文章编号:1004-9037(2014)02-0274-06

# 多流信息融合的集外词检索

熊世富 郭 武

(中国科学技术大学电子工程与信息科学系,合肥,230027)

**摘要**:针对关键词中的集外词检索任务,提出采用音素、音节、词片三种子词单元进行多流信息的联合检索算法,其中对基于音素的语音检索(Spoken term detection,STD)系统使用基于 $n$ 元语言模型-加权有限状态机的完全匹配检索降低漏警,对基于音节、词片的 STD 系统使用模糊匹配检索降低虚警,最后采用线性逻辑回归(Linear logistic regression, LLR)的算法将三个子系统的结果进行融合。在 NIST STD 2006 语音检索评测的英语电话会话语音测试集上的实验结果表明,相对于最好的单流系统,多流信息融合获得了 12% 的实际词项权重值(Actual term weighted value, ATWV)相对提升。

**关键词**:语音检索;集外词;加权有限状态机

**中图分类号**:TN912 **文献标志码**:A

# Multi-Streamed Based out of Vocabulary Terms Detection

*Xiong Shifu*, *Guo Wu*

(Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, 230027, China)

**Abstract**: For out of vocabulary (OOV) terms detection in spoken term detection (STD), we propose a multi-streamed based detection algorithm which makes use of three sub-word units: phone, syllable & fragment. *n*gram-WFST based search is applied in phone-based STD system to reduce miss probability; and fuzzy search is applied in syllable-based and fragment-based STD systems to reduce false alarm probability. The final scores are obtained through linear logistic regression of the three sub-systems scores. Results on the National Institute of Standards and Technology (NIST) STD 2006 English conversational telephone speech (CTS) EvalSet show that the multi-streamed based detection algorithm achieves a 12% relative improvement in actual term weighted value (ATWV) compared to the best single-streamed system.

**Key words**: spoken term detection; out of vocabulary; weighted finite state transducer

## 引 言

语音检索是在大量的语音数据中发现感兴趣的关键词以及主题,其中关键词的检索技术(Spoken term detection,STD)是目前研究的热点。由于 NIST 的推动[1],采用两步骤的关键词检索是主流算法。第一步通过大词汇连续语音识别系统(Large vocabulary continuous speech recognition, LVCSR)将语音文件转化为文本,第二步在识别的文本上查找所关注的关键词。这种算法的优越性在于可以充分利用 LVCSR 的成果,另外关键词还可以动态设置。但是由于 LVCSR 无法识别集外词(Out of vocabulary, OOV),相比于集内词检索,导致集外词检索性能急剧下降,因而如何提高集外词的检索性能是 STD 系统面临的一个主要挑战。

为了解决集外词检索问题,学者们将识别单元投向对集外词具有更强建模能力的子词单元[2],通常为音素:通过音素识别器生成音素网格(Lattices),并将查询词转化为音素序列,最后从音素网格中检索[3]。除音素之外,其他子词单元也被用在语音检索中,如:词片[4](Fragment),音节[5](Syllable)等。这些基于非音素子词单元语音检索的

**收稿日期**:2012-12-13;**修订日期**:2013-03-22

基本思想是创建一个合适的子词列表，该子词列表既能很好地对集外词进行表示，又对语言的上下文约束信息具有较强的捕捉能力。其中词片是基于数据驱动，使用统计方法自动选择的可变长度音素序列，而音节则具有很强的语言学特征。在检索方面，为了满足速度和性能上的要求，完全匹配的 $n$ 元语言模型-加权有限状态机[6]（$n$gram-weighted finite state transducer，$n$gram-WFST）检索和模糊匹配检索[7]分别被提出。

相对而言，基于音素的 STD 系统受语法约束较小，更容易发现集外词，但也更容易在识别中引入虚警；而词片和音节的 STD 系统受语法约束较强，在相同的条件下，对于 OOV 更容易形成漏警。考虑到音素、音节、词片的不同性质和它们之间潜在的互补性，本文分别生成了基于音素、音节、词片的 STD 系统，并将三者进行结果融合。针对音素、音节和词片的不同特点，对基于音素的 STD 系统采用完全匹配的 $n$gram-WFST 检索以降低虚警，对基于音节、词片的 STD 系统则采用模糊匹配检索以减少漏警。最后采用线性逻辑回归[8]（Linear logistic regression，LLR）的算法将三个子系统的结果进行融合，提高检索性能。

## 1　多流信息融合检索系统

多流信息融合的关键词检索系统如图 1 所示。在系统中，包括词片、音节和音素三种不同的识别单元。一般而言，针对 OOV 词的子词 STD 系统基本框架包括语音转写和关键词检索两个模块。对于待检索的任意语音文件，首先通过子词解码器将语音文件转写为子词 Lattices，同时为了方便检索，需要将非音素 Lattices 转化为音素 Lattices，并建立相应的音素倒排索引以加快后端的检索速度。对于待查询的关键词，也需要通过字形到音（Grapheme to phoneme，G2P）的转换得到需查询的音素序列，然后在倒排索引上进行音素匹配，其中置信度的选择是非常重要的。下面将逐一介绍音素、音节、词片子词列表的挑选方法。

### 1.1　音素

为了增加词边界信息，加上特殊符号以标明单词边界，如 alabama：＃ae l ax b ae m ax＃，这样相对于平常英语识别中常用到的 40～50 个左右的音素，本文使用的音素个数相对会多很多，有 171 个带位置信息的音素。在此基础上重新构建字典对应的音素信息，用大量的文本数据训练生成3gram音素语言模型（Language model，LM）用于解码。

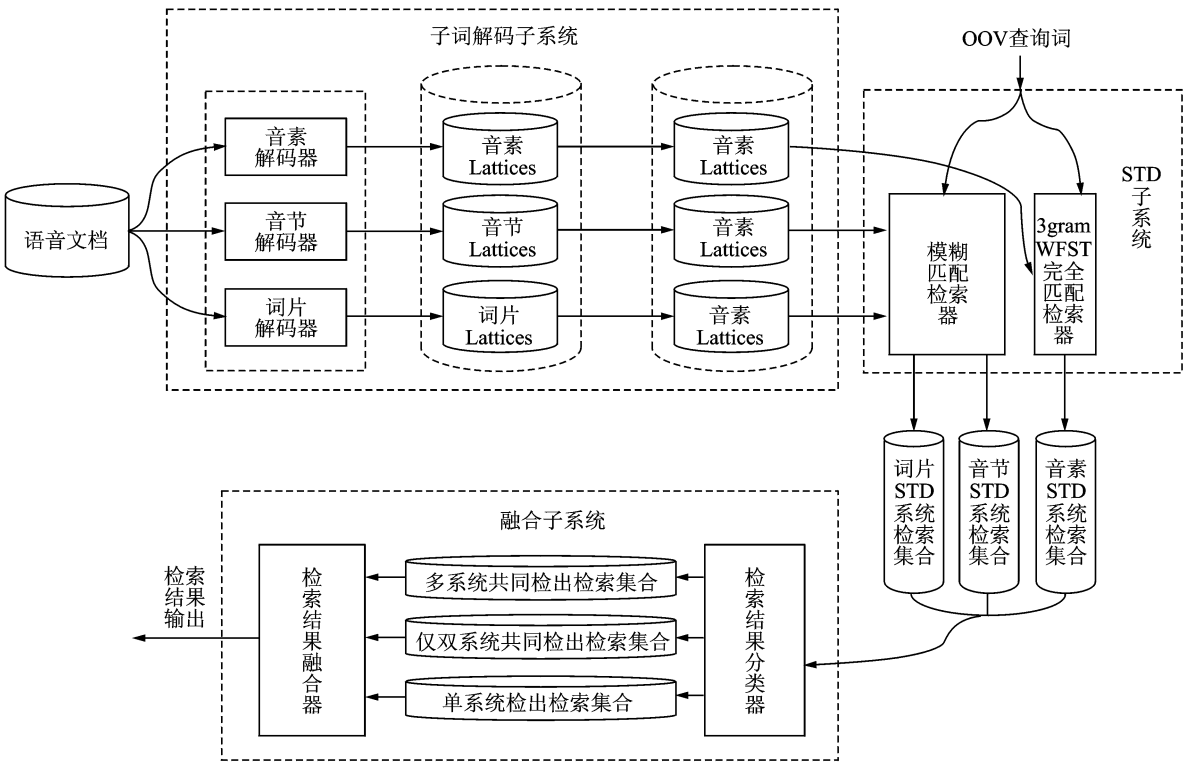


图 1　多流信息融合 STD 系统

Fig. 1　Multi-streamed based STD system

### 1.2 音节

为了获得用于解码的音节列表，首先进行了英语音节化工作。采用基于支持向量机-隐马尔科夫模[9-10]（Support vector machine-hidden markov model，SVM-HMM）的方法对 LVCSR 词典进行音节化，并提取所有不同的音节单元，获得了 21 000个带位置信息的音节，并用于生成 3gram 音节语言模型。

### 1.3 词片

与音节不同，词片为基于数据驱动的。本文通过减值的 5gram 音素语言模型[4]获得了 21 000 个带位置信息的词片，并用这个词片列表生成 3gram 词片语言模型用于解码。

## 2 检索算法

在进行语音识别后，需要建立音素倒排索引。本文在实验中采用 Lattice-tool[11]工具将音素 Lattices 转化为 *n*gram 倒排索引，其中每条 gram 索引 $g$ 包含信息为 gram 音素串 $I_g$、所属语音文件 $ID_g$、发生位置（开始时间-结束时间）$O_g$ 和后验概率得分 $W_g$，以 $g(I_g, O_g, W_g, ID_g)$ 表示。待检索的关键词在转换成音素序列之后，就在 *n*gram 倒排索引中进行检索。

为了提高性能，根据不同子词系统的特点，对基于音素的子系统采用完全匹配的 *n*gram-WFST 检索方法，对基于音节和词片的子系统采用模糊匹配的检索方法。为便于描述，针对固定的语音文件，将 *n*gram 索引 $g(I_g, O_g, W_g, ID_g)$ 简写为 $g(I_g, O_g, W_g)$，定义 $|g|$ 为索引 $g$ 中 $I_g$ 包含的音素个数。

### 2.1 *n*gram-WFST 完全匹配检索

基于 *n*gram-WFST 的检索系统由三部分组成：首先将 *n*gram 倒排索引编译生成索引 FST，其次将查询词发音分段并编译成用于检索的词典 FST，最后将索引 FST 和词典 FST 进行 FST 合成，以达到检索的目的。具体过程如下：

2.1.1 索引 FST

（1）为每条 *n*gram 索引 $g(I_g, O_g, W_g)$ 分配输入状态 $S_g$ 和输出状态 $E_g$，将索引 $g(I_g, O_g, W_g)$ 转化为 FST 弧 $r(S_g, E_g, I_g, O_g, W_g)$，并且新建初始状态和结束状态 $S$，$E$。

（2）添加转移弧 $r(S, S_g, \varepsilon, \varepsilon, 1.0)$ 和 $r(E_g, E, \varepsilon, ID_g, 1.0)$，使所有的 $r(S_g, E_g, I_g, O_g, W_g)$ 与初始状态和结束状态 $S$，$E$ 连通，其中 $\varepsilon$ 为 FST 中的空符号表达。

（3）添加转移弧 $r(E_g, S_{g'}, \varepsilon, \varepsilon, 1.0)$，将满足条件①索引重叠时间 $dist(g, g') < T$ 和②$|g| = N|, g'| <= N$ 的弧 $r(S_g, E_g, I_g, O_g, W_g)$ 和 $r(S_{g'}, E_{g'}, I_{g'}, O_{g'}, W_{g'})$ 相连，生成初始的 FST 索引。

（4）对初始 FST 索引使用 FST 确定化、状态数最小化、ε-移除操作进行优化，生成最终的索引 FST。

2.1.2 词典 FST

（1）将查询词发音进行 *n*gram 分段。以 $n=3$ 为例，对于 alabama 这个词，其 *n*gram 分段发音为 alabama：＃ae-l-ax_ b-ae-m_ ax＃，3gram 分段发音数为 3。

（2）将 3gram 分段发音编译为词典 FST，如图 2 所示。

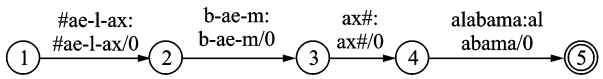


图 2 词典 3gram-WFST

Fig. 2 3gram-WFST of dictionary

2.1.3 检索

由于索引 FST 中 *n*gram 弧 $r(S_g, E_g, I_g, O_g, W_g)$ 均与初始状态和结束状态相连，所以最终的检索过程只需将词典 FST 和索引 FST 进行 FST 合并操作即可。为了降低虚警，对检索返回得分进行长度归一化

$$S(q_{t_s t_e}) = \sqrt[N(q)]{\prod_{i=1}^{M(q)} W_{g_i}} \tag{1}$$

式中：$q_{t_s t_e}$ 为查询词项 $q$ 的一个检索结果，$N(q)$ 为 $q$ 对应发音中的音素个数，$M(q)$ 为 $q$ 的 *n*gram 分段发音数，$W_{g_i}$ 为 $q$ 的 *n*gram 分段发音对应的第 $i$ 条索引 $g_i(I_{g_i}, O_{g_i}, W_{g_i}, ID_{g_i})$ 中的后验概率 $W_{g_i}$。

### 2.2 模糊匹配检索

对于音节和词片子词系统，为了减少漏警，在不过多引入虚警的前提下，使用模糊匹配进行检索。模糊匹配检索系统构建的大致过程为：获得 3gram 倒排索引，其中所有索引 $g$ 满足条件 $|g| = 3$；检索查询词项 $q$ 的 triphone 发音序列，如 alabama：＃ae-l-ax_ l-ax-b_ ax-b-ae_ b-ae-m_ ae-m-ax＃，在相邻 triphone 3gram 索引时间间隔 $dist(g, g')$ 小于一定阈值 $T$ 的条件下，检索到的不同 tirphone 数 $M$ 大于单词总 triphone 发音数 $N(q)$ 的一半时召回并返回如下得分

$$S(q_{t_s t_e})=\frac{M(q_{t_s t_e})}{N(q)}\sqrt[M(q_{t_s t_e})]{\prod_{i=1}^{M(q_{t_s t_e})}W_{g_i}} \quad (2)$$

发音个数，$W_{g_i}$ 为 $q$ 的 triphone 发音对应的第 $i$ 条索引 $g_i(I_{g_i},O_{g_i},W_{g_i},ID_{g_i})$ 中的后验概率 $W_{g_i}$。

## 3　多流信息融合方法

由于本文中有三个子系统，对于同一个关键词，这三个子系统可能给出不同的置信度得分和不同的检索结果。本文在线性回归的基础上，分三种情况对结果进行得分融合。当一个关键词检索结果在三个子系统中都被检出时，对各个系统的得分进行线性加权

$$S_{fu}(q_{t_s t_e})=\sum_{i=1}^{3}\lambda_i\cdot S_i(q_{t_s t_e}),\sum_{i=1}^{3}\lambda_i=1 \quad (3)$$

当一个关键词检索结果只由两个系统检出时，融合得分为这两个系统得分的线性加权

$$S_{fu}(q_{t_s t_e})=\lambda_{ij}^{i}\cdot S_i(q_{t_s t_e})+\lambda_{ij}^{j}\cdot S_j(q_{t_s t_e})$$
$$\lambda_{ij}^{i}+\lambda_{ij}^{j}=1,i,j\in(1,2,3) \quad (4)$$

最后，当一个关键词检索结果仅由单系统检出时，认为它不够可信，对该系统的得分进行惩罚

$$S_{fu}(q,q_e)=p\cdot S(q_{t_s t_e}),0<p<1 \quad (5)$$

式中：$p$ 为惩罚因子。

融合中的关键问题是线性回归参数的选取，本文使用线性逻辑回归融合策略，具体过程为：首先提取开发集中所有三个子系统检索结果中的正例（正确的检索结果）得分和反例（错误的检索结果）得分作为 LLR 的训练数据，训练并获得各系统相应的权重系数 $w_1,w_2,w_3$，然后将这些权重归一化作为式(3)的加权系数和式(5)中对应系统的惩罚因子，最后对 $w_1,w_2,w_3$ 两两归一化作为式(4)相应系统的加权系数，例如：当某个检索结果只由系统 $i$ 和系统 $j$ 检出时，加权系数分别为

$$\lambda_{ij}^{i}=\frac{w_i}{w_i+w_j},\lambda_{ij}^{j}=\frac{w_j}{w_i+w_j} \quad (6)$$

## 4　实验配置

### 4.1　实验数据及基本配置

本文实验是在 NIST STD 2006 英语电话语音数据库上进行的，该数据库包含开发集和测试集两部分，每部分都有大约 3 h 语音。

声学模型训练数据为总计 360 h 语音的 Switchboard 和 CallHome 语料库。语言模型训练采用 Switchboard、CallHome 语料库的标注文件和英语广播新闻数据。

采用 39 维感知线性预测（Perceptual linear prediction，PLP）参数作为声学特征。通过最大似然估计（Maximum likelihood estimation，MLE）训练算法得到 60 高斯的 HMM 模型，然后使用最小音素错误（Minimum phone error，MPE）区分性训练准则对获得的 MLE 参数进行优化。

### 4.2　OOV 词挑选

由于 NIST 任务集中集外词相对较少，只有 2%左右的比例，不适合研究工作的开展，因此需要在 NIST 的任务集上重新挑选一些词汇作为集外词。挑选集外词的原则是：首先保留 NIST 测试任务中已有的集外词，也就是语音识别词典中不包含的词汇；其次挑选具有一定意义的地名、人名，这些词汇的选择是因为它们经常是关键词检索所关注的内容。为了保证关键词检索的稳健性，要求被选择的 OOV 词均最少在开发集和测试集出现过 5 次以上。为保证实验的真实性，对于这些集外词，必须把其对应的原始语音文件从声学模型训练中去除，文本标注从语言模型训练数据中剔除，语音识别词典也要剔除这些 OOV 词。基于以上原则，在开发集上挑选了 313 个集外词，在测试集上挑选了 320 个集外词。

## 5　实验结果与分析

对于 STD 任务，使用 NIST STD 2006 评测计划定义的实际词项权重值[1]（Actual term weighted value，ATWV）作为主要的性能评估尺度。

### 5.1　音素识别率

表 1 给出了 STD 2006 开发集上不同解码单元在集内词区域和集外词区域的音素识别率（Phone recognition accuracy，PACC）。对于集内词识别而言，音素识别系统的 PACC 明显低于音节、词片和词识别系统的 PACC。由于词识别系统对集外词的建模能力较弱，导致词识别系统在集外词和集内词区域的 PACC 反差很大，其在集外词区域上的 PACC 明显低于音节、词片识别系统。

**表 1　不同解码单元在 NIST STD 2006 开发集上的音素识别率**

**Table 1　Phone recognition accuracy using different types of decoding units on NIST STD06 development set**

%

| 解码单元 | 集外词区域 | 集内词区域 |
|---|---|---|
| 音素 | 51.46 | 60.54 |
| 音节 | 59.03 | 78.29 |
| 词片 | 60.64 | 78.63 |
| 词 | 53.79 | 79.62 |

## 5.2 集外词检索性能

(1)单系统检索结果

表 2 分别给出了 *n*gram-WFST 和模糊匹配检索系统在 STD 2006 开发集上的检索结果。由于音节和词片识别系统 PACC 较高,识别混淆低,使用模糊匹配能在引入较少虚警的情况下,降低了漏警,因而模糊匹配检索结果好于 *n*gram-WFST 检索结果。音素识别系统本身混淆就很高,使用模糊匹配,在虚警已很高的情况下又进一步提高了虚警,其模糊匹配性能是不可接受的。

基于表 2 实验结果,对于音素检索系统,使用 *n*gram-WFST 检索,对于音节和词片检索系统,使用模糊匹配检索。同时,为了平衡虚警和漏警,所有系统均使用词项相关置信度归一方法提高系统性能[12]。

**表 2 NIST STD 2006 开发集上 *n*gram-WFST 和模糊匹配检索结果**

**Table 2 ATWV results based on *n*gram-WFST and fuzzy search on NIST STD06 development set**

| 子词单元 | 检索方法 | |
|---|---|---|
| | *n*gram-WFST | 模糊匹配 |
| 音素 | 0.238 | −0.744 |
| 音节 | 0.407 | 0.450 |
| 词片 | 0.391 | 0.455 |

(2)多系统融合结果

为了研究不同子词系统之间的互补性,分别做了音素、音节、词片系统之间的两两融合和三者间的融合,表 3 为对应的 ATWV 值。相对于性能最好的以词片作为识别单元的单系统,多系统融合的性能在开发集和测试集上,分获得了 11%和 12%的 ATWV 相对提升。

**表 3 开发集和测试集上的集外词检索 ATWV**

**Table 3 ATWV results of OOV on development and eval set**

| 系统 | 数据集 | |
|---|---|---|
| | 开发集 | 测试集 |
| 词片(模糊匹配) | 0.531 | 0.493 |
| 音节(模糊匹配) | 0.508 | 0.395 |
| 音素(*n*gram-WFST) | 0.341 | 0.312 |
| 词片(模糊匹配)+音节(模糊匹配) | 0.563 | 0.504 |
| 词片(模糊匹配)+音素(*n*gram-WFST) | 0.566 | 0.523 |
| 音节(模糊匹配)+音素(*n*gram-WFST) | 0.536 | 0.485 |
| 词片(模糊匹配)+音节(模糊匹配)+音素(*n*gram-WFST) | 0.588 | 0.552 |

(3)融合系统检索时间复杂度分析

多流信息融合系统采用三个 STD 子系统进行独立的集外词检索,最后进行三系统的检索结果融合。其中 STD 子系统由两部分构成:子词解码部分和检索部分,子词解码时间依赖于解码器的速度,因此三系统的总解码时间基本上等于单系统的三倍。

而对于本文中使用的检索算法,*n*gram-WFST 检索和模糊匹配检索系统时间复杂度各有不同,模糊匹配由于检索到查询词部分 triphone 发音既可召回,相对于 *n*gram-WFST 完全匹配算法,搜索空间变大,搜索时间更长;具体检索耗费时间如表 4 所示。本文实验中,主机配置为:

Pentium(R) Dual-Core CPU 3.00 GHz,2 GB 内存。

**表 4 开发集上的各子词 STD 系统的检索耗时**

**Table 4 Search time of different sub-word units STD system on development set**

| 子词 STD 系统 | 检索算法 | 开发集检索 313 词所用时间/s |
|---|---|---|
| 词片 | 模糊匹配 | 7.65 |
| 音节 | 模糊匹配 | 7.63 |
| 音素 | *n*gram-WFST | 6.24 |

从表 4 可知,在已经建立好索引的情况下,当采用串行方式时,三系统总计检索时间为三者之和,检索开发集上 313 个词需要耗费 21.52 s。但是值得注意的是,多流融合 STD 系统由三个完全独立的子系统构成,完全可以并行处理,这时融合系统检索速度等同于最慢系统的检索速度,检索开发集上 313 个词只需耗费 7.65 s。

## 6 结束语

虽然词片和音节分别以数据驱动和语言学规则两种不同方式选择,由于两者均为可变长度的音素序列,在一定程度上具有相似性,导致两者的互补性较弱,因而融合之后性能提升不大。由于音素语言模型约束性较弱,无法充分利用上下文信息,因此音素识别器的识别混淆度很大,识别生成的 lattices 中包含很多音节和词片不包含的信息,从而使得音素和音节、词片间互补性较强,融合之后能够显著提高检索性能。

本文首先分别利用音素、音节和词片构建 STD 系统用于集外词检索,接着研究了各子词对集外词的建模能力,并针对各子词 STD 系统的特

性，对音素系统使用完全匹配的 ngram-WFST 检索、对词片和音节进行模糊匹配检索，提高单系统性能，最后利用线性回归得分融合策略，较大程度提高了系统性能。

**参考文献：**


[1] NIST. The spoken term detection (STD) 2006 evaluation plan[EB/OL]. http://www.itl.nist.gov/iad/mig/tests/std/2006/ docs/std06-evalplan-v10.pdf,2006-9-13.

[2] Szoke I, Burget L, Cernocky J, et al. Sub-word modeling of out of vocabulary words in spoken term detection[C]//Proceedings of IEEE Workshop on Spoken Language Technology. Goa, India: IEEE, 2008: 273-276.

[3] Wallace R, Vogt R, Sridharan S. A phonetic search approach to the 2006 NIST spoken term detection evaluation[C]//Proceedings of Interspeech. Antwerp. Belgium: IEEE, 2007: 2393-2396.

[4] Rastrow A, Sethy A, Ramabhadran B, et al. Towards using hybrid word and fragment units for vocabulary independent LVCSR systems[C]//Proc of Interspeech. Brighton, UK: IEEE, 2009: 1931-1934.

[5] Larson M, EickEler S. Using syllable-based indexing features and language models to improve German spoken document retrieval[C]//Proceedings of Eurospeech. Geneva, Switzerland: IEEE, 2003: 1217-1220.

[6] Liu C, Wang D, Tejedor J. N-gram FST indexing for spoken term detection[C]//Proceedings of Interspeech. Portland, Oregon, USA: IEEE, 2012.

[7] Xu Y, Guo W, Shansu, et al. Spoken term detection for OOV terms based on phone fragment[C]//Proceedings of International Conference on Audio, Language and Image Processing. Shanghai, China: IEEE, 2012:1031-1034.

[8] Brummer N, Burget L, Cernocky J, et al. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006[J]. IEEE Trans on Audio, Speech and Language Processing, 2007, 15(7): 2072-2084.

[9] Bartlett S, Kondrak G, Cherry C. On the syllabification of phonemes[C]//Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies. Boulder, Colorado, USA: Association for Computational Linguistics, 2009:308-316.

[10] 刘辉，杨俊安，许学忠. 基于 HMM 和 SVM 串联模型的低空飞行目标声识别方法[J]. 数据采集与处理，2010，25(6)：751-755.
Liu Hui, Yang Junan, Xu Xuezhong. Low altitude passive acoustic target recognition based on HMM and SVM[J]. Journal of Data Acquisition and Processing, 2010, 25(6): 751-755.

[11] Stolcke A. SRILM - An extensible language modeling toolkit[C]// Proceedings of the International Conference of Spoken Language Processing. Denver, Colorado, USA: IEEE, 2002: 901-904.

[12] Wang D, Tejedor J, King S, et al. Term-dependent confidence normalization for out-of-vocabulary spoken term detection[J]. Journal of Computer Science and Technology, 2012, 27(2): 358-375.


**作者简介：**熊世富(1990-)，男，硕士研究生，研究方向：语音识别，语音检索，E-mail: domine@mail.ustc.edu.cn；郭武(1973-)，男，副教授，研究方向：语音识别，说话人识别，语种识别，音频检索。