文章编号：2095-4980(2013)06-0958-06

基于点过程模型连续语音关键词检测

王　勇，张连海

(信息工程大学信息 信息系统工程学院，河南 郑州 450002)

摘　要：提出了基于点过程模型(PPM)的连续语音关键词检测方法。该方法首先利用时态模式(TRAP)特征和多层感知器(MLP)计算每个音素的帧级后验概率，在此基础上，将语音可看作多个相互独立的事件(音素)，利用泊松过程对事件建立点过程模型，最后通过计算似然比达到关键词检测目的。实验结果表明，对 8 kHz 采样语音，关键词平均召回率和准确率分别可达 69.5%和 82%以上。

关键词：关键词检测；音素后验概率；泊松过程；点过程

中图分类号：TN912.34；TP391　　**文献标识码：**A　　**doi：**10.11805/TKYDA201306.0958

Spotting keywords in continuous speech based on Point Process Models

WANG Yong，ZHANG Lian-hai

(School of Information System Engineering，Information Engineering University，Zhengzhou Henan 450002，China)

Abstract：A keyword spotting method is proposed based on Point Process Model(PPM) in continuous speech. Frame-level phone posterior probability is computed by using TempoRAl Patterns(TRAP) and Multiple Layer Perception(MLP). The speech can be considered as independent events(phones), and PPM can be set up by using Poisson process. The likelihood ratio is calculated to estimate whether the keyword is uttered. The experimental results show that the average recall and precision rate of keywords are above 69.5% and 82.0% with 8 kHz sampling frequency for speech, respectively.

Key words：keywords spotting；phone posterior probability；Poisson process；Point Process

关键词检测有着广泛的应用领域，如命令控制、语音监听、语音拨号、数据查询等。目前常用的关键词检测方法有基于垃圾模型关键词检测和基于大词汇量连续语音识别(LVCSR)关键词检测。基于垃圾模型的关键词系统中，利用关键词模型和垃圾模型组成并行网络对关键词进行搜索。对关键词网络加上合适的奖赏，或者给垃圾模型合适的惩罚，使得关键词得分超过垃圾得分，从而得到关键词检测结果[1]。1990 年，Rose 等人[2]提出了基于连续语音识别技术的关键词识别系统，采用垃圾模型和维特比回溯技术，可以检测出连续语音中任意多个关键词。为提高系统鲁棒性，1994 年 Bourlard 等人[3]提出在线垃圾模型，在维特比解码过程中，对输入的每帧语音，计算关键词所含每一个语音单元(音素、状态等)的似然得分，只有当输入语音与关键词匹配度较高时，关键词才可能在与在线垃圾的竞争中胜出。基于 LVCSR 关键词检测方法需对输入语音进行识别，得到基于词的 Lattice，再从 Lattice 中检索出关键词。2006 年，Wallace 等人[4]在音素层进行解码，采用动态匹配技术，在 Lattice 上对关键词进行快速定位，取得了较好的检出效果。2007 年，Thambiratnam 等人[5]提出利用最小编辑距离(Minimum Edit Distance，MED)的方法对关键词进行检出，采用专家先验知识设置最小编辑距离的代价函数，通过不同的惩罚函数进行 MED 搜索，检测效果得到了提升。2010 年，Chang 等人[6]提出了音节误判代价矩阵的方法，综合考虑插入、删除、替换错误，代价函数通过训练数据进行估计，该方法性能相对更好。

以上 2 种关键词检测方法一直在该领域占据主导地位，然而它们仅通过训练模型从概率统计的角度描述语音信号，并没有将语音产生过程中丰富的语音学、语言学等知识加入检测系统。虽然检测性能较以往有了大幅度提高，但仍然不能满足现实的需求，与人耳的语音判别能力相比，仍有较大差距。因此，语言学家将研究目光投向了人类语音识别机制[7]，期望从人对语音的感知过程中找到制约识别性能提高的关键因素。与机器识别不同，人耳并非简单地将语音信号先识别成音素，由音素组合成字、词直至语句，而是充分利用语音信号中蕴含的丰富信息进行综合判决。具体来说，这些信息包括音素/音节的边界、音素类别、韵律学、语种、说话人的性别、情感

收稿日期：2012-12-26；修回日期：2013-02-06

以及周边环境等，学者们称这些信息为语音属性(Speech Attribute)，并将属性在时间上发生突变的点称为“语音事件”(Speech Event)[8]。因此，语音信号可以看作各种事件组成的事件序列。人耳正是通过对各种“语音事件”的正确检测与分辨理解语音内容的。所以，可以模仿人耳的感知过程，建立基于事件检测的语音识别系统，达到语音识别的目的。本文中将构成语音的音素看作独立事件，用基于时间的“点”表示，那么语音可用基于音素事件的多个独立点序列也就是点过程表示，然后利用泊松过程对这些独立点过程建立点过程模型[9]，计算某段时间内似然比函数值，实现关键词检测目的。

1　点过程建立

基于 PPM 关键词检测分为 2 个阶段：点过程建立和关键词检测。首先，计算帧级音素后验概率并建立点过程，其次由关键词模型计算似然比，通过设定阈值判断候选语音是否为关键词。

图 1 为本文使用方法检测关键词结构图。语音信号首先经过信号处理单元 S 得到帧级音素后验概率 X，检测器 D 给定适当的阈值，X 经过检测器 D 后，转化为 n 个独立的点过程 P，那么语音信号就可以用 n 个独立的点过程表示，再由关键词模型计算似然比，最终实现关键词检测。以下分别作详细介绍。

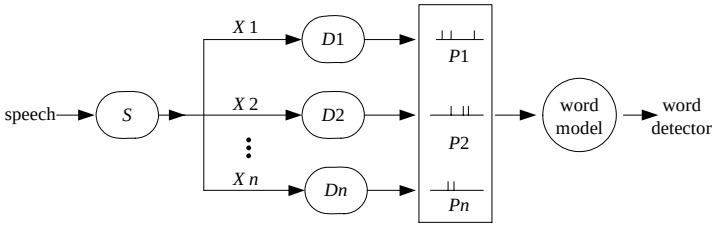

Fig.1 Architecture of keywords spotting framework
图 1 关键词检测结构图

1.1 音素后验概率

1.1.1 TRAP 特征

目前，语音识别声学特征主要使用美尔频率倒谱系数(Mel Frequency Cepstrum Coefficient，MFCC)、感知线性预测系数(Perceptual Linear Predictive，PLP)等频谱参数，但这些参数只使用了 20 ms,30 ms 左右的语音信息，极易受到噪声的影响。TRAP[10]是一种长时属性，反映了长时间特征变化情况，有效地利用语音信号之间的相关性，因此能够提高语音识别的性能。本文将 TRAP 特征引入到音素后验概率的检测之中。

1.1.2 多层感知器

多层感知器(Multiple Layer Perception，MLP)是神经网络的一种，由输入层、输出层和一个或多个隐含层组成。通过隐含层上非线性激活函数将输入映射至非线性空间，从而使模型具有非线性判决能力。图 2 所示为本文使用 3 层 MLP 结构。

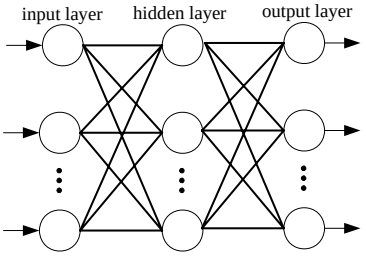

Fig.2 Architecture of MLP
图 2 MLP 结构

1.1.3 音素后验概率

基于 TRAP 结构的音素后验概率检测流程如图 3 所示，具体步骤如下：

1) 预处理：选择帧长与帧移分别为 25 ms 和 10 ms，对语音信号进行预加重、加汉明窗，将频谱转化为梅尔频标后进行三角窗滤波，每帧语音信号输出为 23 个子带能量的一维向量。

2) 拼接加权：将当前帧与其前 n 帧的子带能量拼接成一个长序列，称为左子带序列；将当前帧与其后 n 帧的子带能量拼接组成右子带序列。由于语音信号帧与帧之间距离越近，相关性越强，距离越远，相关性越弱，因此，给距离当前帧较远的帧分配较小的权值，距离当前帧较近的帧分配较大的权值，并且同一帧内的各个子带能量系数分配的权值相同。然后，分别对加权后的序列进行离散余弦变换(Discrete Cosine Transform，DCT)，将变换后的系数规范化后作为低层 MLP 输入特征。

3) 后验概率检测：采用低层 MLP 分别对左、右 2 个子带序列进行音素检测，对输出结果进行非线性变换，然后将低层 2 个 MLP 的输出拼接成新的向量，并作为高层 MLP 的输入特征，最后高层 MLP 的输出为音素后验概率识别结果。

由于 TRAP 结构使用了上下文相关信息，因此最终检测结果准确率相对更高。图 4 所示为词“problem”帧级音素后验概率图，颜色越深表示该帧信号为某个音素的概率越大。

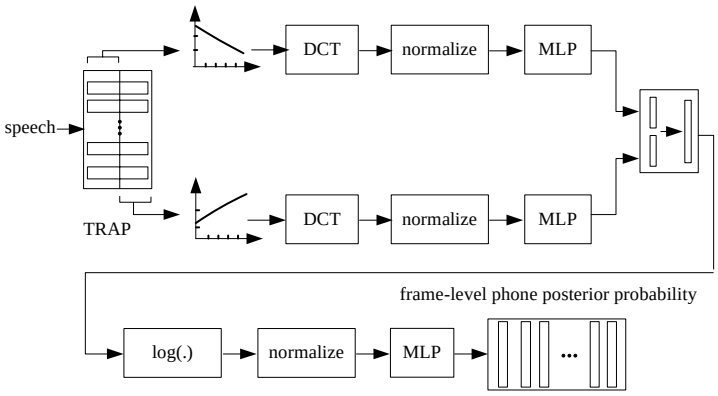

Fig.3 Architecture of detecting phone posterior probability
图 3 音素后验概率检测流程图

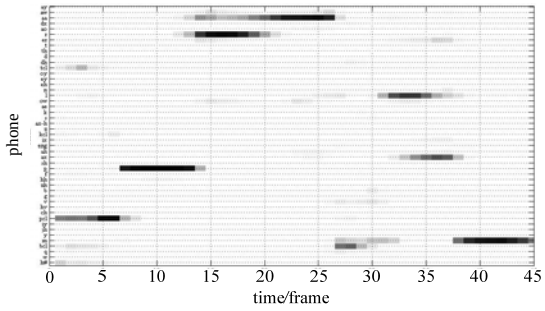

Fig.4 Frame-level phone posterior probabilities of "problem"
图 4 "problem"帧级音素后验概率

1.2 点过程

在计算出帧级音素后验概率的基础上，得到语音信号音素后验概率矩阵。对于后验概率矩阵的每一行，也就是语音信号每一帧，取后验概率最大值，其余后验概率置为 0。然后给定阈值 γ，若该帧信号后验概率最大值大于 γ，则将其置为 1，表示该帧语音信号是某个音素，若小于 γ，则将其置为 0。由此可以将音素后验概率矩阵 0,1 离散化，得到语音信号点过程表示。图 5 所示为词“problem”的点过程表示，其中的点表示该帧信号为“problem”相应的某个音素，点的个数代表音素出现的次数。

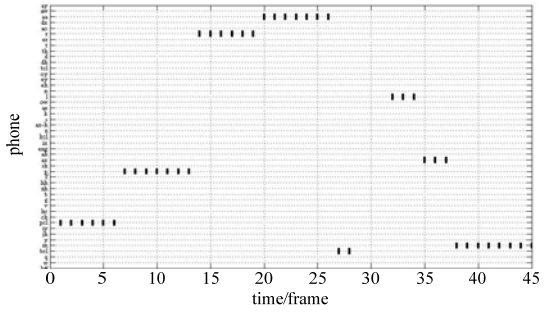

Fig.5 Point process representation of "problem"
图 5 "problem"点过程表示

2 关键词检测

2.1 泊松过程

在得到语音点过程后，还需建立关键词模型。泊松过程是一种时间连续、状态离散的随机过程[11]，本质上描述的是事件在给定时间内发生次数为 n 的概率。它在物理学、生物学、服务系统以及可靠性理论等领域有着广泛的应用。本文使用泊松过程对点过程建模。

设 $N(a_i,b_i]$ 表示半开区间 $(a_i,b_i]$ 时间间隔内事件发生次数，其中 $a_i<b_i\leqslant a_{i+1}$，那么：

$$P\{N(a_i,b_i]=n_i,i=1,2,\cdots,k\}=\prod_{i=1}^{k}\frac{[\lambda(b_i-a_i)]^{n_i}}{n_i!}\mathrm{e}^{-\lambda(b_i-a_i)} \tag{1}$$

上述定义中包含了 3 个条件：1) 每个区间 $(a_i,b_i]$ 内，事件发生次数均服从泊松分布；2) 每个区间 $(a_i,b_i]$ 内，事件发生次数为独立随机变量；3) 事件发生次数满足平稳分布，只依赖于时间间隔长度 b_i-a_i。其中 λ 为泊松过程的速率，本文中为音素到达率。

在相同分布条件下，对于某段时间内事件发生次数有限的泊松过程，可用似然度代替概率描述给定观测时间内事件发生有限次的可能性[12]。假设时间 $(0,T]$ 内事件发生 n 次，时刻为 $t_1,\cdots,t_n$，由式(1)得到，事件在时间 $(t_i-\Delta,t_i]$ 内发生一次而在 $(0,T]$ 的其他时间内没有事件发生的概率为 $\mathrm{e}^{-\lambda T}\prod_{i=1}^{n}\lambda\Delta$。令 $\Delta\to 0$，并除以 Δ^n，得到概率密度，也就是似然度函数为：

$$L_{(0,T]}(N;t_1,\cdots,t_n)=\lambda^n\mathrm{e}^{-\lambda T} \tag{2}$$

本文中使用似然度代替概率描述事件的可能性，为方便描述，仍使用概率和符号 P。

2.2 关键词模型

设有语音点过程 $R=\{N_\varphi\}_{\varphi\in F}$，其中 $N_\varphi=\{t_1,\cdots,t_{n_\varphi}\}$ 为音素 φ 点过程表示，F 为该段语音中所有音素的集合，$t_i\in(0,T]$ 为音素 φ 出现时刻，n_φ 为时间 $(0,T]$ 内音素 φ 出现次数。采用齐次泊松过程描述语音点过程 $R=\{N_\varphi\}_{\varphi\in F}$，那么，由式(2)可得，时间 $(0,T]$ 内音素 φ 出现 n_φ 次的概率为：

$$P(N_\varphi)=\lambda_\varphi^{n_\varphi}\mathrm{e}^{-\lambda_\varphi T} \tag{3}$$

由于语音点过程 $R=\{N_\varphi\}_{\varphi\in F}$ 中所有音素 φ 均相互独立，那么，时间 $(0,T]$ 内，$R=\{N_\varphi\}_{\varphi\in F}$ 中每个音素 φ 相应出现 n_φ 次的概率为：

$$P(R)=\prod_{\varphi\in F}\lambda_\varphi^{n_\varphi}\mathrm{e}^{-\lambda_\varphi T} \tag{4}$$

训练齐次泊松过程模型即估计每个音素 $\varphi\in F$ 的到达率 λ_φ。给定 M 段时长为 T 训练语音，音素 φ 出现总次数为 K_φ，那么到达率 λ_φ 的最大似然估计值为：

$$\hat{\lambda}_\varphi=\arg\max_\lambda K\log\lambda-\lambda MT=\frac{K_\varphi}{MT} \tag{5}$$

然而，在实际情况下，关键词时长 T 内，每个音素 φ 分布并不是平稳的，例如图 5 中，音素 $|pcl|$ 只分布于前几帧，所以并不能直接使用齐次泊松过程对关键词建模。因此考虑对关键词时长 T 进行分段，由于在每一段内音素 φ 的分布是平稳的，可以使用分段连续的齐次泊松过程对关键词建模。将音素点过程 $N_\varphi=\{t_1,\cdots,t_{n_\varphi}\}$ 按照时间 $(0,T]$ 均匀分为 D 段，每段时长为 $\Delta T=T/D$，那么每一段时间 $(d\Delta T,(d+1)\Delta T]$ 内音素 φ 到达率为 $\lambda_\varphi(t)=\lambda_{\varphi,d}$，$t\in(d\Delta T,(d+1)\Delta T]$，其中 $d=\lceil t/\Delta T\rceil$，每一段时间内音素 φ 出现次数为 $n_{\varphi,d}$。时间 T 内，音素 φ 出现总次数 $n_\varphi=\sum_{d=1}^{D}n_{\varphi,d}$。那么，用 D 段独立齐次泊松过程对音素点过程 $N_\varphi=\{t_1,t_2,\cdots,t_{n_\varphi}\}$ 建模，时间 $(0,T]$ 内音素 φ 出现 n_φ 次的概率为：

$$P(N_\varphi)=\prod_{d=1}^{D}P(N_{\varphi,d})=\prod_{d=1}^{D}\lambda_{\varphi,d}^{n_{\varphi,d}}\mathrm{e}^{-\lambda_{\varphi,d}\Delta T} \tag{6}$$

同样，音素 φ 第 d 段到达率参数 $\lambda_{\varphi,d}$ 也可以通过最大似然估计得到，为：

$$\hat{\lambda}_{\varphi,d}=\frac{K_{\varphi,d}D}{MT} \tag{7}$$

图 6 所示为词“problem”分段到达率参数，颜色深浅代表到达率的大小。那么，时间 $(0,T]$ 内，语音点过程 $R=\{N_\varphi\}_{\varphi\in F}$ 中每个音素 φ 相应出现 n_φ 次的概率为：

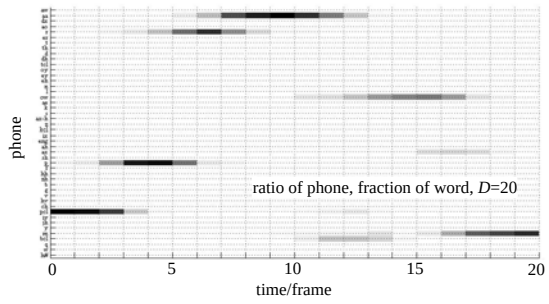

Fig.6 Fraction of word "problem" and ratio of each phone
图 6 "problem"分段及音素到达率

$$P(R)=\prod_{\varphi\in F}\prod_{d=1}^{D}\lambda_{\varphi,d}^{n_{\varphi,d}}\mathrm{e}^{-\lambda_{\varphi,d}\Delta T} \tag{8}$$

2.3 关键词检测

设候选语音时长 T，考虑似然比函数：

$$d_{w,T}=\log\left[\frac{P(R\mid\theta_{w,T}=1)}{P(R\mid\theta_{w,T}=0)}\right] \tag{9}$$

式中：$\theta_{w,T}:\mathbb{R}\to\{0,1\}$ 为指示函数，当候选语音是关键词时，$\theta_{w,T}=1$，其他情况下，$\theta_{w,T}=0$。将式(4)、式(8)带入式(9)，得到：

$$d_{w,T}=\log\left[\frac{P(R_T\mid\theta_{w,T}=1)}{P(R_T\mid\theta_{w,T}=0)}\right]=\log\left[\frac{\prod_{\varphi\in F}\prod_{d=1}^{D}\lambda_{\varphi,d}^{n_{\varphi,d}}\mathrm{e}^{-\lambda_{\varphi,d}\Delta T}}{\prod_{\varphi\in F}\prod_{d=1}^{D}\lambda_\varphi^{n_{\varphi,d}}\mathrm{e}^{-\lambda_\varphi\Delta T}}\right]=\log\left[\prod_{\varphi\in F}\prod_{d=1}^{D}\left(\frac{\lambda_{\varphi,d}}{\lambda_\varphi}\right)^{n_{\varphi,d}}\mathrm{e}^{-\lambda_{\varphi,d}+\lambda_\varphi}\right] \tag{10}$$

式中：$\lambda_{\varphi,d}$，λ_φ 均已由训练得出。当 $\lambda_{\varphi,d}$=0时，用某个极小值代替。给定语音，以时长 T 对语音搜索，计算式(10)似然比函数值，比较其与设定阈值的大小，即可判断该段语音是否为所检测关键词。

2.4 局部语速调整

通常情况下，由于语速对关键词的影响，关键词时长 T 是不定值。因此，需要先估计关键词平均时长 T_{aver}，那么某一段内音素出现的相对次数 $n_{\varphi,d}'=\frac{n_{\varphi,d}T_{\text{aver}}}{T}$，$n_{\varphi,d}$ 是时长为 T 的关键词中某音素的实际出现次数。同时，为

避免个别音素占似然比权重过大，需要给某段内音素出现次数 $n_{\varphi,d}$' 设置上限 $n_{\varphi,d}{}'_{\max}=T_{\text{aver}}/D$。给定似然比阈值，以不同时长 $\tilde{T}$ 对语音进行搜索，计算似然比，当似然比局部极大值大于设定似然比阈值时，即可判定该段时长为 $\tilde{T}$ 的语音为需检测的关键词。

3 实验

3.1 实验配置

本文实验使用 TIMIT 语料库，排除其中用于说话人识别实验的 SA1 和 SA2 中的语句，选择训练集中 3 296 个语句和测试集中 1 344 个语句进行实验，时间共计 3.95 h。由于本文使用方法需训练关键词分段到达率，故选择 TIMIT 语料库中出现频次较高的词进行相关实验。

TIMIT 语料库中共含有 61 个音素单元，其划分较为精细。根据 CMU/MIT 标准，对 TIMIT 中发音类似的音素进行合并，由 61 个音素映射为 47 个[13]，具体对应关系如表 1 所示。

表1 TIMIT 中音素映射关系

Table1 Mapping of phones of TIMIT

TIMIT	experiment of this paper
/epi/ /h#/ /pau/	/h#/
/ux/ /uw/ /w/	/w/
/gcl/ /g/	/g/
/em/ /m/	/m/
/ih/ /y/	/y/
/ah/ /uh/	/ah/
/en/ /n/ /nx/	/n/
/axr/ /er/	/er/
/eng/ /ng/	/eng/
/dcl/ /el/ /l/	/l/

3.2 实验结果

召回率(Recall)和准确率(Precision)是衡量关键词检测性能的 2 项重要指标，可以用来对检测的结果进行量化评估，式(11)、式(12)为二者计算公式。一般而言，召回率和准确率是互相对立的，一个指标的上升伴随着另一个指标的下降。在应用过程中，一般寻找二者的平衡点，使得召回率与准确率均能满足实际的需求。

$$\text{召回率}=\frac{\text{正确检测关键词个数}}{\text{实际关键词个数}}\times 100\% \tag{11}$$

$$\text{准确率}=\frac{\text{正确检测关键词个数}}{\text{正确检测关键词个数}+\text{错误检测关键词个数}}\times 100\% \tag{12}$$

本文选取关键词容错误差为 ±30 ms，表 2 所示为本文使用方法关键词检测结果。

表 2 PPM 关键词检测结果

Table2 Results of keywords spotting based on PPM

keyword	average duration(frame)	recall	precision
keep	20	62.9%	—
take	21	73.5%	—
every	24	77.8%	77.5%
never	28	66.7%	81.5%
cream	29	72.0%	80.0%
place	30	71.7%	73.9%
people	38	52.3%	81.8%
first	43	78.0%	82.0%
system	45	78.3%	87.0%
problem	50	61.5%	92.3%
average	32.8	69.5%	82.0%

由于本文中并未考虑词边界信息，在进行关键词搜索时，若某个词的发音完全包含另一个词的发音，会将该词作为关键词检出。例如搜索关键词“every”(发音为 |*eh v r iy*|)时，由于词“everyone”(发音为 |*eh v r iy w ah n*|) 完全包括词“every”的发音，所以会将“everyone”的前半部分作为关键词检出，本文中未将这种情况作为插入错误进行统计。对于包含音素较少的词，如“take”(发音为 |*t ey kcl*|)，由于英文单词中包含发音 |*t ey kcl*| 情况较多，本文中未统计准确率。

实验中，为提高系统关键词召回率，在准确率允许条件下，可适当将易混淆的音素如 |*iy*|,|*ix*| 等作为同一音素处理。例如某候选语音通过音素后验概率检测发音为 |*eh v r ix*|，可酌情将其作为发音 |*eh v r iy*| 处理。

理论上，当关键词时长越长，包含的音素越多时，建立点过程模型可利用的信息越多，关键词模型的复杂度越高，容易引起的混淆越少，相应的关键词召回率、准确率应该越高。在实验过程中，随着关键词包含音素的增加，关键词检测准确率呈上升趋势，但是由于关键词包含音素的增加，音素后验概率错误也就相应增多，关键词召回率不一定能相应提高。如表 2 中，当关键词包含音素由 4 个(every)上升为 8 个(problem)时，关键词检测准确率由 77.5%上升为 92.3%，但是召回率却由 77.8%下降为 61.5%。由于语料库中，“people”存在较多的发音变体，且尾音 |*l*| 经常脱落不发音，由表 2 可以看出，“people”召回率仅为 52.3%，相较其他词有较大差距。

表 3 所示为在相同条件下，本文方法和基于 HMM 垃圾模型关键词检测结果。由表可以看出本文方法与基于 HMM 垃圾模型关键词检测方法性能接近，但本文方法将复杂语音信号转化为稀疏音素事件点过程表示，算法复杂度有了大幅度

表 3 PPM 与 HMM 关键词检测结果比较

Table3 Comparison of keywords spotting performance between PPM and HMM

	average recall	average precision
PPM	69.5%	82.0%
HMM	70.9%	83.7%

降低。由于本文方法仅使用了音素后验概率信息，后续研究中可以将语言知识(词法、语法、上下文信息等)与本文方法相结合，进一步提高关键词检测性能。

4　结论

基于 PPM 语音关键词检测方法，以音素事件作为基本单元，将复杂语音信号转化为稀疏音素事件点过程表示，采用泊松过程对关键词建模，实验表明，能够有效地检测关键词。在系统性能接近的基础上，大幅降低了系统的复杂度。由于音素后验概率对关键词检测性能具有决定性作用，因此如何提高音素后验概率准确率问题亟待解决。本文在计算音素到达率时，并未考虑语速的影响。在计算关键词分段到达率之前，可先进行语速调整再计算，进一步提高关键词模型的精确度。本文只使用音素事件建立点过程模型，实际上，可以根据其他语音事件建立多个点过程模型，然后将各个点过程进行融合，建立更为复杂精确的关键词模型，此问题值得进一步的关注与研究。

参考文献:

[1] 孙成立. 语音关键词识别技术的研究[D]. 北京:北京邮电大学, 2008:6–7. (SUN Chengli. Research of speech keyword recognition technology[D]. Beijing:Beijing University of Posts and Telecommunications, 2008:6–7.)

[2] Rose R C,Paul D B. A hidden Markov model based keyword recognition system[C]// Proceedings of ICASSP. Albuquerque, NM:IEEE Signal Processing Society, 1990:129–132.

[3] Bourlard H,Bart D. Optimizing recognition and rejection performance in word spotting systems[C]// Proceedings of ICASSP. Adelaide, Australia:IEEE Signal Processing Society, 1994:373–376.

[4] Roy Wallace,Robbie Vogt,Sridha Sridharan. A phonetic search approach to the 2006 NIST spoken term detection evaluation[C]// Proceedings of Interspeech. Antwerp,Belgium:[s.n.], 2007:2385–2388.

[5] Kishan Thambiratmann, Sridha Sridharan. Rapid yet accurate speech indexing using dynamic match lattice spotting[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007:346–357.

[6] Chang Woo Han,Shin Jae Kang,Chul Min Lee. Phone Mismatch Penalty Matrices for Two-Stage Keyword Spotting Via Multi-Pass Phone Recognizer[C]// Proceedings of Interspeech. Makuhari,Japan:[s.n.], 2010:202–205.

[7] Dusan S,Rabiner L R. On Integrating Insights from Human Speech Perception into Automatic Speech Recognition[C]// Proceedings of Interspeech. Lisboa,Portugal:[s.n.], 2005:1233–1236.

[8] Stevens K N. Toward a model for lexical access based on acoustic landmarks and distinctive features[J]. Journal of the Acoustical Society of America, 2002,111(4):1872–1891.

[9] Jansen,Niyogi. Point process models for spotting keywords in continuous speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2009,17(8):1457–1470.

[10] Frantisek Grezl. Trap-Based Probabilistic Features for Automatic Speech Recognition[D]. Czech:BRNO University of Technology, 2007.

[11] 刘次华. 随机过程[M]. 4 版. 武汉:华中科技大学出版社, 2008:27–40. (LIU Cihua. Stochastic Process[M]. 4th ed. Wuhan: Huazhong University of Science and Technology Press, 2008: 27–40.)

[12] Daley D J,Vere-Jones D. An Introduction to the Theory of Point Processes[M]. New York:Springer, 2002:19–26.

[13] Lee K F,Hon H W. Speaker-independent Phone Recognition using Hidden Markov Models[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, 1989,37(11):1641–1648.

作者简介:

王　勇(1987–)，男，江苏省连云港市人，在读硕士研究生，主要研究方向为连续语音关键词检测.email:wyong0609@yahoo.cn.

张连海(1971–)，男，河南省平顶山市人，博士，副教授，主要研究方向为雷达信号处理、医学信号处理、一般性的信号处理.