关键词识别中置信度评估方法的研究

任竹，贾珈，蔡莲红

普适计算教育部重点实验室
清华信息科学与技术国家实验室（筹）
清华大学计算机科学与技术系，北京 100084

摘　要：关键词识别是语音识别的一个重要研究领域，它不仅比连续语音识别的灵活性更好，同时具有很高的应用价值。本文提出了基于多级词表的关键词识别系统，并从声学匹配和语义理解两个层面重点研究了在对语音识别结果进行确认时所采用的置信度评估方法。该系统将传统词表按照单词长度划分为关键词词表和关键短语词表，并在采用模糊匹配的方法检出关键词后通过集合映射的方式进行关键短语的匹配，同时分别提出了对关键词和关键短语进行置信度评估的确认方法。实验结果表明，该系统有效地提高了关键词和关键短语的平均检出率，满足了应用场景的实际需求。

关键词：语音识别；关键词识别；置信度；关键词检出；语音确认

1．引言

关键词识别是语音识别的一个重要研究领域，其目的是在连续无限制的自然语音流中检测并确认出若干的特定关键词。关键词识别主要包括两个方面的基本内容：关键词检出(Keyword Spotting)和关键词确认(Utterance Verification)。关键词检出从无限制语音流中检测出尽可能多的候选关键词，再由关键词确认部分对这些候选关键词进行置信度评估。

在当前有关置信度评估方法的研究中，研究者主要从声学、词图和语义这三个层面来提取置信度的特征。在声学层面，通常使用似然比[1,2]、词候选驻留时间等特征；在词图层面，主要是利用词后验概率[3]、候选词图中同词候选并列的其他候选的个数[4]等特征；在语义层面，则是根据语言理解的结果对整句候选进行确认[5]。

本文提出了基于多级词表的关键词识别系统，并融合了声学匹配和语义理解两个层面的相关信息分别对关键词和关键短语进行置信度的评估。该系统将传统的识别过程分为关键词的提取和关键短语的匹配这两个阶段，同时采用不同的置信度评估方法来实现关键词和关键短语的拒识。本文选用当前流行面较广的智能家居作为基于多级词表的关键词识别系统的应用场景进行测试，该场景的主要目标是利用识别出的关键词来实现智能家居的命令控制。实验结果表明，该系统不仅有效地提高了关键词和关键短语的平均检出率，而且尽可能地降低了二者的平均误识率，能够充分地满足应用场景的实际需求。

资助项目：国家自然科学基金重大计划项目（90920302）
联系作者：任竹，E-mail：bamboo.renzhu@gmail.com

2. 系统框架

为了解决较短单词容易发生错误识别(False Alarm)而较长单词容易发生错误拒绝(False Rejection)的问题，本文按照单词长度对传统词表进行了分级，在该过程中采用的关键词和关键短语的概念为：

关键词(Keyword)：词条规模为识别结果中单个候选元素大小且具有单一词性的单词。关键词作为系统的识别基元，它不仅包含实际应用中经常出现的动词、名词，还包含贯穿于自然对话内的代词、连词、助词、介词等功能词。例如，在智能家居场景中，“打开”、“窗帘”均为关键词，前者描述了命令所实现的动作，后者指明了命令所针对的设备。

关键短语(Key Phrase)：由两个或两个以上关键词组成的单词。每个关键短语都对应着一个构成它的关键词集合，因而关键短语一般都具有比关键词更为丰富的语义信息，能够直接运用于各类应用场景，故其检出率和误识率的高低更为研究者和使用者所重视。例如，在智能家居场景中，控制命令“打开卧室窗帘”为关键短语，它由“打开”、“卧室”、“窗帘”这三个关键词组成。

本文所构建的基于多级词表的关键词识别系统的整体框架如图 1 所示，其中如何实现关键词的拒识和关键短语的置信度评估是本文的研究重点。该系统的前端为利用 Microsoft Speech SDK 软件开发包构建的连续语音识别器，其输入方式为麦克风即时输入或音频文件输入，输出结果为将 N-Best 音节序列进行时间对准后的拼音串，其中每一个识别候选语句是由多个候选元素组成的，而每个候选元素又包含了识别的相关信息。该系统采用模糊匹配的方法检出关键词，在初步拒识掉部分候选关键词后通过集合映射的方式实现关键短语的匹配，在对候选关键短语进行置信度确认后得到最终的识别结果，即控制智能家居场景的相关命令。

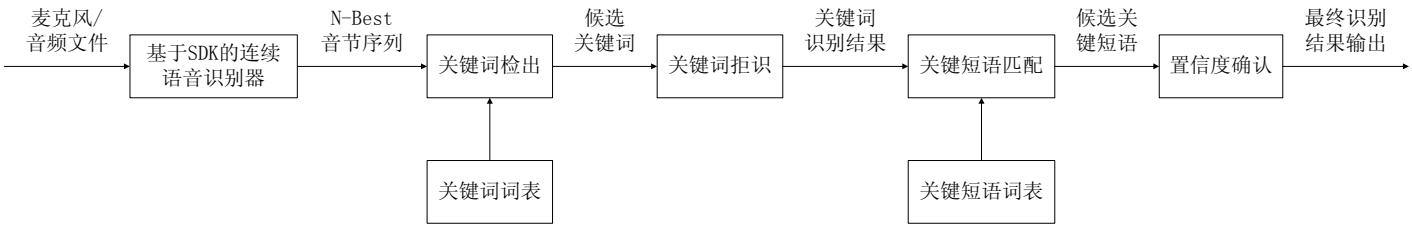

图 1　基于多级词表的关键词识别系统的框架图

3. 置信度评估方法

3.1　声学得分

本文对由前端语音识别平台获得的候选元素的声学得分进行规格化的计算公式为：

$$NA([e;t_b,t_e]) = \frac{A([e;t_b,t_e])}{\frac{1}{t_e-t_b+1}\sum_{t=t_b}^{t_e}\sum_{\substack{[e';t'_b,t'_e]:\\ t'_b\le t\le t'_e}} A([e';t'_b,t'_e])} \tag{1}$$

其中：$[e;t_b,t_e]$代表起始时间为t_b、终止时间为t_e的候选元素e；$A([e;t_b,t_e])$为由前端语音识别平台获取的候选元素e的声学得分；$NA([e;t_b,t_e])$为候选元素e进行规格化后的声学得分。

为了区分关键词的不同候选音节串与其正确读音的匹配准确度，在此给出**子音节单元**(Sub-syllabic Unit)的概念：在汉语中一个汉字就是一个音节，而每个音节是由声母、韵母和声调这三个部分组成的，其中每一个部分都可以看作是一个子音节单元。

在本文中，采用将处于相同时间区间内、能识别出某一关键词的多个候选元素的声学得分进行累加的方式来计算该**关键词的声学得分**，其计算公式为：

$$ACM(k)=\sum_{i=1}^{N}NA(e_i)\cdot\frac{SU(k\mid e_i)}{SU(k)} \tag{2}$$

其中：$NA(e_i)$为第i个候选语句中处于该时间区间内的候选元素e_i的规格化后的声学得分，此处略去了候选元素的起止时间信息；N为候选语句的个数；$SU(k\mid e_i)$为将候选元素e_i识别为关键词k时，二者相匹配的子音节单元数；$SU(k)$为关键词k所包含的子音节单元总数。

类似地，可以得到**关键短语的声学得分**的计算公式为：

$$ACM(p)=\frac{\sum_{j=1}^{M}ACM(k_j)\cdot SN(k_j)}{SN(p)}=\frac{\sum_{j=1}^{M}\sum_{i=1}^{N}NA(e_i)\cdot SU(k_j\mid e_i)}{3SN(p)} \tag{3}$$

其中：组成关键短语p的关键词集合为$\{k_1,k_2,\cdots,k_M\}$；M为该集合所包含的关键词的个数；$SN(k)$为关键词k所包含的音节总数，易知$SU(k)=3SN(k)$。

3.2 音节匹配得分

本文采用所识别出的组成关键短语的关键词包含的音节总数与关键短语所含的音节总数的比值来评价映射效果的优劣，于是给出**关键短语的音节匹配得分**的计算公式为：

$$SCM(p)=\frac{\sum_{j=1}^{M}SN(k_j)\cdot\delta(k_j)}{SN(p)} \tag{4}$$

其中：$\delta(k)$为权值函数，当关键词k被识别出来时$\delta(k)=1$，否则$\delta(k)=0$。

将关键短语的声学得分和音节匹配得分结合起来，得到**关键短语的置信度**计算公式为：

$$CM(p)=\alpha\cdot ACM(p)+(1-\alpha)\cdot SCM(p) \tag{5}$$

其中：$ACM(p)$为关键短语p的声学得分；$SCM(p)$为关键短语p的音节匹配得分；α为调节二者在置信度的计算中所占比例的加权因子，其取值范围为$0\le\alpha\le1$。

3.3 顺序匹配得分

在本文中，考虑到检出的候选关键词之间的时间先后关系在其组成关键短语的过程中所产生的影响，于是给出**关键短语的顺序匹配得分**的计算公式为：

$$SQ(p)=\frac{\sum_{i=1}^{M-1}\sum_{j=i+1}^{M}\varphi(k_i,k_j)}{\frac{M(M-1)}{2}} \tag{6}$$

其中：$\varphi(k_i,k_j)$ 为累计函数，当关键词 k_i 和 k_j 均被识别出来且二者在关键短语 p 中以顺序出现时 $\varphi(k_i,k_j)=1$，否则 $\varphi(k_i,k_j)=0$。

3.4 语义相似度得分

本文采用两个关键词同时出现在关键短语中次数的多少来评价二者之间语义相似度的大小，于是给出**关键词之间的语义相似度**的计算公式：

$$SS(k_j\mid k_i)=\frac{count(k_j,k_i)}{\sum_{n=1}^{N}count(k_n,k_i)}=\frac{\sum_{m=1}^{M}count_m(k_j,k_i)}{\sum_{n=1}^{N}\sum_{m=1}^{M}count_m(k_n,k_i)}\quad(1\le i,j\le N) \tag{7}$$

其中：N 为词表中所包含的关键词总数，即关键词词表为 $\{k_1,k_2,\cdots,k_N\}$；M 为词表中所包含的关键短语总数，即关键短语词表为 $\{p_1,p_2,\cdots,p_M\}$；$count(k_j,k_i)$ 为关键词 k_j 和 k_i 同时出现在组成某一关键短语的关键词集合中的总次数；$count_m(k_j,k_i)$ 为组成关键短语 p_m 的关键词集合里同时包含关键词 k_j 和 k_i 的总次数；$SS(k_j\mid k_i)$ 为关键词 k_j 相对于关键词 k_i 的语义相似度。

考虑到可将某关键词同其时间上前后相邻的那些关键词之间的语义相似度累加起来得到该**关键词的语音相似度得分**，其计算公式为：

$$SS([k;t_b,t_e])=\sum_{\substack{[k';t'_b,t'_e]:\\ t'_e=t_b-1\cup t'_b=t_e+1}}SS([k';t'_b,t'_e]\mid[k;t_b,t_e]) \tag{8}$$

其中：$[k;t_b,t_e]$ 代表起始时间为 t_b、终止时间为 t_e 的候选关键词 k；$SS([k';t'_b,t'_e]\mid[k;t_b,t_e])$ 代表起始时间为 t'_b、终止时间为 t'_e 的候选关键词 k' 相对于关键词 k 的语义相似度。

4．系统实现

实现基于多级词表的关键词识别系统的算法流程如下：

（1）对前端语音识别平台获取的所有候选元素的声学得分进行规格化计算；

（2）从 N 个候选语句中依次取出当前待识别的候选语句后处理其所包含的候选元素，将这些候选元素与关键词词表中的每个词条依次匹配并标记候选语句中匹配成功的音节位置；

（3）将当前候选语句中匹配失败的连续音节片段采用逆向最大匹配分词算法重新切分后，再次进行关键词的匹配过程；
（4）对于当前候选语句中某些匹配失败的单音节，将其同相邻的匹配成功的单音节重组成新的候选元素后再次进行关键词的匹配过程，并刷新候选关键词集合；
（5）若此时还存在待识别的候选语句，则转（2）；
（6）计算各候选关键词的声学得分，合并那些起止时间相同且为词表中同一词条的候选关键词并将它们的声学得分进行累加；
（7）根据识别结果集合中各候选关键词的时间信息来计算其语义相似度得分，把时间上相交的那些候选关键词中声学得分和语义相似度得分均为最低的单词进行迭代删除；
（8）将候选关键词集合依次同关键短语词表中的每个词条进行匹配，若组成某关键短语的关键词集合中有部分或全部关键词均已被识别出来就将该关键短语检出；
（9）计算各候选关键短语的声学得分和音节匹配得分后获得其置信度；
（10）若存在两个不同关键短语所含关键词集合完全相同的情况，则计算两者的顺序匹配得分后把分值较高的那个关键短语的排位提前；
（11）将识别结果集合中置信度低于局部阈值的那些候选关键短语删除；
（12）算法结束，返回关键词和关键短语的识别结果集合。

5．实验结果及分析

实验所用到的词表规模为：关键词词表中包含 78 个关键词，其长度范围为 1 至 3 个音节；关键短语词表中包含 99 个关键短语，其长度范围为 2 至 11 个音节，且由 2 至 7 个关键词组成。

实验所用到的测试语料有两组，且均由 5 个不同的说话人录制而成；其中，语料 1 的总说话时间约为 40 分钟，由 700 句语音组成，包含 2580 个关键词和 680 个关键短语，且每句中只存在 1 个所含关键词集合内各关键词按顺序出现的关键短语；语料 2 的总说话时间约为 30 分钟，由 300 句语音组成，包含 2680 个关键词和 750 个关键短语，且每句中至少存在 2 个关键短语、也出现了组成关键短语的关键词集合内各关键词在句中打乱顺序或缺少部分成分的情况。

本文的基线系统采用的是传统的基于字符串匹配的关键词识别算法，并将 Mitchel Weintraub 在文章[6]中定义的声学似然度得分作为关键词的置信度评价标准。把本文提出的基于多级词表的关键词识别方法在优化前后的性能与基线系统进行对比，得到的实验结果如图 2、图 3 所示，此处的优化指的是算法流程中的（2）、（3）、（4）、（7）这四个步骤。

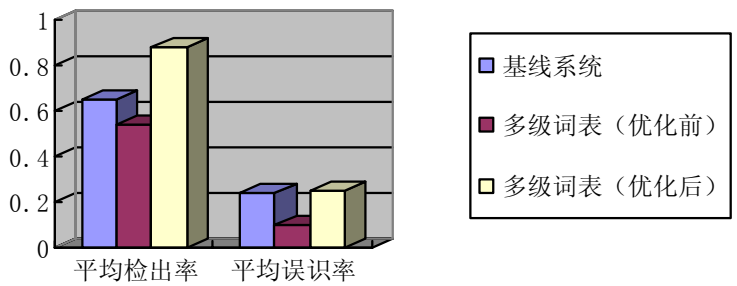

图 2　基于多级词表的方法优化前后与基线系统识别关键词的性能比较

如图 2 所示，基于多级词表的方法在优化前识别关键词的平均检出率要略低于基线系统。这主要是由于基线系统是针对整个候选语句来进行字符串匹配的，而基于多级词表的方法的处理对象则是候选元素。因此，前端语音识别平台在对候选语句进行粗略分词得到候选元素时经常发生的分词错误便会使得关键词的平均检出率下降。可喜的是基于多级词表的方法在优化后关键词的平均检出率有了显著的提升，在两组语料上检测出的值均处于90%左右，然而这是以其平均误识率的增加为代价的。

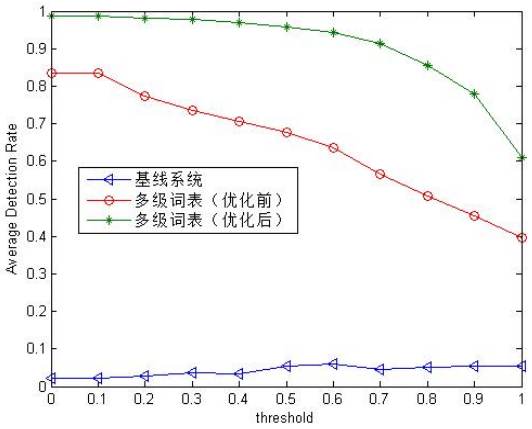

图 3　基于多级词表的方法优化前后与基线系统识别关键短语的性能比较

如图 3 所示，基于多级词表的方法无论优化与否其识别关键短语的平均检出率都远远高于基线系统，几乎为基线系统平均检出率的六倍以上，其差距显而易见。这主要是由于采用了关键词集合到关键短语的映射方式，使得在组成关键短语的部分关键词未被检出时该关键短语仍可能被识别出来。在实际的应用系统中，语音质量不高导致的关键词识别的部分缺失是经常会遇到的情况。因此，本文中提出的基于多级词表的方法能够很好地解决该问题，从而实现系统所需关键短语的正确识别。对比算法进行优化前后系统的性能差异，我们不难发现本文提出的这几种优化方法能够在大幅度提高平均检出率的同时又在一定程度上限制了平均误识率的增长；尤其是在全局阈值 threshold<0.7 的范围内，几乎在每组测试语料上的平均检出率都高于 90%；甚至在没有阈值限定的情况下，平均检出率接近 99%。这已经能够基本满足系统的应用需求了。

6．结束语

本文提出了基于多级词表的关键词识别系统，并重点研究了融合声学匹配和语义理解两个层面的相关信息来分别对关键词和关键短语进行置信度评估的方法。该系统通过先进行关键词提取再实现关键短语匹配的方式，解决了传统的基线系统无法识别出所含关键词乱序出现在语句中的关键短语的问题，并实现了对同一语句中的多个关键短语的拾取。

本文所采用的对关键词和关键短语进行置信度评估的方法虽然使二者的平均检出率有了大幅度的提高，但其平均误识率也随之升高了。在未来的工作中，我们需要考虑增加

其他方面的信息来完善置信度的评估方法，以及实现说话人的自适应技术来使系统具有更高的实用性和鲁棒性。

参考文献

[1] R. Sukkar and C. H. Lee. Vocabulary Independent Discriminative Utterance Verification for Non-keyword Rejection in Sub-word Based Speech Recognition, Proceeding of ICASSP 1998, 4:420-429

[2] Eduardo Lleida, R. C. Rose. Utterance verification in continuous speech recognition: decoding and training procedures, Proceeding of ICASSP2000, Istanbul, Turkey, 2000, 8:126-139

[3] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speed recognition, Proceeding of ICASSP 2001, Salt Lake City, USA, 2001, 9(3):288-298

[4] S. and T. J. Hazen. Word and Phone Level Acoustic Confidence Scoring. Proceeding of ICASSP2000, Istanbul, Turkey, 3:5-9

[5] S. Cox and S. Dasmahapatra. High-level Approaches to Confidence Estimation in Speech Recognition, IEEE Trans. Acoustic, Speech, and Signal Processing, 2002, 10(7):460-471

[6] Weintraub M. LVCSR log-likelihood ratio scoring for keyword spotting, Proceeding of ICASSP 1995, 1:297-300

Research on confidence measures in keyword recognition system

Zhu Ren, Jia Jia, Lianhong Cai

State Key Laboratory on Intelligent Technology and Systems

Tsinghua National Laboratory for Information Science and Technology（TNList）

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

Zhu Ren: E-mail: bamboo.renzhu@gmail.com

Key words: speech recognition, keyword recognition, confidence measure, keyword spotting utterance verification

Abstract: Keyword recognition is an important area of speech recognition, it has more flexibility and higher application value than continuous speech recognition. This paper presents a keyword recognition system based on multi-level vocabulary, and focuses its research on the confidence measures which are used to verify the recognition results from the aspect of acoustic matching and semantic understanding. The system divides the traditional vocabulary into keyword vocabulary and key phrase vocabulary, basing on the length of the words in the vocabulary. And it adopts fuzzy matching method to spot the keyword and then maps the collection of recognized keywords into candidate key phrases. At the same time, it uses different confidence measures of keyword and key phrase for utterance verification. Experiment results show that the proposed approach can effectively improve the average detection rate of keyword and key phrase to meet the actual demand of the applications.