基于矢量泰勒级数的模型自适应算法

吕　勇　　吴镇扬
(东南大学信息科学与工程学院　南京　210096)

摘　要：在实际环境中，由于测试环境与训练环境的不匹配，语音识别系统的性能会急剧恶化。模型自适应算法是减小环境失配影响的有效方法之一，它通过测试环境下的少量自适应数据，将 HMM 模型的参数变换到测试环境下。该文将矢量泰勒级数用于模型自适应，同时对 HMM 模型的均值向量和协方差矩阵进行变换，使其与实际环境相匹配。实验证明，该文算法优于 MLLR 算法和基于矢量泰勒级数的特征补偿算法，在低信噪比环境中性能提高尤为明显。

关键词：语音识别；模型自适应；矢量泰勒级数；隐马尔可夫模型
中图分类号：TN912.34　　**文献标识码**：A　　**文章编号**：1009-5896(2010)01-0107-05
DOI: 10.3724/SP.J.1146.2008.01768

Model Adaptation Algorithm Using Vector Taylor Series

Lü Yong　　Wu Zhen-yang
(*School of Information Science and Engineering*, *Southeast University*, *Nanjing* 210096, *China*)

Abstract: In actual environments the performance of speech recognition system may be degraded significantly because of the mismatch between the training and testing conditions. Model adaptation is an efficient approach that could reduce this mismatch, which adapts model parameters to new conditions by some adaptation data. In this paper, a new model adaptation using vector Taylor series is presented, which adapts the mean vector and covariance matrix of hidden Markov model. The experimental results show that the proposed algorithm is more effective than MLLR and the feature compensation algorithm based on vector Taylor series in various environments, especially in low signal-to-noise ratio environments.

Key words: Speech recognition; Model adaptation; Vector Taylor series; Hidden Markov model

1　引言

在噪声环境下，由于训练环境和测试环境的失配，语音识别系统的性能会急剧恶化。减小环境失配影响的方法主要有两类：特征补偿算法[1−4]和模型自适应算法[5−9]。特征补偿算法在语音识别系统的前端，对噪声环境下提取的特征向量进行补偿，尽可能将其恢复成纯净语音特征向量。模型自适应算法在语音识别系统的后端，通过测试环境下的少量自适应数据，对模型参数进行调整，逐渐将模型参数变换到实际环境，从而达到提高系统识别率的目的。特征补偿算法和模型自适应算法各有优点，一般来说，在较高信噪比下，特征补偿算法能较好地恢复缺失的语音分量，其性能略好于模型自适应算法；在较低信噪比下，由于 MFCC(Mel Frequency Cepstrum Coefficient)中缺失的分量过多，因此特征补偿算法恢复的特征向量误差较大，其性能明显低于模型自适应算法。

MAP(Maximum A Posteriori)[5] 和 MLLR (Maximum Likelihood Linear Regression)[6]是两种最常用的模型自适应算法。MAP 算法将自适应数据插入到 HMM(Hidden Markov Model)的先验参数中，得到测试环境下的模型参数。MAP 算法具有理论上的最优性，随着自适应数据的增加，其性能会逐渐逼近实际环境下重新训练的系统，但它需要大量自适应数据，而且没有自适应数据的模型参数无法更新。MLLR 是一种基于变换的模型自适应算法，它通过一个或多个线性回归矩阵，将全部模型参数变换到测试环境。在只有少量自适应数据时，MLLR 算法的性能优于 MAP 算法。线性假设是 MLLR 算法的主要缺陷，尤其在噪声环境下，其性能不如其它噪声自适应算法。PMC(Parallel Model Combination)[8]是一种有效的抗噪声模型自适应算法，它联合纯净语音 HMM 模型和噪声模型，通过非线性函数，将 HMM 模型的均值和方差变换到测

2008-12-22 收到，2009-09-18 改回
国家 973 计划项目(2002CB312102)和国家自然科学基金(60971098)资助课题
通信作者：吕勇　　lynetwork@gmail.com

试环境。PMC 的抗噪性能相当优秀，但是它需要大量噪声样本去训练噪声模型，也就是只能用于环境噪声已知的情况下。LST(Linear Spectral Transformation)[9]是一种在线性谱域处理加性噪声和卷积噪声的抗噪声模型自适应算法，它将环境改变前后模型参数之间的非线性变换关系直接代入似然函数，通过数值方法求得噪声的均值和方差。LST 算法只需要少量数据即可取得较好的自适应效果，但其计算量巨大，不能实时实现，影响了它在实际环境下的应用。

本文提出了一种基于矢量泰勒级数(Vector Taylor Series, VTS)的模型自适应算法，利用矢量泰勒级数将环境改变前后模型参数之间的非线性变换关系展开，得到近似的一阶线性关系，代入似然函数，通过 EM(Expectation Maximization)算法计算噪声的均值和方差。VTS 算法最初被用于特征补偿[1,2]，通过 EM 算法估得噪声均值后，更新 GMM 模型的均值向量，然后得到特征向量的 MMSE (Minimum Mean Square Error)估计。文献[1,2]没有估计噪声的协方差矩阵，这是因为一方面从含噪语音中提取噪声的方差信息比较困难，另一方面对特征补偿来说，噪声方差对补偿精度的影响比噪声均值小得多。文献[10]将文献[1,2]中环境改变前后模型均值向量之间的矢量泰勒级数关系用于模型自适应，其性能优于 MLLR 算法。但是，实验证明，在噪声环境下仅仅更新 HMM 的均值不能取得比 VTS 特征补偿算法更好的性能。本文利用一阶矢量泰勒级数展开式逼近环境改变前后模型参数之间的非线性变换关系，通过 EM 算法，在含噪语音中实时提取噪声的均值信息和协方差信息，从而同时更新 HMM 模型的均值向量和协方差矩阵。实验证明，本文算法优于 VTS 特征补偿算法[1,2]和 VTS 模型均值自适应算法[10]，尤其在低信噪比环境下，性能提高非常明显。

2 模型参数的矢量泰勒级数关系式

当前，多数语音识别系统以高斯混合连续密度 HMM 为声学模型，以 MFCC 为特征参数。设在对数谱域，含噪语音特征向量 $\boldsymbol{y}$，纯净语音特征向量 $\boldsymbol{x}$，卷积噪声特征向量 $\boldsymbol{h}$ 和加性噪声特征向量 $\boldsymbol{n}$ 的关系如下[1,2]：

$$\boldsymbol{y}=\boldsymbol{x}+\boldsymbol{h}+\ln(1+\exp(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})) \tag{1}$$

在 $\boldsymbol{x}$ 的均值 $\boldsymbol{u}_{x,im}$，$\boldsymbol{h}$ 的均值 $\boldsymbol{u}_h$，$\boldsymbol{n}$ 的均值 $\boldsymbol{u}_n$ 处，将式(1)用一阶矢量泰勒级数展开，得到 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 的近似关系式：

$$\begin{aligned}\boldsymbol{y}=&\boldsymbol{x}+\boldsymbol{h}+\ln(1+\exp(\boldsymbol{u}_n-\boldsymbol{u}_{x,im}-\boldsymbol{u}_h))\\&-\overline{\boldsymbol{U}}_{im}(\boldsymbol{x}-\boldsymbol{u}_{x,im})-\overline{\boldsymbol{U}}_{im}(\boldsymbol{h}-\boldsymbol{u}_h)+\overline{\boldsymbol{U}}_{im}(\boldsymbol{n}-\boldsymbol{u}_n)\end{aligned} \tag{2}$$

$$\overline{\boldsymbol{U}}_{im}=\mathrm{diag}\left(\frac{\exp(\boldsymbol{u}_n-\boldsymbol{u}_{x,im}-\boldsymbol{u}_h)}{1+\exp(\boldsymbol{u}_n-\boldsymbol{u}_{x,im}-\boldsymbol{u}_h)}\right) \tag{3}$$

其中 $\boldsymbol{u}_{x,im}$ 是纯净语音 HMM 第 i 个状态的第 m 个高斯单元的均值向量，diag(·) 表示根据列向量，生成对角矩阵。根据式(2)，可以得到对数谱域含噪语音和纯净语音模型参数之间的变换关系：

$$\boldsymbol{u}_{y,im}=\boldsymbol{u}_{x,im}+\boldsymbol{u}_h+\ln(1+\exp(\boldsymbol{u}_n-\boldsymbol{u}_{x,im}-\boldsymbol{u}_h)) \tag{4}$$

$$\boldsymbol{S}_{y,im}=(\boldsymbol{I}-\overline{\boldsymbol{U}}_{im})\boldsymbol{S}_{x,im}(\boldsymbol{I}-\overline{\boldsymbol{U}}_{im})^{\mathrm{T}}+\overline{\boldsymbol{U}}_{im}\boldsymbol{S}_n\overline{\boldsymbol{U}}_{im}^{\mathrm{T}} \tag{5}$$

其中 $\boldsymbol{S}_n$ 是加性噪声的协方差矩阵，$\boldsymbol{S}_{y,im}$ 是含噪语音的协方差矩阵，式(5)中假设信道噪声为常数，不影响 $\boldsymbol{S}_{y,im}$。噪声的均值和方差 $\boldsymbol{u}_h$，$\boldsymbol{u}_n$ 和 $\boldsymbol{S}_n$ 都是未知数，需通过测试环境下的少量自适应数据来估计。设 EM 算法中上一次(第 k -1 次)迭代得到的噪声均值为 $\boldsymbol{u}_{h,k-1}$ 和 $\boldsymbol{u}_{n,k-1}$，将式(4)和式(5)在 $\boldsymbol{u}_{h,k-1}$ 和 $\boldsymbol{u}_{n,k-1}$ 处展开，得到

$$\begin{aligned}\boldsymbol{u}_{y,im}=&\boldsymbol{u}_{x,im}+\boldsymbol{u}_{h,k-1}+\ln(1+\exp(\boldsymbol{u}_{n,k-1}-\boldsymbol{u}_{x,im}\\&-\boldsymbol{u}_{h,k-1}))+(\boldsymbol{I}-\overline{\boldsymbol{U}}_{im}^{k-1})(\boldsymbol{u}_h-\boldsymbol{u}_{h,k-1})\\&+\overline{\boldsymbol{U}}_{im}^{k-1}(\boldsymbol{u}_n-\boldsymbol{u}_{n,k-1})\end{aligned} \tag{6}$$

$$\boldsymbol{S}_{y,im}=(\boldsymbol{I}-\overline{\boldsymbol{U}}_{im}^{k-1})\boldsymbol{S}_{x,im}(\boldsymbol{I}-\overline{\boldsymbol{U}}_{im}^{k-1})^{\mathrm{T}}+\overline{\boldsymbol{U}}_{im}^{k-1}\boldsymbol{S}_n(\overline{\boldsymbol{U}}_{im}^{k-1})^{\mathrm{T}} \tag{7}$$

$$\overline{\boldsymbol{U}}_{im}^{k-1}=\mathrm{diag}\left(\frac{\exp(\boldsymbol{u}_{n,k-1}-\boldsymbol{u}_{x,im}-\boldsymbol{u}_{h,k-1})}{1+\exp(\boldsymbol{u}_{n,k-1}-\boldsymbol{u}_{x,im}-\boldsymbol{u}_{h,k-1})}\right) \tag{8}$$

设 DCT 矩阵及其逆矩阵分别为 $\boldsymbol{C}$ 和 $\boldsymbol{C}^{-1}$，将式(6)和式(7)变换到倒谱域：

$$\begin{aligned}\boldsymbol{\mu}_{y,im}=&\boldsymbol{\mu}_{x,im}+\boldsymbol{\mu}_{h,k-1}+\boldsymbol{C}\ln(1+\exp(\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{n,k-1}\\&-\boldsymbol{\mu}_{x,im}-\boldsymbol{\mu}_{h,k-1})))+(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})(\boldsymbol{\mu}_h-\boldsymbol{\mu}_{h,k-1})\\&+\boldsymbol{U}_{im}^{k-1}(\boldsymbol{\mu}_n-\boldsymbol{\mu}_{n,k-1})\end{aligned} \tag{9}$$

$$\boldsymbol{\Sigma}_{y,im}=(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\boldsymbol{\Sigma}_{x,im}(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}+\boldsymbol{U}_{im}^{k-1}\boldsymbol{\Sigma}_n(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}} \tag{10}$$

$$\begin{aligned}\boldsymbol{U}_{im}^{k-1}=&\boldsymbol{C}\mathrm{diag}\left(\frac{\exp(\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{n,k-1}-\boldsymbol{\mu}_{x,im}-\boldsymbol{\mu}_{h,k-1}))}{1+\exp(\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{n,k-1}-\boldsymbol{\mu}_{x,im}-\boldsymbol{\mu}_{h,k-1}))}\right)\\&\cdot\boldsymbol{C}^{-1}\end{aligned} \tag{11}$$

其中 $\boldsymbol{\mu}_{y,im}$，$\boldsymbol{\Sigma}_{y,im}$ 和 $\boldsymbol{\mu}_{x,im}$，$\boldsymbol{\Sigma}_{x,im}$ 分别是倒谱域含噪语音和纯净语音的均值向量、协方差矩阵；$\boldsymbol{\mu}_h$，$\boldsymbol{\mu}_n$ 和 $\boldsymbol{\Sigma}_n$ 分别是倒谱域卷积噪声的均值向量，加性噪声的均值向量和协方差矩阵；$\boldsymbol{\mu}_{h,k-1}$ 和 $\boldsymbol{\mu}_{n,k-1}$ 分别是 EM 算法中上一次(第 k -1 次)迭代得到的 $\boldsymbol{\mu}_h$ 和 $\boldsymbol{\mu}_n$ 值。倒谱域参数与对数谱域参数满足以下关系：$\boldsymbol{\mu}_{y,im}=\boldsymbol{C}\boldsymbol{u}_{y,im}$，$\boldsymbol{\Sigma}_{y,im}=\boldsymbol{C}\boldsymbol{\Sigma}_{y,im}\boldsymbol{C}^{\mathrm{T}}$，$\boldsymbol{\mu}_{x,im}=\boldsymbol{C}\boldsymbol{u}_{x,im}$，$\boldsymbol{\Sigma}_{x,im}=\boldsymbol{C}\boldsymbol{\Sigma}_{x,im}\boldsymbol{C}^{\mathrm{T}}$，$\boldsymbol{\mu}_h=\boldsymbol{C}\boldsymbol{u}_h$，$\boldsymbol{\mu}_n=\boldsymbol{C}\boldsymbol{u}_n$，$\boldsymbol{\Sigma}_n=\boldsymbol{C}\boldsymbol{\Sigma}_n\boldsymbol{C}^{\mathrm{T}}$。

3　基于 VTS 的模型自适应算法

3.1 均值估计

噪声参数 $\boldsymbol{\mu}_h$， $\boldsymbol{\mu}_n$ 和 $\boldsymbol{\Sigma}_n$ 通过最大似然准则计算。为求得似然函数的最大值，构建以下辅助函数：

$$Q(\overline{\lambda}\mid\lambda)=\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_{im}(t)\Big[(\boldsymbol{o}_t-\boldsymbol{\mu}_{y,im})^{\mathrm{T}}\cdot\boldsymbol{\Sigma}_{y,im}^{-1}(\boldsymbol{o}_t-\boldsymbol{\mu}_{y,im})-\ln\left|\boldsymbol{\Sigma}_{y,im}^{-1}\right|\Big] \tag{12}$$

$$\gamma_{im}(t)=P(\theta_t=i,k_t=m\mid\boldsymbol{O},\lambda) \tag{13}$$

其中 $\boldsymbol{O}$ 是自适应数据(MFCC 特征向量序列)， $\boldsymbol{o}_t$ 是 $\boldsymbol{O}$ 中第 t 帧 MFCC 特征向量；含噪语音模型参数 $\boldsymbol{\mu}_{y,im}$， $\boldsymbol{\Sigma}_{y,im}$ 与纯净语音模型参数 $\boldsymbol{\mu}_{x,im}$， $\boldsymbol{\Sigma}_{x,im}$ 的关系满足式(9)和式(10)。将式(9)表示成 $\boldsymbol{\mu}_h$ 和 $\boldsymbol{\mu}_n$ 的复合向量 $\overline{\boldsymbol{\mu}}$ 的形式：

$$\boldsymbol{\mu}_{y,im}=\boldsymbol{W}_{im}^{k-1}\overline{\boldsymbol{\mu}}+\boldsymbol{\varphi}_{im}^{k-1} \tag{14}$$

其中

$$\boldsymbol{W}_{im}^{k-1}=\left[\left(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1}\right),\boldsymbol{U}_{im}^{k-1}\right] \tag{15}$$

$$\overline{\boldsymbol{\mu}}=\left[\boldsymbol{\mu}_h^{\mathrm{T}},\boldsymbol{\mu}_n^{\mathrm{T}}\right]^{\mathrm{T}} \tag{16}$$

$$\boldsymbol{\varphi}_{im}^{k-1}=\boldsymbol{\mu}_{x,im}+\boldsymbol{\mu}_{h,k-1}-(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\boldsymbol{\mu}_{h,k-1}-\boldsymbol{U}_{im}^{k-1}\boldsymbol{\mu}_{n,k-1}+\boldsymbol{C}\ln(1+\exp(\boldsymbol{C}^{-1}(\boldsymbol{\mu}_{n,k-1}-\boldsymbol{\mu}_{x,im}-\boldsymbol{\mu}_{h,k-1}))) \tag{17}$$

在求 $\overline{\boldsymbol{\mu}}$ 时，用上次迭代得到的噪声方差 $\boldsymbol{\Sigma}_{n,k-1}$ 取代式(10)中的 $\boldsymbol{\Sigma}_n$，得到 $\boldsymbol{\Sigma}_{y,im}$ 的近似值 $\widetilde{\boldsymbol{\Sigma}}_{y,im}$：

$$\widetilde{\boldsymbol{\Sigma}}_{y,im}=(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\boldsymbol{\Sigma}_{x,im}(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}+\boldsymbol{U}_{im}^{k-1}\boldsymbol{\Sigma}_{n,k-1}(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}} \tag{18}$$

将式(14)和式(18)，代入式(12)，并令 $Q(\overline{\lambda}\mid\lambda)$ 对 $\overline{\boldsymbol{\mu}}$ 的导数等于零，得到以下方程：

$$\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_{im}(t)(\boldsymbol{W}_{im}^{k-1})^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_{y,im}^{-1}(\boldsymbol{o}_t-\boldsymbol{\varphi}_{im}^{k-1})=\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_{im}(t)(\boldsymbol{W}_{im}^{k-1})^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}_{y,im}^{-1}\boldsymbol{W}_{im}^{k-1}\overline{\boldsymbol{\mu}} \tag{19}$$

式(19)给出了估计噪声均值向量 $\overline{\boldsymbol{\mu}}$ 的一般形式。得到 $\boldsymbol{\mu}_h$ 和 $\boldsymbol{\mu}_n$ 后，即可根据式(9)更新 HMM 模型的均值 $\boldsymbol{\mu}_{y,im}$。

3.2 方差估计

由于辅助函数式(12)包含协方差矩阵的求逆运算及协方差矩阵行列式的对数运算，因此更新 HMM 的协方差矩阵比较困难，以往的文献都没有很好地解决这个问题。为了能将多个高斯单元的数据合并估计方差，作如下假设：

(1)HMM 的协方差矩阵为对角阵，即 $\boldsymbol{\Sigma}_{x,im}$ 和 $\boldsymbol{\Sigma}_{y,im}$ 为对角矩阵。在倒谱域，特征向量各维之间的相关性较小，为了提高识别速度，因此目前多数语音识别系统的协方差矩阵取对角阵。

(2)在倒谱域，噪声特征向量各维之间的相关性也较小，其协方差矩阵非对角元素值较小，因此 $\boldsymbol{\Sigma}_n$ 可以近似为对角阵。将 $\boldsymbol{\Sigma}_n$ 近似为对角阵是非常有必要的，如果 $\boldsymbol{\Sigma}_n$ 为满阵，不但计算十分复杂，而且参数太多，在少量数据时难以得到较为准确的估计值。

(3)多个高斯单元的数据合并估计方差时，每个高斯单元加权系数矩阵中的变量可以用上一次迭代得到的估计值替代，从而得到常加权矩阵。

根据上述假设(1)和(2)，式(10)可表示为

$$\boldsymbol{\Sigma}_{y,im}=\mathrm{diag}\left\{(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\boldsymbol{\Sigma}_{x,im}(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}+\boldsymbol{U}_{im}^{k-1}\boldsymbol{\Sigma}_n(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}\right\}=\mathrm{diag}\left\{\left((\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\cdot(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\right)\boldsymbol{\sigma}_{x,im}+\left(\boldsymbol{U}_{im}^{k-1}\bullet\boldsymbol{U}_{im}^{k-1}\right)\boldsymbol{\sigma}_n\right\} \tag{20}$$

其中 $\mathrm{diag}\{A\}$ 有两种含义，如果 A 是方阵，则表示取 A 的对角元素，生成对角矩阵；如果 A 是列向量，则将 A 作为对角元素，生成对角矩阵；符号 $\bullet$ 表示点乘，即矩阵的对应元素相乘； $\boldsymbol{\sigma}_{x,im}$ 是由 $\boldsymbol{\Sigma}_{x,im}$ 对角元素生成的列向量； $\boldsymbol{\sigma}_n$ 是由 $\boldsymbol{\Sigma}_n$ 对角元素生成的列向量。

将式(20)代入式(12)，并令 $Q(\overline{\lambda}\mid\lambda)$ 对 $\boldsymbol{\sigma}_n$ 的导数等于零，得到以下方程：

$$\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_{im}(t)\boldsymbol{G}_{im}\left[(\boldsymbol{o}_t-\boldsymbol{\mu}_{y,im})\cdot(\boldsymbol{o}_t-\boldsymbol{\mu}_{y,im})-\left((\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\cdot(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\right)\boldsymbol{\sigma}_{x,im}\right]=\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_{im}(t)\boldsymbol{G}_{im}\left(\boldsymbol{U}_{im}^{k-1}\cdot\boldsymbol{U}_{im}^{k-1}\right)\boldsymbol{\sigma}_n \tag{21}$$

$$\boldsymbol{G}_{im}=\left((\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}\cdot(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}\right)\left[\mathrm{diag}\left\{\left((\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\cdot(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\right)\boldsymbol{\sigma}_{x,im}+\left(\boldsymbol{U}_{im}^{k-1}\cdot\boldsymbol{U}_{im}^{k-1}\right)\boldsymbol{\sigma}_n\right\}\right]^{-2} \tag{22}$$

式(21)的具体推导过程略。由式(21)可知， $\boldsymbol{G}_{im}$ 在多个高斯单元合并估计方差时，起加权因子的作用。如果某个高斯单元语音能量较小，则 $\boldsymbol{G}_{im}$ 的元素较大；反之， $\boldsymbol{G}_{im}$ 的元素较小。但是， $\boldsymbol{G}_{im}$ 中含有未知变量 $\boldsymbol{\sigma}_n$，为了能将多个高斯单元的数据合并估计方差，用上一次迭代得到的噪声方差 $\boldsymbol{\sigma}_{n,k-1}$ 取代 $\boldsymbol{G}_{im}$ 中的 $\boldsymbol{\sigma}_n$，得到 $\boldsymbol{G}_{im}$ 的近似值：

$$\boldsymbol{G}_{im}\approx\left((\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}\cdot(\boldsymbol{U}_{im}^{k-1})^{\mathrm{T}}\right)\left[\mathrm{diag}\left\{\left((\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\cdot(\boldsymbol{I}-\boldsymbol{U}_{im}^{k-1})\right)\boldsymbol{\sigma}_{x,im}+\left(\boldsymbol{U}_{im}^{k-1}\cdot\boldsymbol{U}_{im}^{k-1}\right)\boldsymbol{\sigma}_{n,k-1}\right\}\right]^{-2} \tag{23}$$

式(21)和式(23)给出了估计噪声方差 $\boldsymbol{\sigma}_n$ 的一般形式。得到 $\boldsymbol{\sigma}_n$ 或 $\boldsymbol{\Sigma}_n$ 后，即可根据式(20)或式(10)更新 HMM 模型的方差 $\boldsymbol{\Sigma}_{y,im}$。

4　实验结果及分析

本文在一个非特定人汉语数字语音识别系统上

进行了一系列实验，验证本文算法的有效性。语音库由 30 个人，由数字 0～9 的发音组成，每个数字重复发音 10 次，共有 3000 个语音数据。其中一半用于训练，另一半在不同信噪比下与噪声混合，用于测试。噪音包括 NoiseX-92 中的 White 噪声，Factory 噪声和 Babble 噪声。语音识别系统的声学模型为高斯混合连续密度隐马尔可夫模型(每个模型的状态数和混合密度数分别为 6 和 4)，特征参数为 MFCC。

本文分别做了 MLLR 均值自适应算法(MLLR-Mean)[6]，MLLR 均值和方差自适应算法(MLLR-Model)[7]，VTS 特征补偿算法(VTS-FC)[1,2]，VTS 均值自适应算法(VTS-Mean)[10]及本文 VTS 模型自适应算法(VTS-Model)的识别实验。用于 VTS 特征补偿算法的 GMM 模型的高斯混合个数为 500。

图 1 是白噪声(SNR: 0 dB)环境下的似然函数收敛曲线，4 种模型自适应算法的收敛特性均较好，只需要 1 到 2 次迭代，即可收敛。在同等条件下，本文算法(VTS-Model)似然函数收敛后的值最大。VTS 特征补偿算法的收敛曲线没有标注在图 1 中，这是因为 VTS-FC 算法是基于 GMM 的，其似然函数值与 4 种模型自适应算法没有可比性。

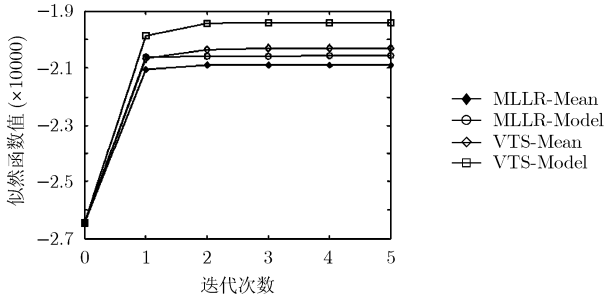

图 1 白噪声(SNR: 0 dB)环境下的似然函数收敛曲线

表 1 是白噪声(SNR: 0 dB)环境下不同数据量的误识率。用于噪声自适应时，MLLR 算法的性能比基于 VTS 的算法差很多。自适应数据较多时(5 个以上)，VTS-FC 算法与 VTS-Mean 算法的性能相近；但在自适应数据小于 5 个时，VTS-FC 算法的误识率较高。从表 1 可以看出基于 VTS 的模型自适应算法，包括 VTS-Mean 和 VTS-Model，受自适应数据数目的影响较小，自适应数据超过 5 个时，误识率基本不再变化。噪声不仅影响 HMM 的均值向量，也影响协方差矩阵，尤其在低信噪比时协方差矩阵受噪声的影响较大，因此本文 VTS-Model 算法取得了更高的识别率，性能优于 VTS-Mean 和其它算法。

表 1 白噪声(SNR: 0dB)环境下不同数据量的误识率(%)

噪声类型 (SNR)	算法	自适应数据个数				
		0	1	2	5	10
白噪声 (0 dB)	MLLR-Mean	82.7	71.4	58.6	56.0	51.2
	MLLR-Model	82.7	69.8	69.4	51.7	47.4
	VTS-FC	82.7	53.9	53.6	35.8	35.8
	VTS-Mean	82.7	38.9	38.2	35.8	36.4
	VTS-Model	82.7	28.7	25.6	23.3	22.6

图 2~图 4 分别是在 White 噪声，Factory 噪声和 Babble 噪声环境下，不同信噪比时的误识率。信噪比较高时(10 dB 以上)，3 种算法的误识率比较接近，VTS-FC 算法略好于其它两种算法。信噪比较低时(5 dB 以下)，VTS-FC 和 VTS-Mean 的性能也比较接近，在白噪声环境下，VTS-FC 优于 VTS-Mean；在 Factory 噪声和 Babble 噪声环境下，VTS-Mean 优于 VTS-FC。本文 VTS-Model 算法同时更新 HMM 模型的均值和方差，在较低信噪比时明显优于 VTS-FC 算法和 VTS-Mean 算法。由式(10)和式(11)可知，HMM 模型的协方差矩阵同时受噪声均值和噪声方差的影响；噪声越强，HMM 协方差矩阵的变化就越大。因此，信噪比越低，VTS-Model 算法的优势就越明显。比如，在−5 dB 时，相对于 VTS-Mean 算法，VTS-Model 算法在 White 噪声，Factory 噪声，Babble 噪声环境下的误识率分别从 56.3%，56.7%，49.8%下降到 35.8%，37.6%，41.2%，

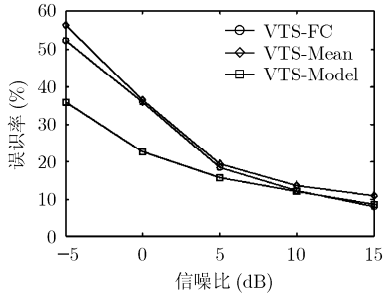

图 2 White 噪声不同信噪比下的误识率(10 个自适应数据)

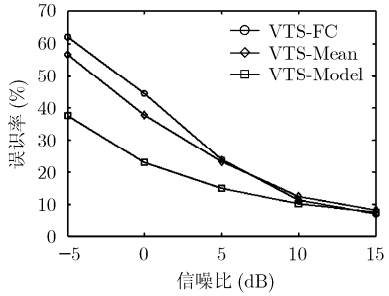

图 3 Factory 噪声不同信噪比下的误识率(10 个自适应数据)

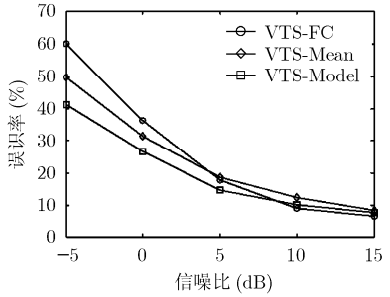

图 4 Babble 噪声不同信噪比下的误识率(10 个自适应数据)

识别率分别提高了 20.5%，19.1%和 8.6%。

综上可知，在各种环境下，基于 VTS 的模型自适应算法利用矢量泰勒级数展开式较好地克服了非线性问题，较为准确地估计了噪声环境下 HMM 的均值向量和协方差矩阵，性能优于 MLLR 算法、基于 VTS 的特征补偿算法和基于 VTS 的均值自适应算法。特别在低信噪比下，性能提高尤为明显。

5　结论

本文针对 MLLR 算法线性假设的限制，利用矢量泰勒级数展开式逼近含噪语音和纯净语音模型参数之间的非线性关系，在含噪语音中提取噪声的均值和方差信息，从而同时更新 HMM 模型的均值向量和协方差矩阵，取得了较好的效果。

参 考 文 献

[1] Moreno P J, Raj B, and Stern R M. A vector Taylor series approach for environment- independent speech recognition[C]. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, Georgia, USA, 7–10 May 1996: 733–736.

[2] Moreno P J. Speech recognition in noisy environments[D]. [Ph.D. dissertation], Carnegie Mellon University, 1996.

[3] Sasou A, Asano F, and Nakamura S, *et al*.. HMM-based noise-robust feature compensation[J]. *Speech Communication*, 2006, 48(9): 1100–1111.

[4] Kim W and Hansen J H L. Feature compensation in the cepstral domain employing model combination[J]. *Speech Communication*, 2009, 51(2): 83–96.

[5] Gauvain J L and Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains[J]. *IEEE Transactions on Speech and Audio Processing*, 1994, 2(2): 291–298.

[6] Leggetter C J and Woodland P C. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models[J]. *Computer Speech and Language*, 1995, 9(2): 171–185.

[7] Gales M J F and Woodland P C. Mean and variance adaptation within the MLLR framework[J]. *Computer Speech and Language*, 1996, 10(4): 249–264.

[8] Gales M J F and Young S J. Robust speech recognition in additive and convolutional noise using parallel model combination[J]. *Computer Speech and Language*, 1995, 9(4): 289–307.

[9] Kim D and Yook D. Linear spectral transformation for robust speech recognition using maximum mutual information[J]. *IEEE Signal Processing Letters*, 2007, 14(7): 496–499.

[10] Li J, Deng L, and Yu D, *et al*.. High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor series[C]. Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan, 9–13 December 2007: 65–70.

吕　勇：　男，1979 年生，博士生，研究方向为语音及听觉信号处理.

吴镇扬：　男，1949 年生，教授，博士生导师，研究方向为听觉及视觉信号处理.