基于 DWT-TEO 的说话人识别[1)]

邱政权 尹俊勋 薛丽萍

(华南理工大学电信学院 广州 510640)
(E-mail: qiuzhengquan168@163.com)

摘 要 针对在噪声环境下的说话人识别系统，做了两点改进. 第一，为了提高系统的鲁棒性，通过不同尺度的小波基，把含有噪声的信号分解于不同频段中，然后在各个频段分别通过 TEO(Teager 能量算子) 去噪. 针对说话人识别的特点，在小波重构时对各小波系数进行了加权处理. 再把各个频段的输出通过小波重构恢复信号. 最后通过 Mel 滤波器组把小波系数转换成 MFCC. 第二，为了进一步提高识别性能和训练速度，在识别阶段采用了改进的 OGMM(正交高斯混合模型), 即把正交变换改到 EM 算法之前进行，这样就不必要在 EM 迭代过程中每次都进行正交运算了. 从实验得出，采用本文提出的 DWT-TEO 参数对于说话人识别的效果较好. 采用改进的 OGMM 进一步提高了识别性能和训练速度.

关键词 小波变换, TEO, DWT-TEO, OGMM
中图分类号 TN912.34

Speaker Recognition Based on DWT-TEO

QIU Zheng-Quan YIN Jun-Xun XUE Li-Ping

(*School of Electronics and Information Engineering, South China University of Technology, Guangzhou* 510640)

(E-mail: qiuzhengquan168@163.com)

Abstract Two modifications for speaker recognition system in noise environment are described. First, in order to improve the robustness of the system, noisy speech is decomposed into various frequency bands and de-noising is carried out by TEO in every frequency band. The wavelet coefficient is weighted according to the characteristics of speaker recognition, and is then transformed into MFCC. Second, in order to improve recognition performance and training speed, a modified OGMM that orthogonal transform is performed before EM arithmetic is applied at the recognition stage. Thus, it is not necessary to do orthogonal operation during every EM iterative process. The experimental results show that the parameters proposed have produced good effect and that modified OGMM can further improve recognition performance and training speed.

Key words Wavelet transforms, TEO, DWT-TEO, OGMM

1 引言

说话人识别很久以来就是一个既有巨大吸引力而又有相当困难的课题[1,2]. 说话人识别中的语音信号特征是信号、噪声的频谱较宽且重叠，噪声传递函数复杂甚至为 IIR 传递

1) 广东省自然科学基金项目 (000872) 资助
Supported by Natural Science Foundation of Guangdong Province (000872)
收稿日期 2005-6-29 收修改稿日期 2006-3-27
Received June 29, 2005; in revised form March 27, 2006

函数，为改变其性能，人们提出了很多解决方案，目前比较有效的方法是把输入信号进行正交化降阶处理，以减少特征值的分散度，如格型滤波器，Gram-Schmidt 正交化[3,4]，离散傅里叶变换．随着小波技术的发展，很多学者都把小波变换应用于小波去噪中[5,6]．

小波去噪是基于小波系数的阈值的一种简单去噪方法．假设小波阈值定义在噪声的小波系数和那些目标信号之间的界限内，对于含噪语音，清音段的语音几乎和噪音相当．对于所有的含噪语音，都使用一致的阈值，不仅压缩了附加噪声，也压缩了部分语音成分，如清音．因此滤掉的语音感知质量会受到极大的影响，可以把小波变换和别的信号处理方法结合起来，用来提高系统的鲁棒性．TEO 是一个由 Kaiser 提出的有力的非线性算子[7]．Teager 能量算子能消除信号的零均值噪声的影响，具有语音增强的能力，同时又保留了倒谱分析方法中的准稳态假设因而更能体现语音信号的复杂性．

本文把小波分析和 TEO 结合起来，形成 DWT-TEO 参数．把输入的含噪语音通过不同尺度的小波基分解于不同频段中，然后在各个频段分别通过 TEO 去噪．针对说话人识别的特点，在小波重构之前对各小波系数进行了加权处理，再把各个频段的输出通过小波重构恢复信号．通过 Mel 滤波器组把小波系数转换成 MFCC. 与目前主流的 ETSI 等抗噪声算法比较，系统的鲁棒性有所增强．为了进一步提高识别性能和训练速度，在识别阶段采用了改进的正交高斯混合模型 (OGMM).

2 离散小波变换

Mallat 于 1989 年提出多尺度分析[8]，通过它可以构造正交小波基，并且在多尺度分析的基础上，产生了有限尺度二进制小波的 Mallat 算法．离散平滑逼近递推公式为

$$x_k^{(j)} = \sum_n h_0(n-2k) \times x_k^{(j-1)} \tag{1}$$

离散细节信号递推公式

$$d_k^{(j)} = \sum_n h_1(n-2k) \times x_k^{(j-1)} \tag{2}$$

其中，$h_0(k)$, $h_1(k)$ 为

$$h_0(k) = \frac{1}{\sqrt{2}} \int \phi\left(\frac{t}{2}\right) \phi^*(t-k)\mathrm{d}t \tag{3}$$

$$h_1(k) = \frac{1}{\sqrt{2}} \int \psi\left(\frac{t}{2}\right) \phi^*(t-k)\mathrm{d}t \tag{4}$$

Mallat 算法的离散小波重建过程的递推公式为

$$x_n^{(j-1)} = \sum_k g_0(n-2k) \times x_k^{(j)} + \sum_k g_1(n-2k) \times d_k^{(j)} \tag{5}$$

针对说话人识别中在高频 (大于 2000Hz) 和在低频 (小于 500Hz) 比中间频带含有更多的说话人信息，因此在小波重构时对各小波系数进行了加权处理，其递推表达式变为

$$x_n^{(j-1)} = \omega_j \times [\sum_k g_0(n-2k) \times x_k^{(j)} + \sum_k g_1(n-2k) \times d_k^{(j)}] \tag{6}$$

式中 $j = 1, 2, \cdots, L$; ω_j 为加权系数.

3　TEO

自从 Maragos 等首先提出了 Teager 能量算子 (TEO) 的定义式后，该算子得到了一系列应用. TEO 能在抑制背景噪声中起到信号增强，同时进行信号特征提取的作用. 在离散时间中，TEO 可以近似为

$$\psi[x(n)] = [x(n)]^2 - x(n-1) \times x(n+1) \tag{7}$$

因为算法涉及到输入信号和信号的错位相乘，该算法对于信号中的噪声可以预见是较为敏感的. 假设 $x(n)$ 为一宽带稳态随机信号，则 $E\{\Psi[x(n)]\} = E\{x^2(n) - x(n-1)x(n+1)\}$ 或者 $E\{\Psi[x(n)]\} = R_x(0) - R_x(2)$. 这里的 $R_x(k)$ 为 $x(n)$ 的自相关函数. 在含噪语音信号中，假设所观察的信号 $x(n)$ 为纯语音信号 $s(n)$ 和零均值加性噪声 $\omega(n)$ 之和，则带噪语音信号 $x(n)$ 的 TEO 为

$$\psi[x(n)] = \psi[s(n)] + \Psi[\omega(n)] + 2\tilde{\Psi}[s(n), \omega(n)] \tag{8}$$

这里的 $\tilde{\Psi}[s(n), \omega(n)] = s(n)\omega(n) - 0.5s(n-1)\omega(n+1) - 0.5s(n+1)\omega(n-1)$, 也称为 $s(n)$ 和 $\omega(n)$ 的互 Teager 能量. 由于 $s(n)$ 和 $\omega(n)$ 是零均值和相互独立的，则 $\tilde{\Psi}[s(n), \omega(n)]$ 的期望值是零，所以 $E\{\Psi[x(n)]\} = E\{\Psi[s(n)]\} + E\{\Psi[\omega(n)]\}$. 这表明算法在噪声，特别是对宽带噪声或者噪声能量大于受干扰的信号能量时有影响.

然而必须注意到：对于许多物理信号，包括语音能量函数 E_n 的带宽远远低于信号本身各分量带宽，因而信号可以通过低通滤波大大地减少这种噪声的影响. 因此大部分情况下，$E\{\Psi[\omega(n)]\}$ 相对于 $E\{\Psi[s(n)]\}$ 是可以忽略的，因而 $E\{\Psi[x(n)]\} \approx E\{\Psi[s(n)]\}$. 这里可以看出 TEO 具有消除零均值噪声的影响的能力.

4　小波去噪

本文把小波分析和 TEO 结合起来，形成 DWT-TEO 参数. 通过不同尺度的小波基把输入含噪信号分解于不同频段中，然后在各个频段分别通过 TEO 去噪. 针对说话人识别的特点，在小波重构时对各小波系数进行了加权处理[2]. 最后再把各个频段的输出通过小波重构恢复信号. 最后通过 Mel 滤波器组把小波系数转换成 MFCC. 方案见图 1.

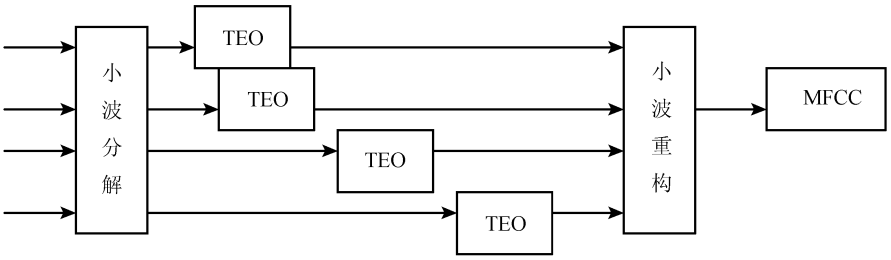

图 1　DWT-TEO 去噪系统

Fig. 1　DWT-TEO de-noising system

5 正交高斯混合模型 (OGMM)

高斯混合模型是 M 个高斯变量的加权和. 假设当前说话人的协方差矩阵是 Σ_x 和变换矩阵 Ω 是由 Σ_x 的特征矢量组成的, 那么在线性变换 ($\boldsymbol{y} = \Omega^{\mathrm{T}}\boldsymbol{x}$) 后, 在 $\boldsymbol{y}$ 空间的协方差矩阵 Σ_y 是对角的. Σ_y 和 Σ_x 有如下的关系:

$$\Sigma_y = \Omega^{\mathrm{T}}\Sigma_x\Omega \tag{9}$$

既然 Ω 是由 Σ_x 的特征矢量组成的, 那么它具有属性 $\Omega^{\mathrm{T}}\Omega = I$. 用 $\Omega\boldsymbol{y}$ 代替 $\boldsymbol{X}$, 有:

$$b_i(\boldsymbol{y}) = \frac{1}{(2\pi)^{D/2}|\Sigma_{yi}|^{1/2}}\exp\{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu}_{yi})^{\mathrm{T}}\Sigma_{yi}^{-1}(\boldsymbol{y}-\boldsymbol{\mu}_{yi})\} \tag{10}$$

其中 $\Sigma_{yi} = \Omega^{\mathrm{T}}\Sigma_{xi}\Omega$ 和 $\mu_{yi} = \Omega^{\mathrm{T}}\boldsymbol{\mu}_{xi}$.

含多个高斯成分的 GMM, 协方差矩阵通常并不真正是对角的, 然而 y 空间的对角 GMM 更能逼近特征矢量的分布. 我们把具有正交的 GMM 叫正交 GMM(OGMM)[1].

图 2 显示了 OGMM 的一个框图. 这个模型是由一个线性变换矩阵和一个对角 GMM 组成. 在训练时我们的方法分两步. 第一步是计算正交变换矩阵. 在获得变换矩阵后, 第二步是用变换矩阵去乘这个模型的所有训练矩阵. 这个变换在训练仅做一次. 然后, 用通用的 EM 算法去估计对角 GMM 的参数. 在测试阶段, 在它们应用到对角 GMM 之前也把测试矢量也变换到新的空间去. 在过去的文献中[9], 正交变换是在 EM 算法中进行, 为了进一步提高系统的运算速度, 我们提出了把正交变换改到 EM 算法之前进行, 这样就不必要在 EM 迭代过程中每次都进行正交运算了.

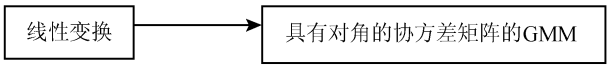

图 2 OGMM 框图

Fig. 2 Block diagram of OGMM

6 实验及结果

与文本无关的说话人辨认是通过鉴定一个说话人发出任何测试文本来鉴别或确认说话人的身份. 实验由 30 个说话人组成, 每个说话人说出 6 个句子, 其中两个是每个说话人都要说的句子, 另外 4 个句子对于每个人不同. 其中每个说话人随机选取 3 个句子组成训练集, 剩下的 3 个句子组成测试集. 采用的语音是在三个月内分三次进行录制的. 采样率为 11025Hz, 帧长为 30ms, 帧移为 15ms, 进行预加重.

为了验证所提出参数性能的优越性, 做了四个试验, 前三个试验, 采用纯净语音附加各种信噪比的高斯白噪声, 用汉明窗加窗. 训练语音时长为 10 秒, 分别用时长为 0.2 秒、1 秒和 2 秒的语音进行辨认. 第四个试验采用的语音把附加噪声改为非平稳的 Babble 噪声, 训练语音时长为 10 秒, 用时长为 2 秒的语音进行辨认.

试验 1. 首先对训练集进行分帧、预加重和加窗, 提取 12 阶 MFCC, 然后对于每一个人, 建立一个 GMM, GMM 中的混合数为 64, 把所得的数据存储下来; 同样对测试集进行分帧、预加重和加窗, 然后提取 12 阶 MFCC, 求出该 MFCC 与训练集中的每个说话人的 GMM 的似然得分, 其中获得似然得分最大的说话人作为正确的辨认人.

试验 2. 首先对训练集进行分帧、预加重和加窗，用 Daubechies 小波进行三尺度分解，把分解后的各高频和低频分量根据说话人识别的特点进行加权，之后进行小波重构，然后提取 12 阶 MFCC(我们称之为 DWT-MFCC 参数)，然后对于每一个人，建立一个 GMM,GMM 中的混合数为 64，把所得的数据存储下来；测试阶段同训练阶段一样提取 12 阶 MFCC，求出该 MFCC 与训练集中的每个说话人的 GMM 的似然得分，其中获得似然得分最大的说话人作为正确的辨认人.

试验 3. 首先对训练集进行分帧、预加重和加窗，用 Daubechies 小波进行三尺度分解，用 Teager 能量算子对分解后的各高频和低频分量进行处理，之后根据说话人识别的特点进行加权，进行小波重构，然后提取出 12 阶 MFCC(我们称之为 DWT-TEO 参数，也就是本文所提出的参数)，然后对于每一个人，建立一个 OGMM，GMM 中的混合数为 64，把所得的数据存储下来；测试阶段同训练阶段一样提取 12 阶 MFCC，求出该 MFCC 与训练集中的每个说话人的 GMM 的似然得分，其中获得似然得分最大的说话人作为正确的辨认人．试验过程见图 3.

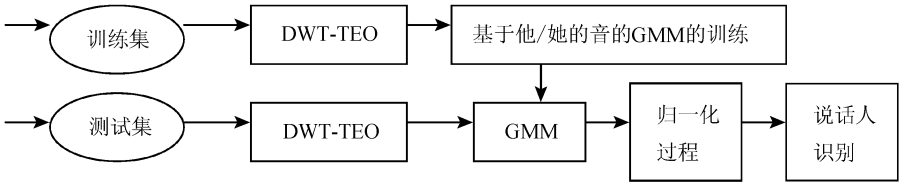

图 3　基于 DWT-TEO 参数说话人辨认框图

Fig. 3　Block diagram of speaker identification based on DWT-TEO parameter

试验 4. 把附加噪声改为非平稳的 Babble 噪声，训练语音时长为 10 秒，用时长为 2 秒的语音进行辨认，再重复前面的三个试验．实验结果见表 1 、表 2 、表 3 和表 4.

表 1　不同 SNR 下的说话人辨认率 (测试语音时长为 0.5 秒)

Table 1　The speaker identification rate under different SNR (the length of the test time is 0.5s)

SNR(DB)	5	10	15	20
MFCC	71.9%	79.1%	84.5%	89.6%
DWT-MFCC	75.5%	81.2%	86.8%	91.3%
DWT-TEO	82.5%	86.8%	90.3%	92.6%

表 2　不同 SNR 下的说话人辨认率 (测试语音时长为 1 秒)

Table 2　The speaker identification rate under different SNR (the length of the test time is 1s)

SNR(DB)	5	10	15	20
MFCC	74.4%	81.6%	87.8%	92.4%
DWT-MFCC	82.2%	87.2%	91.5%	94.6%
DWT-TEO	86.7%	91.5%	94.6%	96.8%

表 3　不同 SNR 下的说话人辨认率 (测试语音时长为 2 秒)

Table 3　The speaker identification rate under different SNR (the length of the test time is 2s)

SNR(DB)	5	10	15	20
MFCC	76.8%	83.6%	89.7%	93.4%
DWT-MFCC	83.5%	88.7%	93.4%	96.6%
DWT-TEO	89.1%	94.6%	97.2%	99.8%

表 4 在噪声为非平稳的 Babble 噪声，测试语音时长为 2 秒时不同的 SNR 时的说话人辨认率
Table 4 The speaker identification rate under different SNR when the length of the test time is 2s and noise is non-stationary Babble

SNR(DB)	5	10	15	20
MFCC	75.5%	83.0%	88.9%	92.2%
DWT-MFCC	80.6%	88.1%	92.7%	96.7%
DWT-TEO	88.5%	94.0%	96.9%	99.5%

从表 1 、表 2 和表 3 可以看出：在噪声环境下，采用本文所提出的 DWT-TEO 参数的说话人辨认率要高于采用其它参数的说话人辨认率，而且，随着信噪比的增大，提出的特征对于说话人识别的性能却表现出更强的鲁棒性. 从表 3 可以看出，随着 SNR 的减少，采用 MFCC 的辨认率减少了 16.6%, 而采用 DWT-MFCC 的辨认率却减少了 13.1%. 当 SNR 为 20DB 时，采用 DWT-MFCC 的辨认率比采用 MFCC 的提高了 3.2%, 可见采用的 DWT-MFCC 参数的鲁棒性和辨认率要比采用 MFCC 参数的好，小波系数进行加权处理能提高说话人的辨认率，而 DWT 则能增强系统的鲁棒性. 采用的 DWT-TEO 参数与前面两个参数比较可知，随着 SNR 的减少，它的辨认率减少了 10.7%. 且当 SNR 为 20DB 时，辨认率达到了 99.8%. 可见把对小波系数进行的加权处理和 TEO 结合起来，系统的鲁棒性进一步增强. 对表 1 、表 2 和表 3 进行比较，可以知道：随着测试语音时长的增长，说话人识别率逐渐升高，这是因为随着语音时长的增长，语音就会含有更多的说话人信息.

从表 4 看出，所提出的 DWT-TEO 参数确实比 MFCC 和 DWT-MFCC 的识别率和鲁棒性要有所提高，因此进一步验证了本文所提出的参数的有效性. 表 4 与表 3 比较，性能有稍微的下降，这是因为噪声环境的不同造成的. 为了比较 OGMM 的优越性，我们测量了三个实验平均所花的时间. 表 4 是当 SNR 为 10 且测试语音时长为 2 秒时三个实验所花的平均时间.

从表 5 可以看出，第二个实验所花的时间比第一个实验要多一些，这是因为采用小波处理要花费更多的时间的缘故；但是第三个实验就比前面两个实验所花的时间明显少很多. 可见， OGMM 确实能够提高计算速度.

表 5 SNR 为 10, 时长为 2 秒时三个实验所花的平均时间
Table 5 The average time when SNR is 10 and the length of the time in three test speech

	MFCC-GMM	DWT-MFCC-GMM	DWT-TEO-OGMM
平均时长	13.56	17.27	7.69

7 结论

从实验结果可以看出，在噪声环境下，采用所提出的参数的说话人辨认率要高于采用其它参数的说话人辨认率，而且随着信噪比的减少，提出的特征参数对于说话人识别的性能却表现出较强的鲁棒性. 可见所提出的 DWT-TEO 参数确实能增强系统的鲁棒性，而小波系数进行加权处理能提高说话人的辨认率. 与目前主流的 ETSI 等抗噪声算法的比较，系统的鲁棒性有所增强. 把正交变换改到 EM 算法之前进行，这样就不必要在 EM 迭代过程中每次都进行正交运算了，对提高系统的运算速度发挥了较大的作用. 实验结果显示：OGMM 确实提高了计算速度.

References

1 Liu L, He J. On the use of orthogonal GMM in speaker recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Phoenix, Arizona, USA, 1999. **2**: 845∼848

2 Sakka Z, Kachouri A, Mezghani A, Samet M. A new method for speech de-noising and speaker verification using sub-band architecture. In: First International Symposium on Control, Communications and Signal Processing, 2004. 37∼40
3 Qu T S, Dai Y S. A new method for adaptive speech noise canceling based on wavelet transform. *Transactions of China Electrotechnical Society*, 2001, **4**(2): 75∼78
4 Lou H W, Hu G R. Speech feature based on Teager energy operator and dyadic wavelet transform. *Journal of Shanghai Jiaotong University*, 2003, (S2): 83∼85
5 Yi Hu, Loizou P C. Speech enhancement based on wavelet thresholding the multitaper spectrum. *IEEE Transactions on Speech and Audio Processing*, 2004, **12**(1): 59∼67
6 Lu C T, Wang H C. Speech enhancement using perceptually-constrained gain factors in critical band wavelet packet transform. *Electronics Letters*, 2004, **40**(6): 394∼396
7 Kaiser J F. Some useful properties of Teager′s energy operators. In: Proceesings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993. **3**: 149∼152
8 Mallat S G. A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence*, 1989, **11**(7): 674∼693
9 Yuo K, Wang H. Gaussian mixture models with common principal axes and their application in text independent speaker identification. In: Proceedings of Eurospeech Conference. Rhodes, Greece, 1997. 2279∼2282

邱政权 博士研究生．研究方向为语音信号处理．

(**QIU Zheng-Quan** Ph. D. candidate at South China University of Technology. His current research interest includes speech signal processing.)

尹俊勋 博士生导师．研究方向为通信与音视频信号处理．

(**YIN Jun-Xun** Professor at South China University of Technology. His current research interests include communications, speech and video signal processing.)

薛丽萍 博士研究生．研究方向为语音信号处理．

(**XUE Li-Ping** Ph. D. candidate at South China University of Technology. Her current research interest includes speech signal processing.)