC 的识别效果比 MFCC 和 LPCC 都好，特别是在低信噪比的条件下，其识别率明显高于 MFCC 和 LPCC。实验结果表明，GFCC 比 MFCC 和 LPCC 具有更强的抗噪声能力。

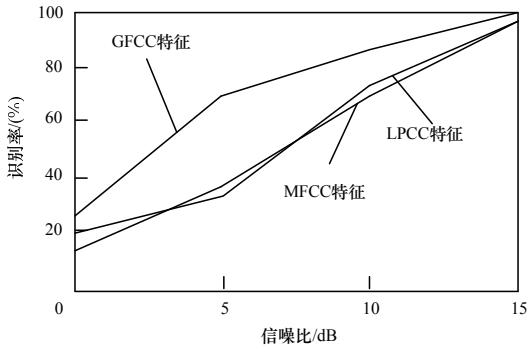

图 4 Babble noise 的识别结果

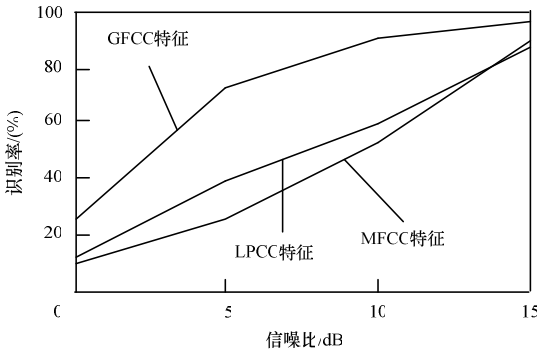

图 5 Airport noise 的识别结果

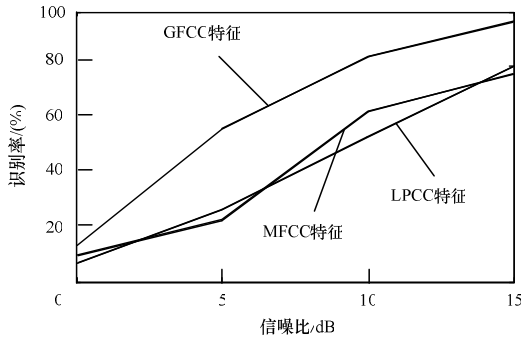

图 6 Car noise 的识别结果

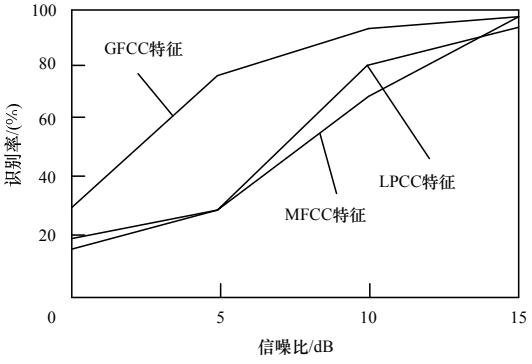

图 7 Restaurant noise 的识别结果

(下转第 174 页)