ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 6,Nov. 2017,pp. 1198－1207
DOI:10.16337/j.1004-9037.2017.06.015

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

基于 NMF 后验特征优化的语音查询样例检测

曹建凯　张连海　李弼昊

(信息工程大学信息系统工程学院,郑州,450001)

摘　要:提出一种基于非负矩阵分解(Nonnegative matrix factorization, NMF)后验特征优化和修正分段动态时间规整(Segmental dynamic time warping, SDTW)检索的无监督语音查询样例检测方法。该方法首先应用频域线性预测(Frequency domain linear prediction, FDLP)声学特征参数代替梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCCs)训练高斯混合模型(Gaussian mixture model, GMM)模型,然后使用 NMF 算法对高斯后验特征矩阵进行分解,将得到的基矩阵作为子空间变换矩阵对原始后验特征投影,投影可以突出特征中主要分量,平滑距离矩阵。在检索阶段,使用多相邻输出得分对最佳匹配得分进行修正,用于代替标准 SDTW 算法的 1-best 输出得分。实验结果表明,在不增加检索时间的情况下,该方法相比应用 MFCCs 和 FDLP 特征的基线系统性能提升明显,检索精度分别相对提升了 18.6%和 18.1%。

关键词:无监督;查询样例检测;后验特征;非负矩阵分解优化;修正分段动态时间规整

中图分类号:TP391　　**文献标志码**:A

Posteriorgram Features Optimization for Query-by-Example Spoken Term Detection Based on NMF

Cao Jiankai, Zhang Lianhai, Li Bohao

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

Abstract: This paper presents the study of posteriorgram features optimization based on nonnegative matrix factorization (NMF) algorithm and modified segmental dynamic time warping (SDTW) detection for unsupervised query-by-example spoken term detection. First, a Gaussian mixture model (GMM) is trained with frequency domain linear prediction (FDLP) acoustics feature parameters instead of Mel-frequency cepstral coefficients (MFCCs). Then the NMF algorithm is applied to the generated Gaussian posteriorgram matrix, and the derived base matrix is used as a subspace transform matrix for projection of raw feature. The projection can highlight the primary component of features and smooth the distance matrix. In the detecting phase, the best matching score is modified by using multi adjacent output scores, instead of the 1-best output score for normal SDTW. Experimental results show that without affecting detection time, the proposed method consistently outperforms the baseline systems with MFCCs and FDLP features with the detection precision improved by 18.6% and 18.1% respectively.

Key words: unsupervised; query-by-example spoken term detection; posterior feature; nonnegative matrix factorization (NMF) optimization; modified segmental dynamic time warping (SDTW)

基金项目:国家自然科学基金(61673395,61403415,61302107)资助项目。
收稿日期:2016-03-31;**修订日期**:2016-06-19

引 言

语音查询样例检测(Query-by-example spoken term detection, QbE-STD)系统能从语音文档中自动定位出查询语音所出现的位置[1]。与传统依赖于大词汇量连续语音识别(Large vocabulary continuous speech recognition, LVCSR)引擎的有监督口语项检测系统(Spoken term detection, STD)系统不同,作为一种无监督系统,其检索过程中不涉及到语音识别,因此它不需要代价高昂的标注语料,同时也避免了集外词(Out of vocabulary, OOV)的问题。当前,QbE-STD 技术已成为低资源、零资源领域的一个研究热点,并且广泛应用于音乐检索、口语文档检索以及语音监听、控制[2-3]等场景。

目前无监督 QbE-STD 系统主要有两种方法:(1)基于模板匹配的方法[4],主要通过一个符号化器[5-7](Tokenizer)将查询样例和测试语句转化为后验特征,然后采用动态时间规整(Dynamic time warping, DTW)算法检测出匹配区域。(2)基于模型的方法,主要采用模式发现技术,为每个模式建立模型,采用语音识别技术将语音转化成模式序列,然后进行基于符号的快速检索[8-9]。前者在检索时存在大量的矩阵运算,因此检索速度慢;后者由于所发现的模式易受说话人、说话环境的影响,缺乏有监督指导,因此检索精度较低。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。

当前提高基于模板匹配的 QbE-STD 系统检索精度的主要方向有:(1)特征层面,选择更具鲁棒性的特征;(2)检索层面,设计高效精准的搜索算法。

在特征层方面,传统符号化器如高斯混合模型(Gaussian mixture model,GMM)是以梅尔频率倒谱系数(Mel-frequency cepstral coefficients,MFCCs)作为输入特征的[5],缺少语音上下文信息,鲁棒性较差。对此,文献[10]提出频域线性预测(Frequency domain linear prediction,FDLP)特征,该特征能够反映语音的长时特性,因而能够捕获语音信号的时域动态属性,可以代替传统频谱参数(如 MFCCs)用于训练声学模型。文献[11]验证了基于 FDLP 估计的高斯后验特征性能要优于基于 MFCC 估计的高斯后验特征。高斯后验特征矢量表示一帧数据属于各个声学单元的后验概率,然而零资源条件下,由于缺乏先验知识,声学单元个数无法预先定义,对声学单元所建的模型之间区分度不足,因此由符号化器生成的后验概率特征会存在一定的冗余信息,此外说话人、说话环境也会对后验概率特征产生干扰,因此有必要对后验概率特征进行优化来提高其鲁棒性。对此,可以应用传统的主成分分析(Principal component analysis,PCA)算法[12]对后验特征作进一步处理,将得到的新特征用于后续的检索,结果表明系统检索精度得到一定的提升。

在检索层方面,文献[5]所提出的分段 DTW(Segmental DTW,SDTW)算法通过一个移动窗限制可以有效降低检索时间,成为当前基于模板匹配的 QbE-STD 系统的主流检索算法。然而该算法一个潜在的缺点是:移动窗有可能分割测试语句中的候选区域,导致匹配得分不能达到全局最优。当前 SDTW 算法输出得分形式为 1-best,而对于连续子段,若其中存在真实匹配子段,则其相邻子段应包含大部分查询样例中的声学单元,对应的失真得分应该随着与最佳子段的距离增加而快速增大。如果考虑最佳匹配子段的相邻子段得分,可以弥补该算法缺陷。

针对基于模板匹配的 QbE-STD 系统检索精度低的问题,本文从前端特征和后端检索两个方面进行了改进,特征层使用后验概率特征进行样例查询,考虑到概率的非负性,本文提出使用非负矩阵分解(Nonnegative matrix factorization,NMF)算法对后验特征矩阵进行分解,用得到的基矩阵作为子空间变换矩阵,然后将原始后验特征投影得到新的特征,并用于后续检索。鉴于 SDTW 检索输出结果为 1-best,在检索层本文提出多相邻输出的修正 SDTW 算法,通过对相邻输出得分进行加权,来改善移动窗分割候选区域的缺陷。

1 系统框架

系统框图如图 1 所示。首先使用 FDLP 代替传统的 MFCCs 作为声学特征训练 GMM 符号化器，然后采用 NMF 算法对生成的高斯后验特征矩阵分解，得到降噪优化后的特征，在该特征上使用修正的 SDTW 算法进行检索，最后给出检索结果。

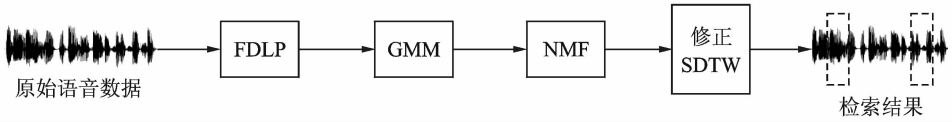

图 1 基于 GMM 符号化器的 QbE-STD 系统框图

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

1.1 FDLP 特征提取

传统基于短时傅里叶变换(Short time Fourier transform，STFT)的特征(如 MFCCs)是对短时帧(大约 20 ms)进行分析，而语音信号中的内容信息分布于长达 200 ms 甚至更长的长时上下文中(如音素时长大约 70～80 ms)。文献[10]提出在长时片段中(1 s 甚至更长)，将全带语音信号划分成若干子带，在每一子带构建 AM-FM 模型进行调制分析，得到子带包络，然后对所有子带包络进行整合加窗并做离散余弦变换(Discrete cosine transform，DCT)，得到 FDLP 特征，并用于各种语音应用中(如语音识别、说话人识别等)。FDLP 特征提取框架如图 2 所示，

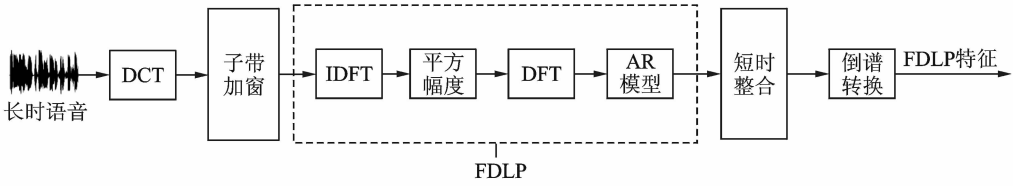

图 2 FDLP 特征提取流程框架

Fig. 2 FDLP feature extracting framework

图 2 中虚线框即是 FDLP 处理过程，具体提取流程如下：

(1)使用 DCT 对语音信号的长时片段(1 s 或者更长)进行分析。

(2)在 DCT 域上对信号进行加窗划分成子带，考虑到人耳的听觉特性，可以在 Mel 域上设计窗的形式。

(3)对所有子带信号作频域线性预测，获得子带能量包络。

(4)将规整后的 FDLP 包络进行整合，然后应用短时窗(25 ms 窗长，10 ms 窗移)进行划分，得到短时能量特征。

(5)通过应用对数和 DCT 变换将短时能量表征转化为倒谱特征，该过程类似于 MFCC 特征的获取。

(6)将提取到的 13 维倒谱系数，以及其一阶和二阶差分分量，组成 39 维特征矢量，称为 FDLP 特征，用于表示该帧语音信号。

1.2 GMM 符号化器

使用无标注的训练语料训练一个包含 K 个高斯分量的 GMM，则

$$G(\boldsymbol{f} \mid \lambda)=\sum_{i=1}^{K}\omega_i N_i(\boldsymbol{f} \mid \lambda_i) \tag{1}$$

$$N_i(\boldsymbol{f} \mid \lambda_i)=\frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma}_i \mid^{1/2}}\exp\left\{-\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{f}-\boldsymbol{\mu}_i)\right\} \tag{2}$$

式中:$\boldsymbol{f}$ 为 D 维声学特征向量,这里使用 39 维的 MFCCs 或者 FDLP;ω_i 为第 i 个高斯分量的权重因子,总和为 1;$\lambda_i=\{\mu_i,\boldsymbol{\Sigma}_i\}$,$i=1,\cdots,K$ 为模型参数集,其中 $\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Sigma}_i$ 分别为第 i 个高斯分量的均值向量和协方差矩阵;$N_i(f|\lambda_i)$为第 i 个高斯分量的分布。使用 K-Means 算法对模型初始化,然后采用基于最大似然估计(Maximum likelihood,ML)的期望最大化(Expectation maximization,EM)算法迭代训练,直至模型参数收敛。依据文献[13],这里设置高斯分量个数 K 为 50。

模型训练完成后,GMM 符号化器对每一帧语音数据生成一个 50 维后验概率特征矢量。假设语音信号序列为 $S=(\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_N)$,其中 N 表示语音帧数量,$\boldsymbol{f}_i$ 表示第 i 帧声学特征,它的后验概率特征矢量 $\boldsymbol{PG}_{f_i}$ 反映了其在 K 个类别$\{C_1,C_2,\cdots,C_K\}$上的分布,该特征向量为

$$\boldsymbol{PG}_{f_i}=[p(C_1 \mid \boldsymbol{f}_i),p(C_2 \mid \boldsymbol{f}_i),\cdots,p(C_K \mid \boldsymbol{f}_i)] \tag{3}$$

式中 $p:(C_j|\boldsymbol{f}_i)$为第 i 帧属于第 j 类的后验概率,计算公式为

$$P(C_j \mid \boldsymbol{f}_i)=\frac{N_j(\boldsymbol{f}_i \mid C_j,\lambda_j)\omega_j}{\sum_{k=1}^{K}N_k(\boldsymbol{f}_i \mid C_k,\lambda_k)\omega_k} \tag{4}$$

1.3 NMF 算法介绍

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,其物理意义是“局部构成整体”[14]。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。对于高斯后验概率矢量,可以认为语音中的每帧声学单元(如音素)都是由其中某一高斯分量生成,而矢量中的每一元素即表示其所代表的高斯分量的概率。利用该假设,同时由于概率是非负的,可以考虑对语音的后验特征矩阵进行非负分解。

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,则可以将矩阵看成含加性噪声的线性混合体模型,则

$$\boldsymbol{X}_{n\times m}=\boldsymbol{W}_{n\times r}\boldsymbol{H}_{r\times m}+\boldsymbol{E}_{n\times m} \tag{5}$$

式中:$\boldsymbol{W}_{n\times r}$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。

依据重建误差均方最小原则,求解分解因子 $\boldsymbol{W}$ 和 $\boldsymbol{H}$,则

$$L_{\mathrm{ED}}(\boldsymbol{W},\boldsymbol{H})=\sum_{ij}[\boldsymbol{X}_{ij}-(\boldsymbol{WH})_{ij}]^2 \tag{6}$$

采用梯度下降法作为迭代算法,则

$$\begin{aligned}\frac{\partial L_{\mathrm{ED}}}{\partial \boldsymbol{W}_{ik}}&=-2[(\boldsymbol{XH}^{\mathrm{T}})_{ik}-(\boldsymbol{WHH}^{\mathrm{T}})_{ik}]\\ \frac{\partial L_{\mathrm{ED}}}{\partial \boldsymbol{H}_{kj}}&=-2[(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})_{kj}-(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH})_{kj}]\end{aligned} \tag{7}$$

可以得到的乘性迭代规则为

$$\begin{aligned}\boldsymbol{W}_{ik}&\leftarrow\boldsymbol{W}_{ik}\frac{(\boldsymbol{XH}^{\mathrm{T}})_{ik}}{(\boldsymbol{WHH}^{\mathrm{T}})_{ik}}\\ \boldsymbol{H}_{kj}&\leftarrow\boldsymbol{H}_{kj}\frac{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X})_{kj}}{(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{WH})_{kj}}\end{aligned} \tag{8}$$

可以看出，由于初始化的系数矩阵与基矩阵是非负的，所以每次乘性迭代过程都能保证两个矩阵的非负性。使用 NMF 矩阵对后验特征矩阵进行分解后，可以有不同的处理方法，一是将权重矩阵作为特征代替原后验矩阵。但考虑到语音具有一定的连续性，并且具有层级结构，而在分解过程中没有考虑到这些性质，因此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，则

$$\boldsymbol{Z}_{r\times m}=\boldsymbol{W}_{n\times r}^{\mathrm{T}}\boldsymbol{X}_{n\times m} \tag{9}$$

式中：$\boldsymbol{Z}_{r\times m}$ 为投影后的特征矩阵，这里设置 $r=n$，即投影后的特征矢量维度保持不变。原因是考虑到 GMM 模型中，每一高斯分量看作一种声学单元（如音素），对于每一语音帧的后验特征矢量，表示的是该帧属于各个声学单元的概率分布，矢量中的每一维表示该帧属于这个维度所代表的声学单元的概率，因此后验特征矢量具有明确的物理意义。投影后的特征矢量，尽管不再表示概率，但依然是非负的，表示的还是该帧属于这个维度所代表的声学单元的程度。如果增加或者降低特征维度，将会改变此物理意义。

假设以 $\boldsymbol{x}_{n\times 1}$ 表示查询样例中的一帧后验特征矢量，$\boldsymbol{y}_{n\times 1}$ 表示测试语句中的一帧特征矢量，选择点积作为特征矢量之间的距离测度，则两个原始后验特征矢量之间距离 d_{raw} 为

$$d_{\mathrm{raw}}=-\log(\boldsymbol{x}_{n\times 1}^{\mathrm{T}}\boldsymbol{y}_{n\times 1}) \tag{10}$$

采用基矩阵 $\boldsymbol{W}_{n\times r}$ 对后验特征矢量投影后，则由式(9)可得投影后的两个特征矢量之间的距离 d_{proj} 为

$$d_{\mathrm{proj}}=-\log((\boldsymbol{W}_{n\times r}^{\mathrm{T}}\boldsymbol{x}_{n\times 1})^{\mathrm{T}}\boldsymbol{W}_{n\times r}^{\mathrm{T}}\boldsymbol{y}_{n\times 1})=-\log(\boldsymbol{x}_{n\times 1}^{\mathrm{T}}\boldsymbol{V}_{n\times n}\boldsymbol{y}_{n\times 1}) \tag{11}$$

式中：$\boldsymbol{V}_{n\times n}=\boldsymbol{W}_{n\times r}\boldsymbol{W}_{n\times r}^{\mathrm{T}}$，图 3 是矩阵 $\boldsymbol{V}_{n\times n}$ 的三维示意图，可以看出它是一个类对角矩阵（即除了对角线，其他矩阵元素几乎为零）。由此，式(11)可以近似为

$$d_{\mathrm{proj}}\approx-\log\Big(\sum_{i=1}^{n}\boldsymbol{V}_{i,i}\boldsymbol{x}_{i,1}\boldsymbol{y}_{i,1}\Big) \tag{12}$$

式中：$\boldsymbol{V}_{i,i}$ 为 $\boldsymbol{V}_{n\times n}$ 中第 i 个对角元素，$\boldsymbol{x}_{i,1}$ 和 $\boldsymbol{y}_{i,1}$ 分别为 $\boldsymbol{x}_{n\times 1}$ 和 $\boldsymbol{y}_{n\times 1}$ 中第 i 个元素。而对 $1\leqslant i\leqslant 50$，均有 $\boldsymbol{V}_{i,i}>1$（事实上 $\boldsymbol{V}_{i,i}$ 中的最小值为 298.1），故可以得出结论：$d_{\mathrm{proj}}<d_{\mathrm{raw}}$，这一结论可以扩展距离矩阵中元素值的下限范围。

同时从图 3 可以看出矩阵 $\boldsymbol{V}_{n\times n}$ 对角线元素相互之间并不相等，如果将 $\boldsymbol{V}_{n\times n}$ 中的对角线元素看作权

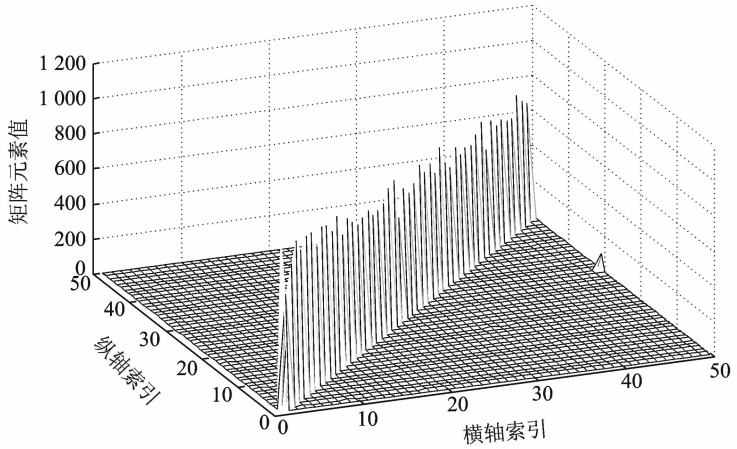

图 3 矩阵 $\boldsymbol{V}$ 三维示意图

Fig. 3 Three-dimensional diagram of Matrix $\boldsymbol{V}$

重,则可以看出式(12)中各个被加权项 $\boldsymbol{x}_{i,1}\boldsymbol{y}_{i,1}$ 在整个距离测度中所占比重不同,而这也正是采用 NMF 算法对后验概率特征的影响之处,即通过对后验特征矩阵进行非负分解,得到基矩阵,然后借助基矩阵改变后验特征矢量各个维度在点积距离测度之中的比重关系,由此可以突出后验概率特征中的主要分量,降低冗余,达到对后验特征矢量去噪优化的目的。

以 TIMIT 数据为例,选取"Popular"作为查询样例,TEST 中的"MDBB0_SI565.WAV"文件作为测试语音(该测试语音的转录文本包含"Popular"),图 4 描述了 TIMIT 语料库中查询样例"Popular"和测试语句"MDBB0_SI565.WAV"投影前后距离矩阵的变化示意图,其中图 4(a)表示投影前,图 4(b)表示投影后,图 4 中的横轴表示测试语句帧索引,纵轴表示查询样例帧索引。可以看出,投影后帧之间的最小距离量级从 5 降到了−5,除去矩阵中的无穷大值,元素之间的差异范围从[5,35]扩展到[−5,35]。而且投影前距离矩阵中出现了许多无穷大值的元素(左图中的纯白色区域),在动态规划检索中,无穷大值所在的区域会武断地直接跳过,这会增加漏警率。相比之下,图 4(b)的无穷大值的元素减少了许多,取而代之的是一些值相对较大的元素(右图中对应的浅色区域),这体现了 NMF 算法对距离矩阵的平滑作用。

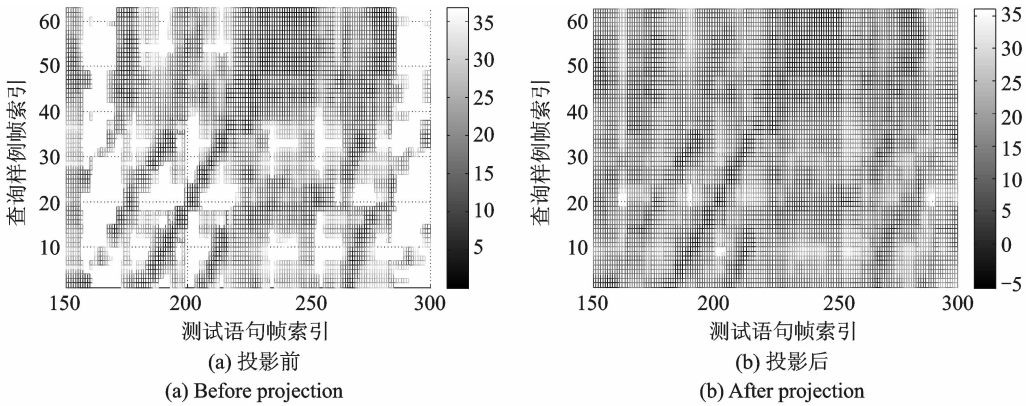

图 4 投影前后距离矩阵变化

Fig. 4 Distance matrix changing before and after projection

综上所述,NMF 算法对后验特征的作用体现在:(1)扩展了距离矩阵中元素值的动态范围。(2)借助基矩阵改变后验概率特征矢量各元素对距离计算中的比重关系,突出主要分量,实现对后验特征的降噪优化。(3)对原始基于后验特征矢量的距离矩阵作平滑处理,减少矩阵中无穷大元素值的数量,有助于降低漏警率。注意到以上 3 点的前提均是以点积作为特征矢量之间的距离测度。

1.4 检索算法

动态时间规整的目的是寻找两个序列之间的最佳对齐,在语音应用中,可以有效解决语速多变的问题。当前 DTW 算法及其变体是基于模板的 QbE-STD 系统的主流检索算法。

1.4.1 分段 DTW 检索

当测试语句持续时间较长时,传统 DTW 会存在大量的冗余计算和匹配,导致检索速度慢。对此文献[5]提出一种 DTW 算法的变体 SDTW。SDTW 是对 DTW 的一种修正,目标是寻找两个输入语句(以特征矢量序列表示)之间的多个局部对齐,它将检索语句通过一个移动窗划分为一系列子片段,然后在每个片段应用 DTW 检索,在所有片段的相对最优路径中确定最终结果。

SDTW 在搜索时定义了两个限制[13]。

(1)调节窗限制 R,避免匹配时出现较大的时差。给定两个特征矢量序列 A 和 B,长度分别为 n 和 m,选择点积作为向量之间的距离测度[15]。假如在 $n\times m$ 测量矩阵上的规整函数为 $D(\cdot)=(i_k,j_k)$,其

中(i_k, j_k)为规整路径上第 k 个坐标，则调节窗限制条件定义为

$$|(i_k - i_1) - (j_k - j_1)| \leqslant R \tag{13}$$

(2)定义动态规划搜索起点坐标的步长。如果固定一条规整路径的起点坐标，那么调节窗条件同样限制了规整路径终点坐标的范围。通过设置规整过程中不同的起点坐标，测量矩阵可以被划分成几个宽度为 $2R+1$ 的连续对角区域。

1.4.2 修正 SDTW 检索算法

标准 SDTW 算法的一个潜在缺点是：移动窗有可能分割测试语句中的候选区域，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句对，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，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。如果考虑最佳匹配子段的相邻子段得分，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。

选取“Popular”作为查询样例，TEST 集中的“MDBB0_SI565.WAV”作为测试语音，搜索算法采用标准 SDTW。表 1 描述了其各子段的失真得分，其中左列是子段序列索引，右列是子段匹配得分(得分越小，表示两段数据越匹配)。

表 1 查询样例“Popular”与测试语句“MDBB0_SI565.WAV”各子段 DTW 失真得分

Tab. 1 Segments distortion scores of DTW between query “popular” and sentence “MDBB0_SI565.WAV”

段数	失真得分
59	157.09
60	−320.32
61	−415.01
62	−397.79
63	−105.31
64	193.64

由表 1 可以看出，第 61 个子段为最佳匹配子段，而其上下两个相邻若干子段的失真得分也比较小，并且随着距离的增加呈现快速上升趋势。当前 SDTW 检索输出结果为 1-best，即只输出最佳匹配子段得分，如果能够考虑最佳子段的相邻若干子段得分，可以使匹配得分更具区分性，降低虚警率。由此本文对每一查询样例测试语句对输出最佳匹配子段得分以及其左右两边各 K 个连续子段得分(本文设置 K 值为 2)，在该 $2K+1$ 个得分内选出 M(本文设置 M 值为 3)个最佳得分进行加权得到最终得分，即

$$F = (1-\omega)F_{\text{best}} + \omega(1-\omega)F_{\text{best}}^{-1} + \omega^2 F_{\text{best}}^{-2} \tag{14}$$

式中：ω 为权重因子(取值在 0 到 0.5 之间)，F_{best} 为最佳匹配子段得分，F_{best}^{-1} 和 F_{best}^{-2} 分别为相邻的第 2 和第 3 最佳匹配子段得分，F 为该查询样例测试语句对的最终得分。

2 实验结果及分析

2.1 实验配置

本文采用 TIMIT 语料库[16]进行 QbE-STD 实验。它共有 6 300 个语句，分为 TRAIN 和 TEST 两个集合。本文实验选择 TRAIN 中 3 296 个语句作为训练集，选择 TEST 中 1 344 个语句作为测试集(实验未采用适合于说话人实验的 SA1 和 SA2 中的语句)。提取的 MFCCs 和 FDLPs 特征参数维度均为 39 维，所训练的 GMM 模型均包含 50 个高斯分量，符号化器训练完成后，输出 50 维高斯后验特征矢量。从测试集中抽取 10 个单词作为查询样例，如表 2 所示(括号内数字为查询样例在测试文档中实际出现的次数)。

表 2 查询项汇总

Tab. 2 Query examples list

help(10)	shoe(9)	contain(9)	popular(15)	abruptly (7)
diseases(7)	breakdown(7)	shampooed(7)	paramagnetic(1)	organization(7)

2.2 评价标准

无监督 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(只统计动态匹配时间,不包括模型训练和特征优化所用的时间)。

2.3 系统性能比较

2.3.1 FDLP 和 MFCCs 检测性能对比

实验首先比较了由两种声学特征参数 FDLP 和 MFCCs 所训练得到的 QbE-STD 系统性能,如图 5 所示,图 5(a)描述了查询系统检索精度与路径限制因子的关系,图 5(b)描述了查询系统平均检索时间与路径限制因子的关系。

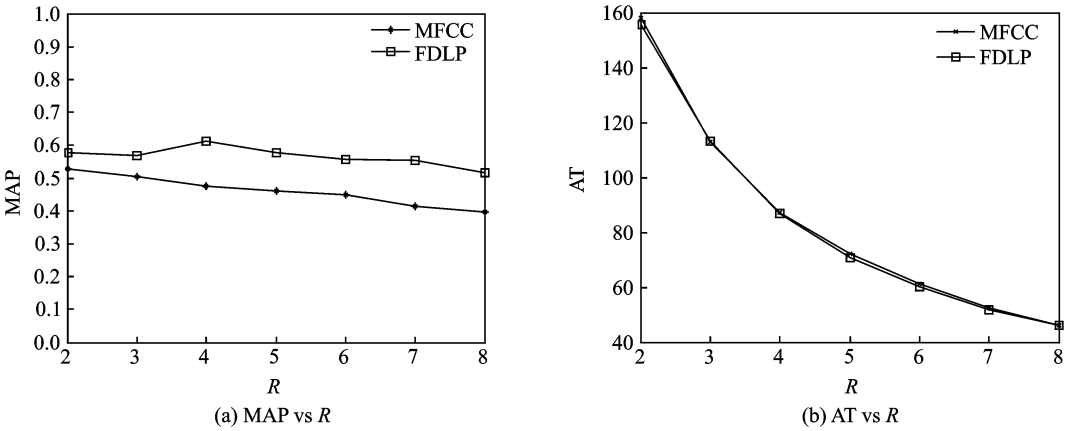

图 5 基于 FDLPs 和 MFCCs 的 GMM 符号化器检索性能对比

Fig. 5 Detection performance of GMM tokenizer based on FDLPs and MFCCs respectively

从图 5 中可以看出,由 FDLP 所训练得到的符号化器比 MFCC 性能更好,在 $R=8$ 时,MAP 相对提升 30.5%。由于两种符号化器所输出的后验特征矢量维数相同,在 SDTW 检索时计算量一致,因此两种系统检索速度大致相同。这一结果验证了 FDLP 声学参数比 MFCC 更适合于 QbE-STD 任务,这是由于在提取 FDLP 特征时,是对语音信号的长时片段分析,能够捕获语音的时域动态属性,并且保持了同 MFCC 一样考虑人耳听觉特性的优点。

接下来的实验中,分别使用基于 MFCCs 和 FDLP 的检测系统作为两个基线系统,并且检测算法均采用标准 SDTW 算法,路径限制因子统一设置为 8。

2.3.2 基于 NMF 的特征优化

使用训练好的 GMM 符号化器对测试语句集输出后验特征,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 50 阶的基矩阵,其中每一列向量可以看作原始矩阵的一个基,然后与每一测试语句的后验特征相乘,得到优化后的特征,在该特征基础上应用标准 SDTW 算法进行检索。表 3 描述了采用 NMF 算法之后各检索系统的系统性能,并与 PCA 算法进行了对比。

从表 3 可以看出,采用 NMF 算法对后验特征优化之后,性能提升明显。对于基于 MFCC 的 GMM 符号化器,采用 PCA 算法 MAP 相对提升了 9.6%,而采用 NMF 算法 MAP 相对提升了 8.1%。对于基于 FDLP 的 GMM 符号化器,采用 PCA 算法 MAP 相对提升了 6.8%,而采用 NMF 算法 MAP 相对提升了 10.8%。由于采用优化算法之后,特征维数不变,不改变检索时的计算量,故检索时间基本保持不变。由上述实验结果可以看出,对于基于 FDLP 的高斯后验特征,NMF 算法的优化性能要优于 PCA 算

法，而对于基于 MFCCs 的后验特征，结果则与之相反。然而考虑基于 FDLP 的高斯后验性能要优于 MFCCs，因此 NMF 算法更加适合于对高斯后验特征的优化。

表 3 不同后验特征优化算法下的各检索系统性能对比

Tab. 3 Performance comparison for different optimization algorithms

算法	模型	MAP	时间/s
MFCC(0.397)	MFCC+PCA	0.435	47.12
	MFCC+NMF	0.429	45.74
FDLP(0.518)	FDLP+PCA	0.553	46.62
	FDLP+NMF	0.574	47.43

2.3.3 基于多相邻输出得分的修正 SDTW 检索

按照式(11)对每一查询样例测试语句对的 SDTW 输出得分进行修正。图 6 描述了修正 SDTW 中不同的权重因子对各个系统检索精度的影响，表 4 描述了各个检索系统下，采用标准 SDTW 算法与修正 SDTW 算法检索性能的对比，括号内是实验中修正 SDTW 算法所设置的权重因子。

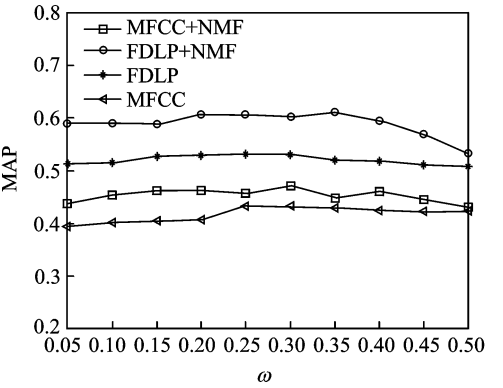

图 6 修正 SDTW 算法的权重因子对各检索系统的影响

Fig. 6 Influence of weight factor of modified SDTW on various systems

从图 6 可以看出 ω 值在 0.3 附近处，基于修正 SDTW 算法的各个系统检索精度较高，当接近 0 或者 0.5 时，检索精度有所下降。这是因为 ω 值过小或者过大，会过于忽视或突出最佳匹配子段的相邻子段得分的影响。而从表 4 可以看出，各个检索系统下，采用修正 SDTW 算法的检索精度相比 SDTW 算法，都有一定的提升，按照表 4 从上到下的顺序，检索精度分别相对提升 9.1%，9.8%，2.7%和 6.6%。这些实验结果验证了猜想，即多相邻输出得分相比 1-best 得分更具区分性。由于修正的 SDTW 算法相比标准 SDTW 算法只增加非常少量的加权运算，因此几乎不影响检索速度。至此，本文方法(即采用 NMF 优化及修正 SDTW 算法)，相对于 MFCCs 基线系统和 FDLP 基线系统，在不影响检索速度的情况下，检索精度分别相对提升了 18.6%，18.1%。

表 4 SDTW 与修正 SDTW 算法下的各检索系统性能对比

Tab. 4 Performance comparison for SDTW and modified SDTW algorithm

系统	SDTW		修正 SDTW	
	MAP	t/s	MAP	t/s
MFCC	0.397	46.10	0.433(ω=0.25)	46.40
MFCC+NMF	0.429	45.74	0.471(ω=0.3)	46.23
FDLP	0.518	46.09	0.532(ω=0.25)	46.18
FDLP+NMF	0.574	47.43	0.612(ω=0.35)	47.64

3 结束语

本文从特征和检索两个层面对基于模板匹配的无监督 QbE-STD 系统进行了改进。首先比较了两种声学特征参数(FDLP 和 MFCCs)所训练的 GMM 符号化器性能，然后提出使用 NMF 算法对高斯后验特征进行优化，最后提出多相邻输出的修正 SDTW 算法。实验结果表明，FDLP 特征参数相比 MFCC

更加适合于 QbE-STD 任务,能够有效利用语音的长时信息。所采用的 NMF 算法能够有效去除后验特征中的冗余信息,达到去噪的目的。基于多相邻输出的修正 SDTW 得分相比原始 SDTW 的 1-best 得分更具区分性,可以提高检索精度。后续可以进一步研究将 NMF 算法应用到其他后验特征中,如 ASM 后验和音素后验,同时可以考虑堆叠 NMF 算法的思想,对后验特征进行多层优化。另外可以考虑不同声学特征的语音属性,寻求不同的检索系统之间的互补性,进一步提升系统性能。

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] Chia T K, Sim K C, Li H, et al. A lattice-based approach to query-by-example spoken document retrieval[C]//Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: [s. n.], 2008:363-370.

[3] Stefan B, Sanu P A, Meinard M. Matching musical themes based on noisy OCR and OMR input [C]// ICASSP 2015. Brisbane, Australia: [s. n.], 2015:703-707.

[4] Hazen T J, Shen W, White C. Query-by-example spoken term detection using phonetic posteriorgram templates[C]//Prco of IEEE Automatic Speech Recognition and Understanding Workshop. Merano, Italy: IEEE, 2009:421-426.

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

[6] Wang Haipeng, 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.

[7] Schwarz P, Matejka P, Cernocky J. Hierarchical structures of neural networks for phoneme recognition[C]// ICASSP 2006. Toulouse, France:[s. n.], 2006: 325-328.

[8] Chung Cheng-tao, Chan Chun-an, Lee Lin-shan. Unsupervised discovery of linguistic structure including two-level acoustic patterns using three cascaded stages of iterative optimization[C]// ICASSP 2013. Vancouver, Canada:[s. n.], 2013: 8081-8085.

[9] Lee Chia-ying, O'Donnell T J. Unsupervised lexicon discovery from acoustic input[J]. Transactions of the Association for Computational Linguistics, 2015,3:389-403.

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

[11] 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.

[12] 彭红星,陈祥光,徐巍. PCA 特征抽取与 SVM 多类分类在传感器故障诊断中的应用[J]. 数据采集与处理, 2010,25(1): 111-116.
Peng Hongxing, Chen Xiangguang, Xu Wei. Application of PCA feature extraction and SVM multi-classification on sensor fault diagnosis [J]. Journal of Data Acquisition and Processing, 2010, 25(1):111-116.

[13] Zhang Yaodong. Unsupervised speech processing with applications to query-by-example spoken term detection[D]. Cambridge, Massachusetts, USA: Massachusetts Institute of Technology, 2013:39-47.

[14] Lee D D, Seung H S. Learning the parts of objects by nonnegative matrix factorization[J]. Nature, 1999, 401:1451-1454.

[15] 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 Data Acquisition and Processing, 2014, 29(2):265-273.

[16] Garofolo J S, Lori F L, William M F, 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(1993-),男,硕士研究生,研究方向:无监督语音关键词检测、模式发现,E-mail: jiankaic @ sina.com。

张连海(1971-),男,副教授,研究方向:语音信号处理、模式识别。

李勃昊(1989-),男,硕士,研究方向:语音识别、模式识别。