维普资讯 http://www.cqvip.com

基于限制模型规模和声学置信度的关键词检出方法*)

郑铁然　张　战　韩纪庆
(哈尔滨工业大学计算机科学与技术学院　哈尔滨 150001)

摘　要　在基于令牌传递算法的关键词检出技术中，为改进实时性，本文首先从限制模型规模的角度，提出了限制上下文相关的词内相关音素模型。针对误识率高的问题，提出了基于声学置信度的关键词确认方法，并实现了多次解码机制，提高了识别性能。其次，从改进解码算法的角度，研究了剪枝和控制最大激活模型数两种策略对识别性能的影响，并结合确认机制进行关键词检出，获得了满意的结果。

关键词　关键词检出，关键词确认，置信度，令牌环，裁剪策略

Keyword Spotting Based on Restricting Model and Acoustic Confidence Measure

ZHENG Tie-Ran　ZHANG Zhan　HAN Ji-Qing
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

Abstract　In order to satisfy the real-time requirement in keyword spotting based on token passing, firstly a restricting context phone model, in-word-dependent phone model, is proposed from the aspect of reducing the size of model, and for reducing the false alarm rate, a keyword verification method based on acoustic confidence measure is explored, and a multi-decoding approach is also realized. Secondly, the influences of decoding effect by the beam pruning and maximum active model pruning are studied from the aspect of improving the decoding algorithm, when they are combined with the verification method, the better experiment results are got.

Keywords　Keyword spotting, Keyword verification, Confidence measure, Token passing, Pruning strategy

1　引言

关键词检出(Keyword Spotting, KWS)技术就是在连续无限制的自然语音流中识别出一组给定词，它涉及如下几方面的内容：首先是选择合适的识别基元，并确定关键词与词表外词的声学建模方式；其次在检出过程中要进行初选和确认工作，前者通过解码找到大量可能的关键词候选，后者通过某种置信测度(confidence measures)来进行确认与拒识。

目前 KWS 主流的方法都是建立在 HMM(Hidden Markov Model)框架下，包括填充(Filler)模板方法和 Two-pass 解码机制的方法。研究者们关注的焦点主要放在填充模板建模和关键词确认等问题上。在填充模板建模方面，经典的方法有：基于子词或整词作为填充模板的方法[1]、N-best 在线垃圾建模方法[2]等。在关键词确认方面，有基于反词模型的拒识方法[3]、基于后验概率的方法[4,5]、采用 One-pass 策略的方法[6]等。

本文在一种基于令牌传递(Token Passing)算法的关键词检出技术基础上，首先从限制模型规模的角度，提出了限制上下文相关的词内相关音素模型，提高了识别速度。为减少误识率，提出了基于声学置信度的关键词确认方法，取得了较好的结果。其次，从改进解码算法的角度，研究了剪枝和控制最大激活模型数两种策略对识别性能的影响，并结合确认机制进行关键词检出，获得了满意的结果。

2　基于令牌传递算法的关键词检出技术

令牌传递算法[7]是一种 Viterbi 帧同步算法的改进算法。由于 Viterbi 算法在搜索过程中需要记录很多相关信息，因此可以加入一种称为令牌的数据结构，用来传递和存储这些信息。

若采用 HMM 模型来表示语音，其从状态 i 到 j 间的转移概率用 P_{ij} 表示，对语音帧序列 $O_1 \cdots\cdots O_T$，每个状态 j 和待测语音帧序列的第 t 帧匹配的相似度值用 d_{jt} 表示。定义一个数据结构令牌，每个令牌可以在状态间按照状态转移规则传递，且自身带有一个累计概率信息的值 $S_j(t)$，它代表该令牌在第 t 帧处于状态 j 所持有的概率累计值。

令牌传递算法如下：

```
初始化：
    在每个模型初始状态生成一个令牌，概率累计初始值为 0；
    所有其他状态的令牌初始值为∞；
递推：
    For t=1 到 T
        For 每个状态 i
            向状态 i 的每一个连接状态 j 发送 i 中令牌的一个复制；
            并计算概率累计值 Sj(t)=Si(t−1)+pij+dj(t)；
        End；
        删除状态 i 的初始令牌；
        For 每个状态 j
            找到状态 j 中概率累计值最小的令牌；
            删除其他令牌；
        End；
    End；
终止：
    检查所有模型的终止状态，拥有最小概率累计值的令牌所在的模型为最终匹配的结果。
```

应用令牌传递算法进行关键词检出，首先要建立关键词模型和垃圾模型的并行识别网络。通过这种网络就可进行关键词识别。关键词模型可由子词级别的上下文相关的音素模型拼接而成。垃圾模型是为了使系统能够正确地识别非关键词，以降低误识率，其模型的选取不固定。垃圾模型首先应包

*)国家自然科学基金项目(No. 60575030)资助。**郑铁然**　讲师，博士生，主要从事语音信号处理方面的研究工作；**张　战**　硕士，主要从事语音信号处理方面的研究工作；**韩纪庆**　博士，教授，博士生导师，主要从事语音信号处理方面的研究工作。

维普资讯 http://www.cqvip.com

括所有或至少大部分的非关键词信息，这样才能预测到非关键词出现的大部分甚至所有情况。其次，垃圾模型的规模不能太过庞大。

3 基于限制模型的关键词检出技术

音节是很合适的垃圾模型的候选，汉语中存在的音节共有1325个，加上1个静音模型，共有1326个垃圾模型。音节由音素拼接而成，每个音素模型都是上下文相关的三音素模型。如选择所有的音素模型作为垃圾模型，包括静音模型在内，则仅有97个音素模型。但初步的实验表明，采用这样的方法识别速度慢。

由于每个音节的前后两个边界音素互相组成上下文，都具有相关性，根据乘法原则，假设共有 N 个音节，则模型规模 C 为

$$C=N^2+N \tag{1}$$

其中 N^2 代表边界音素的可重复全排列，N 为音节中间的音素数。因为会有重复上下文相关音素，这个 C 的计算是近似的。

由于音素模型的上下文相关性导致了模型规模的剧增，因此应该限制其相关性，缩小模型规模。首先，可以完全限制音素的上下文相关，即使用上下文无关的单音素模型，但研究表明性能过低。为此，本文提出一种折衷的限制方式，允许词内音素上下文相关，而限制词边界音素上下文无关。将这种音素模型称为词内相关的音素模型。

词内相关的音素模型比全相关音素模型的规模小很多。对全相关的音素模型，其规模为 1325^2 数量级。在同样的识别网络上，假设音节数为 N，则词内相关的音素模型数 C' 为

$$C'=N+97 \tag{2}$$

其中 N 为词内部音素的三音素模型个数，97为单音素总数。对于全音节的网络而言，$C'=1423$，远小于三音素总数。因此，采用词内相关的音素模型可以改进关键词检出的速度。

4 基于声学置信度的关键词确认技术

关键词确认阶段的工作就是在保证检出率不受较大影响的情况下，拒识错误的候选，降低误识率。置信度是评价正确概率的一种量度。语音识别中，置信度被定义成一个用来衡量模型和观测数据之间匹配程度的函数，且这个函数的值对于不同的观测数据具有可比性。

设模型 W 为类1，除 W 外的所有模型 $\overline{W}$ 为类2，判断语音 O 是否由 W 产生的问题变成判断 O 属于类1还是属于类2的识别问题。设该问题的识别函数为 $D(O)$，满足

$$\begin{cases}\text{若 } D(O)\geqslant 0,\text{则 } O\in W\\ \text{若 } D(O)<0,\text{则 } O\in\overline{W}\end{cases} \tag{3}$$

此时的识别函数就相当于置信度：

$$C(O|W)=D(O) \tag{4}$$

利用置信度可以对识别结果的可靠性进行假设检验，定位识别结果中的错误所在，提高系统的识别率和稳健性。

4.1 基于长度归一化的声学置信度

在令牌传递算法中，通过对概率累计得分最大的令牌进行回溯，可以得到令牌传送路径以及最终的识别结果。这个累计得分值是基于总长度的，如果基于帧数求取一个平均值，则更能体现各个帧的匹配程度。这个平均值就是长度归一化的声学置信度。如该分值超过某一门限，则确认其为关键词，否则拒识之。

假设语音观察序列 $O=o_1,o_2,\cdots,o_T$，t 时刻处在HMM的状态为 q_i，关键词候选为 K，其起始和结束点分别为 b 和 e，对于关键词候选的每一帧计算声学分，并求平均值，得到长度归一化的声学分：

$$S(K|O)=\frac{1}{e-b+1}\sum_{t=b}^{e}\log P(q_i|o_t) \tag{5}$$

利用 $S(K|O)$ 与预先设定的门限值比较，如果大于门限则判断是关键词，否则拒识。

4.2 基于二次解码声学似然比的确认方式

除了可以在初选结果上进行确认，也可以在第一次解码完成后，在确认阶段再次解码，以获得更充分的信息来进行有效的确认。

根据Newman-Pearson准则，假设检验的最优检验为似然比检验，因此置信度估值的中心思想也是进行似然比检验，通过判断基于似然比的声学置信度：

$$CM=LR(O,H_0,H_1)=\frac{P(O|H_0)}{P(O|H_1)}\geqslant Threshold \tag{6}$$

来评价识别结果的可信程度。这里的备选假设模型 H_1 有两种作用：一是减小各种变化因素对置信度的影响。当观测数据被一些系统变化因素影响时，分子和分母中同样包含了该变化，采取似然比检验将消除这些干扰的影响。二是能够更有针对性地表示那些容易与其他模型混淆而被错误识别的语音模型。

这里将给定关键词模型解码的声学似然度作为 $P(O|H_0)$ 的估值，将给定垃圾模型解码的声学似然度作为 $P(O|H_1)$ 的估计值。

设声学似然比为 R_k，$S(K|O)$ 为关键词候选匹配关键词模型的声学得分，$S_{\max}(G|O)$ 为候选匹配垃圾模型的声学置信分中的最大值，即最佳匹配的垃圾模型声学得分，则

$$R_K=\frac{S(K|O)}{S_{\max}(G|O)} \tag{7}$$

如果 R_k 高于某一个门限，则判定该候选是关键词，否则拒识之。

5 解码算法的改进策略

虽然可以通过限制模型规模来实现快速检出，但模型质量的损失也很严重。模型质量是识别性能的基本保障，因此本文考虑不牺牲模型质量而对解码算法进行改进。

5.1 剪枝

剪枝就是在搜索过程中“剪掉”可能性很小的路径。在令牌传递算法中，可以用这样一个简单的办法来实现：在处理第t帧的特征矢量时，找出当前最优部分路径的概率得分，即令牌中概率累计值最大值 $S_{\max}(t)$。

假设在处理第 t 帧特征矢量时，最优概率累计得分为

$$S_{\max}(t)=\max_{\forall s}\{S(t,s)\} \tag{8}$$

对于所有令牌计算它们的概率累计值与这个最高概率累计值的差值。如果这个差值超过某个门限，就将这个令牌抛弃，不再继续传递。这个门限就是剪枝门限。设目前的概率累计值 $S(t,s)$，剪枝门限 $B(s)$，则不满足公式(9)的令牌将被抛弃：

$$S(t,s)<S_{\max}(t)-B(s) \tag{9}$$

可以看出，这种剪枝算法并不是全局最优的，它仅是一种近似算法，有可能在剪枝的过程中就将最佳路径剪掉，产生识别错误。剪枝门限的设置决定着这种错误的概率，同时也影

维普资讯 http://www.cqvip.com

响着识别时间。

5.2 限制最大激活模型数

由于识别网络的规模很大，在解码过程中会存在大量的被激活模型，而且当碰到很多词的边界时往往会出现这种情况。因为在词的边界存在很大的不确定性，可能导致扩展的新模型很多落在裁剪宽度内，使得被激活模型的数目激增。这样，无论从内存耗费，还是时间消耗上看都是相当巨大的。本文试图采用限定最大激活模型数的方法，抛弃一些可能性小的激活模型，以减少内存和计算消耗。

限制最大激活模型数就是事先设定一个最大激活模型个数的门限；在令牌传递过程中，处理每一帧特征矢量之前，考察一下当前被激活模型的数目，这个数目如果超过门限值，则将概率累计值较低的一部分模型抛弃。

设当前模型数为 $C(t)$，最大激活模型数门限为 C_{max}，对某一时刻 t，若

$$C(t)>C_{max} \tag{10}$$

则裁剪概率累计值低的部分模型，使当前活动模型数

$$C'(t)=C_{max} \tag{11}$$

这样，通过限制最大激活模型数可以限制解码的范围，从而实现快速检出。

6 实验与讨论

本文实验中使用 863 语料库作为训练语料，它来自 166 个不同的说话人，男女各半，共有 95928 句，每句 3～6s，总时长超过 100h。测试语料选用的是微软公司发布的测试语料，共 500 句，每句 3～6s，总时长约 40min。实验中的关键词为经济、信息等 10 个常出现的词，它们在测试语料中共出现 99 次。上下文相关的音素模型库三音素集共 37734 个，上下文无关的音素模型库单音素集共 97 个。

对全部语音信号进行 16kHz 的采样，使用 Mel 频率倒谱系数(Mel Frequency Cepstrum Coefficient，MFCC) 作为特征。测试的性能评价标准包括检出率、误识率和识别时间比，它们的定义如下：

检出率＝正确检出的关键词个数/关键词总数 (8)

误识率＝错误检出的关键词数/检出的总关键词数 (9)

识别时间比＝识别所需时间/语料时间 (10)

表 1 给出了基于令牌算法的关键词检出的基本识别结果。可以看出，检出率相对较高，但识别速度很慢，对 3～6s 的一句话，识别时间长达 1～2min，显然无法满足实时的应用要求。

表 1 基于令牌各关键词检出的结果

检出率(%)	误识率(%)	识别时间比
96.0%	12.8%	2150%
87.9%	8.4%	2150%

表 2 采用不同限制模型规模方法的结果

垃圾模型	检出率(%)	误识率(%)	识别时间比
词内相关全音节	98.0%	67.2%	225%
全音素	94.9%	77.7%	25%
部分音节结合全音素	94.9%	68.5%	37.5%

为提高识别速度，采用限制模型规模的方法，分别使用词内相关全音节、全音素、部分音节结合全音素来作为垃圾模型，其结果如表 2 所示。可以看出，采用上述三种模型检出率均较高，误识率也较高。而识别速度显著改善，采用全音素垃圾模型的识别速度最快，全音节垃圾模型最慢。通过部分音节结合全音素垃圾模型的改进，识别时间比略有增加，但降低了误识率。

上述的实验结果尽管检出率较高，但误识率也较高。为降低误识率，进行了关键词确认实验。

采用长度归一化的声学置信度进行实验。当检出率为 63.6%时，误识率为 39.4%。由于确认阶段是基于规则的，因此时间可以不考虑。可以看出，这种确认方法的拒识效果有限，在保证较高检出率的基础上，误识率不能降到很低；而如果着重降低误识率，就会使检出率降低很多，不能达到一个令人满意的结果。

采用二次解码的声学似然比的确认方法，其结果如表 3 所示。

表 3 基于二次解码的声学似然比的确认结果

检出率(%)	误识率(%)	识别时间比
77.8%	34.2%	50%
83.8%	42.0%	50%

可以看出，采用二次解码的声学似然比的方法，在初选的结果上进行确认，能在保证检出率的同时降低误识率。

分别采用剪枝和限制最大激活模型数两种裁剪策略，以及两种策略结合确认机制的实验，结果如表 4 所示。

表 4 两种裁剪策略以及两者结合并加入确认机制的实验结果

裁剪方法	检出率(%)	误识率(%)	识别时间比
剪枝	88.9%	20%	125%
限制最大激活模型数	90.9%	21.7%	150%
两种策略结合确认机制	91.9%	13.3%	100%

从实验结果看，显然结合两种裁剪算法并加入确认机制的情况最好，无论检出率、误识率都比较令人满意，识别时间也能达到实时性要求。

结论 本文在基于令牌传递算法的关键词检出技术基础上，首先从限制模型的角度，提出了限制上下文相关的词内相关音素模型，提高了识别速度；针对误识率高的问题，提出了基于声学置信度的关键词确认方法，并实现了多次解码机制，取得了较好的结果；其次，从改进解码算法的角度，研究了剪枝和控制最大激活模型数两种策略对识别性能的影响，并结合确认机制进行关键词检出，获得了较好的结果。

参考文献

1 Rose R C. Keyword detection in conversational speech utterance using hidden Markov model based continuous speech recognition. Computer ,Speech and Language, 1995,9:309～333

2 Bourlard H, hoore B D, Boite J M. Optimizing recognition and rejection performance in wordspotting system. ICASSP94,1994, Ⅰ:373～376

3 刘俊,朱小燕. 基于动态垃圾评价的语音确认方法. 计算机学报, 2001,24(5):480～486

4 郝杰, 李星. 汉语连续语音识别中的关键词可信度的贝叶斯估计. 声学学报, 2002, 27:393～397

5 Soong F K, Lo W K, Nakamura S. Generalized word posterior probability (GWPP) for measuring reliability of recognized words. In: Proceeding of SWIM2004, 2004. 127～128

6 Chak Shun Lai, Shi B E. A one-pass strategy for keyword spotting and verification. ICASSP2001,2001,1:377～380

7 Hagen A, Pellom B. A Multi-layered Lexical-tree-based Token Passing Architecture for Efficient Recognition of Subword Speech Units. University of Colorado, IEEE, 2005. 1～5