基于声学特征空间非线性流形结构的语音识别声学模型

张文林 [1] 牛 铜 [1] 屈 丹 [1] 李弼程 [1] 裴喜龙 [1]

摘 要 从语音信号声学特征空间的非线性流形结构特点出发, 利用流形上的压缩感知原理, 构建新的语音识别声学模型. 将特征空间划分为多个局部区域, 对每个局部区域用一个低维的因子分析模型进行近似, 从而得到混合因子分析模型. 将上下文相关状态的观测矢量限定在该非线性低维流形结构上, 推导得到其观测概率模型. 最终, 每个状态由一个服从稀疏约束的权重矢量和若干个服从标准正态分布的低维局部因子矢量所决定. 文中给出了局部区域潜在维数的确定准则及模型参数的迭代估计算法. 基于 RM 语料库的连续语音识别实验表明, 相比于传统的高斯混合模型 (Gaussian mixture model, GMM) 和子空间高斯混合模型 (Subspace Gaussian mixture model, SGMM), 新声学模型在测试集上的平均词错误率 (Word error rate, WER) 分别相对下降了 33.1 % 和 9.2 %.

关键词 语音识别, 声学模型, 非线性流形, 混合因子分析

引用格式 张文林, 牛铜, 屈丹, 李弼程, 裴喜龙. 基于声学特征空间非线性流形结构的语音识别声学模型. 自动化学报, 2015, **41**(5): 1024−1033

DOI 10.16383/j.aas.2015.c140399

Feature Space Nonlinear Manifold Based Acoustic Model for Speech Recognition

ZHANG Wen-Lin[1] NIU Tong[1] QU Dan[1] LI Bi-Cheng[1] PEI Xi-Long[1]

Abstract Based on nonlinear manifold structure of acoustic feature space of speech signal, a new type of acoustic model for speech recognition is developed using compressive sensing. The feature space is divided into multiple local areas, with each area approximated by a low dimensional factor analysis model, so that in a mixture of factor analyzers is obtained. By restricting the observation vectors to be located on that nonlinear manifold, the probabilistic model of each context dependent state can be derived. Each state is determined by a sparse weight vector and several low-dimensional factors which follow standard Gaussian distributions. The principle for selection of the dimension for each local area is given, and iterated estimation methods for various model parameters are presented. Continuous speech recognition experiments on the RM corpus show that compared with the conventional Gaussian mixture model (GMM) and the subspace Gaussian mixture model (SGMM), the new acoustic model reduces the word error rate (WER) by 33.1 % and 9.2 % respectively.

Key words Speech recognition, acoustic model, nonlinear manifold, mixture of factor analyzers

Citation Zhang Wen-Lin, Niu Tong, Qu Dan, Li Bi-Cheng, Pei Xi-Long. Feature space nonlinear manifold based acoustic model for speech recognition. *Acta Automatica Sinica*, 2015, **41**(5): 1024−1033

在连续语音识别中, 为了反映同一音素在不同上下文环境中发音的不同, 通常采用上下文相关音素建模方法, 即对每一个音素的不同音位变体, 分别用一个隐马尔科夫模型 (Hidden Markov model, HMM) 进行建模, 其中每一个隐含状态的观测概率分布用高斯混合模型 (Gaussian mixture model, GMM) 或神经网络进行逼近. 这种上下文相关模型的参数数量庞大, 即使采用状态绑定等方法来减少状态个数, 典型的连续语音识别系统参数数量仍然在百万级以上. 为了训练得到一个性能良好的识别系统, 需要大量的训练数据, 而实际中训练数据往往是十分有限的. 因此, 为了减少模型对训练数据量的要求, 需要进一步降低模型的复杂度, 提高参数估计的稳健性.

针对传统的 “隐马尔科夫模型 – 高斯混合模型” 声学模型, 目前常用的解决方案有: 结构化协方差矩阵/精度矩阵建模方法[1], 即假设不同协方差矩阵或其精度矩阵由若干个低秩 (通常是秩为 1 的) 基矩阵的线性叠加得到, 各高斯混元通过某种方式共享一组相同的基矩阵; 本征三音子 (Eigentriphone) 建模方法[2−3], 将上下文相关状态进行聚类, 将每一类状态的均值矢量限定在一个线性子空间中, 通过估计子空间中的低维坐标矢量来重构状态的均值矢量, 从而得到更为精确的参数估计; 子空间高斯混合模型 (Subspace Gaussian mixture model, SGMM)[4], 将高斯混元的均值和权重限制在一个全局参数子空

收稿日期 2014-06-03 录用日期 2015-01-09
Manuscript received June 3, 2014; accepted January 9, 2015
国家自然科学基金 (61403415, 61175017) 资助
Supported by National Natural Science Foundation of China (61403415, 61175017)
本文责任编委 吴玺宏
Recommended by Associate Editor WU Xi-Hong
1. 解放军信息工程大学信息系统工程学院 郑州 450002
1. Institute of Information Systems Engineering, PLA Information Engineering University, Zhengzhou 450002

间中, 因此每一个状态可以用一个或若干个低维参数子空间中的矢量来表示, 从而提高模型参数估计的稳健性. 与传统的高斯混合模型相比, SGMM 声学模型大大压缩了模型尺寸, 并且可以利用集外数据对参数子空间进行估计, 因此特别适用于训练数据量受限条件下的语音识别[5−7]. 前述几种方法可以归结为一大类基于基展开 (Basis expand) 的声学建模方法. 近年来, 基于压缩感知与稀疏表达的方法受到众多学者的青睐, 已被成功应用于语音去噪、稳健性语音识别、声学模型正则化等方面. 2012 年, Saon 等[8] 将压缩感知技术直接应用于连续语音识别声学建模中, 将基表示方法与马尔科夫链相结合, 提出了一种贝叶斯感知隐马尔科夫模型 (Bayesian sensing-HMMs, BS-HMMs), 取得了不错的效果. BS-HMMs 的有效性可以归结为其在声学特征层次上应用压缩感知技术来建立状态模型, 并利用最大后验估计得到了稳健的模型参数. 然而, 与 SGMM 声学模型不同, 其各状态模型之间的参数估计是相互独立的, 需要训练多个状态相关字典, 因此对训练数据量的要求仍较高. 2013 年, Zhang 等[9−10] 提出稀疏精度矩阵建模方法, 即对协方差矩阵的逆矩阵直接施加稀疏约束, 从而间接减少模型参数数量.

上述基展开方法本质上都是寻找模型参数的线性子空间, 事实上, 众多研究表明语音信号存在一个低维的非线性流形结构[11−12], 因此采用线性子空间来对模型参数的相关性进行建模是不精确的. 混合因子分析 (Mixture of factor analyzers, MFA)[13−14] 利用多个低维线性因子分析模型的叠加来逼近流形上的数据分布, 是一种高维数据统计建模方法, 在高维数据可视化、流形学习、无监督聚类、特征降维等方面有着重要应用. 本文从声学特征空间的低维流形结构特点出发, 采用 MFA 来构建语音特征空间非线性流形结构的概率模型, 利用流形上的压缩感知原理[15] 得到声学模型中每个状态的观测概率模型, 从而给出一种基于非线性流形结构的声学建模方法. 该方法的模型假设更为合理, 具有直观的物理意义, 可以得到更为紧凑和稳健的声学模型.

1　基于 MFA 的非线性流形建模

本文所述的非线性流形是一种局部具有欧几里得空间性质的非线性空间, 是欧几里得空间中的曲线、曲面等概念的推广, 具有局部可坐标化的性质. 数学上直接对一个非线性流形进行建模是难以进行的, 然而从几何上看, 对于一个非线性的光滑曲面, 其局部可以用某一个采样点处的切面来近似. 同理, 对于一个非线性流形, 由于其局部具有欧氏空间的性质, 可以通过寻找各局部区域的坐标系来对其进行建模. 只要局部区域划分得足够细致, 就可以任意精度逼近原始的非线性流形. 混合因子分析模型充分利用非线性流形的这一性质, 采用线性因子分析模型来对每个局部区域中的数据分布进行建模, 是一种流形上数据的概率产生模型.

假设声学特征矢量维数为 D, 将其非线性流形划分为 I 个局部区域, 特征矢量 $\boldsymbol{x}$ 落入其中的概率分别为 $w_1, w_2, \cdots, w_I$, 其中 $w_i > 0$ 且 $\sum_{i=1}^{I} w_i = 1$. 对第 i 个局部区域内的特征矢量, 用一个潜在维数为 D_i $(0 < D_i \leq D)$ 的因子分析模型来描述其概率分布, 可得到特征矢量空间的混合因子分析模型, 其数学表达式为

$$p(\boldsymbol{x}|\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_I) = \sum_{i=1}^{I} w_i p(\boldsymbol{x}|\boldsymbol{y}_i) = \sum_{i=1}^{I} w_i \mathcal{N}(\boldsymbol{x}; M_i \boldsymbol{y}_i + \boldsymbol{\mu}_i, \Sigma_i) \tag{1}$$

其中, $\boldsymbol{\mu}_i$、M_i、Σ_i 分别表示第 i 个局部区域内观测矢量的均值矢量、因子负载矩阵 ($D \times D_i$ 维) 与误差矩阵 ($D \times D$ 维对角矩阵), $\boldsymbol{y}_i$ 表示特征矢量 $\boldsymbol{x}$ 在第 i 个局部区域内的潜在因子, 是一个服从标准正态先验分布的 D_i 维矢量, 即

$$p(\boldsymbol{y}_i) = \mathcal{N}(\boldsymbol{y}_i; \boldsymbol{0}, I) \tag{2}$$

图 1 从几何空间角度给出了用混合因子分析模型来对一个低维非线性流形进行建模的示意图.

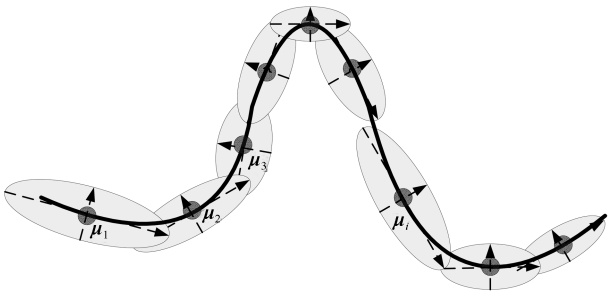


图 1　混合因子分析模型示意图

Fig. 1　Illustration of the mixture of factor analyzers

图 1 中, 曲线所示为一个高维空间中的非线性流形在某个二维平面上的投影. $\{\boldsymbol{\mu}_i\}_{i=1}^{I}$ 为流形上的若干个采样点, 在图中用实心圆点来表示. 椭圆区域表示以采样点为中心的某个局部区域. 权重 w_i 为特征矢量落入第 i 个局部区域的概率. 若采样点的数量足够, 则每一个局部区域可以用采样点处的切平面来进行近似. 对第 i 个局部区域, 以 $\boldsymbol{\mu}_i$ 为原点建立局部坐标系, M_i 的列矢量表示了其坐标轴的方向 (对应图 1 中椭圆的长短轴), $\boldsymbol{y}_i$ 是特征矢量 $\boldsymbol{x}$ 在这

个局部坐标系之下的局部坐标矢量. Σ_i 与流形在采样点 $\boldsymbol{\mu}_i$ 处的曲率有关, 曲率越大, 则用切平面来近似该局部区域的误差就越大.

根据混合因子分析模型来产生观测特征矢量 $\boldsymbol{x}$ 的过程可以归纳为以下步骤:

1) 根据离散概率分布 $\{w_1, w_2, \cdots, w_I\}$ 选择一个局部区域 i;

2) 根据标准正态分布产生局部坐标矢量 $\boldsymbol{y}_i$;

3) 根据 $p(\boldsymbol{n}_i) = \mathcal{N}(\boldsymbol{n}_i; \boldsymbol{0}, \Sigma_i)$ 产生重构误差 $\boldsymbol{n}_i$;

4) 得到观测特征矢量 $\boldsymbol{x} = M_i\boldsymbol{y}_i + \boldsymbol{\mu}_i + \boldsymbol{n}_i$.

事实上, 由式 (1) 和式 (2) 对局部坐标 $\boldsymbol{y}_i$ 进行积分, 可得到特征矢量 $\boldsymbol{x}$ 的边缘概率密度函数[13] 为

$$\begin{aligned} p(\boldsymbol{x}) = &\int_{\boldsymbol{y}_1} \cdots \int_{\boldsymbol{y}_I} p(\boldsymbol{x}|\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_I) \times \\ &\prod_{i=1}^{I} p(\boldsymbol{y}_i) \mathrm{d}\boldsymbol{y}_1 \cdots \mathrm{d}\boldsymbol{y}_I = \\ &\sum_{i=1}^{I} w_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, M_i M_i^{\mathrm{T}} + \Sigma_i) \end{aligned} \tag{3}$$

由式 (3) 可见, 混合因子分析模型本质上是一种退化的高斯混合模型. 给定局部区域个数 I 及各局部区域潜在维数 $\{D_i\}_{i=1}^{I}$, 可以利用所有的训练数据, 采用期望最大化 (Expectation maximization, EM) 算法[13] 估计得到声学特征矢量的混合因子分析模型参数 $\{w_i, \boldsymbol{\mu}_i, \Sigma_i\}_{i=1}^{I}$.

2 上下文相关状态建模

混合因子分析模型 (式 (1) 和式 (2)) 给出了声学特征矢量在特征空间中的一个先验分布概率模型. 对声学模型中的每一个上下文相关状态 j, 将其观测矢量限制在该非线性流形结构上, 令其落入第 i 个局部区域的概率为 w_{ji}, 即

$$p(i|j) = w_{ji} \tag{4}$$

其中, $w_{ji} \geq 0$ 且 $\sum_{i=1}^{I} w_{ji} = 1$.

假设状态 j 的观测矢量在第 i 个局部区域内服从高斯分布, 其均值为 $\boldsymbol{\mu}'_{ji}$, 方差为 Σ'_{ji}. 则给定局部区域 i 和均值矢量 $\boldsymbol{\mu}'_{ji}$, 状态 j 的观测概率为

$$p(\boldsymbol{x}|j, i, \boldsymbol{\mu}'_{ji}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}'_{ji}, \Sigma'_{ji}) \tag{5}$$

设 $\boldsymbol{\mu}'_{ji}$ 在第 i 个局部区域的局部坐标为 $\boldsymbol{y}_{ji}$, 则根据混合因子分析模型的假设式 (1), 给定局部区域 i 及对应坐标 $\boldsymbol{y}_{ji}$, $\boldsymbol{\mu}'_{ji}$ 的先验分布为

$$p(\boldsymbol{\mu}'_{ji}|i, \boldsymbol{y}_{ji}) = \mathcal{N}(\boldsymbol{\mu}'_{ji}; M_i\boldsymbol{y}_{ji} + \boldsymbol{\mu}_i, \Sigma_i) \tag{6}$$

根据贝叶斯公式, 给定局部区域 i 及 $\boldsymbol{y}_{ji}$, 状态 j 的观测概率可以写为

$$p(\boldsymbol{x}|j, i, \boldsymbol{y}_{ji}) = \int p(\boldsymbol{x}|j, i, \boldsymbol{\mu}'_{ji}) p(\boldsymbol{\mu}'_{ji}|i, \boldsymbol{y}_{ji}) \mathrm{d}\boldsymbol{\mu}'_{ji} \tag{7}$$

式 (5) 和式 (6) 构成了一种线性高斯模型, 将其代入式 (7), 整理可得[16]:

$$p(\boldsymbol{x}|j, i, \boldsymbol{y}_{ji}) = \mathcal{N}(\boldsymbol{x}; M_i\boldsymbol{y}_{ji} + \boldsymbol{\mu}_i, \Sigma_i + \Sigma'_{ji}) \tag{8}$$

最终可得到给定各局部区域坐标 $\{\boldsymbol{y}_{ji}\}_{i=1}^{I}$ 的条件下, 状态 j 的观测概率模型为

$$\begin{aligned} p(\boldsymbol{x}|j, \{\boldsymbol{y}_{ji}\}_{i=1}^{I}) = &\sum_{i=1}^{I} p(\boldsymbol{x}|j, i, \boldsymbol{y}_{ji}) p(i|j) = \\ &\sum_{i=1}^{I} w_{ji} \mathcal{N}(\boldsymbol{x}; M_i\boldsymbol{y}_{ji} + \boldsymbol{\mu}_i, \Sigma_i + \Sigma'_{ji}) \end{aligned} \tag{9}$$

为了减少模型参数, 假设不同状态 j 在同一个局部区域 i 内观测矢量分布的方差相同, 即 $\Sigma'_{ji} = \Sigma'_i$, 其中 Σ'_i 是与状态 j 无关的一个非对角矩阵. 令 $\Sigma''_i = \Sigma'_i + \Sigma_i$, 则状态 j 的观测概率模型可以写为

$$b_j(\boldsymbol{x}) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{ji}, \Sigma_{ji}) \tag{10}$$

$$\boldsymbol{\mu}_{ji} = M_i\boldsymbol{y}_{ji} + \boldsymbol{\mu}_i \tag{11}$$

$$\Sigma_{ji} = \Sigma''_i \tag{12}$$

其中,

$$p(\boldsymbol{y}_{ji}) = \mathcal{N}(\boldsymbol{y}_{ji}; \boldsymbol{0}, I) \tag{13}$$

在 HMM 声学模型中, 每个状态本质上是用来描述对应声学建模单元某个平稳段的观测概率分布. 由于每个声学建模单元有其独特的发音方式, 其平稳段对应的观测特征矢量必然分布于流形上的一个或多个局部区域, 不可能覆盖完整的流形结构. 定义权重矢量 $\boldsymbol{w}_j = [w_{j1}, w_{j2}, \cdots, w_{jI}]^{\mathrm{T}}$, 则 $\boldsymbol{w}_j$ 必然是稀疏的, 其大部分的分量为 0. 图 2 给出了状态 j 的观测数据的分布示意图. 其中, 状态 j 的观测数据分布于虚线所示矩形框内, 它只覆盖了声学特征空间流形结构的三个局部区域.

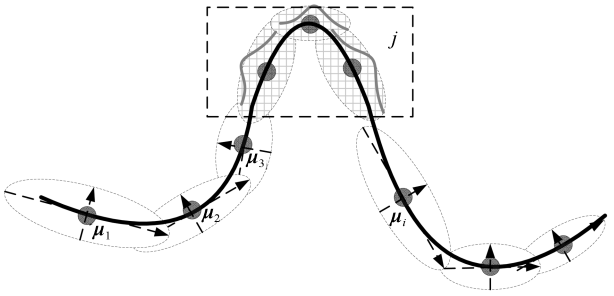


图 2　非线性流形结构上某状态的声学模型示意图

Fig. 2　The acoustic model of some state on the nonlinear manifold

因此, 可以对权重矢量 $\boldsymbol{w}_j$ 显式地引入稀疏约束. 假设最大允许的不为零的权重分量个数为 α $(0<\alpha\ll I)$, 则有:

$$\|\boldsymbol{w}_j\|_0 \le \alpha \tag{14}$$

其中, $\|\boldsymbol{w}_j\|_0$ 表示 $\boldsymbol{w}_j$ 的零范数.

式 (10)∼(14) 即构成了基于 MFA 的上下文相关状态模型.

3　MFA 声学模型的参数估计

在上一节提出的混合因子分析声学模型中, 参数 $\{\boldsymbol{\mu}_i, M_i, \Sigma_i''\}_{i=1}^{I}$ 是所有状态共享的全局参数, 与具体的状态无关. 对于状态 j, 需要估计的参数为权重矢量 $\boldsymbol{w}_j$ 及其非零分量对应局部区域内的坐标矢量 $\{\boldsymbol{y}_{ji}: i\in I_j\}$, 其中指标集 $I_j=\{i: w_{ji}>0\}$. 本节将给出该声学模型的训练过程及各参数的估计算法.

3.1　参数估计流程

为了得到各上下文相关模型, 在估计模型参数之前, 需采用传统的 HMM-GMM 声学建模方法训练得到一个初始声学模型, 具体方法可参见 KALDI 语音识别工具包[17]. 在得到传统 HMM-GMM 声学模型的上下文相关状态的决策树及训练数据的状态对齐信息后, 采用 EM 迭代算法得到模型参数的估计. 假设第 k 次迭代后, 模型参数 Λ 的取值为 $\Lambda^{(k)}$, 则 MFA 声学模型的参数训练流程如算法 1 所示.

算法 1.　混合因子分析声学模型训练流程

1) 选定局部区域个数 I 及每个局部区域的潜在因子维数 D_i;

2) 利用所有训练数据, 采用 EM 算法估计声学特征空间 MFA 模型参数 $\{\boldsymbol{\mu}_i^{(0)}, M_i^{(0)}, \Sigma_i^{(0)}\}_{i=1}^{I}$, 令 $\Sigma''^{(0)}_i=\Sigma_i^{(0)}$;

3) 初始化所有上下文相关模型参数: $w_{ji}^{(0)}=\frac{1}{I}$, $\boldsymbol{y}_{ji}^{(0)}=\boldsymbol{0}$, $1\le j\le J$, $1\le i\le I$, 选定迭代次数 K;

4) **for** $k=1$ to K

5)　**for** $j=1$ to J

6)　根据训练数据的状态对齐信息, 重估第 j 个状态的状态相关参数: $\boldsymbol{w}_j^{(k)}$, $\{\boldsymbol{y}_{ji}^{(k)}: w_{ji}^{(k)}>0\}$;

7)　**end for**

8)　**for** $i=1$ to I

9)　根据训练数据的状态对齐信息, 重估第 i 个局部区域的状态无关 参数: $\boldsymbol{\mu}_i^{(k)}, M_i^{(k)}, \Sigma''^{(k)}_i$;

10)　**end for**

11) **end for**

算法 1 中, J 表示声学模型中上下文相关状态的数量. 算法第 1 行选定声学特征空间中的局部区域个数 I 及每个局部区域的潜在因子维数 D_i, 本文采用一种启发式的方法, 详见第 3.2 节. 第 2 行采用 EM 算法训练得到声学特征空间的 MFA 模型, 第 3 行将所有状态的权重分量 $(\boldsymbol{w}_j)$ 均初始化为 $\frac{1}{I}$, 将局部区域潜在因子 $(\boldsymbol{y}_{ji})$ 初始化为 0. 第 4 行至第 11 行通过 K 次迭代训练得到 MFA 声学模型参数. 其中, 第 5 行至第 7 行估计模型的状态相关参数, 第 8 行至第 10 行估计模型的状态无关参数. 注意, 与混合因子分析模型的局部区域重构误差矩阵 Σ_i 不同, 这里 Σ_i'' 是状态共享的协方差矩阵, 在算法的第 1 行将其初始为 Σ_i, 是一个对角矩阵; 为了提高声学建模的精确度, 在算法的第 8 行重估过程中, 得到的是一个非对角矩阵.

在下文推导各参数的估计公式中, 假设 $\boldsymbol{x}_t$ 表示训练语料中的第 t 帧特征矢量, $\gamma_{ji}(t)$ 表示给定当前模型的参数, $\boldsymbol{x}_t$ 属于第 j 个状态第 i 个高斯混元的后验概率, 定义 $\gamma_{ji}=\sum_t\gamma_{ji}(t)$, $s_{ji}=\sum_t\gamma_{ji}(t)\boldsymbol{x}_t$.

3.2　局部区域个数及潜在因子维数确定

从理论上看, 声学特征空间局部区域个数 I 越大, 其非线性流形结构的划分越细致, 混合因子分析模型的逼近越精确. 然而, 由于训练数据是有限的, 随着局部区域个数的增大, 平均落入每个区域的训练样本将会减少, 从而导致模型参数无法得到稳健的估计. 因此, 实际应用中, 应通过实验来选择合适的局部区域个数 I, 其典型值为 400 ∼ 1 000.

由式 (3) 可见, 混合因子分析模型本质上是一种退化的高斯混合模型. 为了确定每个局部区域的潜在因子维数, 可以借鉴主分量分析 (Principal component analysis, PCA) 的思想, 首先采用训练集的所有特征矢量训练一个含有 I 个高斯混元的、协方差矩阵为满阵的高斯混合模型, 通常称其为统一背景模型 (Universal background model, UBM), 对其协方差矩阵进行特征值分析可以得到其某个低秩近似, 该近似的秩即为潜在因子维数.

假设第 i 个高斯混元的协方差矩阵为 $\tilde{\Sigma}_i$, 将其特征值从大到小依次排列为 $\lambda_{i1}, \lambda_{i2}, \cdots, \lambda_{iD}$, 定义

前 d 个特征值的累计贡献率 (Cumulative contribution rate, CCR) η_{id} 为

$$\eta_{id} = \frac{\sum\limits_{d'=1}^{d} \lambda_{id'}}{\sum\limits_{d''=1}^{D} \lambda_{id''}} \tag{15}$$

对于第 i 个高斯混元, 计算达到 90 % 累计贡献率的最少特征值个数 D_i 为

$$D_i = \min_{d}\{d : \eta_{id} > 90\,\%\} \tag{16}$$

D_i 即为对属于第 i 个高斯混元的特征矢量进行主分量分析后, 保留 90 % 方差所需要的最少主分量 (即特征矢量) 个数, 本文将其取为第 i 个局部区域的潜在因子维数.

在利用 EM 算法训练 MFA 模型时, 采用概率主分量分析 (Probabilistic principal component analysis, PPCA) 对 MFA 的每个局部因子分析模型进行初始化. 即取其局部区域权重为 UBM 中对应高斯混元的权重, 均值矢量 $\boldsymbol{\mu}_i$ 为 UBM 中对应高斯混元的均值矢量, 因子负载矩阵 M_i 的列为 UBM 的协方差矩阵 $\tilde{\Sigma}_i$ 的前 D_i 个特征矢量, 误差矩阵为 $\Sigma_i = \sigma_i I$, 其中 $\sigma_i = \frac{1}{D-D_i}\sum_{i=D_i+1}^{D}\lambda_i$.

3.3 状态相关参数估计

3.3.1 状态相关权重矢量估计

对于状态 j 的权重矢量 $\boldsymbol{w}_j$, 与一般的 HMM-GMM 声学模型类似, 在第 k 次迭代时, 其 EM 算法的辅助函数为

$$Q(\boldsymbol{w}^{(k)}) = \sum_i \gamma_{ji} \log w_{ji}^{(k)} \tag{17}$$

根据模型假设, $\boldsymbol{w}_j^{(k)}$ 必须同时满足以下约束条件:

$$\|\boldsymbol{w}_j^{(k)}\|_1 = 1 \tag{18}$$

$$\|\boldsymbol{w}_j^{(k)}\|_0 < 2 \tag{19}$$

其中, 式 (18) 表示概率分布约束, 式 (19) 表示稀疏约束.

这种带有 l_0 约束的优化问题 (式 (17) 和式 (19)) 难以直接求解. 在压缩感知及正则化理论中, 通常用 l_1 范数作为 l_0 范数的一个凸近似, 从而利用凸优化方法来进行求解. 但是这里权重矢量的 l_1 范数已在约束条件 (式 (18)) 中出现, 因此传统的思路在这里是行不通的.

受压缩感知中迭代收缩算法 (Iterative-shrinkage)[18] 的启发, 本文采用一种启发式的 "权重收缩 (Weight shrinkage)" 求解算法. 其基本思路是在每次迭代中, 先不考虑稀疏约束式 (18), 利用 Langrange 乘子法直接求解得到权重矢量的最大似然估计 $\hat{w}_{ji}^{(k)} = \frac{\gamma_{ji}}{\sum_i \gamma_{ji}}$. 若其非零分量的个数大于 α, 则将其中小于某个门限 β (典型值为 10^{-5}) 的权重分量置为 0, 即

$$\tilde{w}_{ji}^{(k)} = \left(\hat{w}_{ji}^{(k)} - \beta\right)_+ \tag{20}$$

其中, $(x)_+ = \max\{x, 0\}$. 最后根据概率分布约束式 (18) 将权重矢量重新归一化, 得到其估计值:

$$w_{ji}^{(k)} = \frac{\tilde{w}_{ji}^{(k)}}{\sum\limits_i \tilde{w}_{ji}^{(k)}} \tag{21}$$

3.3.2 状态相关因子估计

对于状态 j, 其局部因子矢量 $\boldsymbol{y}_{ji}$ 服从标准正态先验分布, 因此可以采用最大后验准则得到其最大后验估计, 其 EM 算法辅助函数为

$$Q(\boldsymbol{y}_{ji}^{(k)}) = \sum_t \gamma_{ji}(t)\left[\log p(\boldsymbol{x}_t|j,i,\boldsymbol{y}_{ji}^{(k)}) + \log p(\boldsymbol{y}_{ji}^{(k)})\right] \tag{22}$$

将式 (2) 和式 (7) 代入式 (22), 整理可得:

$$Q(\boldsymbol{y}_{ji}^{(k)}) = -\frac{1}{2}\boldsymbol{y}_{ji}^{(k)\mathrm{T}} H_{ji} \boldsymbol{y}_{ji}^{(k)} + \boldsymbol{y}_{ji}^{(k)\mathrm{T}} \boldsymbol{g}_{ji} + \text{const} \tag{23}$$

其中, const 表示与 $\boldsymbol{y}_{ji}^{(k)}$ 无关的常数项.

$$H_{ji} = \sum_t \gamma_{ji}(t)\left(M_i^{\mathrm{T}}\Sigma_i^{-1}M_i + I\right) \tag{24}$$

$$\boldsymbol{g}_{ji} = \sum_t \gamma_{ji}(t) M_i^{\mathrm{T}}\Sigma_i^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_i) \tag{25}$$

令式 (23) 对 $\boldsymbol{y}_{ji}^{(k)}$ 的导数为零, 整理可得:

$$\boldsymbol{y}_{ji}^{(k)} = H_{ji}^{-1}\boldsymbol{g}_{ji} \tag{26}$$

3.4 状态无关参数估计

对于状态无关参数 $\boldsymbol{\mu}_i$, M_i 和 Σ_i, 第 k 次迭代时, 其 EM 算法辅助函数均为

$$Q(\boldsymbol{\mu}_i^{(k)}, M_i^{(k)}, \Sigma_i^{(k)}) = \sum_t \sum_j \gamma_{ji}(t) \log p(\boldsymbol{x}_t|j,i,\boldsymbol{y}_{ji}^{(k)}) \tag{27}$$

将式 (7) 代入式 (27), 分别对各参数求偏导, 并

令导数为 0, 求解可得:

$$\boldsymbol{\mu}_i^{(k)} = \frac{\sum\limits_j \left(s_{ji} - \gamma_{ji} M_i \boldsymbol{y}_{ji}\right)}{\sum\limits_j \gamma_{ji}} \tag{28}$$

$$M_i^{(k)} = \left[\sum_t \sum_j \gamma_{ji}(t)\left(\boldsymbol{x}_t - \boldsymbol{\mu}_i\right)\boldsymbol{y}_{ji}^{\mathrm{T}}\right]^{-1} \times \left(\sum_t \sum_j \gamma_{ji}(t)\boldsymbol{y}_{ji}\boldsymbol{y}_{ji}^{\mathrm{T}}\right) \tag{29}$$

$$\Sigma_i^{(k)} = \frac{\sum\limits_t \sum\limits_j \gamma_{ji}(t)\left(\boldsymbol{x}(t) - \boldsymbol{\mu}_{ji}\right)\left(\boldsymbol{x}(t) - \boldsymbol{\mu}_{ji}\right)^{\mathrm{T}}}{\sum\limits_j \gamma_{ji}} \tag{30}$$

注意, 在式 (24)、(25) 和式 (28)∼(30) 中, 等号右边的表达式中出现的状态相关参数和状态无关参数均取为对应参数的当前值.

4　与现有声学模型的比较

MFA 声学模型本质上是一种特殊的 HMM-GMM 声学建模方法, 与现有方法在某些方面具有一定的相似性, 但其假设条件更为合理, 所得到的声学模型更为精确, 由于合理使用了特征空间的结构性信息, 参数估计更为稳健.

4.1　与半连续隐马尔科夫声学模型的比较

在半连续隐马尔科夫模型 (Semi-continuous HMM, SC-HMM)[19] 中, 各状态的高斯混合模型共享相同一组的高斯混元, 只是不同状态具有不同的权重矢量. 在本文 MFA 声学模型中, 混合因子分析模型可以视为含有 I 个高斯混元的字典. 由于权重矢量 ($\boldsymbol{w}_j$) 的稀疏性约束, 各状态模型可视作是从字典中选取若干个高斯混元来进行构建的. 各状态模型不仅具有不同的稀疏权重矢量 ($\boldsymbol{w}_j$), 其均值矢量 (式 (11)) 也是互不相同的, 是通过最大后验估计得到的. 因此, 本文方法比 SC-HMM 具有更精确的声学建模能力.

4.2　与子空间高斯混合模型 (SGMM) 的比较

在 SGMM 声学模型中, 每一个上下文相关状态 j 的观测概率模型可以用下式表示:

$$b_j(\boldsymbol{x}) = \sum_{i=1}^{I} w_{ji}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{ji}, \Sigma_{ji}) \tag{31}$$

$$\boldsymbol{\mu}_{ji} = M_i \boldsymbol{v}_j \tag{32}$$

$$\Sigma_{ji} = \Sigma_i \tag{33}$$

$$w_{ji} = \frac{\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)}{\sum\limits_{i'=1}^{I}\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_j)} \tag{34}$$

对比式 (10)∼(14) 与式 (31)∼(34), 可见 MFA 声学模型与 SGMM 声学模型的区别如下:

1) 在 SGMM 声学模型中, 每个状态 j 均由一个坐标矢量 $\boldsymbol{v}_j$ 决定, 所有高斯混元的均值矢量 $\boldsymbol{\mu}_{ji}$ 及状态权重的对数值 $\log w_{ji}$ 共享相同的坐标, 因此其前提假设是存在一个模型参数的全局线性子空间. 这一点与语音特征空间的非线性流形结构特点是不相吻合的. 在 SGMM 模型中, 可以通过引入 "子状态" 来提高模型表达能力, 得到如下模型:

$$b_j(\boldsymbol{x}) = \sum_{m=1}^{M_j}\sum_{i=1}^{I} w_{jmi}\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{jmi}, \Sigma_{jmi}) \tag{35}$$

$$\boldsymbol{\mu}_{jmi} = M_i \boldsymbol{v}_{jm} \tag{36}$$

$$\Sigma_{jmi} = \Sigma_i \tag{37}$$

$$w_{jmi} = \frac{\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_{jm})}{\sum\limits_{i'=1}^{I}\exp(\boldsymbol{w}_i^{\mathrm{T}}\boldsymbol{v}_{jm})} \tag{38}$$

其中, M_j 表示状态 j 的子状态个数, m 表示子状态的序号, 第 j 个状态的第 m 个子状态由一个坐标矢量 $\boldsymbol{v}_{jm}$ 决定. 在上述基于子状态的 SGMM 模型中, 对于每个子状态来说, 参数全局线性子空间的本质没有改变. 在 MFA 声学模型中, 每个状态的高斯混元具有不同的局部坐标矢量 $\boldsymbol{y}_{ji}$, 体现了声学特征空间的非线性结构特点. 因此, MFA 声学模型的假设条件更为合理, 模型参数的物理意义也更为明显.

2) 高斯混元的权重 w_{ji} 本质上是一个概率值, 而均值矢量 $\boldsymbol{\mu}_{ji}$ 是特征空间中的一个点, 两者具有完全不同的物理意义. 因此在 SGMM 中, 两者共享相同的坐标矢量 $\boldsymbol{v}_j$ 是不合理的. 且权重矢量与均值矢量的相互耦合造成参数估计过程异常复杂, 只能通过近似算法来进行求解[4]. 在 MFA 声学模型中, 由于假设每个状态的观测矢量只能落入声学特征空间的若干个局部区域, 其权重矢量 $\boldsymbol{w}_j$ 服从稀疏约束, 一方面, 这种假设与声学空间非线性流形结构特点相吻合; 另一方面, 由于权重矢量的估计与均值矢量相互独立, 使得二者的估计算法都更为简单.

3) 在 SGMM 中, 每个状态均含有 I 个高斯混元, 因此, 在识别过程中, 状态观测概率的计算量十

分巨大, 需要在解码过程中引入 “高斯预选 (Gaussian pre-selection)” 机制[4] 来减少参与计算的高斯混元数量. 在 MFA 声学模型中, 由于权重矢量的稀疏性, 每个状态均仅有少量的高斯混元参与状态观测概率的计算, 使得解码过程大大简化.

4.3 与贝叶斯压缩感知隐马尔科夫模型 (BS-HMM) 的比较

在 BS-HMM[8] 中, 对每个上下文相关状态 j 都要建立一个特征矢量字典 Φ_j, 通过压缩感知原理得到状态模型, 各状态模型之间的参数是相互独立的. 在 MFA 声学模型中, 通过混合因子分析 (MFA) 对整个声学特征空间建立了非线性流形模型, 可以将其视为一个模型字典, 其中特征空间每个局部区域的因子分析模型为一个字典项. 因此, 状态模型的建立过程可以认为是一种基于非线性流形结构的压缩感知过程[15]. 相比于 BS-HMM, 在 MFA 声学模型中, 所有状态共享相同的非线性流形结构, 有效利用了状态之间的相关性信息, 因此需要的训练数据量更少, 模型结构更为紧致, 参数估计也更为稳健.

5 实验结果及分析

为了验证混合因子分析声学模型的性能, 针对国际上常用的 Resource Management (RM) 语料库[1]进行了连续语音识别实验. RM 语料库是一个国际上广泛采用的小词汇量连续语音识别语料库, 其中包含约 1 000 个英语词汇, 由 TI 公司在安静环境下用头戴式噪声消除麦克风录制. 该语料库由说话人无关 (Speaker independent, SI) 语料和说话人相关 (Speaker dependent, SD) 语料两部分组成, 本文实验针对其中的说话人无关语料进行. 训练集由 78 个男性说话人和 31 个女性说话人组成, 共 109 个说话人, 总时长约 200 分钟, 通常将其标识为 SI-109. 测试集分为 6 个部分, 各录制于 1987 年 3 月, 1987 年 10 月, 1989 年 2 月, 1989 年 10 月, 1991 年 2 月和 1992 年 9 月, 分别标识为 Mar87, Oct87, Feb89, Oct89, Feb91 和 Sep92, 共有 60 个说话人, 约 80 分钟的语料.

实验中, 利用开源的 Kaldi 语音识别工具箱[17] 搭建了三套声学模型: 传统 HMM-GMM 声学模型、SGMM 声学模型与 MFA 声学模型. 声学特征矢量均采用 13 维的美尔频率倒谱系数 (Mel-frequency cepstrum coefficients, MFCC) 及其一阶和二阶差分, 总的特征维数 (D) 为 39 维. 对每一个说话人的所有语音数据, 采用倒谱均值方差归一化 (Cepstrum mean and variance normalization, CMVN) 对其特征矢量序列进行预处理. 采用上下文相关三音子 (Triphone) 作为声学建模单元. 利用 Kaldi 工具箱自动生成的问题集进行三音子状态聚类, 最终三种声学模型中均含有 2 011 个不同的绑定状态 (Tied states).

1) 训练阶段. 首先在训练集上训练得到传统的 HMM-GMM 声学模型, 总的高斯混元数为 17 233. 然后, 采用 Kaldi 中的高斯混元聚类算法, 对该声学模型中的高斯混元进行聚类, 生成含有 $I = 400$ 个高斯混元的统一背景模型 (UBM), 其每一个高斯混元的协方差矩阵均为 39×39 维的满阵. 该 UBM 一方面将用于对 SGMM 声学模型进行初始化, 另一方面还用于对声学特征空间的混合因子分析模型进行初始化. 将 SGMM 声学模型的音子矢量维数设置为 40, 采用 Kaldi 工具箱中的 “s5” 脚本训练含有子状态的扩展 SGMM 声学模型 (式 (35) ∼ (38)), 最终声学模型中总的子状态个数为 7 495. 最后, 采用 EM 算法训练声学特征空间中的混合因子分析模型, 根据第 3.1 节中的算法 1 迭代 15 次, 得到 MFA 声学模型的上下文相关状态模型.

2) 测试阶段. 采用 RM 语料库自带的词对语法 (Word pair grammar) 文件构建二元语法模型, 采用 Kaldi 中基于加权有限状态机 (Weighted finite state transducer, WFST) 的解码器构建静态解码网络对测试集进行解码识别. 最后, 统计测试集上的平均词错误率 (Word error rate, WER), 作为连续语音识别评价指标. 在比较不同声学模型的性能时, 为了减少测试语料随机性的影响, 采用 NIST 公布的开源工具包 SCTK 进行显著性水平测试 (Significance test), 以检验三种声学模型识别结果之间的差异在统计上是否显著.

5.1 低维非线性流形的存在性及其潜在维数

对高斯混元聚类算法得到的 UBM 的 400 个高斯混元, 采用第 3.2 节中的 “90 %” 累计贡献率准则, 分别计算其对应的局部区域潜在因子维数 D_i, 其分布的直方图如图 3 所示.

由图 3 可见, 大部分高斯混元的 D_i 值在 $10 \sim 14$ 之间, 其平均值为 11.8. 这意味着 UBM 中大部分高斯混元的协方差矩阵都可以近似认为是退化的, 平均只需要 11.8 个特征矢量即可对其进行近似. 这就证明了声学特征空间中低维非线流形结构的存在性.

5.2 权重迭代收缩算法实验结果

MFA 声学模型与 SGMM 声学模型的另一个重要区别在于, 在 SGMM 声学模型中, 各子状态的高斯混元数是相同的, 等于 UBM 中的高斯混元数;

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S3A

而在 MFA 声学模型中, 权重矢量服从稀疏约束, 导致各状态的高斯混元数各不相同, 且远远小于局部区域的个数 (即 UBM 中高斯混元数). 本文实验中, 训练得到的 SGMM 声学模型中共有 7 495 个子状态, 每个子状态均有 400 个高斯混元, 因此 SGMM 声学模型中实际共有 2 998 000 个高斯混元. 而对于 MFA 声学模型, 在初始化阶段, 每个状态的高斯混元数都等于 400, 其后采用第 3.1 节中给出的权重迭代收缩算法进行权重训练, 迭代过程中将每个状态高斯混元数的上限设置为 $\alpha = 10$, 门限 β 取为 10^{-5}. 图 4 给出了迭代 30 次过程中, 状态的平均混元数的变化曲线.

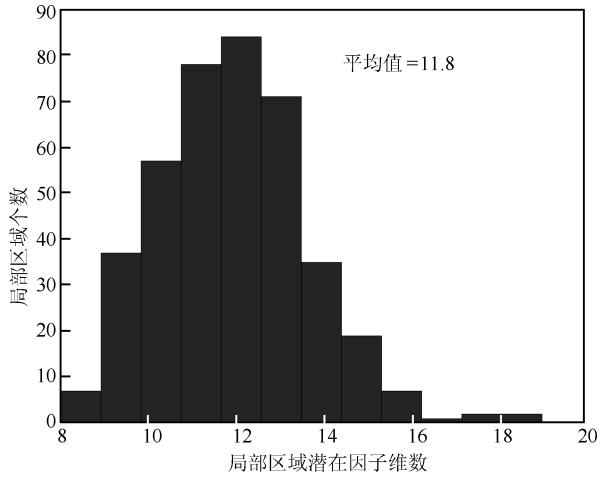


图 3　局部区域潜在因子维数分布直方图

Fig. 3　The histogram of the latent dimensions of different local area

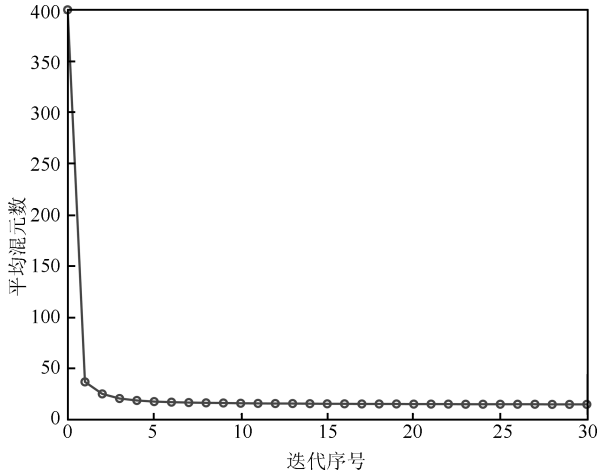


图 4　状态平均混元数随迭代次数的变化

Fig. 4　The average count of Gaussian components changing with the iteration number

由图 4 可见, 权重矢量迭代收缩算法可以有效地减少每个状态的高斯混元数, 算法快速达到收敛. 在迭代 10 次之后, 平均混元数量基本保持不变. 图 5 给出了 30 次迭代之后, 状态的高斯混元数量 (即权重矢量的 l_0 范数) 分布直方图.

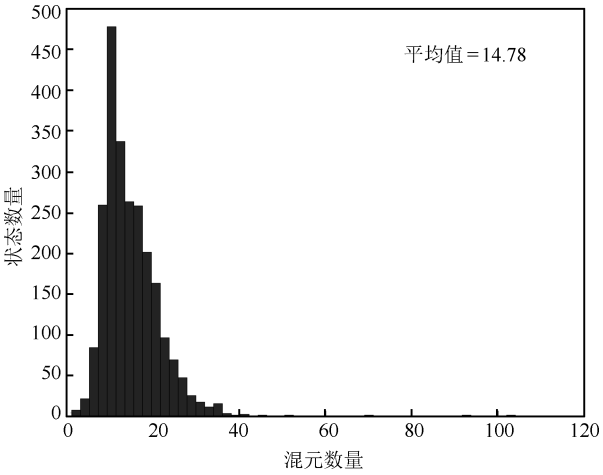


图 5　迭代 30 次后状态的高斯混元数量分布直方图

Fig. 5　The histogram of the number of Gaussian components after 30 iterations

由图 5 可见, 大部分的状态只有约 10 个高斯混元, 高斯混元数的平均值为 14.78, MFA 声学模型中总的高斯混元数为 29 723. 这一数量远远小于 SGMM 声学模型的高斯混元数 (2 998 000), 却大于传统的 HMM-GMM 声学模型的高斯混元数 (17 233). 因此, MFA 声学模型比 SGMM 声学模型更为紧致, 解码识别中不需要 “高斯预选择过程”, 解码器的构造更为简单. 由于不同状态的高斯混元数是根据训练数据自动确定的, 不需要预先设定或通过实验调整, 因此所得到声学模型更为精确.

5.3　不同声学模型参数数量比较

实验中, 传统的 HMM-GMM 声学模型共有 17 233 个高斯混元, 每个高斯混元需要一个权重参数、一个 39 维的均值矢量和一个 39×39 维的对角协方差矩阵, 因此共需 $17\,233 \times (1 + 39 + 39) = 1\,361\,407$ 个状态相关参数. 在 HMM-GMM 声学模型中, 没有状态无关参数.

在 SGMM 声学模型中, 共有 7 495 个子状态, 每个子状态需要一个 40 维的音子矢量, 共需 299 800 个状态相关参数. 对于子空间参数, 分别有 400 个音子矩阵 M_i (维数为 39×40) 和协方差矩阵 Σ_i (维数 39×39), 因此共需 $400 \times (39 \times 40 + 39 \times 39) = 1\,232\,400$ 个状态无关参数.

而在 MFA 声学模型中, 共有 29 723 个高斯混元, 每个高斯混元需要一个权重参数 w_{ji} 及一个潜在因子矢量 $\boldsymbol{y}_{ji}$, 统计得到总的状态相关参数数量为 $29\,723 + 285\,368 = 315\,091$. 对于全局 MFA 模型, 共有 400 个均值矢量 (维数为 39 维)、

局部因子负载矩阵 M_i (维数为 $39 \times D_i$, D_i 的平均值为 11.8) 和协方差矩阵 Σ_i (39×39), 共需 $400 \times (39 + 39 \times 11.8 + 39 \times 39) = 808\,080$ 个状态无关参数.

由上述分析可见, 与 HMM-GMM 声学模型相比, 在 SGMM 和 MFA 声学模型中, 大部分的模型参数是状态无关的, 而状态相关参数相对较少, 因此可以得到更稳健的声学模型. 而 MFA 声学模型的状态无关参数远少于 SGMM 声学模型的状态无关参数, 因此在相同训练数据量下, 前者的参数估计将更为稳健.

5.4 不同声学模型识别性能比较

表 1 给出了利用三种声学模型对测试集数据进行解码识别后的平均词错误率 (WER).

表 1 三种声学模型在测试集上的 WER (%)
Table 1 WERs of the three acoustic models on the test set (%)

声学模型	WER
HMM-GMM	3.26
SGMM	2.40
MFA	2.18

由表 1 中结果可见, MFA 声学模型的词错误率更低, 相比于传统 HMM-GMM 声学模型和 SGMM 声学模型, 平均词错误率分别相对下降了 33.1 % 和 9.2 %.

由于三种声学模型的平均词错误率比较接近, 为了得出更为可靠的比较结论, 利用 NIST 提供的 SCTK 工具包对其进行了三种显著性测试: 成对句子分段词错误 (Matched pair sentence segment word error) 测试 (简称 MP 测试)、符号成对比较说话人词准确率 (Signed paired comparison speaker word accuracy) 测试 (简称 SI 测试) 和 Wilcoxin 符号秩说话人词准确率 (Wilcoxin signed Rank speaker word accuracy) 测试 (简称 WI 测试). 三种显著性测试结果均表明, 在 5 % 的显著性水平之下, 三种声学模型的识别结果存在显著差异. 因此, SGMM 声学模型优于传统的 HMM-GMM 声学模型, 而 MFA 声学模型优于 SGMM 声学模型.

6 结论

本文利用多个局部线性的因子分析模型对声学特征空间的非线性流形结构进行逼近, 得到特征矢量基于混合因子分析的先验概率模型. 进而利用流形上的压缩感知原理, 建立语音识别系统中上下文相关状态的观测概率模型. 由于各状态共享相同的流形结构, 大大减少了模型参数; 借助压缩感知和贝叶斯原理, 提高了参数估计的稳健性. 文中给出了声学特征空间中 MFA 模型的训练算法, 并详细推导了各状态参数的最大似然估计公式. 实验表明, 相比于传统的 HMM-GMM 声学模型和 SGMM 声学模型, 在训练数据量有限的条件下, 该声学模型的模型参数估计更为稳健, 识别性能更好. 近年来, 基于深度神经网络 (Deep neural network, DNN) 的声学模型取得了很好的效果. 在 DNN 中, 每一层神经网络的输出均可看作是对其输入特征的一次抽象, 因此 DNN 的良好的性能可归结为其强大的特征提取能力. 作为本文的进一步研究方向, 可以考虑将 DNN 用于高层特征提取, 在其基础上构建 MFA 声学模型, 从而将二者的优点相结合.

References

1 Olsen P A, Gopinath R A. Modeling inverse covariance matrices by basis expansion. *IEEE Transactions on Speech and Audio Processing*, 2004, **12**(1): 37−46

2 Ko T, Mak B. Eigentriphones for context-dependent acoustic modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(6): 1285−1294

3 Ko T, Mak B. Eigentrigraphemes for under-resourced languages. *Speech Communication*, 2014, **56**: 132−141

4 Povey D, Burget L, Agarwal M, Akyazi P, Kai F, Ghoshal A, Glembek O, Goel N, Karafiàt M, Rastrow A, Rose R C, Schwarz P, Thomas S. The subspace Gaussian mixture model — a structured model for speech recognition. *Computer Speech & Language*, 2011, **25**(2): 404−439

5 Qi J, Wang D, Tejedor J. Subspace models for bottleneck features. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013. 1746−1750

6 Motlìcek P, Imseng D, Garner P N. Crosslingual tandem-SGMM: exploiting out-of-language data for acoustic model and feature level adaptation. In: Proceedings of the 14th Annual Conference of the International Speech Communication Association. Lyon, France: ISCA, 2013. 510−514

7 Lu L, Ghoshal A, Renals S. Cross-lingual subspace Gaussian mixture models for low-resource speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(1): 17−27

8 Saon G, Chien J T. Bayesian sensing hidden Markov models. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, **20**(1): 43−54

9 Zhang W B, Fung P. Sparse inverse covariance matrices for low resource speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2013, **21**(3): 659−668

10 Zhang W B, Fung P. Discriminatively trained sparse inverse covariance matrices for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, **22**(5): 873−882

11 Jansen A, Niyogi P. Intrinsic Fourier analysis on the manifold of speech sounds. In: Proceedings of the 2006 International Conference on Acoustics, Speech, and Signal Processing. Toulouse: IEEE, 2006. **1**: 241−244

12 Lu X G, Dang J W. Vowel production manifold: intrinsic factor analysis of vowel articulation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, **18**(5): 1053−1062

13 Ghahramani Z, Hinton G. The EM Algorithm for Mixtures of Factor Analyzers, Technical Report CRG-TR-96−1, Department of Computer Science, University of Toronto, Toronto, Canada, 1996.

14 Carin L, Baraniuk R G, Cevher V, Dunson D, Jordan M I, Sapiro G, Wakin M B. Learning low-dimensional signal models. *IEEE Signal Processing Magazine*, 2011, **28**(2): 39−51

15 Chen M H, Silva J, Paisley J, Wang C P, Dunson D, Carin L. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: algorithm and performance bounds. *IEEE Transactions on Signal Processing*, 2010, **58**(12): 6140−6155

16 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer Science+Business Media, 2006. 90−93

17 Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y M, Schwarz P, Silovsky J, Stemmer G, Vesely K. The Kaldi speech recognition toolkit. In: Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding. Hawaii, US: IEEE, 2011.

18 Zibulevsky M, Elad M. $L1$-$L2$ optimization in signal and image processing. *IEEE Signal Processing Magazine*, 2010, **27**(3): 76−88

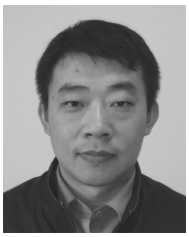
19 Riedhammer K, Bocklet T, Ghoshal A, Povey D. Revisiting semi-continuous hidden Markov models. In: Proceedings of the 2012 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Kyoto: IEEE, 2012. 4721−4724

张文林 中国人民解放军信息工程大学信息系统工程学院讲师. 2013 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理, 语音识别, 机器学习等. 本文通信作者.
E-mail: zwlin_2004@163.com
(**ZHANG Wen-Lin** Lecturer at the Institute of Information Systems Engineering, PLA Information Engineering University. His research interest covers speech signal processing, speech recognition, and machine learning. Corresponding author of this paper.)

牛 铜 中国人民解放军信息工程大学信息系统工程学院博士研究生. 主要研究方向为语音增强, 语音识别.
E-mail: niutong0072@gmail.com
(**NIU Tong** Ph. D. candidate at the Institute of Information Systems Engineering, PLA Information Engineering University. His research interest covers speech enhancement and speech recognition.)

屈 丹 中国人民解放军信息工程大学信息系统工程学院副教授. 2005 年获解放军信息工程大学博士学位. 主要研究方向为语音信号处理与模式识别.
E-mail: qudanqudan@sina.com
(**QU Dan** Associate professor at the Institute of Information Systems Engineering, PLA Information Engineering University. Her research interest covers speech signal processing and pattern recognition.)

李弼程 中国人民解放军信息工程大学信息系统工程学院教授. 主要研究方向为文本分析与理解, 语音处理与识别, 图像/视频处理与识别, 信息融合.
E-mail: lbclm@163.com
(**LI Bi-Cheng** Professor at the Institute of Information Systems Engineering, PLA Information Engineering University. His research interest covers text analysis and understanding, speech/image/video processing and recognition, and information fusing.)



裴喜龙 中国人民解放军信息工程大学信息系统工程学院助教. 主要研究方向为智能信息处理, 信息融合.
E-mail: 13838173693@139.com
(**PEI Xi-Long** Teaching assistant at the Institute of Information Systems Engineering, PLA Information Engineering University. His research interest covers speech signal processing and information fusing.)