文章编号:1003-0530(2017)05-0703-08

采用 HDPHMM 符号化器的语音查询样例检测方法

曹建凯 张连海
(解放军信息工程大学信息系统工程学院, 河南郑州 450001)

摘 要: 提出一种基于层级狄利克雷过程隐马尔科夫模型(HDPHMM)符号化器的无监督语音查询样例检测(QbE-STD)方法。该方法首先应用一个双状态层隐马尔科夫模型, 其中顶层状态用于表示所发现的声学单元, 底层状态用于建模顶层状态的发射概率, 通过对顶层状态假设一个层级狄利克雷过程先验, 获得非参贝叶斯模型 HDPHMM。使用无标注语音数据对该模型进行训练, 然后对测试语音和查询样例输出后验概率特征矢量, 使用非负矩阵分解算法对后验概率进行优化得到新的特征, 然后在此基础上, 应用修正分段动态时间规整算法进行检索, 构成 QbE-STD 系统。实验结果表明, 相比于基于高斯混合模型符号化器的基线系统, 本文所提出的方法性能更优, 检索精度得到显著提升。

关键词: 无监督; 语音查询样例检测; 层级狄利克雷过程; 非负矩阵分解
中图分类号: TP391 **文献标识码**: A **DOI**: 10.16798/j.issn.1003-0530.2017.05.007

Query-by-example Spoken Term Detection by Applying the HDPHMM Tokenizer

CAO Jian-kai ZHANG Lian-hai
(Institute of Information Systems Engineering, PLA Information Engineering University, Zhengzhou, Henan 450001, China)

Abstract: This paper presents a study of hierarchical Dirichlet processing hidden Markov model (HDPHMM) approach for unsupervised query-by-example spoken term detection (QbE-STD). First a hierarchical hidden Markov model is applied, in which the top layer states are used for representing the finding acoustic units, bottom layer states are used for modeling the emission probability of top layer states. We can get a nonparametric Bayesian model HDPHMM when imposing a hierarchical Dirichlet processing prior on the top layer states. After the model is trained by unlabeled speech data, it outputs posteriorgram feature vector for test utterance and query term. The posteriorgram feature is optimized by non-negative matrix factorization algorithm. Then the detection is performed by modified SDTW algorithm. Experimental results show that the proposed method outperforms the baseline system based on Gaussian mixture model tokenizer, and improve the detection precision obviously.

Key words: unsupervised; query-by-example spoken term detection; hierarchical Dirichlet processing; non-negative matrix factorization

1 引言

语音是人类交流的重要媒介,随着自动语音识别(automatic speech recognition,ASR)技术的发展,人机之间的交互在人类生活中的应用越来越广泛。口语项检测(spoken term detection,STD)是一种常见的 ASR 任务,目的是从语音文档库中检索出给定的查询口语样例[1],该技术目前已广泛应用于如音乐检索,口语文档检索[2-3]等现实应用中。

当前,对于解决 STD 问题主要有两种研究方向。一是依赖于大词汇量连续语音识别引擎的有监督方法[4],该方法首先依靠训练好的语音识别器

收稿日期: 2016-09-13; 修回日期: 2016-11-16
基金项目: 国家自然科学基金资助项目(61673395, 61403415, 61302107)

将查询样例和测试语音都转化成词或者子词序列，然后进行基于文本的匹配。可以看出该种方法不仅需要预先知道语音构成以及发音字典，还需要收集标注语料来训练识别器。相对应地，另一种方向即是无需标注语料和发音字典的无监督方法。尤其地，当查询样例以语音形式给出时，该 STD 任务又称为语音查询样例检测（query-by-example STD，QbE-STD）。

目前无监督 QbE-STD 系统主要有两种方法。一是基于模板匹配的方法[5]，基本思想是通过一个符号化器[6-7]将查询样例和测试语音转化为后验概率特征，然后采用动态时间规整（dynamic time warping，DTW）算法检测出匹配区域。二是基于模型的方法，基本思想是采用模式发现技术，建模构成语音的声学单元，并将查询样例和测试语音转化成模式序列，然后进行基于文本的符号检索[8-9]。与有监督系统相比，当前无监督 QbE-STD 系统检索性能还存在较大差距。本文主要针对第一种方法研究新的符号化器模型对无标注数据进行建模，生成更具鲁棒性的后验概率特征，以提高检索精度。

本文应用一个层级狄利克雷过程隐马尔科夫模型（hierarchical Dirichlet processing hidden Markov model，HDPHMM）[10]符号化器来代替传统的高斯混合模型[6]（Gaussian mixture model，GMM）和声学分段模型[7]（acoustic segmental model，ASM）符号化器，有三个优势：一是 HDPHMM 的基本拓扑结构是 HMM，相比 GMM 中的单一高斯分量，HMM 对具有一定持续时间的声学单元拥有更强的可塑性，比如可以用不同的状态来表示同一声学单元持续时间内的不同阶段，而且每一状态都是由一个 GMM 模型构成，能够更灵活地刻画特征分布。二是相比于 ASM 模型中各个 HMM 之间并无关联，HDPHMM 中的层级隐马尔科夫模型（hierarchical hidden Markov model，HHMM）将各个子 HMM 以状态转移概率联系起来，这就相当于在声学模型中构造了一个 2-gram 语言模型。三是对于 HHMM 的参数，使用一个基于 HDP 的 Gibbs 采样过程来训练得到[10-11]，这种非参方法相比于 GMM、k-means 等参数聚类能够更加自适应于无标注数据。

针对现有无监督符号化器在 QbE-STD 系统中检索精度低的问题，提出一种基于 HDPHMM 模型的检索系统，该方法首先应用一个双状态层 HHMM，其中顶层状态用于表示所发现的声学单元，底层状态用于建模顶层状态的发射概率，并对顶层状态假设一个 HDP 先验，获得非参贝叶斯模型 HDPHMM。使用无标注语音数据和 Gibbs 采样算法对该模型进行训练获得模型参数。训练完成后，模型对测试语音和查询样例输出后验概率特征矢量，为了去除后验规律特征所包含的冗余信息，使用非负矩阵分解（non-negative matrix factorization，NMF）算法对其进行优化得到新的特征，然后在此基础上，应用修正分段 DTW（segmental DTW，SDTW）算法进行检索，整个流程构成了基于 HDPHMM 符号化器的无监督 QbE-STD 系统。

2 系统框架

图 1 描述了本文基于 HDPHMM 符号化器的 QbE-STD 系统框图。首先对原始语音数据提取频域线性预测[11]（frequency domain linear prediction，FDLP）特征，然后使用训练数据的声学特征训练 HDPHMM 模型，模型训练完成后，对测试语音和查询样例输出后验概率特征，使用 NMF 算法对这些后验概率特征进行优化得到新的特征，最后在新的特征上执行修正 SDTW 算法，获得检索结果。

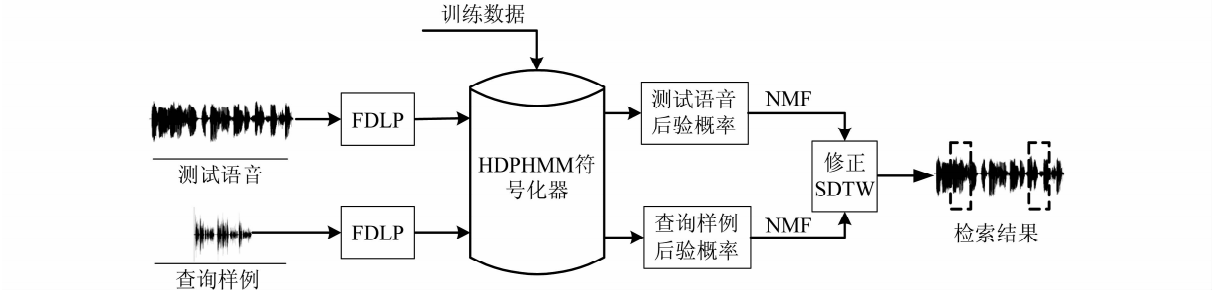

图 1 基于 HDPHMM 符号化器的 QbE-STD 系统框图

Fig. 1 QbE-STD system framework based on HDPHMM tokenizer

2.1　声学特征

传统符号化器模型通常采用梅尔频率倒谱系数(mel-frequency cepstral coefficients,MFCCs)作为输入特征[6],缺少语音上下文信息,鲁棒性较差。对此,文献[12]提出 FDLP 特征,通过分析语音的长时特性来捕获语音信号的时域动态属性,并可以代替传统频谱参数(如 MFCCs)用于训练声学模型。文献[13]已证明基于 FDLP 的高斯后验性能要优于基于 MFCCs 的高斯后验。因此,本文采用 13 维的 FDLP 及其一阶和二阶分量所组成的 39 维特征矢量作为整个 QbE-STD 系统的输入声学特征,每一句子的特征按维度进行 z-norm 规整。

2.2　层级狄利克雷过程隐马尔科夫模型

2.2.1　层级隐马尔科夫模型

HHMM 是一种包含两层状态 HMM 的扩展[10]。图 2 描述了 HHMM 的模型结构,其中顶层的每一状态($S_1,S_2,\cdots,S_K$)表示每一类聚类数据,也即一种声学单元,K 表示所定义的聚类数量。描述每一顶层状态的发射概率的结构为一个常规 HMM(本文中每一常规 HMM 包含 3 个状态,每一状态包含 8 个高斯分量),而不是传统的 GMM,比如顶层状态 S_1 的发射概率由包含 3 状态($S_{1,1},S_{1,2},S_{1,3}$)的常规 HMM 构成。所有常规 HMM 中包含的状态构成了 HHMM 的底层状态结构。顶层状态所表示的 HMM 用于生成长度为 $d_i(i=1,2,\cdots,M)$ 的声学特征序列 $x_{i,1},x_{i,2},\cdots,x_{i,d_i}$,其中 M 表示该语音片段所包含的声学单元数量,每一声学特征序列表示原始语音中的一个声学单元。图中实线箭头表示转移概率,虚线箭头表示控制关系。注意到顶层的状态转移为全转移,即任意两个状态之间以及状态自身都存在转移关系,这与语音中任意两个声学单元都可能会相邻出现相对应,而且转移概率越大,表示这两个声学单元相邻出现的概率就越大,反之则越小。

2.2.2　层级狄利克雷过程

层级狄利克雷过程(hierarchical Dirichlet processing,HDP)是 DP 的一种扩展形式[14],用于解决在 DP 聚类中,当基分布连续时,所获得的采样参数以概率 1 不等的情况。HDP 在基分布上又定义了一个先验分布,通过一个 DP 获得基分布的采样,这样就保证了基分布的离散性。HDP 的定义形式如下:

$$\begin{aligned} G_0 \mid \gamma, H &\sim \mathrm{DP}(\gamma, H) \\ \mathbf{G}_j \mid \alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_{ji} \mid \mathbf{G}_j &\sim \mathbf{G}_j \\ x_{ji} \mid \theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{1}$$

其中 H 表示参数 θ_{ji} 的先验分布,超参数 γ 用于控制从 H 中采样获得的基分布 G_0 集中于 H 的程度,超参数 α 用于控制从 G_0 中采样获得的参数集 $\mathbf{G}_j$ 集中于 G_0 的程度,$\mathbf{G}_j$ 是所有 θ_{ji} 参数的集合,x_{ji} 表示由参数 θ_{ji} 所控制的函数 $F(\cdot)$ 生成的观察量。

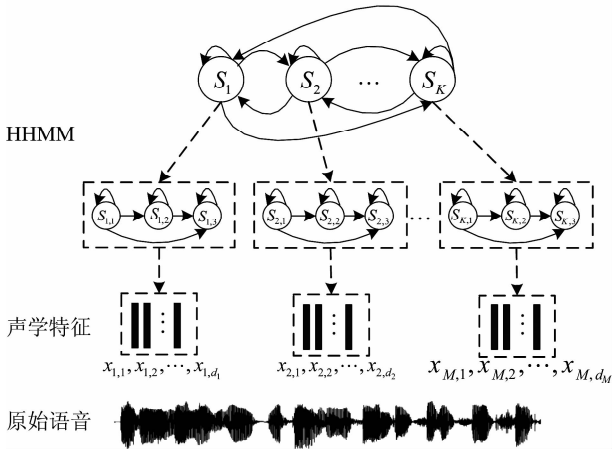

图 2　HHMM 模型结构

Fig. 2　The structure of HHMM model

通过对 HHMM 的顶层状态转移概率假设一个 HDP 先验,即可以获得一个 HDPHMM 模型,并由 Gibbs 采样获得 HDPHMM 的参数[10]。注意到 HDP 本质上属于非参聚类,即聚类数量是不固定的,公式(1)中的 θ_{ji} 个数可能无限大。然而在实际应用中,构成语音的声学单元数量通常会有一个上限,比如在 TIMIT 数据库中,共定义了 61 个单音子,而在一些文献中这些单音子通过组合数量会进一步降低。鉴于此,可以对 HDP 的聚类数量也即 HHMM 中所包含的顶层状态数设置一个限制值 K,该值要大于构成语音的声学单元数量上限(本文中 K 值设置为 100)。这种应用参数化的视角构建非参模型可以大大降低模型训练的复杂度。为了进一步简化模型推导,将先验分布设为一个等概分布,公式(1)可以简化成如下形式:

$$G_0 \sim \mathrm{Dir}(\boldsymbol{\gamma})$$
$$\theta_k \sim \mathrm{Dir}(\boldsymbol{\alpha} G_0)$$
$$x_k \mid \theta_k \sim F(\theta_k) \quad (2)$$

其中 Dir 表示 Dirichlet 分布,这里超参数 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\alpha}$ 均为长度为 K 的矢量,并且各自矢量中的所有元素值相等。

2.2.3 生成性过程

HDPHMM 的概率图模型结构如图 3 所示,该图引用于文献[10]。假设一段语音识别成 M 个片段,其中π 表示 HHMM 顶层状态的初始概率,也即该段语音首个声学单元的概率,服从由超参数 $\boldsymbol{\eta}$ 控制的 Dirichlet 分布,ϕ_k 表示顶层状态 k 的转移概率,是一个 K 维矢量,使用 ϕ_{kj}来表示状态 k 与 j 之间的转移概率,β 表示 ϕ_k 的先验分布,服从由超参数 $\boldsymbol{\gamma}$ 控制的 Dirichlet 分布,β 和 $\boldsymbol{\alpha}$ 共同控制 ϕ_k 的 Dirichlet 分布;$c_1, c_2, \cdots, c_K$ 表示语音片段的聚类标签,用于区分不同得分声学单元;超参数 θ_0 表示由 HHMM 底层状态所构成的常规 HMM 参数的先验分布,θ_k 表示由第 k 个顶层状态所关联的底层常规 HMM 参数(包含底层状态转移概率参数和发射概率参数);$d_1, d_2, \cdots, d_M$ 分别表示各个语音片段所包含的帧数量,$\boldsymbol{x}_{i,t}$表示第 i 个语音片段的第 t 帧特征矢量。可以看出,该图从上到下描述了一段语音的生成过程,具体生成性过程描述如下:

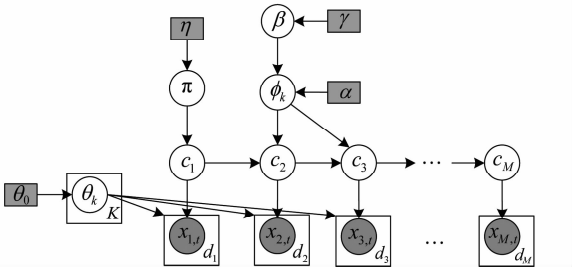

图 3 HDPHMM 概率图模型

Fig. 3 The graphical representation of HDPHMM

1. 首先获得初始状态概率分布,形式如下,

$$\pi \sim \mathrm{Dir}(\boldsymbol{\eta}) \quad (3)$$

2. 然后获得转移概率的先验分布,形式如下,

$$\beta \sim \mathrm{Dir}(\boldsymbol{\gamma}) \quad (4)$$

3. 通过对 β 采样,获得每一顶层 HHMM 状态的转移概率分布,

$$\phi_k \sim \mathrm{Dir}(\boldsymbol{\alpha}\beta) \quad (1 \leqslant k \leqslant K) \quad (5)$$

4. 通过对先验分布 θ_0 采样,获得所有底层常规 HMM 的参数,

$$\theta_k \sim \theta_0 \quad (1 \leqslant k \leqslant K) \quad (6)$$

5. 通过初始状态概率分布选择一个声学单元作为该段语音的首个单元,

$$c_1 \sim \pi \quad (7)$$

6. 通过转移概率分布,对第 i 个语音片段标签进行采样,

$$c_i \sim \phi_{c_{i-1}} \quad (2 \leqslant i \leqslant M) \quad (8)$$

7. 对于每一语音片段,由 θ_{c_i}所控制的底层常规 HMM 生成特征矢量序列,其中每一特征矢量对应该 HMM 中的一个状态。

$$\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,d_i} \sim \theta_{c_i} \quad (1 \leqslant i \leqslant M) \quad (9)$$

注意到,上述生成性过程语音片段的总数以及片段边界是假设已知的,但在实际中,这些条件是通过 Gibbs 抽样推测得到的,即同时实现语音的分段,聚类和建模过程,具体 Gibbs 抽样算法请参照文献[10-11]。模型训练完成后,输入为声学特征矢量,输出则为后验概率特征矢量。本文中模型的超参数设置如表 1 所示。

表 1 HDPHMM 模型中超参数值设置

Tab. 1 The values of the hyperparameters of HDPHMM

$\boldsymbol{\eta}$	$\boldsymbol{\gamma}$	$\boldsymbol{\alpha}$	K
$\langle 1 \rangle_K$	$\langle 50 \rangle_K$	$\langle 1 \rangle_K$	100

与公式(2)相对应,表 1 中$\langle a \rangle_K$ 表示一个 K 维矢量,并且矢量中的每一元素值均为 a。

2.3 后验概率特征矢量

HDPHMM 模型训练完成后,每一顶层状态代表一类声学单元。对每帧声学特征,通过计算所有顶层状态的发射概率,可以输出一个 K 维后验概率特征矢量,矢量中的每一元素表示该帧语音属于该元素位置所代表的声学单元的后验概率。后验概率特征矢量通过 Bayesian 准则计算得到,形式如下,

$$\boldsymbol{PG}_{x_t} = [p(c_1 \mid x_t), p(c_2 \mid x_t), \ldots, p(c_K \mid x_t)] \quad (10)$$

其中 x_t 表示语音帧,c_k 表示第 k 个顶层状态或者称第 i 类,K 为限制因子,也即聚类数量。式中 $p(c_k \mid x_t)$表示

第 t 帧属于第 k 类的后验概率,计算公式如下,

$$P(c_k \mid x_t)=\frac{\sum_{i=1}^{3}\sum_{j=1}^{8}\omega_{kij}\mathrm{N}_{kij}(x_t \mid \boldsymbol{\lambda}_{kij})}{\sum_{m=1}^{K}\sum_{i=1}^{3}\sum_{j=1}^{8}\omega_{mij}\mathrm{N}_{mij}(x_{mti} \mid \boldsymbol{\lambda}_{mij})} \quad (1 \leqslant k \leqslant K) \tag{11}$$

式(11)中 k 与 m 表示声学单元的类别索引,也即顶层状态索引,i 表示 HMM 状态索引,j 表示高斯单元索引,ω 表示高斯单元权重,例如 ω_{kij} 表示第 k 个顶层状态下第 i 个 HMM 状态的第 j 个高斯单元的权重,$\mathrm{N}_{kij}(x_t \mid \boldsymbol{\lambda}_{kij})$ 表示其高斯分布,$\boldsymbol{\lambda}_{kij}$ 是包含均值和协方差矩阵的参数集。

2.4　NMF 特征优化

由于在训练 HDPHMM 模型时,参数 K 的值要大于训练语音数据中所包含的实际声学单元数量。在模型训练完成后,对输入的每帧语音所生成的 K 维后验概率特征矢量中,必然会包含冗余信息。

NMF 算法的基本思想是将一个非负矩阵分解成两个非负矩阵的乘积[15],其物理意义是“局部构成整体”。NMF 算法的本质是寻找样本数据的特征子空间,然后将高维数据投影到低维子空间中,从而在子空间上获得样本的本质特征。对于后验概率矢量,其物理含义是描述特征帧属于各个声学单元的后验概率。通过对后验特征矩阵进行非负分解,可以突出后验概率矢量中的主要分量,抑制无关分量,从而达到去除冗余信息的目的。

假设处理 m 个 n 维空间的样本数据,用 $\boldsymbol{X}_{n\times m}$ 表示。该数据矩阵中各个元素都是非负的,则可以将矩阵分解看成如下含加性噪声的线性混合体模型,

$$\boldsymbol{X}_{n\times m}=\boldsymbol{W}_{n\times r}\boldsymbol{H}_{r\times m}+\boldsymbol{E}_{n\times m} \tag{12}$$

其中 $\boldsymbol{W}_{n\times r}$ 称为基矩阵,$\boldsymbol{H}_{r\times m}$ 为系数矩阵,$\boldsymbol{E}_{n\times m}$ 为噪声矩阵。若选择 $r<n$,用系数矩阵代替原数据矩阵,就可以实现对原矩阵的降维。

使用 NMF 矩阵对后验特征矩阵进行分解后,直接使用系数矩阵会带来原始信息的分割破坏。所以这里采用另一种方法,使用基矩阵对原始矩阵做空间变换。将基矩阵看作子空间变换矩阵,即原始矩阵中的每一样例数据可以看作由基矩阵的所有列向量线性加权得到,则相乘的结果即为数据在基矩阵上的投影长度所组成的矢量,

$$\boldsymbol{Z}_{r\times m}=\boldsymbol{W}_{n\times r}^{\mathrm{T}}\boldsymbol{X}_{n\times m} \tag{13}$$

其中 $\boldsymbol{Z}_{r\times m}$ 为投影后的特征矩阵,具体实验过程为:使用训练好的 HDPHMM 符号化器对测试语音集输出后验特征矢量序列,将所有后验特征拼接到一起组成一个大的矩阵,采用 NMF 算法对其进行分解,得到一个 $n\times r$ 的基矩阵,其中每一列向量可以看作原始矩阵的一个基,然后与每一测试语音的后验特征相乘,得到优化后的特征,在新的特征上执行检索算法。

2.5　修正 SDTW 算法

传统 DTW 算法在匹配时存在大量冗余计算,对此文献[6]提出分段 DTW(segmental DTW,SDTW)算法。相比 DTW,SDTW 增加了一个长度为 R 的移动窗,用于将测试语音划分为一系列子片段,然后在每个子片段应用 DTW 检索。由于限制了回溯路径长度,因此能够有效提升检索速度。然而这会导致 SDTW 算法的一个潜在缺点:移动窗有可能分割测试语音中的候选区域,导致匹配得分不能达到全局最优。当前 SDTW 算法为单输出形式,即对一个查询样例测试语音对,只输出一个最佳匹配得分。对于基于 SDTW 检索的真实匹配子段对,一种合理性假设是,则其相邻子段应包含部分查询样例中的声学单元,并且数量随着与最佳子段的距离增加而减少,而对应的失真得分则越来越大。如果能够考虑最佳匹配子段的若干相邻子段得分,则可以缓解该算法缺陷,使得修正后的得分相比 1-best 得分更具区分性。由此本文对每一查询样例测试语音对输出最佳匹配子段得分以及其左右两边各 L 个连续子段得分(本文设置 L 值为 2),在该 $2L+1$ 个得分内选出 N(本文设置 N 值为 3)个最佳得分进行加权得到最终得分,即

$$F=(1-\lambda)F_{\mathrm{best}}+\lambda(1-\lambda)F_{\mathrm{best}}^{-1}+\lambda^2F_{\mathrm{best}}^{-2} \quad (0 \leqslant \lambda \leqslant 0.5) \tag{14}$$

其中 λ 表示权重因子,F_{best} 表示最佳匹配子段得分,F_{best}^{-1} 和 F_{best}^{-2} 分别表示相邻的第二和第三最佳匹配子段得分,F 表示该查询样例测试语音对的最终得分。

3　实验结果及分析

3.1　实验数据

本文采用 TIMIT 语料库[16]执行 HDPHMM 模型训练以及 QbE-STD 实验。TIMIT 共有 6300 条语句,分为

TRAIN 和 TEST 两个集合。本文实验选择 TRAIN 中 3296 条语句作为训练集,选择 TEST 中 1344 个语句作为测试集(不包含 SA1 和 SA2 中的语句)。从测试集中抽取 10 个单词作为查询样例。

3.2 评价标准

QbE-STD 系统常用评价指标有平均正确率均值(mean average precision, MAP)和平均检索时间(average detection time, AT),前者用于衡量检索精度,后者用来衡量检索速度。

MAP 定义为对所有查询样例的平均精确度(average precision, AP)求均值,平均检索时间定义为查询样例完成检索所平均消耗的时间(不包括特征提取和模型训练所用的时间)。

3.3 系统性能分析

3.3.1 NMF 特征优化算法性能分析

公式(12)中选用不同的 r,得到的新的特征矢量会有不同的维度,表 2 描述了不同 r 下 HDPHMM 检索精度的对比,检索算法采用标准 SDTW 算法。可以看出,$r=60$ 是该符号化器检索精度的一个转折点,后验概率特征矢量经 NMF 算法降维后得到的新的特征矢量,其维度若小于 60,随着维度的下降其检索精度亦急剧下降,说明由于降维所导致的后验概率有效信息损失开始增大,而当其维度大于 60 时,随着维度的增加,其检索精度趋于稳定,不会有剧烈变化,表明降维主要去除的是后验概率中所包含的冗余信息,而不会损失过多的有效信息。若考虑到实验所用的 TIMIT 数据库中英语发音构成恰好包含 61 个单音子,这意味着对训练数据,最佳的聚类数量或者称所发现的声学单元数量应为 60 个左右。而表 2 所表明的实验现象恰与该先验知识相符合,这说明了 NMF 算法对后验概率特征降维的有效性。

表 2 不同 NMF 降维维度下检索性能比较

Tab. 2 Detection performance comparison of different NMF dimensions

R	40	50	60	70	80	90	100
MAP	0.5765	0.6136	0.6370	0.6388	0.6386	0.6388	0.6410

由于该 HDPHMM 符号化器是针对于无监督环境提出的,而在实际应用中,对于无标注的训练数据,可能并不知道其属于何种语言,以及所包含的声学单元数量,因此并不能确定降维数量的大小。而上述实验表明,应用 NMF 算法对后验概率特征在降噪的同时保持特征矢量维度不变,同样可以获得较好的效果。在接下来的实验中,所有经 NMF 算法处理后所得特征矢量其维度均与原后验概率矢量保持一致。

3.3.2 修正 SDTW 算法性能分析

针对有无 NMF 后验特征优化步骤的两种 HDPHMM 检索系统,图 4 描述了公式(14)中权重因子对其检索性能的影响,其中 $\lambda=0$ 时为标准 SDTW 算法,可以看出在权重因子 $\lambda=0.25$ 时,两种系统的检索性能均达到最好,当接近 0 或者 0.5 时,检索精度有所下降,这是因为权重因子过小或者过大,会过于忽视或者突出最佳匹配子段的相邻子段得分的影响。由于修正 SDTW 算法只增加少量的加权运算,因此几乎不影响检索速度。

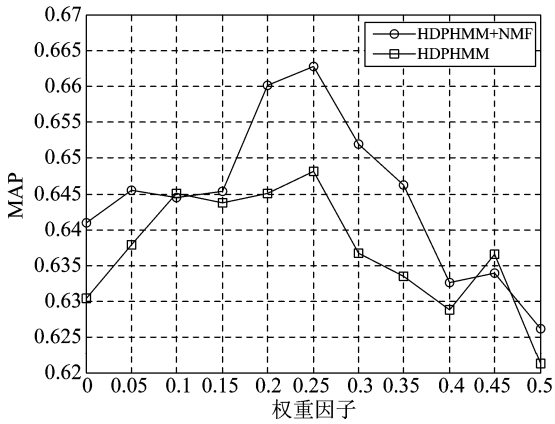

图 4 不同权重因子下的 MAP 曲线

Fig. 4 MAP curves for different weight factors

从图 4 可以看出,对于 HDPHMM 符号化器,采用标准 SDTW 检索算法,即 $\lambda=0$ 时,其检索精度为 0.630,而采用 NMF 后验特征优化和修正 SDTW 算法后,检索精度达到 0.663,相对提升 5.24%。

3.3.3 HDPHMM 符号化器性能分析

本文比较了 HDPHMM 与其他传统符号化器的 QbE-STD 性能,包括文献[6]和[7]所提出的 GMM 和 ASM 符号化器(为保证实验条件的一致性,这里的 GMM 和 ASM 训练同样是以 FDLP 作为输入声学特征),以及语种匹配的英语音素识别器(phoneme recognition, PR)和语种失配的俄罗斯语音素识别器,搜索算法均采用修正 SDTW 算法。表 3 描述了

各种系统的实验结果,其中“EN”表示英语,“RU”表示俄罗斯语。

表 3　不同符号化器检测性能对比

Tab. 3　Detection performance comparison of different tokenizers

系统	GMM	ASM	PR(EN)	PR(RU)	HDPHMM
MAP	0.532	0.556	0.799	0.208	0.648

从表 3 可以看出,有监督条件下训练的英语音素识别器其 QbE-STD 性能最好,MAP 要远优于无监督条件下训练的符号化器,而语种失配的音素识别器性能较差。无监督条件下,HDPHMM 符号化器性能最好。接下来,本文以常用的 GMM 符号化器作为基线系统对 HDPHMM 模型的性能进行分析。

表 4　不同系统检测性能对比

Tab. 4　Detection performance comparison of different systems

系统	MAP	时间/s
GMM	0.532	46.1
GMM+NMF	0.612	47.8
HDPHMM	0.648	51.0
HDPHMM+NMF	0.663	52.7

表 4 描述了 HDPHMM 检索系统与 GMM 基线系统检索性能的对比。为了保证比较的公正性,所有检索系统均使用修正 SDTW 检索算法,并且表中所列出的数据均为对应系统性能在最优权重因子下的检索结果。同时为了更好地体现 HDPHMM 符号器的性能优势,对有无 NMF 后验特征优化的两种情况均作了比较。并且两种符号化器后验概率特征经 NMF 优化后,均保持优化后的特征与原特征维度保持一致。可以看出,在应用 NMF 算法情况下,相比基线系统,HDPHMM 在检索精度方面相对提升 8.3%。而在不应用 NMF 算法情况下,检索精度相对提升 21.8%。而在检索速度方面,GMM 基线系统要略好于 HDPHMM 系统,这是因为根据文献[6],GMM 高斯分量的个数设置为 50,而本文中 HDPHMM 的 K 值设置为 100,故两者所输出的后验概率特征矢量维度分别为 50 和 100,又由于修正 SDTW 算法所采用的特征矢量之间的距离测度为点积,因此在距离矩阵的计算上,HDPHMM 的计算量是 GMM 的 2 倍,故检索耗时较长。上述实验结果表明 HDPHMM 作为符号化器,对 QbE-STD 系统的检索精度提升明显,体现了 HDPHMM 对无标注数据更好的建模效果。

4　结论

本文提出一种基于 HDPHMM 符号化器的无监督 QbE-STD 系统,首先使用 HHMM 能够建模语音数据中不同声学单元之间的转移关系,类似于在声学模型中构造一个 2-gram 语言模型,使得所生成的后验概率更具鲁棒性。其次通过对 HHMM 顶层状态假设一个 HDP 先验,可以通过非参聚类算法更加自适应于无标注数据。最后实验结果表明,相比传统无监督符号化器,HDPHMM 符号化器检索精度提升明显,但与有监督的音素识别器相比,检索精度还存在较大的差距,检索速度亦有待提高。由于该模型是以 HMM 为架构的,未来可以研究将该模型转化成一个声学单元识别器,对测试语音和查询样例识别出所发现的声学单元符号,建立 lattice 索引,构建基于文本符号的快速检索系统。

参考文献

[1] Shen W, White C M, Hazen H T. A comparison of query-by-example methods for spoken term detection[C]// Interspeech 2009. Brighton, United Kingdom, 2009:2143-2146.

[2] Balke S, Arifi-Muller V, Lamprecht L, Miiller M. Retrieving Audio Recordings using musical themes[C]// ICASSP 2016. Shanghai, China, 2016: 281-285.

[3] Zhang Y C, Duan Z Y. IMISound: an unsupervised system for sound Query by vocal imitation[C]// ICASSP 2016. Shanghai, China, 2016: 2269-2273.

[4] David R. H. Miller, Michael Kleber, Chia-lin Kao, et al. Rapid and accurate spoken term detection[C]// Interspeech 2007. Antwerp, Belgium, 2007: 314-317.

[5] Xu H H, Hou J Y, Xiao X, et al. Approximate search of audio queries by using DTW with phone time boundary and data augmentation[C]// ICASSP 2016. Shanghai, China, 2016: 6030-6034.

[6] Zhang Y D, Glass J. Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams[C]// In Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop. Merano, Italy, 2009:398-403.

[7] Wang H P, Lee T, Leung C C, et al. Acoustic segment modeling with spectral clustering methods[J]. IEEE

Transactions on Audio Speech & Language Processing, 2015,23(2):264-277.

[8] Chung C T, Chan C A,Lee L S. Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization[C]// ICASSP 2013. Vancouver, Canada, 2013: 8081-8085.

[9] Chung C T, Hsu W N, Lee C Y, et al. Enhancing automatically discovered multi-level acoustic patterns considering context consistency eith applications in spoken term detection[C]// ICASSP 2015. Brisbane, Australia, 2015: 5231-5235.

[10] Lee C Y. Discovering linguistic structures in speech: models and applications[D]. [Ph. D. dissertation], Massachusetts Institute of Technology, 2014: 105-119.

[11] Lee C Y, Glass J. A nonparametric Bayesian approach to acoustic model discovery[C]// In Proceedings of ACL, 2012: 40-49.

[12] Ganapathy S. Signal analysis using autoregressive models of amplitude modulation [D]. Baltimore, Maryland, USA: Johns Hopkins University, 2012:60-68.

[13] Mantena G, Achanta S, Prahallad K. Query-by-example spoken term detection using frequency domain linear prediction and non-segmental dynamic time warping[J]. IEEE Transactions on Audio Speech and Language Processing, 2014, 22(5): 946-955.

[14] Harati A, Picone J. A doubly hierarchical Dirichlet process hidden Markov model with a non-ergodic structure[J]. IEEE Transactions on Audio Speech and Language Processing, 2016, 24(1): 174-184.

[15] 胡永刚,张雄伟,邹霞,等. 改进的非负矩阵分解语音增强算法[J]. 信号处理,2015,31(9): 1117-1123.
HuYonggang, Zhang Xiongwei, Zou Xia, et al. Improved Nonnegative Matrix Factorization based Speech Enhancement Algorithm[J]. Journal of Signal Processing, 2015, 31(9): 1117-1123. (in Chinese)

[16] Garofolo J, Lamel L, Fisher W, et al. TIMIT acoustic-phonetic continuous speech (MS-WAV version)[J]. Journal of the Acoustical Society of America, 1990, 88(88):210-221.

作者简介

曹建凯 男,1993 年生,河南周口人。解放军信息工程大学信息系统工程学院硕士研究生,研究方向为无监督语音关键词检测、模式发现。
E-mail: jiankaic@ sina. com

张连海 男,1971 年生,山东菏泽人。解放军信息工程大学信息系统工程学院副教授,研究方向为语音信号处理、模式识别。
E-mail: lianhaiz@ sina. com