# 关键词检测系统中基于音素网格的置信度计算

张鹏远　　韩　疆　　颜永红
(中国科学院声学研究所中科信利语音实验室　北京 100080)

**摘　要**：该文提出了一种基于音素网格的置信度计算方法。与传统的基于整个声学模型的置信度不同的是，这种方法在解码器生成的音素网格上计算关键词的置信度，从而具有更好的拒识能力。另外，针对两种置信度取值范围的不同，该文采用权重因子的方法综合利用两种置信度，取得了较好的效果。在自然对话的电话数据测试中，与传统的置信度计算方式相比，混和置信度的 FOM(Figure Of Merit)值相对提高了 17.0%。

**关键词**：语音识别；关键词检测；置信度；后验概率；网格
**中图分类号**：TP391.42　　**文献标识码**：A　　**文章编号**：1009-5896(2007)07-2063-04

# Phoneme Lattice Based Confidence Measures in Keyword Spotting

Zhang Peng-yuan　　Han Jiang　　Yan Yong-hong
(*Zhongke Xinli Speech Lab*, *Institute of Acoustics*, *Chinese Academy of Sciences*, *Beijing* 100080, *China*)

**Abstract**: Phoneme lattice based Confidence Measure (CM) is proposed in this paper. It makes use of phoneme lattices generated by a phoneme recognizer. Acoustic Model (AM) based CM is also introduced. For a decoded speech frame aligned to an HMM state, the CM based on AM is calculated. These two confidence measures are combined using a weighting factor to obtain a hybrid CM as they had different dynamic scales. On spontaneous conversational telephone database, the Figure Of Merit (FOM) achieves 17.0% relative improvement comparing to AM based CM.

**Key words**: Speech recognition; Keyword spotting; Confidence Measure (CM); Posterior probability; Lattice

## 1　引言

关键词检测已成为语音识别的一个重要分支。在实际应用中，期望识别引擎能够处理所有的发音是十分困难的[1−3]。同时，可靠的置信度计算在很多应用场合十分有效。例如，在噪音或背景音乐环境下，识别结果会出现一些错误，对于这些错误的识别结果，应该分配较低的置信度加以拒识，从而有效地降低虚警。

目前，语音识别中有很多置信度计算方法[4]。基于声学模型的置信度计算就是常用的一种。在关键词检测系统中，这种方法常被用来在整个声学空间上计算关键词的置信度。对于每一帧观测序列，首先计算出其后验概率，然后通过计算对数域的算术均值就可以得到整个关键词的置信度[5]。通过合适的域值，那些错误的识别结果就可以被拒识掉。但是这种方法的缺点就是局部的辨别能力不够。最近，基于词格的置信度计算方法成为研究的热点。在文献[6]中，置信度是在识别过程中产生的词格上计算的。这种方法在大规模连续语音识别中取得了很好的效果。

在此基础上，本文提出了一种关键词检测系统中的置信度计算方法。首先，我们在识别器产生的音素网格上计算每一个音素的置信度。关键词的置信度通过取所包含音素的对数域算术均值获得。在计算置信度时，这种方法不仅考虑了最优路径，其他路径的信息也同时考虑进来，因而具有更好的辨别能力。最后，结合传统的置信度，本文提出了一种混和置信度计算方法，取得了较好的效果。

本文的其它部分是这样安排的：第 2 节首先介绍一下传统的置信度计算方法。在第 3 节里，重点叙述基于音素网格的置信度。第 4 节给出实验结果。最后是结束语。

## 2　基于声学模型的置信度计算

在这一节里，先介绍一下基于声学模型的置信度计算。一般来说，这种基于后验概率估计的计算方法都是一个后处理的过程。置信度的计算分两个阶段：音素级的置信度和词级的置信度。下面对其进行分别介绍。

### 2.1 音素的置信度计算

在基于隐马尔可夫模型的语音识别系统中，每一个上下文相关的音素都是由隐马尔可夫模型表示的。在关键词检测系统中，每一个关键词由若干个音素组成。为了计算关键词的置信度，首先要计算每一个音素的置信度得分[7]。

音素的置信度是由帧级的后验概率得到的。计算音素 $\mathrm{ph}_i$ 的置信度的公式可以表示为

$$\begin{aligned}\mathrm{CM}(\mathrm{ph}_i) &= \frac{1}{e[i]-b[i]+1}\sum_{n=b[i]}^{e[i]}\log p(q^{(n)} \mid o^{(n)}) \\ &= \frac{1}{e[i]-b[i]+1}\sum_{n=b[i]}^{e[i]}\log\frac{P(o^{(n)} \mid q^{(n)})P(q^{(n)})}{P(o^{(n)})} \quad (1)\end{aligned}$$

这里 $b[i]$ 和 $e[i]$ 分别是 $\mathrm{ph}_i$ 的起始帧和结束帧，$o^{(n)}$ 为观测序

2006-02-22 收到，2007-07-31 改回
国家 973(2004CB318106)资助课题

列，$q^{(n)}$ 表示相应的状态序列。

### 2.2 关键词的置信度计算

在得到音素的置信度后，就可以计算关键词的置信度了。计算关键词的置信度的方法有很多种，本文采用对数域算术均值的方法。计算公式定义为

$$\mathrm{CM}_{\mathrm{pos}}(w)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{CM}(\mathrm{ph}_i) \tag{2}$$

这里 $m$ 是关键词 $w$ 所包含的音素个数。

## 3 基于音素网格的置信度计算

文献[6]采用基于词格的方法计算置信度。与其不同的是，本文采用基于音素网格的方法计算关键词的置信度，并对计算方法进行了改进。音素网格是在解码过程中产生的。网格的产生质量对于置信度的计算有着很大的影响。因此，本文采用的紧束的方法来控制生成的网格的大小。另外，在计算置信度时，本文没有考虑语言模型的概率得分。

### 3.1 音素网格中边的后验概率

每一个网格都是由很多边和节点组成的，如图 1 所示。图 1 是一个包含关键词“软件”的网格，出于描述的方便，本文对这个网格作了适当的简化，真实的网格要复杂的多。在音素网格中，每一条边都由音素名称和声学概率两部分信息标识。每一个节点的值代表了到达这个节点的时间帧信息。这样，每一条边的起止时间也就是其对应的两个节点的时间。要计算每一条边的后验概率，必须首先计算其声学似然得分：

$$p(l)=p_{al}(O_{t_s}^{t_e}\mid \mathrm{ph}_i)^{\frac{1}{\gamma}} \tag{3}$$

这里 $t_s$ 和 $t_e$ 分别代表边 $l$ 的起始帧和结束帧。$\gamma$ 为调节因子，其值通过实验得到，本文中取 14。

文献[6]介绍了计算后验概率的详细算法。每一条边的后验概率 $p(l\mid O)$ 采用前向后向算法获得，这种方法与训练 HMM 的算法相似。对于网格中的每一个节点 $n$，分别计算其前向概率 $\alpha(n)$ 和后向概率 $\beta(n)$。$\alpha(n)$ 是从网格的开始节点到达 $n$ 的所有路径的累计概率。而 $\beta(n)$ 是从 $n$ 到达网格结束节点的所有路径的累积概率。它们的计算可以通过递归算法实现：

$$\left.\begin{aligned}\alpha(n)&=\sum_{l,e(l)=n}p(l)\alpha(s(l))\\ \beta(n)&=\sum_{l,s(l)=n}p(l)\beta(e(l))\end{aligned}\right\} \tag{4}$$

这里 $s(l)$ 表示边 $l$ 的起始节点，$e(l)$ 表示边 $l$ 的结束节点。在此基础之上，边的后验概率可以定义为

$$p(l\mid O)=\frac{\alpha(s(l))p(l)\beta(e(l))}{p(O)} \tag{5}$$

其中 $p(O)$ 是整个网格上所有音素的联合概率的和：

$$p(O)=\sum_{\mathrm{PH}}p(\mathrm{PH},O) \tag{6}$$

显然，$p(O)$ 的值等于结束节点的前向概率 $\alpha(n_e)$。

### 3.2 基于时间信息的置信度

得到每一条边的后验概率后，就可以在网格中计算音素的后验概率。文献[6]提出了一种基于时间的后验概率计算方法。对于音素 $\mathrm{ph}_i$ 的每一时间帧 $t$，对网格中符合特定条件的边的后验概率求和：

$$\mathrm{CM}_{\mathrm{lat1}}(\mathrm{ph}_i\mid t,O)=\sum_{\{l\mid t_s\le t\le t_e,\mathrm{Title}(l)=\mathrm{ph}_i\}}p(l\mid O) \tag{7}$$

由式(7)可以看出，符合条件的边必须包含时间帧 $t$，且所代表的音素为 $\mathrm{ph}_i$。最后音素 $\mathrm{ph}_i$ 的置信度通过对所有帧求对数域上的算术均值获得：

$$\mathrm{CM}_{\mathrm{lat1}}(\mathrm{ph}_i\mid t_s,t_e,O)=\frac{1}{t_e-t_s+1}\sum_{t=t_s}^{t=t_e}\log(\mathrm{CM}(\mathrm{ph}_i\mid t,O)) \tag{8}$$

### 3.3 基于重叠比率的置信度

在两条边所代表音素相同的情况下，如果它们持续时间重叠的很短，文献[6]的方法也会包含进去。这样显然是不合理的。针对这种情况，本文提出了一种基于边界的置信度计算方法。这种方法充分考虑每一条边的边界信息，只有重叠达到一定程度后才会计算。这里采用一个重叠因子 $\delta$，用来调节两条边的重叠程度。置信度的计算公式：

$$\mathrm{CM}_{\mathrm{lat2}}(\mathrm{ph}_i\mid t_s,t_e,O)=\log\left(\sum_{\{l\mid\mathrm{OverlapRatio}\ge\delta,\mathrm{Title}(l)=\mathrm{ph}_i\}}p(l\mid O)\right) \tag{9}$$

$\delta$ 的值通过实验获得。

得到音素的置信度得分后，词的置信度计算可通过下式求得：

$$\mathrm{CM}_{\mathrm{lat2}}(w)=\frac{1}{n}\sum_{i=1}^{n}\mathrm{CM}_{\mathrm{lat2}}(\mathrm{ph}_i\mid t_s,t_e,O) \tag{10}$$

这里 $n$ 是关键词 $w$ 中音素的个数。

### 3.4 置信度的融合

如上所述，在新的系统框架下，本文分别介绍了两种置信度的计算方式：基于整个声学模型的置信度和基于网格的置信度。这两种置信度计算方式都有各自的优点。相对于基于网格的置信度的计算，在整个声学模型上计算置信度具有更好的全局性。而基于网格的置信度计算则具有更好的局部

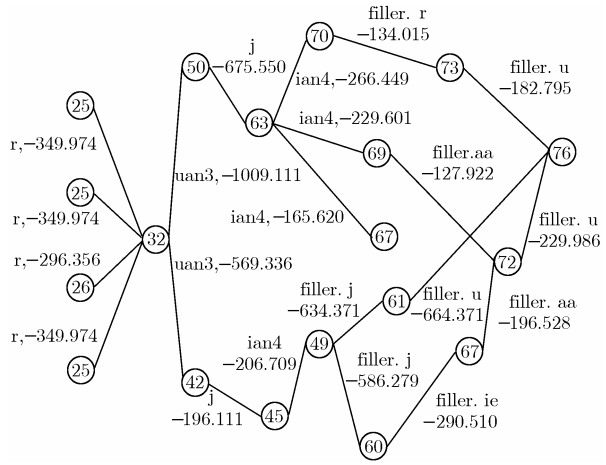


图 1 关键词“软件”的音素网格

最优性。这是由于每一个网格都是对应与一个特定的语音片段，在计算置信度时，我们不仅考虑到最优路径，其他路径的信息也同时考虑进来。这样，可以期望两种置信度计算方式有一定的互补性。本文尝试综合利用两种置信度计算的优点，构造一个混合的置信度计算方式，提高关键词检测系统的性能。

两种置信度得分有不同的变化范围，因此直接将他们求和是行不通的。本文采用了一个加权因子，用来平衡两种置信度得分。混合的置信度计算可以定义为

$$\mathrm{CM}_{\mathrm{hyb}}(w) = (1-\lambda)\mathrm{CM}_{\mathrm{pos}}(w) + \lambda \mathrm{CM}_{\mathrm{lat2}}(w) \qquad (11)$$

实验表明这种置信度合并方式取得了良好的效果。

## 4　实验

为了评估本文提出的算法，我们在中国科学院声学研究所中科信利语音实验室的关键词检测系统中进行了实验。该系统为基于电话语音的 1.5 倍实时系统。接下来我们将介绍实验以及相关结果。

### 4.1 实验数据描述

实验数据采用的是国家高技术研究开发项目(HTRDP)提供的电话信道的自然对话语音。这些语音是在真实噪音环境下由电话信道实际采录的，所有语音均为普通话，部分带有口音，采样率为 8kHz, 16bit。本文采用的测试集的语音长度为 1 个小时，包含 14 个说话人，关键词个数为 100。在 100 个关键词中，二字词居多，约占 80%，其余的为三字词。

在我们的关键词检测系统中，声学模型是由 150 个小时的电话朗读数据训练出来的。由于缺乏足够的自然对话语音，声学模型与测试数据之间存在着明显的不匹配。因此，在测试集上关键词的识别率只有 54.77%。然而，本文的主要目的是验证所提出的置信度的有效性。识别率的提高可以通过更好的声学模型的获得。

### 4.2 实验结果

为了验证置信度的性能，将识别的结果与参考答案进行比较，将识别出来的关键词分成正确的和错误的两类。然后，找出一个合适的域值，判断该关键词是被接受还是被拒绝。本文采用 ROC(Receiver Operating Characteristics)曲线来评估关键词检测系统的性能，该曲线的横轴为每个关键词每小时的虚警个数(FA/KW/HR)，纵轴为关键词的识别率。同时，为了方便比较，我们计算出 FOM(Figure of Merit)的值。根据 NIST 的定义，FOM 为 0 到 10 FA/KW/HR 下识别率的均值。

表1 描述了基于边界信息的置信度计算的性能($\mathrm{CM}_{\mathrm{lat2}}$)。显然，当$\delta$取 0.5 时，可以得到最佳的 FOM 值。表 2 给出了与基于时间的置信度($\mathrm{CM}_{\mathrm{lat1}}$)的比较结果。可以看出，基于边界的置信度的计算要优于基于时间的方法。

**表 1　$\mathrm{CM}_{\mathrm{lat2}}$ 在不同 $\delta$ 值下的性能对比**

| $\delta$ | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|
| FOM(%) | 23.17 | 23.50 | 22.86 | 22.25 |

**表 2　$\mathrm{CM}_{\mathrm{lat1}}$ 和 $\mathrm{CM}_{\mathrm{lat2}}$ 比较**

| 置信度计算方法 | FOM(%) |
|---|---|
| $\mathrm{CM}_{\mathrm{lat1}}$ | 23.10 |
| $\mathrm{CM}_{\mathrm{lat2}}$ ($\delta$=0.5) | 23.50 |

在训练声学模型的时候，缺乏足够的电话信道的自然对话数据，因此存在着声学模型和测试集之间的不匹配。这也是基于声学模型的置信度性能不好的主要原因。基于音素网格的置信度在某种程度上克服了这方面的不足。由于这种置信度是在网格上计算出来的，它主要强调了最优路径相对于网格中其他路径的一种得分，从而忽略了信道和数据不匹配的影响。另一方面，基于音素网格的置信度也有其自身的不足，其在网格内部具有较好的辨别性，但在整个声学空间上的辨别能力不足。因此，我们考虑把这两种置信度进行加权相加，相信能够得到更好的效果。实验结果也证明了这一点。表 3 给出了混合置信度 $\mathrm{CM}_{\mathrm{hyb}}$ 的性能。当 $\lambda$=0.8 时，混合置信度方式取得了最佳的 FOM 值。与 $\mathrm{CM}_{\mathrm{pos}}$ 比较，$\mathrm{CM}_{\mathrm{hyb}}$ 的 FOM 值相对提高了 17.0%。

**表 3　3 种置信度方法的 FOM 值**

| 置信度计算方法 | FOM(%) |
|---|---|
| $\mathrm{CM}_{\mathrm{pos}}$ | 22.43 |
| $\mathrm{CM}_{\mathrm{lat2}}$ ($\delta$=0.5) | 23.50 |
| $\mathrm{CM}_{\mathrm{hyb}}$ ($\lambda$=0.8) | 26.25 |

图 2 描述了 3 种置信度计算方式的 ROC 曲线。显然，在同样的虚警下，$\mathrm{CM}_{\mathrm{lat2}}$ 优于 $\mathrm{CM}_{\mathrm{pos}}$。而 $\mathrm{CM}_{\mathrm{hyb}}$ 是 3 种方式中最好的。

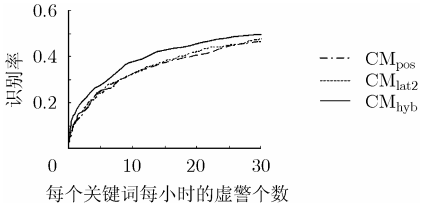


图 2　电话测试集上的 ROC 曲线

## 5　结束语

本文提出了一种改进的音素网格的置信度计算方法，FOM 相对提高了 1.7%。另外，结合基于声学模型的置信度，本文提出了一种混合的置信度计算方法，在实验中取得了较好的效果。与声学模型置信度相比，FOM 相对提高 17.0%。在其他的电话测试集上，我们进行了同样的实验，取得了类似的实验结果。因此可以说，本文提出的算法是有效的，基于网格的置信度与基于声学模型的置信度具有一定的互补性。下一步的工作重点是研究更加有效的音素网格的质量控制策略。

## 参 考 文 献

[1] Williams G and Renals S. Confidence measures for hybrid HMM/ANN speech recognition. Proceedings of Eurospeech-97, Rhodes, Greece, 1997: 1955–1958.

[2] Sankar A and Wu Su-Lin. Utterance verification based on statistics of phone-level confidence scores. Proceedings of IEEE ICASSP-2003, Hong Kong, 2003: 584–587.

[3] Guo Gang, Huang Chao, Jiang Hui, and Wang Renhua. A comparative study on various confidence measures in large vocabulary speech recognition, ISCSLP 2004, Hong Kong, 2004: 9–12.

[4] Rivlin Z, Cohen M, Abrash V, and Chung T. A phone dependent confidence measure for utterance rejection. Proceedings IEEE International Conference on Acoustics Speech and Signal Processing, Atlanta, USA, 1996: 515–517.

[5] Kamppari S O and Hazen T J. Word and phone level acoustic confidence scoreing, Proceedings of IEEE ICASSP-2000, Istanbul, Turkey, 2000: 1799–1802.

[6] Evermann G. Minimum word error rate decoding. [MPhil thesis], Cambridge University, 1999.

[7] Abdou S and Scordilis M S. Beam search pruning in speech recognition using a posterior probability-based confidence measure. *Speech Communication*, 2004: 409–428.

张鹏远： 男，1978 年生，博士生，研究方向为关键词检测、置信度计算等.

韩 疆： 男，1969 年生，副研究员，主要研究方向为大词表非特定人连续语音识别、关键词检测、置信度计算等.

颜永红： 男，1967年生，研究员，博士生导师，主要研究方向为大词表非特定人连续语音识别、多模口语系统、嵌入式系统、系统自适应和快速搜索算法.