ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 2,Mar. 2016,pp. 407-414
DOI:10.16337/j.1004-9037.2016.02.023

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:+86-025-84892742

基于声学分段模型的无监督语音样例检测

李勃昊　张连海　郑永军

(解放军信息工程大学信息系统工程学院,郑州,450001)

摘　要:提出一种基于声学分段模型的无监督语音样例检测方法。该方法首先利用高斯混合模型(Gaussian mixture model, GMM)将训练数据频谱参数转换为后验概率特征向量,采用层次聚类算法确定后验概率的边界信息,得到声学分段;然后通过 k-means 算法将片段聚类并添加标签,构建基于后验概率的声学分段模型。检索时以模型对查询样例与检索文档的解码序列代替测量矩阵以降低检索时间,通过基于最小编辑距离的动态匹配检索查询项,最小编辑距离的代价函数由模型相似度距离矩阵修正。实验结果表明,相比 GMM 及传统声学分段模型,本文提出的方法性能更好,检索速度得到显著提升。
关键词:声学分段模型;语音样例检测;后验概率特征;无监督
中图分类号:TP391　　**文献标志码**:A

Unsupervised Query-by-Example Spoken Term Detection Based on Acoustic Segment Models

Li Bohao, Zhang Lianhai, Zheng Yongjun

(Institute of Information Systems Engineering, PLA Information Engineering University, Zhengzhou, 450001, China)

Abstract: A study of acoustic segment models(ASMs) for unsupervised query-by-example spoken term detection is presented. Firsty, a Gaussian mixture model(GMM) is trained without any transcription information to label speech frames with Gaussian posteriorgram. Hierarchical agglomerative clustering is used to decompose the posterior features into acoustically exhibiting segments. A label is assigned to each result segment by k-means clustering, then posteriorgram is faciltitated to train ASMs. In query matching phase, Viterbi decode is prosed to represent query and test posteriorgrams as ASM sequences. Dynamic match lattice spotting based on minimum edit distance is used to locate possible occurrences of the query term. Experimental results show that the proposed method outperforms traditional GMM and ASMs tokenizers.
Key words: acoustic segment models; query-by-example spoken term detection; posterior features; unsupervised

引　言

口语词汇检测(Spoken term detection, STD)[1]的目的是从声学文档中自动检索出口语查询项,相比传统关键词检测(Key word spotting, KWP)技术,STD 中查询项可以是文本形式也可以是语

基金项目:国家自然科学基金(61175017)资助项目。
收稿日期:2014-11-27;**修订日期**:2015-1-14

音样例,可以是单一关键词也可以是连续短句。当前的 STD 系统主要采用大规模连续语音识别(Large vocabulary continuous speech recognition, LVCSR)技术,需要大量标注语料训练 HMM,同时还面临集外词(Out-of-vocabulary, OOV)性能不高的问题[2]。如果缺乏标注信息,现有的检测系统将无法使用。因此无监督的语音查询样例检测(Query-by-example STD, QbE-STD)成为当前研究的热点之一[3]。

当前无监督 QbE-STD 主要采用基于后验概率特征的模板匹配框架[4],通过无监督方法训练模型分类器,将查询样例和检索文档转换为相应的后验概率特征向量矩阵,通过分段动态时间规整(Segment dynamic time warping, SDTW)算法进行检索。这一框架摆脱了系统对标注语料的依赖,不需要语言模型和发音字典,也就不存在 OOV 问题。框架的核心是如何通过无监督的训练方法得到鲁棒性与区分性能良好的模型分类器。常用的方法有:文献[5]利用标注语料训练音素识别器,将语音信号转换为音素后验概率,该方法性能优于无监督方法,但系统性能受标注信息影响;文献[6]利用高斯混合模型(Gaussian mixture model, GMM)的分类特性,以高斯子分布作为基本语音单元,通过计算数据在每一个高斯下的后验概率表示数据的声学特征分布;文献[7]提出声学分段模型(Acoustic segment models, ASMs)概念,与传统有监督条件下训练的隐马尔可夫模型(Hidden Markov model, HMM)具有相似的拓扑结构,通过无监督的聚类算法得到声学分段边界信息并人工添加标注,拥有相同标注的声学分段统一训练一个 HMM,HMM 相互间的转换关系由一个转移列表表示,求取数据在模型中每一个 HMM 下的后验概率并通过 SDTW 检索查询项。

文献[5]虽然利用了标注信息,但是可用于零资源条件下的语音检索,以资源丰富的语种训练音素识别器,检索另一种零资源语种,这种方法与文献[6]的 GMM 方法性能接近,而 ASMs 性能相对最优[8],这是由于 GMM 的训练基于帧级数据,假设数据帧之间相互独立,没有考虑数据间的前后联系,而 ASMs 是建立在声学分段基础上,声学分段内部具有相近的语音信息,考虑了时间信息;另一方面,GMM 将每一个高斯子分布视作基本语音单元,而 ASMs 中每一个声学单元由 HMM 表示,HMM 能够更加精确灵活地描述语音信号特征分布。但现有 ASMs 由频谱参数训练得到,是对最底层声学特征的直接描述,包含大量冗余信息,提取时大多采用短时信息,易受噪声影响,鲁棒性和区分性较差。

SDTW 算法需计算测量矩阵并回溯最优路径,检索时间慢,不满足系统实时性要求。为提高检索速度,文献[9]在 SDTW 中融入下界估计算法确定分段边界信息,计算查询样例和候选分段的 DTW 得分,并运用 K 近邻算法检索查询项;文献[10]通过聚类算法得到声学分段,通过减少计算量提升速度。现有的方法主要在 SDTW 基础上将基于帧级数据计算简化为基于声学分段的计算,但本质上仍属于 DTW 范畴。

针对基于频谱参数提取的后验概率特征检测精度低和检索时间长的不足,根据后验概率鲁棒性和 ASMs 中 HMM 分布特点,本文提出一种基于后验概率的改进 ASMs 无监督 QbE-STD。该方法首先将声学信号频谱参数输入 GMM,转换为鲁棒性与区分性更为良好的高斯混元后验概率,然后根据聚类算法得到声学分段的边界信息,并自定义标签添加标注,最后根据标注数据训练 ASMs 模型,利用 ASMs 模型对待检测数据解码。本文以模型解码序列替代后验概率特征向量矩阵,把 DTW 测量距离矩阵计算转换为文本字符串动态匹配,有效提高了检索速度,并对无监督条件下的最小编辑距离代价进行了讨论。

1 后验概率特征向量的无监督语音样例检测

图 1 所示为基于后验概率特征向量的无监督语音样例检测系统框架,首先利用无标注语料训练模型分类器,分类器将查询样例和测试语音文档频谱参数转换为模型后验概率特征向量序列,运用 DTW 算法(或其相应改进算法)对测试文档的特征向量矩阵进行匹配,检索出样例特征向量矩阵所在区域。

1.1 后验概率特征提取

给定一段长度为 n 帧的观测向量序列 $\boldsymbol{S}=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_n)$,其中 $\boldsymbol{f}_i$ 为第 i 帧数据的语音特征参数,

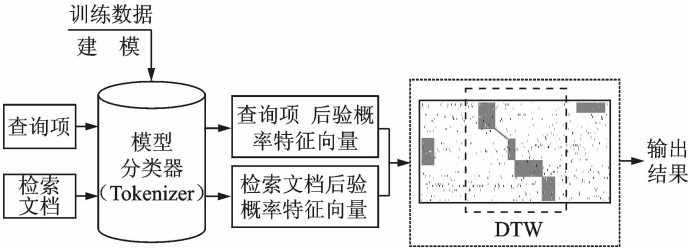

图 1 无监督语音样例检测

Fig. 1 Unsupervised query-by-example spoken term detection

其后验概率特征向量定义为

$$\boldsymbol{PG}(s)=(\boldsymbol{q}_1,\boldsymbol{q}_2,\cdots,\boldsymbol{q}_n) \tag{1}$$

$\boldsymbol{PG}(s)$反映了观测向量序列在模型分类器的 M 个类别 $\{C_1,C_2,\cdots,C_M\}$ 中的后验分布情况,每一个类别对应一种基本声学单元,其中 $\boldsymbol{q}_i$ 定义为

$$\boldsymbol{q}_i=\{P(C_1\mid\boldsymbol{f}_i),P(C_2\mid\boldsymbol{f}_i),\cdots,P(C_M\mid\boldsymbol{f}_i)\} \tag{2}$$

模型中的类别可以是任意形式的语音单元,如音素、音位属性或其他具有相似声学特征的语音片段等。后验概率特征向量可视作一种特征参数,是模型分类器对频谱参数优化处理后得到的信息,去除了冗余与噪声,具有良好的鲁棒性与区分性[11],后验概率特征向量可通过无监督的模型分类器,采用贝叶斯准则计算得到。

1.2 分段动态时间规整的检索

DTW 是无监督模板匹配框架中应用最广泛的检索技术,它将时间归整和距离测度计算相结合,通过计算查询样例与检索文档的后验概率特征向量之间的测量距离矩阵,寻找累计距离最小的一条路径,即匹配程度最高的区域。DTW 算法的检索时间较长,实际中通常采用基于 DTW 的改进算法——分段动态时间规整[6,12]算法,该算法通过增加限制条件窗长 R。将测量矩阵划分为一系列相互交叠的子矩阵,每一个子矩阵内分别进行 DTW 匹配,找出相对最优路径,最后在所有子矩阵的最优路径中选出全局最优结果,这样可以有效降低在全局的回溯时间,同时避免两个匹配子段在时域上相差过大,提高检索性能。由于处理的特征向量为后验概率,因此常将测量矩阵定义为在对数空间的内积距离[13]

$$D_{\mathrm{IP}}(\boldsymbol{x},\boldsymbol{y})=-\log(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{y}) \tag{3}$$

SDTW 算法虽然降低了路径回溯时间,但仍需大量计算,在系统的实时性方面仍存在不足。

2 声学分段模型的检索方法

为了提高检索速度,本文提出了基于 ASMs 的检索方法。ASMs 由一系列的 HMM 构成,每一个 HMM 代表了一类声学片段的分布特性,类与类之间的相互关系由一个转移列表表示,因而 ASMs 与有监督情况下训练的 HMMs 拥有相似的拓扑结构。不同在于标注条件中,音素边界信息和标注已明确给定,训练时直接将具有相同标签的片段集中训练,而在无监督中则缺乏指导信息,需通过算法获得。因此无监督建模实质是一个聚类过程,通过聚类算法将具有相同声学信息的片段归为一类,得到边界信息并添加标签,对每一类分别建立 HMM,就可以构建 ASMs。

2.1 高斯混元后验概率

本文采用高斯混元后验概率代替频谱参数训练 ASMs,这是由于后验概率特征向量的鲁棒性与区分性都优于频谱参数。GMM 是一种由多个高斯函数加权得到的多维概率密度函数,理论上语音的频谱特征分布可以由若干不同高斯分布的加权组合得到,给定一帧语音的频谱参数,可以计算每个高斯混

元后验概率，这样通过计算一帧语音的频谱参数在每一个子高斯中的后验概率表示该帧语音的特征，即高斯混元后验概率特征向量。一个 K 阶 GMM 可以表示为

$$G(\boldsymbol{x} \mid \lambda)=\sum_{i=1}^{K} w_i N_i(\boldsymbol{x} \mid \lambda_i) \qquad \sum_{i=1}^{K} w_i=1 \tag{4}$$

式中：$\boldsymbol{x}$ 为 D 维观测向量；$N_i(\boldsymbol{x} \mid \lambda_i)$ 为第 i 个高斯子分布。则 $\boldsymbol{x}$ 属于第 k 个高斯子分布的概率为

$$p(k \mid \boldsymbol{x})=\frac{N(\boldsymbol{x} \mid k, \lambda_i) w_k}{\sum_{l=1}^{K} N(\boldsymbol{x} \mid l, \lambda_i) w_l} \tag{5}$$

2.2 层次聚类的后验概率分割

对训练数据的高斯混元后验概率进行聚类分割处理，目的是找到训练语句中具有相似声学特征的语音片段边界信息，将训练语句分割成一系列不同声学特征片段，而片段内部呈现出相对稳定的特征。本文采用基于最小误差平方和(Minimum sum of squared error, MSSE)的层次聚类算法(Hierarchical agglomerative clustering, HAC)对后验概率特征向量进行分割处理。HAC 是一种常见的层次聚类方法，它自下而上对目标数据层层迭代聚类，每个目标数据被看成一个片段，每次迭代找出使 SSE 之差最小的两个连续片段，通过阈值 λ 判断是否合并(见图 2)。

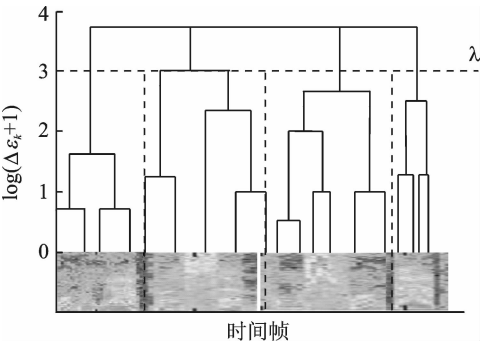

图 2 HAC 分段聚类

Fig. 2 Signal segmentation using HAC

已知训练数据的高斯混元后验概率特征向量序列 $\boldsymbol{GP}=(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N)$，$\boldsymbol{p}_j$ 为第 j 帧后验概率向量，初始时，每帧可看作一个分段。第 i 迭代得到分段边界 $\boldsymbol{B}_i=(t_1, t_2, \cdots, t_{Li})$，且 $1 \leqslant t_1 < \cdots < t_{Li} \leqslant N$，$L_i$ 为分段总数，则其 SSE 定义为

$$\varepsilon(\boldsymbol{B}_i)=\sum_{l=1}^{L_i} \sum_{j=t_{l-1}+1}^{t_l} \boldsymbol{p}_j-\boldsymbol{m}_l{}^2 \tag{6}$$

式中：$\boldsymbol{m}_l$ 为第 l 个分段的平均向量

$$\boldsymbol{m}_l=\frac{1}{t_l-t_{l-1}} \sum_{j=t_{l-1}+1}^{t_l} \boldsymbol{p}_j \tag{7}$$

在第 i 次迭代中需计算相邻两段的误差平方和，以第 k 段与第 $k+1$ 段为例，其边界分别为 $(t_{k-1}+1, t_k)$，(t_k+1, t_{k+1})，则其误差平方和为

$$se_k=\sum_{n=t_{k-1}+1}^{t_{k+1}} \| \boldsymbol{p}_n-\boldsymbol{m}_k \|^2 \tag{8}$$

式中：$\boldsymbol{m}_k$ 为 k 与 $k+1$ 两个分段的平均向量。找到使 se_k 最小的两个连续分段，使其合并为一个分段，完成此次迭代。此时的代价损失函数 $\Delta\varepsilon_i$ 最小

$$\Delta\varepsilon_i=\min \mid \varepsilon(\boldsymbol{B}_i)-\varepsilon(\boldsymbol{B}_{i-1}) \mid \tag{9}$$

阈值 λ 定义为

$$\lambda=\operatorname{mean}(\boldsymbol{M}_\Delta)+\beta \cdot \operatorname{std}(\boldsymbol{M}_\Delta) \tag{10}$$

式中：$\boldsymbol{M}_\Delta=(\Delta\varepsilon_1, \Delta\varepsilon_2, \cdots, \Delta\varepsilon_{k-1})$，$\operatorname{std}(\boldsymbol{M}_\Delta)$ 为 $\boldsymbol{M}_\Delta$ 的标准差；β 为阈值控制因子。当 $\Delta\varepsilon_i$ 超过阈值 λ 时，迭代停止，聚类结束。算法总结如下：

算法 1 基于 MSSE 的 HAC 算法

输入 $\boldsymbol{GP}=(\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots, \boldsymbol{p}_N)$，阈值 λ

输出 $\boldsymbol{B}=(t_1, t_2, \cdots, t_L)$，$1 \leqslant t_1 < \cdots < t_L \leqslant N$

计算相邻分段的误差平方和 se_k

定义堆栈列表 H,用于存放第 i 次迭代的 $(N-i)$ 个 se,选取 H 中的最小值 M
while H 非空 do
if $M<\lambda$
合并 M 代表的两个分段,更新 H 中的数值,更新 M
else
清空 H,break
end
end while

2.3 k-means 的聚类标注

聚类标注的目的是找到具有相似属性的片段集合,通过无监督的方法构建 ASMs 模型。经过 HAC 处理,已经将训练语句分割成一系列片段,接下来需要通过聚类算法,将具有相似特征信息的片段划为一类,对每一类分别建立模型,就可以得到 ASMs。现有的基于 ASMs 模型的无监督框架,在聚类标注阶段,主要采用基于 GMM 的分类算法,该方法与求取数据的高斯混元后验概率的过程类似,首先训练一个 M 阶 GMM,每一个高斯子分布对应 ASMs 中的一个基本单元,将所有的声学分段的频谱参数送入 GMM,计算与每一个高斯的相似度,就可以将其分类处理。这种方法通过计算与高斯的最大似然度进行分类,而单一高斯的区分性能不高,常需要拟合多个高斯,因此本文采用 k-means 算法[14],k-means 的性能一般要优于 GMM 算法[15],k-means 聚类算法通过计算分段边界进行聚类。k-means 需要预先设定聚类个数,这里的类对应 ASMs 模型中的基本单元。设定聚类数目为 N,聚类结束后可以人工设定各组标签为 $c_1,c_2,\cdots,c_N$,这样就完成了无监督数据的自动标注。

2.4 ASMs 模型训练和解码序列检索

得到每一类的数据后,就可以建立每类的声学模型,这里用 HMM 为每类数据建模,即 ASMs。因此 ASMs 由若干 HMM 构成,用于表征无标注训练数据的声学特征分布。各类声学片段之间的相互关系则通过一个转移列表表示。训练过程如下:

(1) 对于每一个标注为 c_i 的类声学分段类,通过最大期望 (Expectation maximization,EM)迭代算法分别训练其 HMM;统计声学片段的出现次数计算各类片段之间的转移概率。(2) 利用得到的 ASMs 对训练数据 Viterbi 解码,生成新的标注序列。(3) 利用解码的新标注序列修正边界信息与标注,重新训练 HMM 并修正转移概率列表。(4) 重复步骤(2,3)直至模型收敛。

这种方法广泛用于训练 ASMs 模型,初始标注序列通过聚类得到,而新的标注由 ASMs 解码得到,两者都通过无监督的方法得到。在训练 HMM 过程中,单个 HMM 已经收敛,增加训练 ASMs 的迭代过程对其性能提升有限。因此本文训练 ASMs 时,直接采用第 1 次解码结果。

2.5 模型相似度的最小编辑距离

得到 ASMs,可以将查询项与检索文档的高斯混元后验概率特征向量转换为混合模型后验概率向量,直接通过 SDTW 检索结果。为提高检索速度,本文采用 Viterbi 算法将查询项与检索文档转换为解码序列,采用基于最小编辑距离(Minimum edit distance, MED)的动态匹配(Dynamic match lattice spotting, DMLS)[16]算法。DMLS 是一种基于 Lattice 的动态匹配技术,能够快速准确地实现关键词检测,本文将该方法引入基于无监督解码序列的检索,能够有效提升检索速度。

检索时需计算查询项与检索文档序列之间的距离,MED 包含匹配、替换、插入和删除 4 种错误,常见的主要错误是替换错误,因此本文 MED 计算仅考虑替换错误代价。在有监督中预先设定执行相应操作的原始代价均为 1,通过先验知识或者混淆矩阵对其进行修正。而在无监督条件下则缺乏指引信息,因此本文通过计算模型相似度距离矩阵修正 MED 的代价函数矩阵,模型距离越近,混淆的可能性越大。

定义 A,B 为 HMM 中具有相同高斯拓扑结构的两个状态,它们之间的距离为

$$\mathrm{dist}(A,B)=\sum_{i=j=1}^{H} w_{Ai}w_{Bj}d_{A,B}(i,j) \tag{11}$$

式中:H 为状态 A,B 中的高斯个数;w_{Ai} 为状态 A 中第 i 个高斯的权重;$d_{A,B}(i,j)$ 为两个多维高斯的距离。常用的距离计算方法有 K-L 距离和 Bhattacharyya 距离

$$d_{\mathrm{K\text{-}L}}(i,j)=\frac{1}{2}(\boldsymbol{u}_i-\boldsymbol{u}_j)^{\mathrm{T}}(\boldsymbol{\Sigma}_i^{-1}+\boldsymbol{\Sigma}_j^{-1})^{-1}(\boldsymbol{u}_i-\boldsymbol{u}_j)+\frac{1}{2}\mathrm{tr}[\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j+\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j^{-1}-2\boldsymbol{I}_d] \tag{12}$$

$$d_{\mathrm{Bhat}}(i,j)=\frac{1}{8}(\boldsymbol{u}_i-\boldsymbol{u}_j)^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right)^{-1}(\boldsymbol{u}_i-\boldsymbol{u}_j)+\frac{1}{2}\ln\frac{\left|\dfrac{\boldsymbol{\Sigma}_i+\boldsymbol{\Sigma}_j}{2}\right|}{|\boldsymbol{\Sigma}_i|^{1/2}|\boldsymbol{\Sigma}_j|^{1/2}} \tag{13}$$

式中:u_i,$\boldsymbol{\Sigma}_i$ 为高斯的均值和协方差矩阵;$\boldsymbol{I}_d$ 为单位矩阵。对于拓扑结构相同的两个 HMM 模型 H_1 和 H_2,其模型距离为

$$\mathrm{dist}(H_1,H_2)=\frac{1}{S}\sum_{i=1}^{S}\mathrm{dist}(H_{1i},H_{2i}) \tag{14}$$

式中:S 为模型中状态数。得到模型相似度矩阵后,需定义阈值修正距离矩阵,进而得到代价函数。本文定义阈值 δ,其含义为在第 i 个模型中,除模型本身外与其距离最近的 δ 个模型,这些模型构成了一个集合 Φ_i,因此替换代价矩阵定义在给定观察值 i 的条件下,替换为 j 的代价为

$$C(i,j)=\begin{cases}0 & j=i\\ 1 & j\neq i,\ j\in\Phi_i\\ 2 & j\neq i,\ j\notin\Phi_i\end{cases} \tag{15}$$

3 实验结果及分析

3.1 实验配置与评价指标

实验采用 TIMIT 语料库,选择 TRAIN 中 3 696 个语句作为训练集,首先提取训练集的 39 维 MFCCs 参数训练 GMM,GMM 模型具有 50 个高斯分布。得到 GMM 后,利用该模型处理训练集的 MFCCs 参数,得到其高斯混元后验概率,对后验概率进行 HAC 分段处理,找到边界信息,得到一系列声学分段,阈值控制因子 $\beta=0.15$[17];对所有声学分段聚类处理,k-means 中的聚类个数设置为 50,对应着 ASMs 中的类别数。每一类分别构建一个从左到右的 5 个状态、32 个高斯数的 HMM,进而得到 ASMs 模型。选择 TEST 中 1 344 个语句作为测试集,并从测试集中选取 10 个词作为查询样例。

本文采用信息检索领域常见的评价指标评测提出的方法:(1) 平均正确率均值(Mean average precision, MAP),用于描述检索精度;(2)平均检索时间,每个查询项的评价检索时间用于描述检索速度。

3.2 系统性能比较

实验首先比较了基于 SDTW 检索的 3 种不同模型的无监督 QbE-STD 系统性能(见表 1)。采用相同的训练数据分别训练了 GMM、基于频谱参数的 ASMs (MFCC+ASMs)和基于高斯混元后验概率的 ASMs (GMM+ASMs),并将检索文档与查询项转换为相应后验概率特征向量矩阵,通过 SDTW 检索算法比较三者的 MAP 与平均检索时间。其中 GMM 由 39 维 MFCCs 训练得到,高斯个数为 50;ASMs 的模型数为 50,与 GMM 中的高斯个数相同,每一个 HMM 由 5 个状态、32 个高斯数组成。与 GMM 相比,MFCC+ASMs 在 MAP 上提升 7.6%,这一结果验证了 ASMs 模型的性能优于 GMM,这是由于 GMM 是基于帧级数据训练而成,忽略了相邻数据帧之间的关系,而 ASMs 是在声学分段基础上训练得到,增加了时间信息。

表 1 SDTW 检索的 3 种方法性能比较

Tab. 1 Performances of SDTW algorithm in alternative tokenizers

模型	MAP	时间/s
GMM	0.407	263.624
MFCC+ASMs	0.438	268.748
GMM+ASMs	0.481	266.948

本文提出的基于后验概率的 GMM+ASMs 模型性能优于 GMM 和 MFCC+ASMs 两种模型。与 GMM 和 MFCC+ASMs 模型相比,GMM+ASMs 在 MAP 上分别提升 18.2%和 9.8%。基于后验概率的 GMM+ASMs 集中了 GMM 与 ASMs 两种模型的优点,通过训练 GMM 得到训练集的高斯混元后验概率,与频谱参数相比,后验概率的鲁棒性与区分性更好,利用后验概率训练的 ASMs,其性能较传统的 ASMs 有明显提升,而三者的检索时间差异不大,但难以胜任实时性需求。

基于 SDTW 的检索算法需要计算后验概率的测量距离矩阵并回溯最优路径,耗费大量的时间。表 2 所示为针对 ASMs 的解码序列,采用基于 MED 的动态匹配算法检索查询项,其中的惩罚代价函数设定初值为 1。结果表明:针对 ASMs 解码序列的动态匹配,查询项的平均检索时间从几百秒降至零点几秒,有效地提升了检索速度。这是由于该方法将原有的大数据量矩阵运算转换为基于字符串匹配的过程,省去了大数据的运算,但同时,检索精度与 SDTW 方法相比,也有一定程度的下降,MFCC+ASMs 的 MAP 为 0.340,GMM+ASMs 的 MAP 为 0.378,对解码序列作平滑处理,剔除孤立点后再采用 DMLS 方法,性能得到一定的提升,但仍低于 SDTW 方法。这主要是由于 MED 的代价函数缺乏指导信息,原始代价设定过于严格,因此本文通过计算模型相似度修正代价函数。

表 3 所示为 MED 代价阈值 δ 对检索性能的影响,表中的解码序列经过了平滑处理,分别采用通过两种距离矩阵修正的代价函数进行检索,其中 $\delta=0$ 表示直接以距离矩阵作为代价函数,其余项表示通过 δ 修正后的代价函数。结果表明:本文提出的基于模型相似度的 MED 检测能够有效提升检索性能,对于 ASMs 平滑后的解码序列,在 $\delta=4$ 时,采用 K-L 距离时 MAP 相对提升 10.2%,在 $\delta=2$ 时,采用 Bhattacharyya 距离时 MAP 提升 13.7%;改用高斯混元后验概率训练的 ASMs 得到的解码序列,在 $\delta=4$ 时,采用 Bhattacharyya 距离时 MAP 提升 5.5%。

表 2 DMLS 检索性能比较

Tab. 2 Performances of in alternative tokenizers

算法	模型	MAP	时间/s
DMLS	MFCC+ASMs	0.340	0.333
	GMM+ASMs	0.378	0.337
平滑+DMLS	MFCC+ASMs	0.343	0.335
	GMM+ASMs	0.396	0.336

表 3 δ 对检索性能的影响

Tab. 3 Performances comparison under various δ

		δ	0	1	2	3	4	5	6	7
MFCC+ASMs (0.343)	K-L	MAP	0.302	0.346	0.347	0.361	0.378	0.372	0.371	0.361
		时间/s	0.423	0.423	0.422	0.422	0.422	0.423	0.425	0.423
	Bhat	MAP	0.358	0.381	0.389	0.360	0.358	0.347	0.325	0.341
		时间/s	0.427	0.422	0.430	0.425	0.423	0.424	0.425	0.422
GMM+ASMs (0.396)	K-L	MAP	0.260	0.397	0.379	0.367	0.375	0.370	0.340	0.342
		时间/s	0.424	0.425	0.424	0.428	0.425	0.425	0.426	0.422
	Bhat	MAP	0.336	0.395	0.404	0.406	0.418	0.415	0.407	0.403
		时间/s	0.421	0.423	0.423	0.427	0.424	0.423	0.423	0.423

与 SDTW 方法相比,经过模型相似度测量矩阵修正后的最优结果(0.418)已优于传统的 GMM 方法(0.407),且检索时间大幅度降低(由 263.624 s 降至 0.424 s),但与 MFCC+ASMs 和本文提出的 GMM+ASMs 相比在检索精度上还存在一定差距(0.438,0.481),下降的原因是 ASMs 中的一个 HMM 代表了一类基本语音单元,但与实际的音素相比还存在一定差距,事实上一个音素通常由 ASMs 中的多个 HMM 拟合构成,而解码后结果只有一个,因此导致检索精度的下降。

4 结束语

本文提出一种基于后验概率 ASMs 解码序列的检索方法,首先利用高斯混元后验概率特征向量训

练 ASMs 模型，相比频谱参数，后验概率的鲁棒性与区分性更为良好，能较好地提升 ASMs 性能；同时利用该模型对数据解码得到解码序列，通过基于 MED 的动态匹配方法检索查询项解码序列，MED 的代价函数由 ASMs 中 HMM 之间的相似度距离修正。本文方法有效解决了基于后验概率和 SDTW 方法检索速度慢的不足，保证了检索系统的实时性要求。相比 SDTW 下各种模型识别器，本文方法优于 GMM 但低于 ASMs，通过模型相似度构建的代价函数与有监督混淆矩阵相比还存在一定差距，因此后续的研究可以融合解码序列和后验概率，进一步提高检索精度。

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: [s. n.], 2009: 2143-2146.

[2] Chelba C, Hazen T J, Saraclar M. Retrieval and browsing of spoken content [J]. IEEE Signal Processing Magazine, 2008, 25 (3): 39-49.

[3] Jansen A, Dupoux E, Goldwater S. A summary of the 2012 JHU CLSP workshop on zero resource speech technologies and models of early language acquisition [C]// ICASSP 2013. Vancouver, Canada: [s. n.], 2013: 8111-8115.

[4] Park A S, Glass J R. Unsupervised pattern discovery in speech [J]. IEEE Transactions on Audio, Speech and Language Processing, 2008, 16(1): 186-197.

[5] Hazen T J, Shen W, White C. Query-by-example spoken term detection using phonetic posteriorgram templates [C] // Automatic Speech Recognition and Understanding 2009. Merano/ Meran, Italy: [s. n.], 2009: 421-426.

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

[7] Wang H P, Lee T, Leung C. Unsupervised spoken term detection with acoustic segment model [C]// Int Conf Speech Database and Assessments. Hsinchu, China: [s. n.], 2011: 106-111.

[8] Wang H P, Leung C, Lee T, et al. An acoustic segment modeling approach to query-by-example spoken term detection [C]// ICASSP 2012. Kyoto, Japan: [s. n.], 2012: 5157-5160.

[9] Zhang Y D, Glass J. A piecewise aggregate approximation lower-bound estimate for posteriorgram-based dynamic time warping[C]// Interspeech 2011. Florence, Italy: [s. n.], 2011: 1909- 1912.

[10] Chan Chunan, Lee Linshan. Model-based unsupervised spoken term detection with spoken queries [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1330-1342.

[11] Aradilla G, Vepa J, Bourlad H. Using posterior-based features in template matching for speech recognition [C] // Interspeech 2006. Pittsburgh, Pennsylvania, USA: [s. n.], 2006: 1186-1189.

[12] 冯志远，张连海. 基于分段动态时间规整的语音样例快速检索 [J]. 数据采集与处理，2014，29(2)：265-273.
Feng Zhiyuan, Zhang Lianhai. Fast query-by-example spoken term detection using segmental dynamic time warping [J]. Journal of Date Acquisition and Processing, 2014, 29(2): 265-273.

[13] Michael A C, Thomas S, Jansen A, et al. Rapid evaluation of speech representations for spoken term discovery[C]// Interspeech 2011. Florence, Italy: [s. n.], 2011: 821-824.

[14] Mantena G, Anguera X. Speed Improvements to information retrieval-based dynamic time warping using hierarchical K-means clustering [C]// ICASSP 2013. Vancouver, Canada: [s. n.], 2013: 8515-8519.

[15] Wang H P, Lee T, Leung C. Unsupervised mining of acoustic subword units with segment-level Gaussian posteriorgrams [C]// Interspeech 2013. Lyon, France: [s. n.], 2013: 2297-2301.

[16] Thambirtnam K, Sridharan S. Rapid yet accurate speech indexing using dynamic match lattice spotting [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(1): 346-357.

[17] Chan Chunan, Lee Linshan. Unsupervised spoken term detection with spoken queries using segment-based dynamic time warping [C]// Interspeech 2010. Makuhari, Chiba, Japan: [s. n.], 2010: 693-696.

作者简介：

李勃昊(1989-)，男，硕士生，研究方向：无监督语音关键词检测，E-mail: beichenyouxiang@163.com。

张连海(1971-)，男，副教授，研究方向：语音信号处理和模式识别，E-mail: lianhaiz@sina.com。

郑永军(1984-)，男，硕士研究生，研究方向：语音识别和模式识别，E-mail: banjiu123cool@163.com。