基于 KLT 算法和 MMCE 的说话人识别

范小春,邱政权
FAN Xiao-chun,QIU Zheng-quan

湖南科技大学 信息与电气工程学院,湖南 湘潭 411201
School of Information and Electrical Engineer,Hunan University Science and Technology,Xiangtan,Hunan 411201,China

FAN Xiao-chun,QIU Zheng-quan.Speaker recognition based on KLT and MMCE.Computer Engineering and Applications, 2009,45(15):194-195.

Abstract: In this paper,a kind of novel speech de-noising method based on the combination of wavelet de-noising and adapting KLT is proposed.The speech de-noising arithmetic of adapting KLT need not white processing.It is adapted to tract KLT matrix, as well as effectively harmonized the contradiction between signal tone after de-noising and intelligibility.At the recognition stage, Modified MCE is adopted.The experiment is showed that the hybrid system can indeed enhance the robust and the speaker recognition rate of speaker recognition.

Key words: adapted KLT;modified MCE;robust

摘 要: 利用自适应 KLT 提出了一种新的语音去噪方法。自适应 KLT 的语音去噪算法,不需要白化处理,既可以自适应跟踪 KLT 阵,又能够有效地协调去噪后信号的音质与可懂度之间的矛盾。在说话人识别阶段采用改进的 MCE。实验表明,该混合系统确实能够增强说话人辨认的鲁棒性和识别率。

关键词: 自适应 KLT;改进的 MCE;鲁棒性
DOI: 10.3778/j.issn.1002-8331.2009.15.056 **文章编号:** 1002-8331(2009)15-0194-02 **文献标识码:** A **中图分类号:** TN912.34

说话人识别研究有几十年了，有些技术已经非常成熟,尤其是与文本相关的说话人识别已经商品化。但与文本无关的说话人识别,因为事先不知道文本的内容,就会困难得多,但这也正是其吸引人的地方,所以成为研究热点。虽然说话人识别在纯净语音环境中取得了令人满意的结果,但是一旦到了噪声环境下,说话人的识别率就会急剧地下降[1]。

传统的线性滤波方法存在着保护信号局部特征与抑制噪声之间的矛盾,小波变换由于具有时频局部化特点及小波基选择的灵活性,为解决这一问题提供了一个有效的工具。然而,基于阈值的小波域语音增强算法是有潜力的算法,阈值的选择和如何处理阈值是这一算法的关键,但常规阈值方法具有很大局限性,在非平稳噪声条件下效果并不理想[2-6]。光有小波降噪处理,往往还不能达到最佳的语音去噪效果,因此有必要在小波非平稳降噪的基础上进一步进行噪声处理,并且和其他方法结合起来,增强噪声处理的效果。

结合小波降噪和自适应 KLT 提出了一种新的语音去噪方法，有效地解决了小波去噪的最佳的语音去噪效果的问题。在说话人识别阶段采用改进的 MCE。实验表明,该混合系统确实能够增强说话人辨认的鲁棒性和识别率。

1 自适应 KLT 算法

在噪声干扰下的语音增强系统中,带噪语音为:

$$\boldsymbol{z}=\boldsymbol{y}+\boldsymbol{w} \tag{1}$$

其中 $\boldsymbol{z}$、$\boldsymbol{y}$ 和 $\boldsymbol{w}$ 分别表示 k 维带噪语音向量、$\boldsymbol{k}$ 维纯净语音向量和 $\boldsymbol{k}$ 维有色噪声向量。令 $\boldsymbol{H}$ 为 $k\times k$ 维估计阵,$\hat{y}$为 y 的线性估计,有:

$$\hat{y}=\boldsymbol{H}\boldsymbol{z} \tag{2}$$

残余误差信号:

$$\varepsilon=\hat{y}-\boldsymbol{y}=\boldsymbol{H}(\boldsymbol{y}+w)-\boldsymbol{y}=(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{y}+\boldsymbol{H}\boldsymbol{w}=\varepsilon_y+\varepsilon_w \tag{3}$$

其中 $\varepsilon_y=(\boldsymbol{H}-\boldsymbol{I})\boldsymbol{y}$ 表示信号畸变,$\varepsilon_w=\boldsymbol{H}\boldsymbol{w}$ 表示残余噪声。令 R_z、R_y 和 R_w 分别表示 $\boldsymbol{z}$、$\boldsymbol{y}$ 和 $\boldsymbol{w}$ 的自相关函数。且 R_y 的特征分解表示为:

$$R_y=\boldsymbol{U}_y\boldsymbol{\Lambda}_y\boldsymbol{U}_y^{\mathrm{T}} \tag{4}$$

$\boldsymbol{U}_y=(u_{y1},u_{y2},\cdots,u_{yk})$,$\boldsymbol{\Lambda}_y=diag(\Lambda_{y1},\Lambda_{y2},\cdots,\Lambda_{yk})$为特征值的对角阵。$\boldsymbol{U}=[U_1,U_2]$,$U_1$ 为 $\boldsymbol{\Lambda}_y$ 中正特征值对应特征向量组成的 $K\times M$ 阵,M 为子空间维数值的个数。称由 U_1 的列张成的子空间为信号子空间;U_2 为 $\boldsymbol{\Lambda}_y$ 中零特征值对应特征向量组成的 $(K-M)\times M$ 阵,称由 U_2 的列张成的子空间,为与信号子空间正交的子空间。若干扰为白噪声,称为噪声子空间。定义信号畸变的能量为:

$$\xi_y=tr(E(\varepsilon_y\boldsymbol{\varepsilon}_y^{\mathrm{T}})) \tag{5}$$

残余噪声 ε_w 的第 k 个频谱分量表示为 $\boldsymbol{u}_{yk}^{\mathrm{T}}\varepsilon_w$。为了保证每个残余噪声的频谱相似于纯净语音的频谱，同时保证信号畸变 ε_y 的能量最小,提出了一种约束最优化问题,用以求解估计滤波

作者简介: 范小春(1970-),工学硕士,主要从事工业控制组态软件研究。
收稿日期: 2008-03-27 **修回日期:** 2008-06-19

器 H，即：

$$\min_{H}\xi_y,\begin{cases}E\left|\boldsymbol{u}_{yk}^{\mathrm{T}}\varepsilon_w\right|^2\leqslant\alpha_k\delta_{uk}^2,k=1,\cdots,M\\E\left|\boldsymbol{u}_{yk}^{\mathrm{T}}\varepsilon_w\right|^2=0,k=M+1,\cdots,K\end{cases}\tag{6}$$

其中 δ_{uk}^2 为 $\boldsymbol{U}_y^{\mathrm{T}}R_wU_y$ 的第 k 个对角线元素值，也表示沿纯净语音 KLT 特征向量上噪声的能量，$0\leqslant\alpha_k\leqslant1$。设 H 具有如下形式：

$$H=U_yQU_y^{\mathrm{T}}\tag{7}$$

其中 $Q=diag(q_{11},q_{kk},\cdots,q_{KK})$。由式(10)和式(11)可得到：

$$q_{kk}=\begin{cases}\alpha_k^{1/2} & k=1,\cdots,M\\0 & k=M+1,\cdots,K\end{cases}\tag{8}$$

$$\hat{y}=Hz=U_yQU_y^{\mathrm{T}}z\tag{9}$$

假设噪声和信号不相关，则：

$$R_z=R_y+R_w\tag{10}$$

一种有色噪声的近似模型[7]，既描述了有色噪声的特性，又包含了白噪声的情况。

$$E(\boldsymbol{U}_y^{\mathrm{T}}w\boldsymbol{w}^{\mathrm{T}}U_y)=\boldsymbol{U}_y^{\mathrm{T}}R_wU_y=\boldsymbol{\Lambda}_w\tag{11}$$

其中 $\boldsymbol{\Lambda}_w=diag(\delta_{w1},\cdots,\delta_{wK})$，所以：

$$R_z=U_y\boldsymbol{\Lambda}_y\boldsymbol{U}_y^{\mathrm{T}}+U_y\boldsymbol{\Lambda}_w\boldsymbol{U}_y^{\mathrm{T}}=U_y\boldsymbol{\Lambda}_z\boldsymbol{U}_y^{\mathrm{T}}\tag{12}$$

其中 $\boldsymbol{\Lambda}_w=diag(\lambda_{w1},\cdots,\lambda_{wK})$ 为 R_z 特征值组成的对角阵。R_z 和 R_y 的特征向量近似相等，且 R_z 的特征值等于 R_y 特征值与噪声方差之和。为了实现估计滤波器 H，需要估计出每帧纯净语音的 U_y，还要估计出 $\boldsymbol{\Lambda}_y$ 和 $\boldsymbol{\Lambda}_w$。由上面的讨论可知：U_y 和 $\boldsymbol{\Lambda}_z$ 可由 R_z 的特征分解求出；$\boldsymbol{\Lambda}_y$ 可由 $\boldsymbol{\Lambda}_z-\boldsymbol{\Lambda}_w$ 得到；$\boldsymbol{\Lambda}_w$ 由 VAD 单元[7]检测到的噪声样本在当前语音帧的 KLT 特征向量上的投影得到。在求解估计滤波器 H 时，需要准确地估计 R_z 的特征值和特征向量。由于语音信号为非平稳过程，所以可利用自适应跟踪算法来改进 KLT 阵的估计精度。这里采用投影逼近子空间跟踪方法[8]，运用 RLS 算法跟踪协方差阵的特征向量。

代价函数 $J(u(n))=\sum_{i=1}^{N}\beta^{n-i}\parallel z(i)-u(n)\boldsymbol{u}^{\mathrm{T}}(n)z(i)\parallel^2$，其中 n 为帧数，$\boldsymbol{u}(n)$ 为 k 维向量，$0\leqslant\beta\leqslant i$ 为遗忘因子。可以证明 $J(u(n))$ 只有全局最小点，无局部最小点，而且，全局最小点位于指数加权协方差阵的主特征向量上。定义协方差阵 $\boldsymbol{R}_z(n)=\sum_{i=1}^{N}\beta^{n-i}z(i)\boldsymbol{z}^{\mathrm{T}}(i)$ 和另一个代价函数 $J'(u(n))=\sum_{i=1}^{N}\beta^{n-i}\parallel z(i)-u(n)\boldsymbol{u}^{\mathrm{T}}(i-1)z(i)\parallel^2$。$J'(u(n))$ 与 $J(u(n))$ 的不同在于使用了 $\boldsymbol{u}^{\mathrm{T}}(i-1)$ 代替了 $\boldsymbol{u}^{\mathrm{T}}(n)$。$J'(u(n))$ 能够近似地表示为 $J(u(n))$。因为当 $i<<n$ 时，β^{n-i} 很小，减小了两个代价函数的差异；当 i 接近 n 时，由于语音信号的统计特性随时间变化较慢，所以 $u(i-1)\approx u(n)$。另外新的代价函数 $J'(u(n))$ 可以用 RLS 算法最小化，自适应的最小化 $J'(u(n))$ 可以渐进地跟踪 R_z 的主特征向量，这一向量与 R_z 的主特征向量一致。为了跟踪余下的特征向量，采用了压缩技术：首先，主特征向量由 $J'(u(n))$ 的最优化得到；然后，去除 z 在这一主特征向量上的投影。此时，$R_z(n)$ 的第二个特征向量变为主特征向量，可采用相同的方法得到。重复上述过程，可以递推出所有特征向量。KLT 自适应跟踪算法如下：(1)$z_1(n)=z(n)$；(2)$v_i(n)=\boldsymbol{u}_i^{\mathrm{T}}(n-1)z_i(n)$；(3)$d_i(n)=\beta d_i(n-1)+|v_i(n)|^2$；(4)$E_i(n)=z_i(n)-u_i(n-1)v_i(n)$；(5)$u_i(n)=u_i(n-1)+E_i(n)v_i(n)/d_i(n)$；(6)$z_{i+1}(n)=z_i(n)-u_i(n)v_i(n)$；(7)$U_E(n)=[u_1(n),\cdots,u_K(n)]$，$n=1,2,\cdots$。

2　改进的 MCE 模型

采用上述的 MCE 方法，在说话人中得到了成功应用[9]，但是相比于传统的 HMM 方法，MCE 方法的一个问题是训练时间的大量增加，这其中有一部分是由于对于有 K 个说话人的系统而言，每一类别的分类错误都需要计算 $K-1$ 类的判别函数，随着 K 的增加，使得计算量大量增加。于是提出了一种改进的 MCE 方法。

对应于每一个说话人定义一组错误函数来代替一个函数。改进的分类错误函数的形式为：

$$d_{kj}(O;\Lambda)=-\log(P(O;\Lambda_j))+\log(P(O;\Lambda_k)),k\neq j$$

其中 $P(O;\Lambda_k)$ 表示特征序列 O 是由模板 Λ_k 产生的概率。

对于每一个说话人，可以定义 $K-1$ 个 d_{kj}。按照 d_{kj} 的大小，对于每一个说话人 K 可以定义一个数值从小到大的 d_{kj} 序列，代表除了 K 以外的相似度。从而，可以定义一个参数集 $K(O,K,N)$，用来求分类函数，其中 $N<K-1$ 代表对于每一个说话人，用 N 个和他最近的说话人的模型来生成模型参数。此时，相应的代价函数为：

$$l_{kj}(d_{kj}(O,\Lambda))=\frac{1}{1+\exp(d_{kj}(O,\Lambda))}\tag{13}$$

对应一个特征值的代价函数为：

$$l(O,\Lambda)=\sum_k\sum_{j\in K(O<K<N)}l_{kj}(d_{kj}(O,\Lambda))\tag{14}$$

同 MCE 模型一样采用梯度下降算法来实现函数 $l(O,\Lambda)$ 的最小化。

3　实验结果

实验由 80 个说话人组成，由 40 个男的和 40 个女的组成，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。对训练集进行分帧、预加重和加窗，提取 12 阶 MFCC。对于所有的例子，用 128 个元素的分析帧，并分为 3 个子帧。对于所有仿真，用它的采样相关矩阵取代每个需要的相关矩阵。然后基于重叠子帧的 KLT 应用于整个训练数据库去减少原始的 128 维数到更少的维数空间。利用自适应的 KLT 进行去噪；识别阶段采用改进的 MCE。

做了几个实验区评估不同的 KLT，特别计算 KLT 所需的时间和考虑计算量。利用特征分解方法，计算被第 P 个子帧数据矢量所扩展的子空间的主要计算成本是 $12kp^3$ 计算量(CM)。基于所提出的技术，获得 G_x 需 $12k_p^3+11pM^3+M^2k$。因此，用所提出的技术所需的计算量比传统的 KLT 所需的 $12k^3$ 的计算量要少。

为了比较小波去噪的效果，做了三个实验，即第一个采用 GMM 参数，第二个采用 KLT-GMM 参数和第 3 个是采用 KLT-MCE 参数。测试采用混合数为 64 的 GMM 模型。实验结果见表 1。

3　结论

从表 1 可以看出，系统的辨认率都随着 SNR 的增加而增

(下转 215 页)