噪声环境下说话人识别的TEO-CFCC特征参数提取方法

李晶皎 安 冬 杨 丹 王 骄
(东北大学信息科学与工程学院 沈阳 110819)

摘 要 针对广泛应用于说话人识别的MFCC特征参数在低信噪比环境下识别正确率急剧下降的问题,提出了一种TEO-CFCC特征参数提取方法。在CFCC特征参数的基础上,首先通过信号相位匹配的方法消除语音噪声,然后在CFCC特征参数的求取过程中加入Teager能量算子,将语音的能量作为说话人的特征参数之一,得到TEO-CFCC特征参数。实验表明,提出的TEO-CFCC特征参数在信噪比为-5dB的汽车噪声条件下,识别正确率可达到83.2%。

关键词 听觉变换,相位匹配,能量算子,说话人识别
中图法分类号 TP391.42 **文献标识码** A

TEO-CFCC Characteristic Parameter Extraction Method for Speaker Recognition in Noisy Environments

LI Jing-jiao AN Dong YANG Dan WANG Jiao
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China)

Abstract Considering the sharp decline in the recognition accuracy of MFCC characteristic parameter for speaker recognition in low SNR environments, this paper proposed TEO-CFCC characteristic parameter extraction method. Signal phase matching is applied to eliminate speech noise on the basis of CFCC characteristic parameter, and then teager energy operator is added to the acquisition of CFCC characteristic parameter. In this way TEO-CFCC characteristic parameter is obtained and the energy of speech becomes one of the characteristic parameters for speaker recognition. Experiment results show that the recognition accuracy can reach to 83.2% in a -5dB SNR of vehicle interior noise environment by using TEO-CFCC characteristic parameter.

Keywords Auditory transform, Phase matching, Energy operator, Speaker recognition

1 引言

语音特征参数的准确提取在语音音调识别[1]、语音情感识别[2]、语音识别[3]和说话人识别领域中,都是最为关键的一步。目前应用最为广泛的说话人个性特征参数仍然为Mel倒谱系数(Mel Cepstrum Coefficients, MFCC)[4],但是在低信噪比环境下,MFCC的性能急剧下降[5]。提取一种有效且具有很好的抗噪声性的说话人个性特征参数,一直是说话人识别研究的重点问题之一。耳蜗倒谱系数(Cochlear Filter Cepstral Coefficients, CFCC)是由贝尔实验室的Peter Li博士于2011年首次提出的用于说话人识别的特征参数[6]。与基于傅里叶变换的MFCC特征参数不同,CFCC是基于听觉变换(Auditory Transform, AT)的特征参数[7],在信噪比为6dB的白噪声和汽车噪声条件下,MFCC特征参数的识别正确率下降到了41.2%,而CFCC特征参数的识别正确率却可以达到88.3%[6],由此可见,CFCC特征参数具有很好的抗噪声能力,其性能优于MFCC。但是当信噪比在-6dB的白噪声情况下,CFCC特征参数的识别正确率却下降到了20%。

基于以上分析,本文将相位匹配的思想引入到CFCC特征参数中,并在CFCC特征参数中加入了Teager能量算子(Teager Energy Operator, TEO),提出了噪声环境下基于相位匹配的TEO-CFCC说话人特征参数提取的方法。

2 CFCC特征参数

2.1 听觉变换原理

在信号处理领域,傅里叶变换被广泛地应用于信号的时频转换过程。然而傅里叶变换在处理线性信号时显现的优势却在处理非线性信号时受到了很大的限制,离散傅里叶变换和快速傅里叶变换虽然很好地弥补了傅里叶变换在处理非线性变换时的不足,但是仍然不能满足语音信号处理的需要。

在文献[7]中,Peter Li首次提出了听觉变换的概念。听觉变换模拟了人耳的听觉原理,其首先定义了一个耳蜗滤波函数 $\psi(t)\in L^2(R)$,要求 $\psi(t)$ 满足式(1)-式(3)的条件:

$$\int_{-\infty}^{\infty}\psi(t)\mathrm{d}t=0 \tag{1}$$

$$\int_{-\infty}^{\infty}|\psi(t)|^2\mathrm{d}t<\infty \tag{2}$$

到稿日期:2012-02-26 返修日期:2012-05-15 本文受国家自然科学基金项目(60970157),辽宁省博士启动基金项目(20081019),中央高校基础科研经费(N100304008)资助。

李晶皎(1964-),女,教授,博士生导师,主要研究方向为嵌入式系统、模式识别、语言处理,E-mail:lijingjiao@ise.neu.edu.cn;安 冬(1984-),男,博士生,主要研究方向为自然语音处理,E-mail:249350656@qq.com(通信作者);杨 丹(1979-),女,博士,讲师,主要研究方向为自然语音处理;王 骄(1978-),男,副教授,主要研究方向为嵌入式计算、机器学习、机器博弈。

$$\int_{-\infty}^{\infty}\frac{|\Psi(\omega)|^2}{\omega}\mathrm{d}\omega=C \tag{3}$$

式中，$0<C<\infty$，并且

$$\Psi(\omega)=\int_{-\infty}^{\infty}\psi(t)\mathrm{e}^{-j\omega t}\mathrm{d}\omega \tag{4}$$

设 $f(t)$ 为任意一个平方可积的函数，则对 $f(t)$ 的听觉变换定义为：

$$T(a,b)=\int_{-\infty}^{\infty}f(t)\frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a})\mathrm{d}t \tag{5}$$

式中，a,b 为实数。式(5)可以简写为：

$$T(a,b)=\int_{-\infty}^{\infty}f(t)\psi_{a,b}(t)\mathrm{d}t \tag{6}$$

其中

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}\psi(\frac{t-b}{a}) \tag{7}$$

将式(6)写成离散表达式为：

$$T[a_i,b]=\sum_{n=0}^{N}f[n]\frac{1}{\sqrt{|a_i|}}\psi[\frac{n-b}{a_i}] \tag{8}$$

2.2 CFCC 特征参数提取方法

听觉变换是一个处理非线性信号的新方法，与傅里叶变换相同之处在于，听觉变换也相当于一个滤波器组，完成了信号由时域到频域的转换，听觉变换首次提出便在说话人特征提取领域得到了很好的应用。文献[6]基于听觉变换，提出了说话人 CFCC 特征参数的提取方法，Peter Li 定义了一个典型的耳蜗滤波函数的表达式为：

$$\psi_{a,b}(t)=\frac{1}{\sqrt{|a|}}(\frac{t-b}{a})^{\alpha}\exp[-2\pi f_L\beta(\frac{t-b}{a})]\times\cos[2\pi f_L(\frac{t-b}{a})+\theta]u(t-b) \tag{9}$$

式中，$\alpha>0$ 和 $\beta>0$，θ 的取值应该满足式(1)，$u(t)$ 为单位步进函数，b 为随时间可变的实数，a 为尺度变量，一般情况下可由滤波器组的中心频率 f_c 和最低中心频率 f_L 决定：

$$a=f_L/f_c \tag{10}$$

由式(10)可知，a 的取值范围为 $0<a\leqslant 1$。而 α 和 β 一般取经验值为 $\alpha=3,\beta=0.2$。

设 $f(t)$ 为一段语音信号，将式(9)代入式(5)便得到 $f(t)$ 经过 AT 变换得到的 $T(a,b)$，但耳蜗滤波器组只是模拟了人耳耳蜗的冲击响应，人耳耳蜗的内毛细胞将经过时频变换后的语音信号转变为人脑可分析的电信号。Peter Li 用式(11)－式(13)来模拟这个过程：

$$h(a,b)=T(a,b)^2;\forall\, T(a,b) \tag{11}$$

$$S(i,j)=\frac{1}{d}\sum_{b=l}^{l+d-1}h(i,b),l=1,L,2L,\cdots;\forall\, i,j \tag{12}$$

$$y(i,j)=S(i,j)^{\frac{1}{3}} \tag{13}$$

式中，$d=\max\{3.5\tau_i,20\mathrm{ms}\}$，$\tau_i$ 为第 i 个子带中心频率的周期，$L=10\mathrm{ms}$。最后将 $y(i,j)$ 通过离散余弦变换得到 CFCC 特征参数。

3 TEO-CFCC 特征参数的提取方法

一个完整的语音信号包含了频率信息和能量信息[8]，能量是语音信号的一个最基本的参数，它代表了一帧语音信号能量的大小，即使对同一内容的文本，不同说话人在相同环境下表述出的语音信号中的能量值也是不同的[9]；此外，由一帧语音求出的短时能量是一个标量值，表征了语音的时域特征，而 CFCC 参数是人耳听觉感知特征，二者反映了语音信号的不同特征，将二者结合得到的说话人特征参数更能代表说话人的个性特征。文献[10]提出了基于 Teager 能量算子和小波变换的语音识别特征参数，成功地将 Teager 能量算子应用于语音识别的特征参数提取，提高了语音识别的性能。本文同样选取 Teager 能量算子来表征语音信号的能量特征。TEO 是一种非线性算子，能在抑制背景噪声中起到增强信号同时进行特征提取的作用。对于信号采样点 $x(n)$，TEO 的离散表达式为：

$$T[x(n)]=x^2(n)-x(n+1)x(n-1) \tag{14}$$

在噪声环境下，假设观察到的语音信号 $x(n)$ 为纯语音信号 $s(n)$ 和非零均值加性噪声 $w(n)$ 之和，即：

$$x(n)=s(n)+w(n) \tag{15}$$

如果不加处理，将 $x(n)$ 直接代入式(14)求取 TEO，必然会因为噪声而影响 TEO 结果的准确性。为了消除噪声对语音信号的影响，文献[10]提出了能量估计的方法来消除噪声，这种能量的估计方法在零均值加性噪声的条件下可以很好地估算出语音的 TEO，但是在噪声为非零均值加性噪声的条件下不适用。为了解决这个问题，本文将信号相位匹配思想引入 TEO 的求取过程中，利用三元阵的信号相位匹配原理[11]来消除语音信号中非零均值加性噪声对 TEO 的影响。首先将式(15)写成模与相位的形式：

$$|X(j\omega)|\mathrm{e}^{j\psi}=|S(j\omega)|\mathrm{e}^{j\alpha}+|W(j\omega)|\mathrm{e}^{j\varphi} \tag{16}$$

式中，$|X(j\omega)|$，$|S(j\omega)|$ 和 $|W(j\omega)|$ 为谱幅度，ψ、α 和 φ 为相位角，它们都是 ω 的函数。如图 1 所示，用 3 个传感器组成的线列阵来接收信号，θ 为信号方向与线列阵的法线方向的夹角，那么 3 个传感器输出信号的频域形式为：

$$|X_1(j\omega)|\mathrm{e}^{j\psi_1}=|S(j\omega)|\mathrm{e}^{j\alpha}+|W_1(j\omega)|\mathrm{e}^{j\varphi_1} \tag{17}$$

$$|X_2(j\omega)|\mathrm{e}^{j\psi_2}=|S(j\omega)|\mathrm{e}^{j(\alpha-\omega\tau)}+|W_2(j\omega)|\mathrm{e}^{j\varphi_2} \tag{18}$$

$$|X_3(j\omega)|\mathrm{e}^{j\psi_2}=|S(j\omega)|\mathrm{e}^{j(\alpha-2\omega\tau)}+|W_3(j\omega)|\mathrm{e}^{j\varphi_3} \tag{19}$$

式中，$\tau=\frac{d}{c}\sin\theta$，$d$ 为阵元间距，c 为波的传播速度。

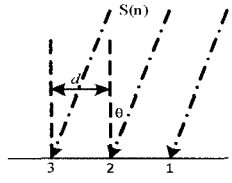

图 1 三元阵的信号相位匹配原理示意图

对式(18)、式(19)两边分别乘以 $\mathrm{e}^{j\omega\tau}$ 和 $\mathrm{e}^{j2\omega\tau}$，得到：

$$|X_1(j\omega)|\mathrm{e}^{j\psi_1}=|S(j\omega)|\mathrm{e}^{j\alpha}+|W_1(j\omega)|\mathrm{e}^{j\varphi_1} \tag{20}$$

$$|X_2(j\omega)|\mathrm{e}^{j(\psi_2+\omega\tau)}=|S(j\omega)|\mathrm{e}^{j\alpha}+|W_2(j\omega)|\mathrm{e}^{j(\varphi_2+\omega\tau)} \tag{21}$$

$$|X_3(j\omega)|\mathrm{e}^{j(\psi_2+2\omega\tau)}=|S(j\omega)|\mathrm{e}^{j\alpha}+|W_3(j\omega)|\mathrm{e}^{j(\varphi_3+2\omega\tau)} \tag{22}$$

用文献[11]的方法求出期望信号的解为：

$$\mathrm{Re}(S)=\frac{EA-FB}{2(CA-DB)} \tag{23}$$

$$\mathrm{Im}(S)=\frac{FC-ED}{2(CA-DB)} \tag{24}$$

式中，$A=\mathrm{Im}(X_3-X_1)$，$B=\mathrm{Im}(X_2-X_1)$，$C=\mathrm{Re}(X_2-X_1)$，$D=\mathrm{Re}(X_3-X_1)$，$E=|X_2|^2-|X_1|^2$，$F=|X_3|^2-|X_1|^2$，Re 表示取实部，Im 表示取虚部。

这样，由 3 个传感器组成的线列阵接收到的信号 X_1、X_2 和 X_3 就可以更加精准地估计语音信号 $s(n)$。在低信噪比的情况下，通过这样的处理不但可以求取更具有代表说话人特征的 TEO，而且可以提高低信噪比情况下 CFCC 特征参数的识别正确率。

本文提出的 TEO-CFCC 特征参数提取方法的原理图如图 2 所示。

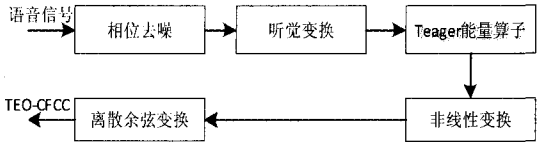

图 2　TEO-CFCC 特征参数提取方法原理图

首先对语音信号进行相位去噪处理，减少噪声信号对语音特征参数的影响；再对语音信号进行听觉变换处理，计算功率谱值 $P(i)$；然后根据式(14)对功率谱各点计算 TEO；对经过 TEO 变换后的谱值通过式(11)－式(13)的非线性变换得到 $y(i,j)$；最后通过离散余弦变换去除各维信号之间的相关性，将信号映射到低维空间得到说话人的 TEO-CFCC 特征参数。

4　实验及结果分析

本文将 TIMIT 语音数据库 Train 目录下的 90 个说话人作为训练和测试数据，其中女性说话人 30 人，男性 60 人，采用与文本无关的高斯混合模型(Gaussian Mixture Model，GMM)为语音的声学模型。实验中，三元阵所选用的传感器为麦克风阵列，将其放置于大小为(6m×4m×2.5m)的房间中，其中第一个麦克风位于房间(1m，2m，1.5m)处，相邻麦克风之间的距离为 d=10cm；声源位于房间(4m，1m，1.5m)处。

为了验证本文提出的 TEO-CFCC 特征参数的有效性，分别在纯净环境下、5dB、0dB、－5dB 和－10dB 的噪声环境下进行对比测试，加入的噪声为标准噪声库 noisex-92 中的汽车噪声(Vehicle interior noise)、多路重合噪声(Babble noise)和白噪声(White noise)。图 3－图 5 分别是在以上 3 种噪声情况下得到的测试结果。

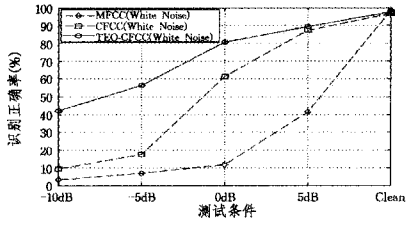

图 3　白噪声情况下 MFCC、CFCC、TEO-CFCC 特征参数识别正确率测试对比图

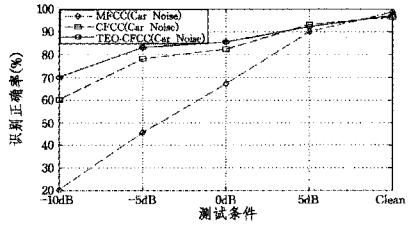

图 4　汽车噪声情况下 MFCC、CFCC、TEO-CFCC 特征参数识别正确率测试对比图

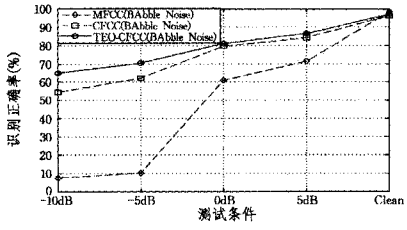

图 5　多路重合噪声情况下 MFCC、CFCC、TEO-CFCC 特征参数识别正确率测试对比图

由图 3 可知，在纯净的测试条件下，3 种特征参数都达到了 96%以上的识别正确率，但随着白噪声逐渐增强，MFCC 特征参数的识别正确率急剧下降，在信噪比为 5dB 的情况下，MFCC 特征参数的识别正确率不到 50%，而 CFCC 特征参数和本文提出的 TEO-CFCC 特征参数的识别正确率都接近 90%；当信噪比为 0dB 时，MFCC 特征参数的识别正确率不到 10%，CFCC 特征参数的识别正确率虽然有所下降，但也达到了 60%，本文提出的 TEO-CFCC 特征参数的识别正确率在 0dB 的条件下可以达到 80%；在－5dB 条件下，CFCC 的识别正确率不到 20%，而本文提出的 TEO-CFCC 特征参数的识别正确率却可以达到 57.8%。

同样，分析图 4 和图 5 可知，在汽车噪声和多路重合噪声的情况下，MFCC 特征参数在信噪比低于 5dB 的情况下，其识别正确率急剧下降，而 CFCC 特征参数和本文提出的 TEO-CFCC 特征参数均表现出良好的抗噪性，尤其在信噪比低于－5dB 的情况下，TEO-CFCC 特征参数的识别正确率明显优于 CFCC 特征参数的识别正确率，在汽车噪声信噪比为－10dB 时，依然可以达到 70%的识别正确率。

结束语　本文在 CFCC 特征参数的基础上，结合相位匹配方法和 Teager 能量算子，提出了表征说话人个性特征的 TEO-CFCC 参数的提取方法。通过统一的 GMM 说话人识别模型验证可知，结合能量算子后的 TEO-CFCC 特征参数的识别正确率明显优于 MFCC 和单纯的 CFCC 参数，在－5dB 的汽车噪声条件下的识别识别正确率可以达到 83.2%。

参 考 文 献

[1]　倪崇嘉，刘文举，徐波. 基于多空间概率分布的汉语连续语音声调识别研究[J]. 计算机科学，2011，38(9)：224-227

[2]　罗宪华，杨大利，徐明星. 面向非特定人语音情感识别的 PCA 特征选择方法[J]. 计算机科学，2011，38(8)：212-214

[3]　韩志艳，王健，王旭. 基于正交实验设计的语音识别特征参数优化[J]. 计算机科学，2010，37(1)：214-216

[4]　Vijayasenan D，Valente F，Bourlard H. Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features [J]. Speech Communication，2012，54(1)：55-67

[5]　Wang L，Minami K，Yamamoto K，et al. Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions [J]. IEICE Transactions on Information and Systems，2010，E93D(9)：2397-2406

[6]　Li Q，Huang Y. An Auditory-Based Feature Extraction Algorithm for Robust Speaker Identification Under Mismatched Conditions [J]. IEEE Transactions on Audio Speech and Language Processing，2011，19(6)：1791-1801

[7]　Li Qi. An auditory-based transfrom for audio signal processing [C]// 2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. New Paltz，NY，United states，Oct. 2009：181-184

[8]　Dimitriadis D，Maragos P，Potamianos A. On the Effects of Filterbank Design and Energy Computation on Robust Speech Recog-nition[J]. IEEE Transactions on Audio Speech and Language Processing，2011，19(6)：1504-1516

[9]　Tu C-C，Juang C-F. Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments [J]. Expert Systems With Applications，2012，39(3)：2479-2488

[10]　楼红伟，胡光锐. 基于 Teager 能量算子和小波变换的语音识别特征参数 [J]. 上海交通大学学报，2003，37：83-85

[11]　孙进才，朱维杰，孙铁源. 利用小尺度阵的波达方向估计[J]. 西北工业大学学报，2003，21(2)：152-155