计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):227−232 [doi: 10.15888/j.cnki.csa.006106]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

弯折滤波器在说话人识别的鲁棒特征提取中的应用①

邓　蕾, 高　勇

(四川大学 电子信息学院, 成都 610065)

摘　要: 针对噪声环境中说话人识别性能急剧下降的问题. 提出了一种用于说话人识别的鲁棒特征提取的方法. 采用弯折滤波器组 (Warped filter banks, WFBS) 来模拟人耳听觉特性, 将立方根压缩算法、相对谱滤波技术 (RASTA)、倒谱均值方差归一化算法 (CMVN) 引入到鲁棒特征的提取中. 在高斯混合模型 (GMM) 下进行仿真, 实验结果表明该方法提取的特征参数在鲁棒性和识别性能上均优于 MFCC 特征参数和 CFCC 特征参数.

关键词: 说话人识别; 弯折滤波器组; 鲁棒性

引用格式: 邓蕾,高勇.弯折滤波器在说话人识别的鲁棒特征提取中的应用.计算机系统应用,2017,26(12):227–232. http://www.c-s-a.org.cn/1003-3254/6106.html

Warped Filter Banks Applied in Robust Feature Extraction Method for Speaker Recognition

DENG Lei, GAO Yong

(College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China)

Abstract: The performance of the speaker recognition system degrades drastically in the noisy environment. A robust feature extraction method for speaker recognition is proposed in this paper. Warped filter banks(WFBS) are used to simulate the human auditory characteristics. The cubic root compression method, relative spectral filtering technique(RASTA) and the cepstral mean and variance normalization algorithm(CMVN) are introduced into the robust feature extraction. Subsequently, simulation experiment is conducted based on Gaussian mixes model(GMM). The experimental results indicate that the proposed feature has better robustness and recognition performance than the mel cepstral coefficients(MFCC) and cochlear filter cepstral coefficients(CFCC).

Key words: speaker recognition; warped filter banks; robustness

1　引言

说话人识别又称为声纹识别, 即提取语音波形中反映说话人的生理和行为特征的语音特征参数来自动确定说话人身份的技术. 随着识别技术的研究不断深入, 说话人识别在实验室环境中已经能获得较高的识别率, 而在实际应用中, 由于噪声的影响, 识别性能有恶化的趋势. 其根本原因在于噪声的影响引起了语音的畸变, 导致了训练环境和测试环境的不匹配, 因此, 训练数据所获得的语音信息无法正确表达测试环境的数据. 鲁棒性语音识别的根源是为了消除噪声引起的训练环境和测试环境之间的不匹配. 解决鲁棒性语音识别问题的主要方法有以下四种[1]: 1) 抗噪特征参数提取: 寻求对噪声不敏感的语音特征. 2) 人耳听觉特性研究: 人耳的听觉特性有较强的噪声鲁棒性. 3) 语音增强: 从带噪语音中恢复出干净语音, 消除噪声的影响, 增强语音. 4) 模型补偿: 根据环境噪声特性, 对纯净语音模型的参数进行修正, 补偿训练和测试环境间的不匹配. 本文主要研究抗噪特征参数的提取方法.

人耳具有较强的噪声鲁棒性, 在低信噪比条件下具有较好的识别能力. 耳蜗是人耳听觉系统的重要器

① 收稿时间: 2017-03-13; 修改时间: 2017-04-05; 采用时间: 2017-04-07

官, 耳蜗内有一个重要的部分叫基底膜, 其作用相当于一个频谱分析仪, 它能够把传入人耳的声信号在频域上按频带进行分解, 就像一个带通滤波器组. 基底膜作为滤波器组, 具有在低频处频率分辨率较高, 高频处频率分辨率较低的特性[2], 因此, 耳蜗基底膜不同位置对应滤波器带宽是不一样的; 单个滤波器的频率响应呈非对称分布, 特征频率的左侧斜率比较平缓, 而右侧斜率较为陡峭. 目前, 考虑人耳听觉特性来提取的语音特征参数主要有利用 Mel 滤波器组提取的 Mel 频率倒谱系数 (Mel frequency cepstrum coefficient, MFCC)[3]和利用耳蜗滤波器组提取的耳蜗倒谱系数 (Cochlear filter cepstral coefficients, CFCC)[4]. 其中 MFCC 已部分考虑到了人耳的听觉特性[5], MFCC 在纯净语音的识别率可达到 98%, 但当信噪比为-10dB 的噪声条件下, 识别率下降到了 5%. 不同于 MFCC, CFCC 是基于听觉变换的说话人特征参数, 具有良好的识别效果和鲁棒性. 在文献[4]中, 当信噪比为 6dB 时, MFCC 的识别率为 42.1%, 而 CFCC 的识别率为 90.3%. 然而, 在 white 噪声-6dB 条件下时, MFCC 的识别率分别为 5.8%, 而 CFCC 识别率下降到了 16.6%. Mel 滤波器组和耳蜗滤波器组的频率响应关于中心频率呈对称分布, 并不满足基底膜的非对称特性. 为了充分利用人耳的听觉特性, Zhang X, Huang L 等人[6]利用弯折滤波器组 (Warped filter banks) 提取语音特征参数, 然后再将特征参数运用到语音识别中, 提高了语音识别系统的识别率.

本文在文献[6]的基础上, 将弯折滤波器组用于说话人识别中, 并融合了以下三种技术: 立方根压缩技术[7]、相对谱滤波技术 (RASTA)[8]和倒谱均值方差归一化技术 (CMVN)[9], 提出了基于弯折滤波器组的 C-R-C-WFCC 特征参数.

2 说话人识别系统构成

说话人识别系统包括训练阶段和识别阶段[10,11]. 其系统框图如图 1 所示.

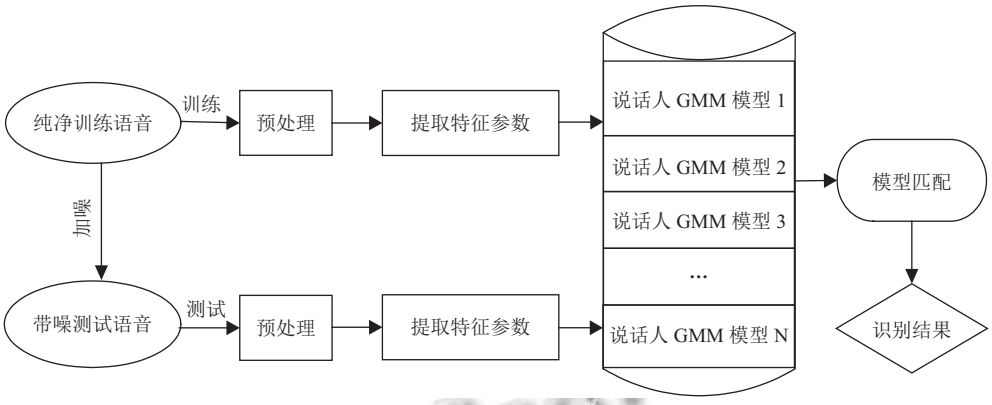

图 1 说话人识别系统框图

训练阶段, 选取 N 个说话人的纯净语音, 对输入的纯净语音信号先进行预处理, 再提取 N 个说话人的语音特征参数, 并将其作为 GMM 模型的输入, 最后训练出 N 个说话人的 GMM 模型.

测试阶段将 N 个说话人的纯净语音分别加入不同信噪比 (dB) 的噪声得到带噪语音, 将每个人的带噪语音分成 M 段, 形成 $N\times M$ 段带噪语音, 将带噪语音经过预处理后, 再提取特征参数, 并将其作为 GMM 模型输入, 训练出 $N\times M$ 个带噪语音的说话人 GMM 模型, 最后将训练阶段和测试阶段的 GMM 模型进行匹配, 输出识别结果.

3 MFCC 特征参数和 CFCC 特征参数提取

3.1 MFCC 特征参数提取

MFCC 特征参数是基于 *Mel* 滤波器组的基础上实现的, *Mel* 滤波器组的频率响应如图 2 所示, 由图 2 可以看出 *Mel* 滤波器组的频率响应关于中心频率对称, 且中心频率附近幅值较陡峭.

MFCC 特征参数提取流程[3]如图 3 所示.

3.2 CFCC 特征参数提取

耳蜗倒谱系数 (Cochlear filter cepstral coefficients, CFCC)[12]是利用耳蜗滤波器组提取的, 具有较好的识别效果和鲁棒性. 耳蜗滤波器的频率响应如图 4 所示,

从图 4 中可以看出, 耳蜗滤波器组的频率响应关于中心频率对称.

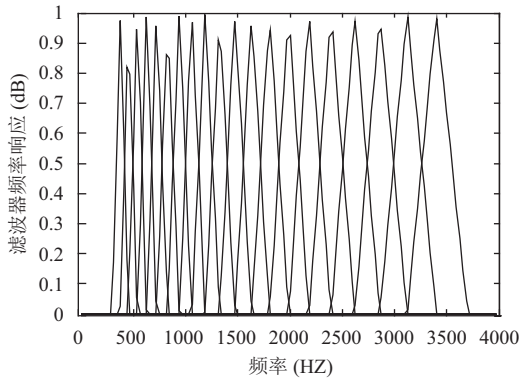

图 2 Mel 滤波器组的频率响应

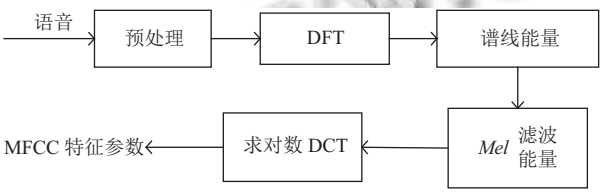

图 3 MFCC 特征参数提取流程

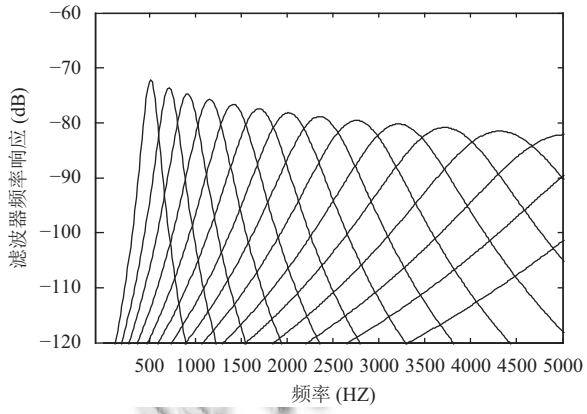

图 4 耳蜗滤波器组的频率响应

CFCC 特征参数的提取方法[4,12]如图 5 所示.

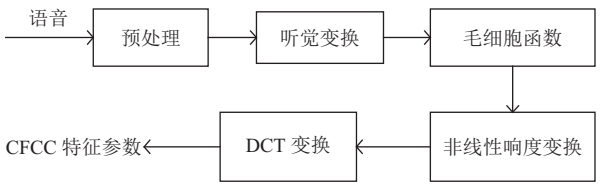

图 5 CFCC 特征参数提取流程

4 C-R-C-WFCC 特征参数提取

4.1 36 通道弯折滤波器组的设计

一组 m 通道的均匀滤波器组[13]的传递函数如式 (1) 所示:

$$H_m(z)=\sum_{n=0}^{N-1}h(n)z^{-n}e^{j\frac{2\pi}{M}mn},m=0,1,....,M-1 \quad (1)$$

其中, $h(n)$ 为长度为 N 的序列, M 为滤波器组的通道个数. 文献[11]中将一阶全通变换代替式 (1)$\zeta^{-1}=\frac{-\alpha+z^{-1}}{1-\alpha z^{-1}}$ 中的 z^{-1}, 得到式 (2).

$$H_m(z)=\sum_{n=0}^{N-1}h(n)\left(\frac{-\alpha+z^{-1}}{1-\alpha z^{-1}}\right)^n e^{j\frac{2\pi}{M}mn} \quad (2)$$

令 $z=e^{j\omega}$, 则弯折滤波器组的频率响应如式 (3) 所示.

$$H_m(e^{j\omega})=\sum_{n=0}^{N-1}h(n)\left(\frac{-\alpha+e^{-j\omega}}{1-\alpha z^{-j\omega}}\right)^n e^{j\frac{2\pi}{M}mn} \quad (3)$$

其中 α 为弯折因子, 公式 (4) 和 (5) 分别给出了采用一阶全通变换模拟 Bark 和 ERB 频率尺度[14]时对应的弯折因子.

$$\alpha=0.7446\left[\frac{2}{\pi}\arctan(0.1418f_s)\right]^{\frac{1}{2}}+0.03237 \quad (4)$$

$$\alpha=1.0674\left[\frac{2}{\pi}\arctan(0.06583f_s)\right]^{\frac{1}{2}}-0.1916 \quad (5)$$

在式 (4) 和式 (5) 中, 当采样频率 f_s=8 kHz 时, α=0.58 和 α=0.40 分别模拟 ERB 频率尺度和 Bark 频率尺度. 当 α=0 时, 弯折滤波器组的频率响应则为 m 通道均匀滤波器组的频率响应. 式 (2) 中, 取 $h(n)$ 为 20 样点的汉明窗序列, M、α=0.58 和 α=0.40 时, 则得到分布在$[0,f_s]$上的 36 通道滤波器. 36 通道弯折滤波器组的频率响应如图 6 和图 7 所示.

从图 6 和图 7 中可以看出, 弯折滤波器的分布在低频处比较密集, 高频处较宽松, 并且滤波器的带宽关于中心频率是非对称分布的, 符合基底膜作为滤波器的特性. α=0.58 比 α=0.40 特性更加明显. 由于语音信号的频率主要集中在[200, 4000] Hz 范围, 因此本文在设计滤波器组时, 保留了第 3 通道到第 20 通道, 取 18 通道滤波器组分布在[200, 5500] Hz 范围内.

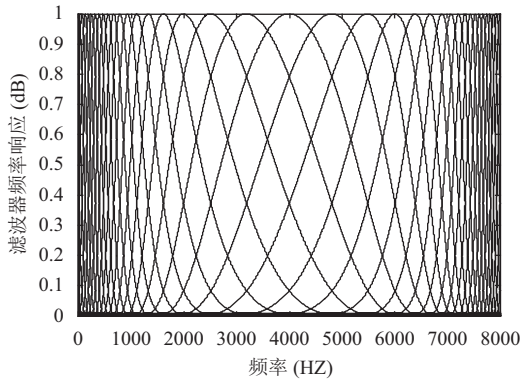

图 6 弯折滤波器组的频率响应, α=0.58

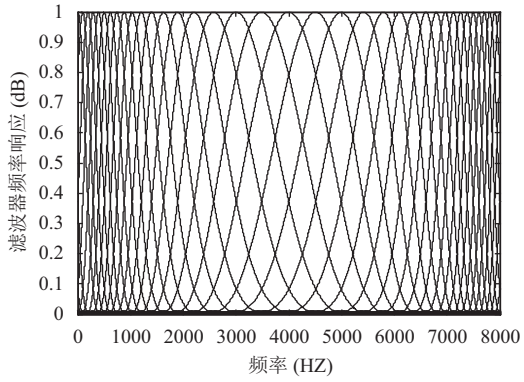

图 7 弯折滤波器组的频率响应, α=0.40

4.2 C-R-C-WFCC 特征参数提取方法

C-R-C-WFCC 特征参数的提取步骤如图 8 所示.

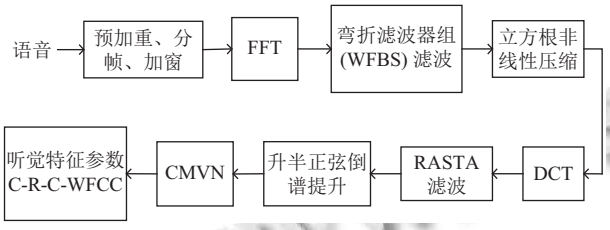

图 8 C-R-C-WFCC 特征参数提取过程

Step 1. 将语音信号预处理之后得到一帧帧的语音信号, 用表示$x_w(n)$. 将$x_w(n)$进行 FFT(快速傅里叶变换) 后得到信号频谱 $X(k)$.

Step 2. 对 $X(k)$ 取平方得到短时能量谱, 然后用 WFBS 滤波器组滤波处理, 滤波器的输出如式 (6) 所示:

$$X_p=\sum_{k=0}^{\frac{N}{2}-1}\left|X(k)^2\times H_p(k)\right| \tag{6}$$

其中, N 为在每一帧语音信号中, 快速傅里叶变换的点数; p 为滤波器的个数, 本文取 20; $H_p(k)$为第 p 个 WFBS 滤波器的频率响应.

Step 3. 为模拟人耳听觉模型处理信号的非线性, 对每个滤波器的输出做立方根压缩, 得到一组能量谱 $Y_1,Y_2,\cdots,Y_p$. 其计算公式如式 (7) 所示:

$$Y_p=\left(X_p\right)^{\frac{1}{3}} \tag{7}$$

Step 4. 对所有滤波器输出经过立方根压缩后, 再经 DCT(离散余弦变换) 得到倒谱, 其计算公式如下:

$$C-WFCC(i)=\sqrt{\frac{2}{N}}\sum_{j=1}^{p}Y_p\cos[\frac{\pi i}{p}(j-0.5)],i=1,2,\cdots M \tag{8}$$

其中, M 为特征参数的维数; p 为滤波器的个数.

Step 5. 对上一步输出进行 RASTA 滤波. 将 RASTA 滤波技术用于特征参数提取过程中, 不仅可以参数的识别率, 还可以使参数具有较高的稳健性. 它的传输函数为:

$$H(z)=0.1\times\frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}\times(1-0.98z^{-1})} \tag{9}$$

进行 RASTA 滤波. 其计算公式如下:

$$R-C-\overline{WFCC}(i)=H(z)\times(C-WFCC(i)) \tag{10}$$

Step 6. 升正弦倒谱提升后得到$\overline{WFCC}(i)$.

$$C-R-C-\overline{WFCC}(i)=\mathrm{R}-\mathrm{C}-\overline{WFCC}(i)\times w(i) \tag{11}$$

其中$w(i)=0.5+0.5\times\sin(\pi i/N),1\le i\le N$为半升正弦窗函数.

Step 7. 最后将进行 CMVN(倒谱均值方差归一化) 得到特征参数 C-R-C-WFCC.

5 仿真实验

5.1 实验设计

本文采用 18 通道的弯折滤波器组, 进行语音特征参数的提取 (以下简称为 C-R-C-WFCC 特征参数). 当采样频率 f_s=8 kHz 时, α=0.58 和 α=0.40. 本文采用的分类器模型为高斯混合模型 (GMM)[15].

语料库为不含噪声的普通话语音数据库 (采样频率 f_s=8 kHz), 从中选取 36 人 (男 22 人, 女 14 人), 每个说话人包含大约 1 min 的语句, 作为训练语音, 共 36 条. 测试阶段每个人包含 5 条 5 s 的语句, 作为测试语音, 共 180 条.

实验 1. 测试 C-R-C-WFCC 特征参数在纯净语音条件下的有效性.

为了验证本文提取的特征参数在纯净语音条件下对说话人识别的有效性, 将本文提取的特征参数在 α=0.58 和 α=0.40 的条件下进行测试. GMM 混合度分别选取 8 阶、16 阶、32 阶和 64 阶.

实验 2. 测试 C-R-C-WFCC 特征参数的抗噪声能力.

为了测试本文提出的 C-R-C-WFCC 特征参数的抗噪声能力. 实验 2 将 C-R-C-WFCC 特征参数与 MFCC 特征参数和 CFCC 特征参数在同等噪声条件下得出识别结果. 采用 noise-92 标准噪声库. 分别在 f16 座舱噪声 (f-16 cockpit noise)、白噪声 (white noise) 和粉红噪声 (pink noise) 条件下进行实验. 含噪语音的信噪比 (SNR) 分别为-10 dB、-5 dB、0 dB、5 dB、10 dB. 实验 2 的 GMM 混合度为 64 阶.

5.2 实验结果及分析

本文将弯折滤波器用于 C-R-C-WFCC 语音特征参数提取过程, 在 α=0.58 和 α=0.40 两个不同的弯折因子上得出了对应的识别效果.

实验 1 的识别率见图 9. 从图 9 中可得知, 在纯净语音条件下, 当 α=0.40 时, 系统的识别率总体上要高于 α=0.58. 同时, 从图 9 中可看出当 α=0.58 时, 识别率仅在 GMM 混合度为 32 阶时识别率能达到 95.56%, 在 GMM 混合度为 8 阶、16 阶、32 阶和 64 阶时识别率呈降低趋势, 由此可得知, 当 α=0.58 时, GMM 混合度的阶数对识别率有较大的影响. 然而, 当 α=0.40 时, 识别率在 GMM 混合度为 8 阶、16 阶、32 阶和 64 阶的条件下具有相同的识别率, 在 GMM 混合度为 64 阶的条件下, 识别率高达 96.11%.

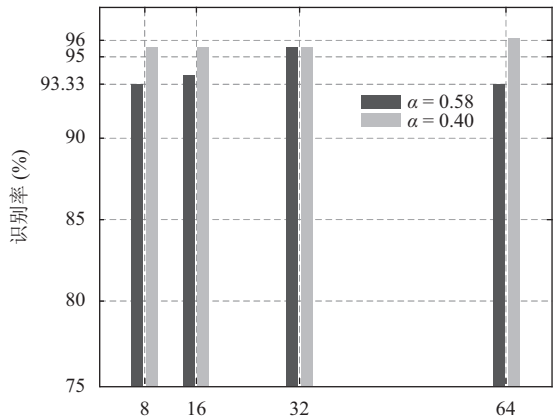

图 9 纯净语音条件下, 不同 GMM 混合度系统识率

实验 2 的识别结果见图 10~图 12. 从图 10~图 12 中可以看出, 在三种不同的噪声环境下, 本文所提取的 C-R-C-WFCC 特征参数随着信噪比的增加而升高. 在不同 α 的值条件下, 系统识别率差异并不大. 如图 10 中, 系统识别率仅在信噪比为-10 dB 至-5 dB 有低于 2% 的差异. 随着信噪比的升高, C-R-C-WFCC 特征参数的识别率均高于 MFCC 特征参数和 CFCC 特征参数. 实验结果表明, 本文提出的特征参数具有更强的抗噪声能力.

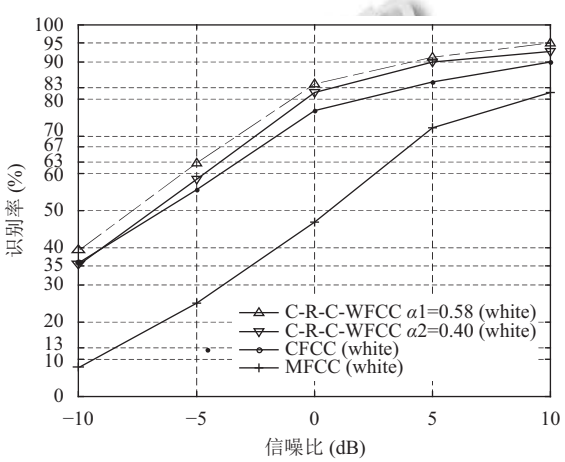

图 10 white 噪声识别结果

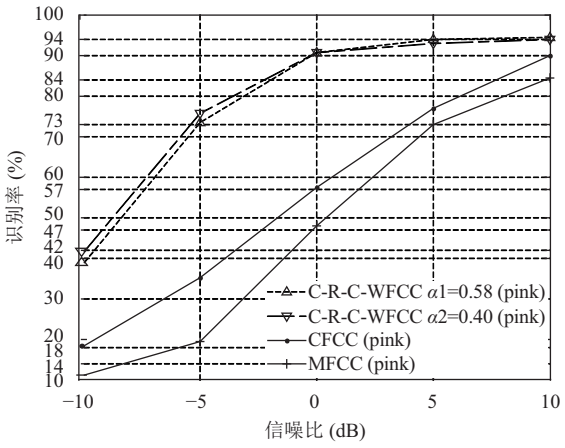

图 11 pink 噪声识别结果

6 结语

本文将弯折滤波器组用于说话人特征提取过程, 并引入了立方根压缩、RASTA 滤波、倒谱均值方差归一化 (CMVN)3 种技术, 得出了不同弯折因子 α 对应的识别效果. 实验仿真结果表明, 在纯净语音条件下, 弯折因子 α=0.40 的总体识别效果比 α=0.58 更好; 在噪声条件下, 本文提出的 C-R-C-WFCC 特征参数具有较好的识别效果, 均高于 MFCC 特征参数和 CFCC 特征

参数, 且弯折因子 α=0.58 和 α=0.40 的识别效果相差不大. 然而弯折因子并不是影响实验结果的唯一因素, 滤波器的通道个数也是影响实验结果的重要因素. 在将来的实验中将致力于这方面的研究.

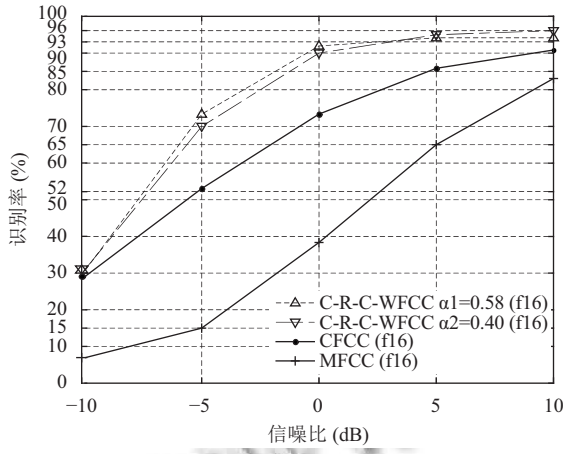

图 12 f16 噪声识别结果

参考文献

1 Jin Q. Robust speaker recognition[Ph. D. thesis]. Pittsburgh: Carnegie Mellon University, 2007: 276–279.

2 曹龙涛, 李如玮, 鲍长春, 等. 基于噪声估计的二值掩蔽语音增强算法. 计算机工程与应用, 2015, (17): 222–227. [doi: 10.3778/j.issn.1002-8331.1312-0396]

3 Muda L, Begam M, Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. Journal of Computing, 2010, 2(3): 138–143.

4 Li Q, Huang Y. An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions. IEEE Trans. on Audio, Speech, and Language Processing, 2011, 19(6): 1791–1801.

5 Li L, An D, Zhao D, *et al*. TEO-CFCC characteristic parameter extraction method for speaker recognition in noisy environments. Przeglad Elektrotechniczny, 2013, 89(2): 118–121.

6 Zhang XY, Huang LX, Evangelista G. Warped filter banks used in noisy speech recognition. Proc. of the 2009 Fourth International Conference on Innovative Computing, Information and Control. Kaohsiung, China. 2009. 1385–1388.

7 Jawarkar NP, Holambe RS, Basu TK. Effect of nonlinear compression function on the performance of the speaker identification system under noisy conditions. Proc. of the 2nd International Conference on Perception and Machine Intelligence. Kolkata, West Bengal, India. 2015. 137–144.

8 Nidhyananthan SS, Kumari RSS. Text independent voice based students attendance system under noisy environment using RASTA-MFCC feature. Proc. of the International Conference on Communication and Network Technologies. Sivakasi, India. 2014. 182–187.

9 Prasad NV, Umesh S. Improved cepstral mean and variance normalization using Bayesian framework. Proc. of 2013 IEEE Workshop on Automatic Speech Recognition and Understanding. Olomouc, Czech Republic. 2013. 156–161.

10 Geng Y, Liang RZ, Li W, *et al*. Learning convolutional neural network to maximize Pos@Top performance measure. Computer Vision and Pattern Recognition. arXiv: 1609.08417. 2017.

11 Li QF, Zhou XF, Gu AH, *et al*. Nuclear norm regularized convolutional Max Pos@Top machine. Neural Computing & Applications, 2016: 1–10. [doi: 10.1007/s00521-016-2680-2]

12 Raikar A, Gandhi A, Patil HA. Combining evidences from mel cepstral and cochlear cepstral features for speaker recognition using whispered speech. Král P, Matoušek V. Text, Speech, and Dialogue. Cham, Germany. 2015. 405–413.

13 黄丽霞. 非特定人鲁棒性语音识别中前端滤波器的研究[博士学位论文]. 太原: 太原理工大学, 2011.

14 Chavan MS, Chougule SV. Speaker identification in mismatch condition using warped filter bank features. International Journal of Circuits, Systems and Signal Processing, 2015, 9: 88–93.

15 Chakroun R, Zouari LB, Frikha M. An improved approach for text-independent speaker recognition. International Journal of Advanced Computer Science and Applications, 2016, 7(8): 343–348.