ISSN 1000-0054 清华大学学报（自然科学版） 2013 年 第 53 卷 第 8 期 2/23
CN 11-2223/N J Tsinghua Univ (Sci & Technol), 2013, Vol. 53, No. 8 1082-1086

语音识别中听觉特征的噪声鲁棒性分析

李银国[1]， 欧阳希子[1,2]， 郑 方[2]

（1. 重庆邮电大学，重庆 400065；2. 清华大学 语音和语言技术中心，北京 100084）

摘 要：自动语音识别系统在噪声环境下的性能通常会显著下降，这成为制约语音识别技术广泛应用的一个重大障碍。该文在他人的基于 Gammatone 的听觉特征（GFCC 特征）研究基础上，进一步对 GFCC 与基于 Mel 频率的倒谱系数（MFCC）在不同噪声环境下的性能表现进行分析研究。选择 5 种人工和自然噪声进行比较试验：白噪声、粉红噪声、褐色噪声、背景说话人噪声、汽车噪声。通过混合不同类型和不同强度的噪声，系统地研究了基于听觉特性的 GFCC 特征的特性和抗噪能力；特别地，用不同频段的正弦波噪声与纯净语音混合，分析了 GFCC 和 MFCC 在各个频带上的噪声鲁棒性。研究发现，与传统的 MFCC 相比，GFCC 对低频噪声具有更高的鲁棒性，而对中高频噪声相对敏感。由于人类发音通常在较低频率（300～700 Hz），这一特性使得 GFCC 在语音识别任务中具有良好的抗噪能力。实验结果表明，GFCC 在多种常见噪声环境下都取得了比 MFCC 更好的识别效果，特别是在低信噪比的情况下表现出更大的优势。

关键词：语音识别；Gammatone 滤波器；基于 Gammatone 的听觉特征（GFCC）；鲁棒性

中图分类号：TN 912.3 **文献标志码**：A

文章编号：1000-0054(2013)08-1082-05

Analysis of noise robustness of auditory features in speech recognition

LI Yinguo[1], OUYANG Xizi[1,2], Thomas Fang ZHENG[2]

(1. Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
2. Center for Speech and Language Technologies, Tsinghua University, Beijing 100084, China)

Abstract: A particular difficulty of automatic speech recognition in real applications involves significant performance degradation in noisy environment. Based on the research on gammatone-based auditory features (GFCCs) proposed by other researchers, an additional comparative study on the GFCC and the MFCC was presented for various noise conditions. Particularly, the behavior of GFCC/MFCC features with noise in different frequency bands was analyzed by mixing the test speech with sine noises to show that the GFCC is more robust against low-frequency noises than the MFCC while more sensitive to noises at middle and high frequencies. This property is desirable for speech recognition since most of the information of human speech resides in the low frequency band of 300—700 Hz. Experimental results demonstrate that the GFCC exhibits significant advantages over the MFCC for various noise conditions, especially when the SNR is low.

Key words: speech recognition; gammatone filters; gammatone-based auditory feature (GFCC); robust

当前自动语音识别系统（automatic speech recognition, ASR）面临的一个重大挑战是在噪声环境下识别性能的急剧下滑，这极大制约了 ASR 技术在实际应用中的推广。因此，增强识别系统的抗噪性能一直是语音识别领域的重要研究方向。近年来提出的比较重要的 ASR 鲁棒性的方法包括通道归一化、信号增强、模型自适应等。

在特征层，人们试图通过模拟人类听觉系统的结构和响应特性以提高语音特征对噪声的抗干扰能力，其中最通用的是基于 Mel 频率的倒谱系数（Mel frequency cepstrum coefficient, MFCC）及其衍生特征，其他有感知线性预测（perceptual linear prediction, PLP）和线性预测倒谱系数（linear prediction cepstrum coefficient, LPCC）等[1]。上述几种特征中，MFCC 和 PLP 是基于人类听觉的特征，LPCC 基于人类的发声机理。

与 MFCC 类似，本文中研究的基于 Gammatone 的倒谱系数（gammatone frequency cepstrum coefficient, GFCC）也是一种模拟人类听觉系统响应特性的语音特征提取方法。人类的听觉系统是一个高度复杂敏感的系统，对不同频率的信号分量有不同形式的响应，这种响应是非线性的，这种非线性可以通过一组 Gammatone 滤波器实现[2]。在文[3]

收稿日期：2013-04-19

作者简介：李银国（1955—），男（汉），湖北，教授。
E-mail: liyg@cqupt.edu.cn

中，作者提出了时域 GFCC 的实现方法，并对 GFCC 和 MFCC 的识别性能做了分析比较。

本文在上述工作基础上，对 GFCC 和 MFCC 在各种噪声环境下的识别性能进行补充性对比分析，并对 GFCC 和 MFCC 在不同频率区间内的敏感性进行了对比分析。选择 5 种噪声进行噪声对比实验：白噪声，粉红噪声，褐色噪声，背景说话人噪声，汽车噪声。通过混合不同类型和不同强度的噪声，分析 GFCC 和 MFCC 在不同噪声环境下的优劣和对不同噪声的抗干扰能力。同时，基于正弦噪声的能量在频率域上分布的单一性，用不同频率的正弦噪声对纯净的语音信号各频段进行混合，从而可以分析 GFCC 和 MFCC 对不同频率区间的敏感性和对不同能量分布的噪声的鲁棒性。

1 时域 GFCC 特征提取

首先介绍时域 GFCC 特征的提取方法。时域 Gammatone 滤波首先出现在文[3]中。

1.1 Gammatone 滤波

考虑人耳基底膜的滤波特性，假设有一组滤波器，每个滤波器的中心频率 f_c 和等效矩形带宽(equivalent rectangular bandwidth，ERB)各不相同，它们之间的关系为

$$\mathrm{ERB}(f_c) = f_c/Q + B_0. \tag{1}$$

其中：Q 为渐进因子，B_0 为最小带宽。

Gammatone 滤波(GF)时域阶跃响应可以表示为

$$g(t) = at^{n-1}\mathrm{e}^{2\pi bt}\cos(2\pi f_c t + \phi). \tag{2}$$

其中：f_c 为滤波器的中心频率；ϕ 为相位，通常取 $\phi=0$；a 为增益常数；n 为滤波器的阶数，通常设为 $n\leqslant 4$；b 为带宽相关的参数，它与 ERB 的关系为

$$b = 1.019\,\mathrm{ERB}(f_c). \tag{3}$$

参照文[5]取 $Q=9.264\,49$，$B_0=24.7$ Hz，根据式(1)和式(3)，b 与 f_c 的关系可表示为

$$b = 1.019\times 24.7\times(4.37\times f_c/1\,000+1). \tag{4}$$

多个不同中心频率的 Gammatone 滤波器构成了一个 Gammatone 滤波器组，经过 Gammatone 滤波器组的信号代表了原始信号在不同频率分量上的响应特征。观察式(2)，它由两部分组成：波形包络 $at^{n-1}\mathrm{e}^{-2\pi bt}$ 和频率 f_c 的调幅 $\cos(2\pi f_c t+\phi)$。通过 Fourier 分析，$g(t)$ 的频率域表达如下：

$$G(f) = \frac{a(n-1)!}{2(2\pi b)^n}\left\{\left(\frac{\mathrm{j}(f-f_c)}{b}+1\right)^{-n} + \left(\frac{\mathrm{j}(f+f_c)}{b}+1\right)^{-n}\right\}. \tag{5}$$

当 f_c/b 足够大时，$[\mathrm{j}(f+f_c)/b+1]^{-n}$ 可以被忽略。令 $s=\mathrm{j}2\pi f$，GF 的 Laplace 变换表示为

$$G(s) = \frac{a(n-1)!}{2}[s-(\mathrm{j}2\pi f_c - 2\pi b)]^{-n}.$$

其 Z 变换为

$$G(z) = \frac{a(n-1)!}{2}(1-\mathrm{e}^{\mathrm{j}2\pi f_c - 2\pi b})^{-n}.$$

令 $A(z)$ 为

$$A(z) = \frac{1}{1-\mathrm{e}^{\mathrm{j}2\pi f_c/f_s - 2\pi b/f_s}z^{-1}}. \tag{6}$$

则 $G(z)$ 可以看作是 n 个 $A(z)$ 递归应用的串联。$A(z)$ 与中心频率 f_c 有关，因此 $G(z)$ 也与 f_c 有关。考虑 $n=4$ 的情况，此时基本变换表示如下：

$$\hat{G}(z) = \frac{3a}{1-4mz^{-1}+6m^2z^{-2}-4m^3z^{-3}+m^4z^{-4}}. \tag{7}$$

其中 $m=\mathrm{e}^{-2\pi b/f_s}$。

1.2 GFCC 特征提取

为了取得更好的实验效果，在特征提取之前，首先对每一通道的 Gammatone 滤波信号进行预加重并分帧。在本文的实验中，信号采样率为 16 kHz，帧长为 400 个点，相邻帧重叠 160 个点，即每帧长度为 25 ms，帧移为 10 ms。对每一帧中的信号进行平均，得到该通道的平均帧能量。

对每一帧时刻，Gammatone 滤波器在各个通道上的平均帧能量组成该帧的向量表达，并通过离散余弦变换(DCT)以去除相关性，得到 GFCC。为使数值处理更加稳定，帧向量在进行 DCT 之前首先通过对数压缩。GFCC 特征向量可表达为如下公式：

$$F(n,v) = \left(\frac{2}{M}\right)^{0.5}\sum_{i=1}^{M}\left\{\frac{1}{3}\ln[\bar{y}(n,i)]\cos\left[\frac{\pi v}{2N}(2i-1)\right]\right\}.$$

其中：M 为通道数，本文中取值为 32；n 为通道号，范围从 0 到 31。当 $v>13$ 时，$F(\cdot,v)$ 的大多数值接近 0，因此本文选取前 13 个元素作为特征向量(含 C0 分量)，即 GFCC 的静态特征。

在静态特征的基础上进行差分运算，将静态特征与差分运算后生成的向量拼接成新的特征，能够有效提高语音识别的性能[4]，这些拼接的特征又称作动态特征。本文中选取一阶差分和二阶差分作为动态特征，最终得到的 GFCC 特征向量为 39 维。

2 噪声分析

为分析不同噪声的特异性,将纯净语音混入5种噪声并观察混合前后的频谱变化。这5种噪声为:白噪声、粉红噪声、褐色噪声、背景说话人噪声,汽车噪声。前3种噪声是人工噪声,通过SoX工具生成,后2种是自然噪声,来自NoiseX-92噪声库[5]。信号混合通过SoX工具完成。图1给出了在SNR=0 dB时,混合各种噪音后的频谱。可以发现,白噪声和粉红噪声的能量在各个频率段都有较强分布,因此对原始语音信号的影响最大,这会引起识别性能的急剧下降;背景说话人噪声在中低频段的干扰比较大,会对识别造成比较大的影响;褐色噪声和汽车噪声的能量主要分布在低频部分,这意味着在信噪比SNR较低的情况下,相对于其他几种噪声,这2种噪声对语音信号的影响相对较小,对语音识别系统的影响也相对有限。

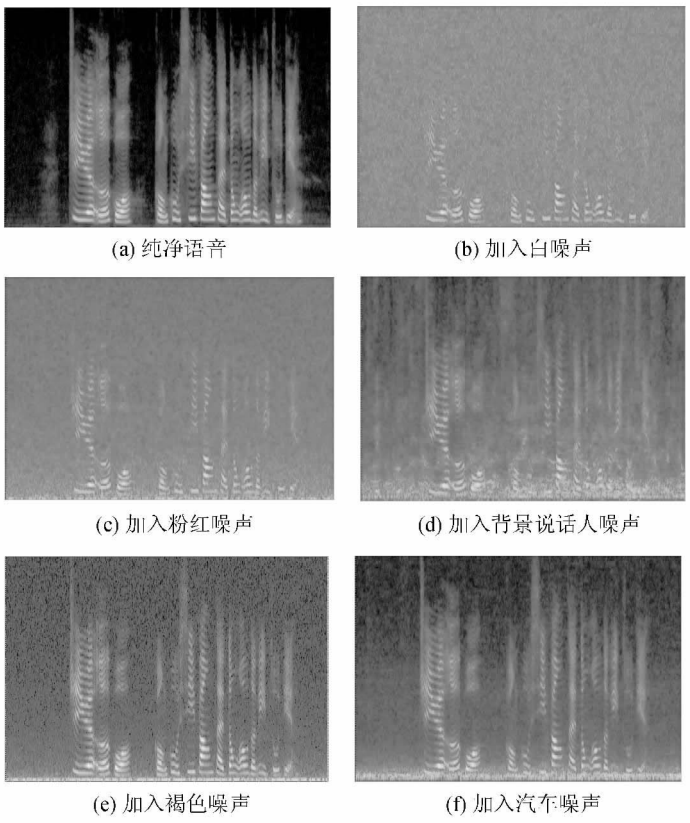

图1 加入各种噪声后的语音信号频谱图

3 实验设计

本文实验测试分为2个部分:第1部分比较了MFCC和GFCC在不同噪声和信噪比条件的大规模连续语音识别性能;第2部分比较了MFCC和GFCC对不同频段正弦噪声的敏感程度,以期为研究适合于GFCC的降噪方法提供实验依据。首先介绍实验所用的数据库及实验设置,然后报告2种特征的比较结果。

3.1 测试数据

本次实验所有的数据均来自863大规模连续语音识别计划所录制的标准普通话语音数据集。该数据集有38名女性和38名男性共76个说话人,采样频率为16 kHz,采样精度为单声道16 bits。选取其中的32名女性32名男性共64人42 h语音数据作为训练集,2名女性2名男性共4人作为测试集。噪声数据中的人工噪声由SoX工具生成,背景说话人噪声和汽车噪声来自标准噪声库NoiseX-92,并用Cooledit Pro进行预处理,使其采样率和采样精度与863数据相同。

3.2 实验设置

选择基于3状态HMM的上下文相关音素模型(tri-phone)作为声学模型。模型中含有218个单音素(包括静音),其中元音音素对应带音调韵母。采用基于最大互信息量的区分性训练(bMMI)方法进行建模,最终模型中含有约4 000个共享状态、15 000个Gauss分布。将前后相临9帧进行拼接以描

述上下文相关信息。线性区分分析(LDA)用于对拼接特征进行降维，最大似然线性变换(MLLT)用于消除降维后特征的各维间相关性。

语言模型为2万词表的back-off 3元文法模型(tri-gram)。模型训练基于中文Gibbytes数据库，Witten-Bell discounting及插值技术用于模型平滑。原始模型以最小概率10^{-7}进行剪枝，经过压缩后的模型大小为25 MB。

实验的主要目的是研究GFCC和MFCC在各种噪声环境下的性能。为了体现对比性，MFCC和GFCC提取均采用相同的25 ms帧长和10 ms帧移，特征向量包括13维的倒谱系数和一阶、二阶差分，共计39维。

整个实验在Linux环境下进行，SoX工具用于人工噪音合成及语音/噪声信号混合。GFCC特征提取采用时域GFCC提取工具包(http://homepages.inf.ed.ac.uk/v1dwang2/public/tools/index.html)，MFCC特征提取、声学模型训练和解码过程采用Kaldi工具包[6]。SRILM工具包用于语言模型训练。实验结果以词错误率(WER)作为评价标准。

3.3 实验结果

实验分为2组。在第1组实验中，对比MFCC和GFCC在不同噪音种类和SNR级别上的识别性能；在第2组实验中，用不同频段的正弦噪声来分析GFCC和MFCC对不同频段噪声的敏感度。

3.3.1 MFCC/GFCC抗噪对比实验

将5种不同噪声与原始的纯净语音进行混合，并对混合后的噪声语音进行识别。这5种噪声包括：白噪声、粉红噪声、褐色噪声、背景说话人噪声、汽车噪声。测试噪声等级以SNR为标准分为7级：30、25、20、15、10、5、0 dB。以WER为标准的对比实验结果如表1所示。

表1 各种条件下的语音识别结果

噪声	特征	WER/%						
		SNR=30 dB	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
白噪声	MFCC	18.29	19.65	27.07	34.46	49.14	73.51	93.81
	GFCC	17.22	18.12	21.20	29.25	45.55	71.37	91.27
粉红噪声	MFCC	17.37	18.10	19.56	24.17	35.56	64.37	91.54
	GFCC	16.36	17.08	18.62	23.33	35.29	63.41	87.57
褐色噪声	MFCC	17.06	17.09	17.16	17.20	17.42	18.40	21.88
	GFCC	15.75	15.79	15.83	16.29	16.97	18.12	22.26
背景说话人噪声	MFCC	17.26	17.78	18.93	22.20	33.98	61.88	87.83
	GFCC	16.07	16.35	17.62	20.80	30.04	55.00	83.53
汽车噪声	MFCC	17.24	17.41	17.61	18.08	20.64	26.48	37.88
	GFCC	15.85	16.20	16.57	17.30	18.44	21.73	29.81

首先观察到，不论是GFCC还是MFCC，噪声的引入都会严重影响语音识别的正确率，特别是白噪声和粉红噪声，由于其频带分布宽，语音的谐振信息在各个频段都受到干扰和破坏(见图1)，因而对语音识别性能影响极为显著。当SNR值低到0时，语音信息几乎完全被噪声所淹没，语音识别的错误率接近100%。相比而言，褐色噪声、背景说话人噪声、汽车噪声等多集中于某一较低的频率范围内，对语音信号的破坏具有局部性，对语音识别的影响也相对有限。如汽车噪声，即使SNR降到0，依然可以得到60%以上的识别率。

其次，可以观察到，在所测试的所有噪声环境下，GFCC都有明显优于MFCC的识别性能，说明GFCC较MFCC具有更强的噪声鲁棒性。特别是对白噪声、粉红噪声等严重破坏语音信息的宽带噪声，GFCC表现出特别明显的优势。这一发现证明GFCC具有在破坏性噪声环境中保证识别性能的宝贵性质。

3.3.2 MFCC/GFCC不同频段抗噪对比实验

在第2组实验中，将不同频段的正弦噪声混入纯静语音中，以分析GFCC和MFCC对不同频率噪音的抵抗能力。考虑到特征提取中所采用滤波器组的频率范围为80 Hz～5 kHz，选择低频(0.5 kHz)、中频(2 kHz)、高频(4 kHz)3种具有代表性的正弦噪音对纯静语音进行“破坏”。对这些混入噪声的语音进行识别，得到结果如图2所示。可以发现，当存在较低频的噪声时，GFCC较MFCC可得到更好的识别性能，而当噪音频率比较高时，GFCC与MFCC相比有较大的差距。结合表1数据，作

者认为这可能要归因于人说话频率主要在 300～700 Hz 之间,而人耳对 2～5 kHz 的频率范围感受力最强[7],由于 GFCC 倾向于模拟人耳听觉系统的频率响应,因而中高频的噪音对特征的破坏相对比较严重,对识别结果的影响较大。这提示在特征提取时,对 MFCC 和 GFCC 可能需要设计不同的前端滤波器以提高各自的抗噪性能。

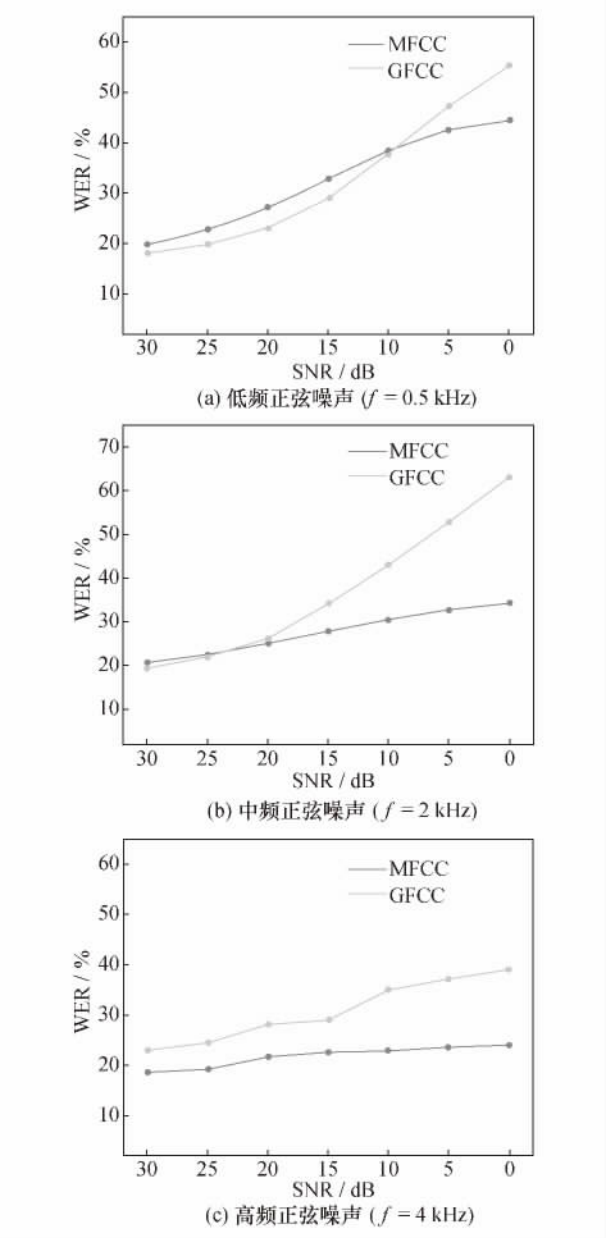

图 2 不同频带正弦噪声下的语音识别结果

4 结束语

本文对 MFCC 和 GFCC 两种特征在不同噪声环境下和不同频率区间的语音识别性能进行了对比研究。发现 GFCC 对低频噪声具有更强的抵抗能力,并对白噪声、粉红噪声等宽带噪声具有更强的鲁棒性,特别是在低信噪比条件下,GFCC 的优势更为明显。GFCC 的这些特点使其具有普遍的实际应用价值。

未来工作将研究各种抗噪算法(如 Winner 滤波)对 GFCC 性能的影响,研究提高 GFCC 对中频段噪声的鲁棒性的方法。同时,将尝试将 GFCC 应用到其他场景中,如基于 GFCC 的端点检测技术及噪声环境下的说话人识别技术等。

参考文献 (References)

[1] Huang X D, Acero A, Hon H W. Spoken Language Processing [M]. Upper Saddle River, NJ: Prentice Hall PTR, 2000.

[2] Patterson R D, Moore B C J. Auditory filters and excitation patterns as representations of frequency resolution [M]// Moore B C J. Frequency Selectivity in Hearing. London: Academic Press, 1986: 123-177.

[3] Qi J, Wang D, Jiang Y, et al. Auditory features based on gammatone filters for robust speech recognition [C]//ISCA. 2012.

[4] Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum [J]. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 1986, **34**(1):, 52-59.

[5] Varga A P, Steeneken H J M, Tomlinson M, et al. The NOISEX-92 study on the effect of additive noise on automatic speech recognition [R]. Speech Research Unit, Defense Research Agency, Malvern, UK, 1992.

[6] Povey D, Ghoshal A, Boulianne G, et al. The Kaldi speech recognition toolkit [C]//Proc ASRU. 2011.

[7] 韩纪庆, 张磊, 郑铁然. 语音信号处理[M]. 北京: 清华大学出版社, 2004.