基于 DTW 的语音关键词检出*

侯靖勇[1]，谢磊[1]，杨鹏[1]，肖雄[2]，梁祥智[3]，徐海华[2]，
王磊[3]，吕航[1]，马斌[3]，CHNG EngSiong[2,4]，李海洲[2,3,4]

（1. 西北工业大学 计算机学院 陕西省语音与图像信息处理重点实验室，西安 710129
2. 南洋理工大学 Temasek 实验室，新加坡
3. 新加坡科技局资讯通信研究院 人类语言技术部，新加坡
4. 南洋理工大学 计算机工程学院，新加坡)

摘　要：近年来，针对少资源语言的语音关键词检出技术受到了国内外研究机构的广泛关注。 本文在基于动态时间规整（Dynamic Time Warping，DTW）的关键词检出框架下，提出了基于音素边界的局部匹配策略，用以解决基于样例的语音关键词检出任务中的近似查询问题。在 QUESST 2014 评测数据上采用多种特征进行了实验验证。实验结果显示，基于音素边界的局部匹配策略不仅在近似查询（T2 和 T3）任务上的检出效果明显提升，在精确查询（T1）任务上也获得了有效提升。随后的系统融合实验表明，基于音素边界的局部匹配策略能够大幅提升融合系统的性能。

关键词：语音关键词检出；少资源语言；动态时间规整；局部匹配；近似查询

中图分类号：TP 391

Spoken Term Detection Technology Based on DTW

HOU Jingyong[1], XIE Lei[1], YANG Peng[1], XIAO Xiong[2], LEUNG Cheung-Chi[3], XU Haihua[2], WANG Lei[3], LV Hang[1], MA Bin[3], CHNG EngSiong[2,4], LI Haizhou[2,3,4]

（1. Shaanxi Provincial Key Laboratory of Speech and Image Information Processing (SAIIP) School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
2. Temasek Lab@NTU, Nanyang Technological University, Singapore
3. Institute for Infocomm Research, A*STAR, Singapore
4. School of Computer Engineering, Nanyang Technological University）

Abstract: Spoken term detection (STD) for low resource languages has drawn much interest recently. In this paper, we study the problem of query-by-example spoken term detection (QbE-STD) using dynamic time warping (DTW). Specifically, we propose a partial matching strategy based on phoneme boundaries, which is used to perform fuzzy match subtasks in the MediaEval 2014 QUESST Evaluation. Experimental results on QUESST show that the proposed approach is quite effective for both fuzzy match tasks and the exact match task. In the fusion experiments, our approach improves the performance significantly.

Key words: Spoken Term Detection; Low Resource Languages; Dynamic Time Warping; Partial Matching;

1 引言

语音关键词检出（Spoken Term Detection, STD）是从给定的语料数据中查询指定关键词是否出现的任务。该任务在语音检索、电话监控、舆情分析等领域具有广泛的应用。根据关键词的输入形式，该任务可以分为基于文本[1-2]和基于语音样例两种类型。传统的基于文本的关键词检出技术又称为关键词搜索（Keyword Search, KWS），通常基于一个大词汇量连续语音识别系统，在识别的 NBest 结果或者网格上进行搜索。近年来，基于样例的语音关键词检出（Query-by-Example Spoken Term Detection, QbE-STD）成为新的研究热点。MediaEval 多媒体评测组织从 2011 年起，已经连续四年进行了 QbE-STD 方面的评测，瞄准跨语种、少资源或零资源（low or zero-resource）场景下的关键词检出任务。在没有任何针对待检语种的专家知识（如音素、词、发音字典）和标注数据的情况下，无法建立一个有效的语音识别系统，传统基于文本的关键词检出技术无法直接应用。

QbE-STD 目前通常有三种实现方法[4-5]。一种是借鉴传统语音关键词检出的思路，称之为声学关键词检出（Acoustic Keyword Spotting, AKWS）[7]。另外一种是借助其他语种成熟的语音识别系统进行解码，将查询样例和待检语音转换成音素序列或音素网格，而后进行字符串匹配，该类方法称之为字符搜索（Symbolic Search，SS）[4,7,14]。此类方法通常采用加权有限状态机（Weighted Finite State Transducer, WFST）建立索引并进行快速查找。还有一种方法是基于经典的模板匹配的思路，采用动态时间规整（Dynamic Time Warping，DTW）[16, 9-13,4]将查询样例和待查语句进行匹配。由于查询语句的时长通常远远大于样例时长，因此需要在语句上进行滑动查找，常用策略包括 Segmental DTW[11] 、Subsequence DTW [12]、Segmental Local Normalized DTW(SLN-DTW)[13]等。

本文以西北工业大学、新加坡南洋理工大学、新加坡资讯通讯研究院联合组队参加的 MediaEval 2014 少资源语言关键词检出评测（QUESST）提交的系统为例，介绍如何运用 DTW 算法有效提升关

*基金项目：国家自然科学基金项目（61175018）
作者简介：侯靖勇，1992.03.22，男，汉族，陕西榆林，研究生。
通讯联系人：谢磊，教授，xielei21st@gmail.com

键词检出的效果。该系统融合 DTW 模板匹配和字符搜索两种策略，在评测中获得综合评测第二名、第三种复杂评测(T3)第一名的成绩。本文主要介绍其中采用的基于 DTW 的语音关键词检出方法。我们采用 SLN-DTW 算法[13]进行关键词匹配；针对评测中“模糊查询”的复杂评测子任务，提出了基于滑动固定窗的局部匹配和基于音素边界的局部匹配两种策略。实验表明，基于音素边界的局部匹配策略可以明显提升关键词检出效果；将使用该策略的 DTW 子系统与参加评测的 DTW 系统进行融合后，检测效果得到进一步提升。

2 MediaEval QUESST 2014 评测介绍

QUESST 的全称为 Query by Example Search on Speech，是由欧美几所大学和研究机构共同组织的 MediaEval2014 多媒体评测中的一项评测任务[3]。该评测任务往年称为 Spoken Web Search，从 2011 年开始已经连续举办了四届。该任务瞄准少资源语言的关键词检出，评测数据中包含多种少资源语言的未标注语音数据，语音文件来自不同录制环境和多种说话风格，部分语音文件带有较强的背景噪声。

QUESST2014 任务提供了包含 23 小时的检索语料库，共 12492 句话。用于算法调试的发展集关键词 560 个，最终评测集关键词 555 个。语料涉及斯洛伐克语、罗马尼亚语、阿尔巴尼亚语、捷克语、巴斯克语和英语 6 种语言，其中英语部分多来自非母语说话人。在评测任务期间，语料库不提供任何语种信息。

评测任务分为三种不同类型的查询[3]：

- 精确匹配（T1）：在语料库中找出与查询关键词精确匹配的地方。
- 近似查询（T2）：允许前后缀不同的匹配，例如给定的关键词是 friend，可以匹配到语料库中包含 friendly，friends 单词的句子。
- 近似查询（T3）：在 T2 类型的基础上允许填充词和次序颠倒的匹配，例如关键词为 white house，在语料库中可以找到包含 house is whiter 的句子。

评测结果采用最小交叉熵（Cnxe）和 TWV（Term Weighted Value）来衡量[3]。为了计算这两项指标，参加评测的队伍需要提交每个关键词在语料库中每个句子上的打分，用以表示该关键词与该句子的匹配分数。

3 系统框架与特征提取

如图 1 所示，基于 DTW 的 QbE-STD 系统通常包括特征提取和关键词匹配两个部分。常用特征包括谱特征和后验特征。为了提高匹配精度并提高计算效率，在进行关键词匹配之前，通常经过语音激活检测（VAD）来去除关键词与待检语句中的静音部分。

短时谱特征，例如美尔域倒谱系数（MFCC），是目前语音识别中广泛采用的语音特征。然而该类特征在基于模板匹配的 QbE-STD 任务中效果欠佳。和语音识别中通过大量样本建立的语音统计模型（如 HMM）不同，DTW 算法进行的是关键词—语句的“一对一”模板匹配，因此对说话人、信道和环境不具备鲁棒性。研究表明，MFCC 特征在关键词和待检语句来自同一说话人的情况下，检出效果尚可，但是跨说话人检测效果较差[4-5]。因此，具有统计特性的语音特征具有更好的鲁棒性。

后验特征（Posterior）是一类常用的统计语音特征，在 QbyE-STD 任务上表现出良好的效果。给定一帧语音的谱特征向量 $\boldsymbol{a}$，后验特征定义为：在 K 个类别（$C_1,C_2,\ldots,C_K$）上的后验概率分布 P_a:

$$P_a = [P(C_1 \mid a), ..., P(C_K \mid a)] \quad (1)$$

其中 $P(C_k \mid \boldsymbol{a})$是谱特征向量 $\boldsymbol{a}$ 在第 k 类上的后验概率。如类别定义为音素，训练基于 HMM 的音素识别器，对语音数据进行解码，得到在每个音素上的打分，即为音素后验特征。与此类似，针对语音数据训练具有 K 个成分混合高斯模型（GMM），然后对每帧语音数据在 K 个高斯成分上打分，即可得到高斯后延特征（GMM Posterior）。一般来说，音素后验特征在基于 DTW 的 QbE-STD 任务中的效果最好。但是对于少资源语言来说，在没有该语种的专家知识和标注数据的情况下，通常无法训练音素识别器。借用其他语种的音素识别器（不匹配音素识别）来获得音素后验特征，是一种普遍采用的做法[18,20,26]。此外，采用机器学习方法自动发现某种语言中的“类音素”子词单元，获得类音素后验特征也是一种可行的方法[8, 21]。

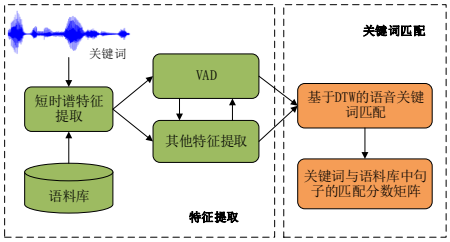

图 1 关键词检出系统框架

4 基于 DTW 模板匹配方法

本文采用 Segmental Local Normalized DTW (SLN-DTW) 算法进行关键词与语句的匹配[13]。和 Segmental-DTW[11]、Subsequence-DTW[12]等相比，SLN-DTW 在效率和效果上具有明显优势[4-5]。

4.1 SLN-DTW

如图 2 所示，现给定一个以语音特征序列组成的语音关键词 **q**={$q_1, \ldots q_m$}，其中 m 为语音关键词的帧数；语料库中待检的以语音特征序列组成的某个句子 **s**={$s_1, \ldots, s_n$}，其中 n 为语音句子的帧数。我们首先计算关键词 **q** 和句子 **s** 各帧之间的距离矩阵 ***dist***，其中 ***dist****(i, j)*代表关键词的第 *i* 帧与句子的第 *j* 帧之间的距离。对于 MFCC、Stacked Bottleneck[25]等特征，通常使用余弦距离:

$$dist(\mathrm{i},\mathrm{j}) = 1 - \frac{q[\mathrm{i}]^T \cdot s[\mathrm{j}]}{|q[\mathrm{i}]|*|s[\mathrm{j}]|}。 \quad (2)$$

对于后验特征，通常使用内积距离:

$$dist(\mathrm{i},\mathrm{j}) = -\log(\mathrm{q}[\mathrm{i}]^T \cdot \mathrm{s}[\mathrm{j}])。 \quad (3)$$

得到距离矩阵后，需对距离矩阵进行归一化处理以使距离取值在 0 到 1 之间：

$$dist_{norm}(\mathrm{i},\mathrm{j}) = \frac{dist(\mathrm{i},\mathrm{j}) - \min\limits_{j=1,\ldots,n}(\mathrm{dist}(\mathrm{i},\mathrm{j}))}{\max\limits_{j=1,\ldots,n}(\mathrm{dist}(\mathrm{i},\mathrm{j})) - \min\limits_{j=1,\ldots,n}(\mathrm{dist}(\mathrm{i},\mathrm{j}))}。 \quad (4)$$

SLN-DTW 算法的目的是在距离矩阵中找到一条使平均累积距离 *cost(i,j)=a(i,j) / l(i,j)*最小的匹配路径。其中 *a(i,j)*代表从某个起点*(1, e)*到达*(i, j)*所经历的累计距离，而 *l(i, j)*表示从某个起点*(1, e)*开始到达*(i, j)*所经历的路径长度。为了得到最小匹配路径，通常采用动态规划（DP）方法 [9-10,13]：

1、 初始化 *a* 和 *l*：

$$\begin{cases} a(\mathrm{i},1) = \sum_{k=1}^{i} dist_{norm}(\mathrm{k},1) \\ l(\mathrm{i},1) = i \end{cases} \quad (5)$$

$$\begin{cases} a(1,\mathrm{j}) = dist_{norm}(1,\mathrm{j}) \\ l(1,\mathrm{j}) = 1 \end{cases} \quad (6)$$

2、 迭代计算，对于 *i>0* 且 *j>0* 的部分从 *{(i-1, j), (i, j-1), (i-1, j-1)}*中选取一个点*(u, v)*,使：

$$\frac{a(\mathrm{u},\mathrm{v}) + dist_{norm}(\mathrm{i},\mathrm{j})}{l(\mathrm{u},\mathrm{v}) + 1} \quad (7)$$

最小，则：

$$\begin{cases} a(\mathrm{i},\mathrm{j}) = \mathrm{a}(\mathrm{u},\mathrm{v}) + dist_{norm}(\mathrm{i},\mathrm{j}) \\ l(\mathrm{i},\mathrm{j}) = l(\mathrm{u},\mathrm{v}) + 1 \end{cases} \quad (8)$$

3、找到 $\min\limits_{j=1,\ldots,n}(\mathrm{cost}(\mathrm{m},\mathrm{j}))$，即为该关键词对当前句子的匹配分数，分数越低匹配度越高。

4.2 基于滑动窗的局部匹配

SLN-DTW 算法在解决 QUESST 评测中的精确匹配(T1)任务时表现出绝对的优势，然而在近似查询任务(T2、T3)上效果不佳，原因在于上述算法并没有考虑局部匹配。为此，我们提出了两种解决模糊查询的局部匹配策略。

（1）基于滑动固定窗的局部匹配：在上述 SLN-DTW 算法的基础上，在语音关键词上加一个可以滑动的固定大小的窗口[14]。窗口的大小和每次的滑动步长均可以调整，即每次只用关键词的语音特征序列的一部分与语料库中的每一个句子做匹配。每次窗口滑动的时候计算出一个最佳匹配分数，最后取窗口滑动过程中得到的匹配分数最小值作为该语音关键词对该句子的最终匹配分数。该局部匹配算法策略简单、直观。

（2）基于音素边界的局部匹配：滑动固定窗策略虽然实现了局部匹配，但是对于固定窗长度的选取以及每次窗口的滑动步长都需要调节已达到最优。不仅如此，对于不同关键词，最佳窗长往往是不同的，这使得该策略在实际解决近似查询问题时效果欠佳。为了使每次取得的关键词匹配区域更“有意义”，我们借助其他语种的音素识别器对关键词进行解码，得到每个关键词所包含的“音素”边界信息。如图 3 所示，以音素边界为基础，每次我们取固定音素长度的窗口（例如取 6 个音素）与语料库中的句子做匹配。在匹配中，每次移动 1-2 个音素的距离，最后依然取窗口滑动过程中得到的最小匹配分数作为该语音关键词对该句子的最终匹配分数。如图 3 所示，假设现在匹配的关键词是 bomb the white house，待匹配语句是 I will attack the white house at next Saturday afternoon。采用该策略，可以方便的找到部分匹配区域“the white house”。

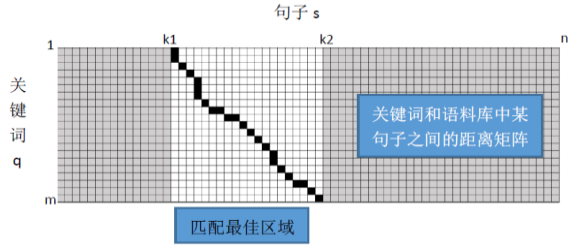

图 2 普通全匹配 SLN-DTW 算法示意图

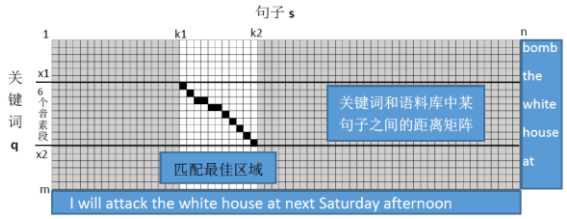

图 3 基于边界的局部匹配 SLN-DTW 算法示意图

5 实验设置

在 QUESST 发展集和评估集上对上述三种策略进行了关键词检出测试，采用的语音特征包括：

- MFCC：39 维经过 VTLN 和 CMVN 预处理的 MFCC 特征，包括一阶和二阶差分。
- EN Stacked Bottleneck：使用 100 个小时的 SwitchBoard（SWB）数据训练基于深度神经网络的英文音素识别器，在 QUESST 数据上获得 Bottlenectk 特征，其中音素个数为 42。
- CZ Posterior： BUT 提供的捷克语音素识别器[20]在 QUESST 数据上解码获得的音素后验特征，音素个数为 45。
- HU Posterior：BUT 提供的匈牙利语音素识别器[20]在 QUESST 数据上解码获得的音素后验特征，音素个数 61。
- RU Posterior：BUT 提供的俄语音素识别器[20]在 QUESST 数据上解码获得的音素后验特征，音素个数 52。

其中只有 CZ、HU 和 RU 三种音素识别器可以获得音素边界信息，因此基于音素边界的局部匹配策略在这三种特征上进行了实验。

对于使用基于滑动固定窗的局部匹配策略的子系统，经调试，我们最终设置固定窗口的大小为 60 帧，窗移为 5 帧。对于使用基于音素边界局部匹配策略的子系统，经调试，设置窗口的大小为 6 个音素的长度，窗口移动的步长为 1 个音素。

实验结果以 MinCnxe 和 MaxTWV 衡量。其中，MinCnxe 最初用于衡量说话人辨认评测结果，它的取值越小越好，具体可以参考文献[24]；MaxTWV 是指给定一个最优阈值时，系统漏检率（Miss）和虚警（False alarm）在所有检测关键词上的一个加权平均值，取值越大越好[23]。

6 实验结果

表 1 给出了全匹配、基于滑动固定窗的局部匹配和基于音素边界局部匹配策略在 QUESST 评测任务中的结果，其中包括在发展集上每种子任务的结果和在评估集上的综合结果。从实验结果中我们可以看出，在 T1 全匹配任务中，En Stacked Bottleneck 特征与 BUT 的 HU 和 RUS 音素后验特征的效果接近，但当音素后验特征加入音素边界信息后，在所有查询任务上，音素后验特征均显著好于 En Stacked Bottleneck 特征。这说明借用其他语种成熟的音素识别器，在少资源语言上获得音素后验特征，进行基于 DTW 的关键词检出，是一种可行的方法。

综合在所有特征上的结果，还可以看出基于滑动固定窗的局部匹配策略在三种匹配策略中表现最差，此策略在 QUESST 任务的 T2 和 T3 查询任务上的效果不如普通全匹配策略。经分析，原因主要有两个：1）窗口大小和窗口滑动步长选择没有较为合适的方法，此外对于不同关键词，窗口大小应该灵活设置；2）固定窗口在滑动时，从语音关键词中选择的匹配区域，没有明显的物理意义，可能将完整的音素截断。相比之下，使用基于音素边界的局部匹配策略在解决 T2 和 T3 查询任务上，实验结果在 MinCnxe 和 MaxTWV 两项指标上均有显著改善，尤其在 MaxTWV 指标上的提升更加显著。而在 T1 精确查询任务上，除俄语音素后验特征（RU Posterior）外，基于音素边界的局部匹配使匹配效果也有所提升。经过对匹配结果的分析，我们发现，这种策略可以明显提升查询的命中率，同时使虚警率略微升高。

我们将 3 个使用基于音素边界局部匹配策略的 DTW 子系统与参加 QUESST 评测的 9 个 DTW 子系统（9-DTW-Fusion）[14]融合，该 9-DTW-Fusion 系统仅使用基于滑动固定窗的方法解决近似查询的问题。融合后得到 12-DTW-Fusion 的融合系统，融合前后的效果对比如表 2 所示。其中，基于符号搜索（SS）的融合系统（SS-Fusion）[7]也列入其中进行对比。融合采用 Focal 工具包进行[22]。

由表 2 可以看出，加入 3 个基于音素边界局部匹配的子系统产生的 12-DTW-Fusion 系统，相比原 9-DTW-Fusion 系统，MinCnxe 和 MaxTWV 两项指标在三种查询类型上均有显著改善。同时可以看到，两个 DTW 融合系统的性能显著好于 SS-Fusion 融合系统的性能。

最后我们将 12-DTW-Fusion 和 SS-Fusion[7]系统进行进一步融合，得到新的融合系统，与 BUT 参加评测的系统[18]、SPL-IT 参加评测的系统[19]进行了比较，结果如表 3 所示。我们发现，最终融合系统在 T1 精确查询任务上效果较差，在 T2 查询

任务上与 BUT 和 SPL-IT 系统效果接近，而在 T3 查询任务上效果显著好于其他两个系统。经分析，效果提升主要归功于 3 个基于音素边界局部匹配的 DTW 子系统。

表 1 全匹配、基于滑动固定窗的局部匹配、基于音素边界局部匹配在 QUESST 任务的发展集和评估集上的结果
（实验结果形式：minCnxe/maxTWV）

系统		T1	T2	T3	T1+T2+T3 综合	T1+T2+T3 评估集
特征	策略					
MFCC	全匹配	0.875/0.112	0.933/0.038	0.964/0.017	0.915/0.052	0.924/0.062
	固定窗局部匹配	0.844/0.130	0.946/0.036	0.922/0.040	0.906/0.085	0.889/0.098
EN Stacked Bottleneck	全匹配	0.753/0.305	0.883/0.129	0.951/0.031	0.847/0.186	0.840/0.175
	固定窗局部匹配	0.772/0.196	0.875/0.107	0.878/0.087	0.840/0.141	0.796/0.157
CZ Posterior	全匹配	0.690/0.361	0.874/0.125	0.911/0.017	0.814/0.201	0.786/0.178
	固定窗局部匹配	0.875/0.112	0.933/0.038	0.947/0.017	0.915/0.051	0.892/0.056
	音素边界局部匹配	**0.583/0.370**	**0.778/0.240**	**0.765/0.164**	**0.708/0.267**	**0.650/0.290**
HU Posterior	全匹配	0.763/0.307	0.889/0.123	0.923/0.014	0.846/0.172	0.831/0.151
	固定窗局部匹配	0.910/0.086	0.930/0.043	0.957/0.012	0.927/0.043	0.911/0.053
	音素边界局部匹配	**0.633/0.326**	**0.789/0.208**	**0.722/0.148**	**0.728/0.238**	**0.691/0.245**
RU Posterior	全匹配	0.764/**0.317**	0.882/0.139	0.933/0.015	0.844/0.180	0.825/0.167
	固定窗局部匹配	0.908/0.087	0.933/0.047	0.961/0.005	0.930/0.043	0.901/0.063
	音素边界局部匹配	**0.647**/0.260	**0.647/0.171**	**0.759/0.121**	**0.741/0.190**	**0.704/0.212**

表 2 DTW 融合系统与 SS 融合系统在评估集数据的对比

系统	T1	T2	T3	T1+T2+T3 评估集
minCnxe				
12 DTW-fusion	**0.408**	**0.540**	**0.630**	**0.511**
9 DTW-fusion [14]	0.597	0.719	0.793	0.682
SS-fusion [7]	0.662	0.731	0.793	0.724
maxTWV				
12 DTW-fusion	**0.596**	**0.396**	**0.382**	**0.472**
9 DTW-fusion [14]	0.447	0.241	0.167	0.297
SS-fusion [7]	0.368	0.245	0.236	0.283

表 3 最终融合系统与 BUT 系统、SPL-IT 系统在评估集数据的对比

系统	T1	T2	T3	T1+T2+T3 评估集
minCnxe				
BUT system[18]	**0.310**	0.513	0.624	**0.461**
Our System	0.390	**0.505**	**0.613**	0.492
SPL-IT system[19]	0.367	**0.501**	0.734	0.508
maxTWV				
BUT system[18]	**0.674**	0.436	0.171	**0.473**
Our System	0.606	0.407	**0.361**	0.472
SPL-IT system[19]	0.622	**0.447**	0.201	0.442

7 总结

本文提出在 DTW 框架下的基于音素边界的局部匹配策略，用以解决语言独立的基于样例的语音关键词检出任务中近似查询的问题。在 QUESST 2014 评测数据上分别进行单系统和融合系统的实验。实验结果显示，提出的基于音素边界的局部匹配策略在解决 QUESST 2014 任务的近似查询（类型 2、3）任务上有显著的效果

参考文献 (References)

[1] Intelligence Advanced Research Projects Activity (IARPA), “Babel program,” in*http://www.iarpa.gov/index.php/researchprograms/babel*

[2] NIST, Openkws14 keyword search evaluation plan, in *http://nist.gov/itl/iad/mig/upload/KWS14-evalplan-v11.pdf*

[3] Anguera, Rodriguez-Fuentes, Szoke, Buzo, and Metze, Query by example search on speech at mediaeval 2014, in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, Oct. 16-17.

[4] Javier Tejedor, et al., Comparison of methods for language-dependent and language-independent query-by-example spoken term detection, ACM Transactions on Information Systems, 2012.

[5] 杨鹏，谢磊，张艳宁，低资源语言的无监督语音关键词检测技术综述《中国图像图形学报》vol.20，no.2, pp 211-218, Feb 2015.

[6] Vikram Gupta, Jitendra Ajmera, Arun Kumar, and Ashish Verma, A

language independent approach to audio search, in *Proc. INTERSPEECH*, 2011.

[7] Haihua Xu., et al., Language Independent Query-by -Example Spoken Term Detection Using N-Best Phone Sequences and Partial Matching. ICASSP, 2014.

[8] Haipeng Wang, Tan Lee and Cheung-Chi Leung, Unsupervised Spoken Term Detection with Acoustic Segment Model, 2011 International Conference on Speech Database and Assessments, Oriental COCOSDA 2011 – Proceedings, 2011.

[9] Peng Yang, Cheung-Chi Leung, Lei. Xie, Bin Ma, and Haizhou Li, Intrinsic spectral analysis based on temporal context feature for query by example spoken term detection, in *Proc. INTERSPEECH*, 2014.

[10] Gupta, V, et al., A language independent approach to audio search. in Proc. INTERSPEECH, 2011.

[11] Zhang Y, Glass J R, Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams[C]//Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on. IEEE, 2009: 398-403.

[12] Anguera X. Speaker independent discriminant feature extraction for acoustic pattern-matching[C]//Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on. IEEE, 2012: 485-488.

[13] Muscariello A, Gravier G, Bimbot F, Audio keyword extraction by unsupervised word discovery[C]//INTERSPEECH 2009: 10th Annual Conference of the International Speech Communication Association. 2009.

[14] Peng Yang et al., The nni query-by-example system for mediaeval 2014, in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, Oct. 16-17.

[15] Schwarz et al., Hierarchical structures of neural networks for phoneme,in Proc. ICASSP, 2006.

[16] Müller M. Dynamic time warping [J]. Information retrieval for music and motion.2007: 69–84.

[17] Szöke I, Schwarz P, Matějka P, et al. Phoneme based acoustics keyword spotting in informal continuous speech [J]. Lecture Notes in Computer Science. 2005, 2005 (3658): 302–309.

[18] Igor Szöke et al., BUT QUESST 2014 System Description, in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, Oct. 16-17.

[19] Jorge Proença et al., The SPL-IT Query by Example Search on Speech system for MediaEval 2014. in Working Notes Proceedings of the Mediaeval 2014 Workshop, Barcelona, Spain, Oct. 16-17.

[20] Schwarz et al., Hierarchical structures of neural networks for phoneme, in Proc. ICASSP, 2006.

[21] Haipeng Wang et al.,A Graph-based Gaussion Component Clustering Approach to Unsupervised Acoustic Modeling,INTERSPEECH, 2014.

[22] Brummer, Focaltoolkit, in *https://sites.google.com/site/nikobrummer/focal.*

[23] Fiscus, J., Ajot, J., Garofolo, J., Doddingtion, G. Results of the 2006 Spoken Term Detection Evaluation. In Proceedings of ACM SIGIR 2007 Workshop on Searching Spontaneous Conversational Speech. Amsterdam, Netherlands, 2007.

[24] Luis J. Rodriguez-Fuentes, M. Penagarikano, MediaEval 2013 Spoken Web Search Task: System Performance Measures, n.TR-2013-1, Department of Electricity and Electronics, University of the Basque Country, 2013

[25] Yaodong Zhang, Ekapol Chuangsuwanich, James Glass，Extracting deep neural network bottleneck features using low-rank marix factorization，ICASSP，2014.

[26] 张卫强，宋贝利，蔡猛，等. 基于音素后验概率的样例关键词检测算法[J]. 天津大学学报：自然科学与工程技术版，2015,48（9）：757-760.