计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800269

基于端到端语音识别的关键词检索技术研究

杨润延[1,2]　程高峰[1]　刘　建[1]
1 中国科学院声学研究所　北京 100190
2 中国科学院大学　北京 100049
(yangrunyan@hccl.ioa.ac.cn)

摘　要　近十年来,端到端的语音识别框架发展迅速。区别于传统的基于隐马尔可夫模型的语音识别框架,端到端语音识别拥有众多新特性,而且可以达到相同或更优秀的性能。因此,端到端语音识别吸引了越来越多的关注,已经成为了与传统语音识别并列的第二类主流框架。针对端到端语音识别无法提供关键词检索所需的关键词准确时间起止点与可靠置信度的问题,提出了一种基于端到端语音识别和帧级别对齐的关键词检索框架,并在越南语数据集上进行了实验验证。首先,使用端到端语音识别模型解码待测语句,得到 *N*-最佳假设;然后,从一个与上述识别模型联合训练的音素分类器中获得逐帧音素概率,使用一个基于动态规划的对齐算法为检出的 *N*-最佳假设和逐帧音素概率进行对齐,进而得到 *N*-最佳假设中各个单词的时间起止点和置信度;最后,在 *N*-最佳假设中匹配关键词,并利用时间起止点和置信度合并重复匹配的关键词,得到最终检索结果。在一个越南语自由交谈数据集上的实验表明,提出的关键词检索系统的 *F*1 值可以达到 77.6%,相对于传统的基于隐马尔可夫模型的关键词检索系统的 *F*1 值提升了 7.8%,而且可以提供可靠的关键词置信度。

关键词:关键词检索;语音识别;端到端;帧级别对齐
中图法分类号　TP391

Study on Keyword Search Framework Based on End-to-End Automatic Speech Recognition

YANG Run-yan[1,2],CHENG Gao-feng[1] and LIU Jian[1]
1 Institute of Acoustics,Chinese Academy of Sciences,Beijing 100190,China
2 University of Chinese Academy of Sciences,Beijing 100049,China

Abstract　In the past decade,end-to-end automatic speech recognition (ASR) frameworks have developed rapidly. End-to-end ASR has shown not only very different characteristics from traditional ASR based on hidden Markov models (HMMs),but also advanced performances. Thus,end-to-end ASR is being more and more popular and has become another major type of ASR frameworks. A keyword search (KWS) framework based on end-to-end ASR and frame-synchronous alignment is proposed for solving the problem that end-to-end ASR cannot provide accurate keyword timestamps and confidence scores,and experimental verification on a Vietnamese dataset is made. First,utterances are decoded by an end-to-end Uyghur ASR system,obtaining *N*-best hypotheses. Next,a dynamic programming-based alignment algorithm is implemented on each of these ASR hypotheses and per-frame phoneme probabilities,which are provided by a phoneme classifier jointly trained with the ASR model,to compute time stamps and confidence scores for each word in *N*-best hypotheses. Then,final KWS result is obtained by detecting keywords within N-best hypotheses and removing duplicated keyword occurrences according to time stamps and confident scores. Experimental results on a Vietnamese conversational telephone speech dataset show that the proposed KWS system achieves an F1 score of 77.6%,which is relatively 7.8% higher than the F1 score of the traditional HMM-based KWS system. The proposed system also provides reliable keyword confidence scores.

Keywords　Keyword search,Speech recognition,End-to-end,Frame-synchronous alignment

1　引言

关键词检索(Keyword Search,KWS)任务指在一批语音数据中定位到特定若干个关键词的全部出现位置,广泛应用于音频文档索引、音频片段截取等下游任务。关键词检索框架可分为基于语音识别的和语音识别无关的两大类,在基于语音识别的关键词检索方面,文献[1-2]使用语音识别作为前端,在其识别结果的基础上进行关键词的检索,具有泛化性

到稿日期:2021-08-31　返修日期:2021-10-15
基金项目:国家重点研发计划(2020AAA0108002)
This work was supported by the National Key Research and Development Program(2020AAA0108002).
通信作者:刘建(liujian@hccl.ioa.ac.cn)

强、关键词列表可自由变换的优势，且其性能和特性与其语音识别前端联系密切。而在语音识别无关的关键词检索方面，文献[3-4]依据发音特征等直接判断待测语音中是否包含关键词，其特点是架构简单、资源占用少。本文主要针对基于语音识别的关键词检索展开研究。

传统的基于深度神经网络(Deep Neural Network，DNN)和隐马尔可夫模型(Hidden Markov Model，HMM)的语音识别框架在关键词检索中已经得到了成熟的应用[5]。该框架主要分为3个模型：声学模型、发音模型和语言模型。其中，声学模型建模语音特征向量与音素概率之间的联系；发音模型以发音字典的形式存在，建模音素与单词之间的对应关系；语言模型对句子中单词与单词之间相互关联的概率进行建模。由于声学模型采用帧同步输出的模式，DNN-HMM语音识别能够为检出的关键词提供准确的时间起止点。同时它针对音素的发音进行直接建模，在发音层面给出的置信度分数较为可靠，因此，基于DNN-HMM语音识别的关键词检索系统能够做到通过调整关键词置信度阈值来进行精度和召回率之间的平衡调节。

端到端的语音识别框架近十年来发展迅速。其由于训练简单、无需发音字典等语言学知识，已经与基于HMM的语音识别并列成为业界主流的框架。端到端语音识别框架主要分为两类：基于注意力机制的编码器-解码器架构[6-7]以及基于联结时间分类(Connectionist Temporal Classification，CTC)的架构[8-9]。在上述两者的基础上，有研究提出了联合CTC/注意力架构[10]的混合架构。该混合架构在训练期间利用附加在编码器网络上的CTC目标函数作为正则化手段，辅助基于注意力的编码器-解码器网络的训练。在解码过程中，使用联合CTC/注意力解码方法，同时使用解码器分数和CTC前缀分数进行波束搜索，从而利用CTC的对齐性质排除基于注意力机制的解码器给出的对齐不佳的假设。通过同时利用CTC和注意力机制的优势，联合CTC/注意力架构能达到更优的性能，并且已经在语音识别任务上得到了广泛的应用。

然而，现有的端到端语音识别框架由于模型训练方法的限制，自身无法为其识别结果提供准确的时间起止点和可靠的置信度，因此难以被直接应用于关键词检索任务。为了解决上述困难，本文提出了一套基于端到端语音识别的关键词检索框架。基于这一框架的关键词检索过程可分为3步：第1步，使用联合CTC/注意力端到端语音识别模型解码待测语句，得到N-最佳假设；第2步，利用一个与上述识别模型联合训练的音素分类器得到待测语句的逐帧音素概率，使用一个基于动态规划的帧级别对齐算法进行N-最佳假设与逐帧音素概率之间的对齐，为N-最佳假设中的各个单词计算时间起止点和置信度；第3步，在N-最佳假设中对关键词列表中的关键词逐个进行搜索匹配，并利用第2步中得到的时间起止点和置信度合并重复匹配的关键词，得到最终的检索结果。我们在一个越南语自由交谈数据集上进行了实验，结果表明，本文提出的基于端到端语音识别的关键词检索系统的$F1$值达到77.6%，相比基于传统DNN-HMM语音识别的关键词检索系统的$F1$值提升了7.8%。此外，精度-召回率曲线证明了本文系统能够提供可靠的关键词置信度，因此可以有效地针对不同的下游应用场景进行检索结果的精度-召回率调节。

本文第2节介绍了本文采用的端到端语音识别系统；第3节描述了提出的关键词检索框架；第4节给出了实验配置及实验结果；最后总结全文。

2 端到端CTC/注意力语音识别

本节将介绍关键词检索框架中的端到端语音识别系统。我们采用基于Transformer[11]神经网络结构的联合CTC/注意力架构[12]作为语音识别的基本框架。

2.1 多头注意力

Transformer作为一种完全基于注意力机制的编码器-解码器神经网络结构，不依赖于任何循环或卷积操作。Transformer的核心运算为多头注意力(Multi-Head Attention，MHA)，即使用h组不同的注意力，从而综合利用输入序列中不同方面的特征。

$$MHA(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=concat(head_1,\cdots,head_h)\boldsymbol{W}^O \tag{1}$$

其中：

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V) \tag{2}$$

其中，3个输入$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$分别代表查询(query)、键(key)和值(value)；$concat$表示拼接操作；$\boldsymbol{W}_i^Q,\boldsymbol{W}_i^K,\boldsymbol{W}_i^V,\boldsymbol{W}^O$表示线性投影矩阵。多头注意力中使用的注意力机制是缩放点乘注意力：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

可以理解为，注意力的输出由$\boldsymbol{V}$加权求和得到，权重由$\boldsymbol{Q},\boldsymbol{K}$的内积决定。$d_k$表示键的维度，用于归一化，避免$\boldsymbol{Q},\boldsymbol{K}$内积过大导致softmax函数过于接近0或1，梯度趋近于0。

2.2 编码器-解码器

与基于深度循环神经网络等其他结构的编码器-解码器网络相同，Transformer的编码器和解码器也分别由若干个相同的层堆叠而成，在编码器顶端是一个softmax全连接输出层，用于预测文本标签后验概率。编码器的输入为语音特征序列；解码器的输入为文本标签的嵌入向量序列。在语音特征输入编码器前，会通过一个卷积网络进行降采样，以减少后续神经网络的计算量。在训练时，解码器以上一个真实答案标签作为输入，最小化解码器输出当前标签概率与真实答案标签之间的交叉熵损失。在推理时，以上一步预测的输出标签作为输入，以自回归的形式预测当前标签。由于Transformer不包含任何循环神经网络结构，因此，在编码器和解码器输入序列中需要加入位置编码，从而使模型能够利用时序关系信息。

每个编码器层均由多头自注意力层和前馈网络两个子层构成，其中自注意力的输入$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$均来自前一子层的输出；每个子层前面都对输入进行了层归一化操作，而且每个子层后面添加了残差连接。解码器层结构与编码器层结构相似，但有两个区别：1)在自注意力中加入了掩盖操作，以保证训练时仅能看到在当前及之前时刻的输入；2)在自注意力和前馈网络之间增加了一个多头注意力子层，其输入$\boldsymbol{Q}$来自前一子层的输出，而$\boldsymbol{K},\boldsymbol{V}$来自编码器的最后一层输出，从而形成编

码器-解码器之间的注意力机制。

2.3 联合 CTC/注意力框架

在语音识别任务中，语音与文本间通常有明显的单调对齐关系。然而在基于注意力机制的编码器-解码器网络中，解码器输入与编码器输出之间并没有明确的对齐限制，因此这类架构难以学习语音与文本间的对齐关系。CTC 与之不同，通过动态规划可以保证输入与输出单调对齐。我们按照 Watanabe 等提出的联合 CTC/注意力框架[10]，在解码器顶端附加 CTC 进行联合训练和推理。

在解码器的预测标签集合之外，CTC 的预测标签额外包含一个特殊的“空白”标签。给定输入声学特征序列 $\boldsymbol{x}$，基于 CTC 的神经网络逐帧预测出 CTC 标签序列 $\boldsymbol{\pi}$，该序列包含“空白”标签，且允许连续多帧重复预测同一个标签。之后通过合并 $\boldsymbol{\pi}$ 中连续的相同输出标签、移除“空白”，得到坍缩后的最终标签序列。由于在坍缩过程中，多个不同的 CTC 标签序列可能会被映射为同一个最终标签序列，因此，记由所有映射到最终标签序列 $\boldsymbol{y}$ 的 CTC 标签序列构成的集合为 $\Phi(\boldsymbol{y})$。CTC 的目标是最大化真实答案标签序列 $\boldsymbol{y}^*$ 的输出后验概率：

$$P(\boldsymbol{y}^*|\boldsymbol{x})=\sum_{\pi\in\Phi(\boldsymbol{y}^*)}P(\boldsymbol{\pi}|\boldsymbol{x}) \tag{4}$$

联合 CTC/注意力架构利用 CTC 的对齐性质来辅助编码器-解码器的训练和推理。在训练中，CTC 损失函数 L_{CTC} 作为多任务学习中的一个目标，与基于注意力机制的解码器交叉熵损失函数 L_{ATT} 进行线性插值，得到总的损失函数：

$$L_{\mathrm{ASR}}=\alpha L_{\mathrm{CTC}}+(1-\alpha)L_{\mathrm{ATT}} \tag{5}$$

其中，α 为插值系数。在推理过程中，结合 CTC 分数和解码器分数，进行 CTC/注意力联合解码。

3 关键词检索

图 1 为本文提出的基于端到端语音识别的关键词检索框架示意图。图 1 中右侧虚线框的内容是第 2 节所述的联合 CTC/注意力的端到端语音识别前端。在该语音识别前端的解码结果的基础上可以直接进行关键词的搜索，但是无法得到可靠的关键词时间起止点和置信度。

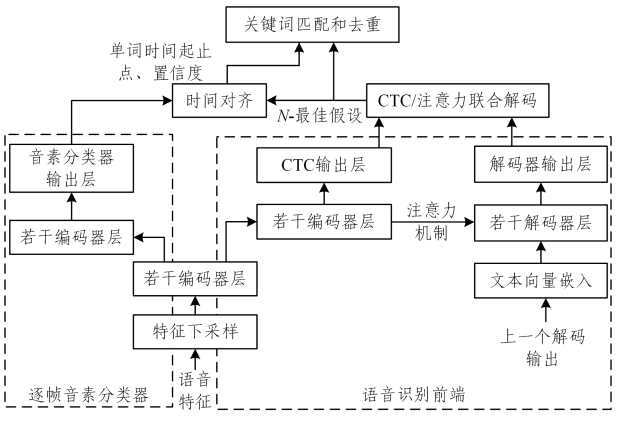


图 1 关键词检索框架示意图

Fig. 1 Diagram of the proposed keyword search framework

为了实现基于端到端语音识别的关键词检索，本文提出了一个和语音识别前端共同训练的逐帧音素分类器网络，以及一个基于解码结果和逐帧音素概率的帧级别对齐算法，来提供准确的时间起止点并改善置信度。在语音识别的 N-最佳假设中进行关键词匹配，并利用各个单词的时间起止点和置信度去除被重复匹配的关键词，获得最终的检索结果。下面将分别介绍逐帧音素分类器、帧级别对齐算法以及 N-最佳假设中的关键词匹配和去重方法。

3.1 逐帧音素分类器

图 1 中左侧虚线框给出了关键词检索框架中的逐帧音素分类器。该分类器采用与上一节所述的端到端语音识别中的编码器相同的基本网络结构。原始语音特征经由下采样输入由若干层堆叠构成的编码器网络，之后经过一个 softmax 全连接输出层，得到逐帧的音素后验概率。除有实际发音的音素之外，音素分类器的输出还包含特殊的“静音”音素，用于建模静音。我们通过使用交叉熵损失，利用逐帧的音素标注训练该音素分类器。

逐帧音素分类器与语音识别编码器两者的网络结构几乎完全相同，仅输出层维度存在差异。因此，在如图 1 所示的关键词检索架构中，两者共享下采样层以及最初的若干编码器层的参数，从而提高参数的利用效率。但为了避免两者在训练时互相干扰，我们将两者的较高层的若干解码器层隔离开，不进行参数共享。两者以多任务学习的方式联合训练，总损失函数由音素分类器损失函数 L_{PC} 与语音识别前端损失函数 L_{ASR} 线性插值得到：

$$L=\beta L_{\mathrm{PC}}+(1-\beta)L_{\mathrm{ASR}} \tag{6}$$

其中，β 为插值系数。

3.2 帧级别对齐

音素分类器的输出为单词每一帧语音上每一个音素的后验概率。在推理阶段，为了利用这一输出，给语音识别前端的识别结果提供准确的时间起止点和更可靠的置信度，我们使用如下的帧级别对齐算法。

第一步　将语音识别的解码结果单词序列映射为音素序列，并在句首、句尾以及相邻单词间插入“静音”音素。记由这一步得到的音素序列为 $(\varphi_1,\cdots,\varphi_M)$。

第二步　对于每个音素 φ_m，从逐帧音素分类器的输出中获取 φ_m 在每一语音帧上的后验概率。记 φ_m 在第 n 帧上的后验概率为 $p_n(\varphi_m)$，总语音帧数为 N，那么按照上述方法取出的音素后验概率构成了一个 $M\times N$ 矩阵 $\boldsymbol{P}$：

$$\boldsymbol{P}_{m,n}=p_n(\varphi_m) \tag{7}$$

第三步　得到该矩阵后，使用时间复杂度为 $\mathrm{O}(MN)$ 的动态规划算法，找到从矩阵左上角元素 $\boldsymbol{P}_{1,1}$ 到右下角元素 $\boldsymbol{P}_{M,N}$ 的最大累积后验概率路径。该路径满足：路径只向右或右下方前进，每一帧对应且仅对应一个音素，每个有实际发音的音素至少对应一帧，但“静音”音素可以跳过。

第四步　基于此最佳路径，我们追溯各个音素所对应的语音帧，进而得出语音识别解码结果中各个单词的起止帧，并按照模型的帧率计算出起止时间点。最后我们可以推算出单词 w 的帧平均音素后验概率置信度（简称为对齐置信度）：

$$score(w)=\frac{\sum_{(m,n)\in\Pi_w}\boldsymbol{P}_{m,n}}{len(\Pi_w)} \tag{8}$$

其中，Π_w 为最佳路径中单词 w 对应的部分，其长度为 $len(\Pi_w)$。我们使用解码器输出的标签后验概率均值置信度

(简称为解码器置信度)与对齐置信度进行线性插值得到混合置信度,并作为最终输出的关键词置信度。

3.3 N-最佳假设中的关键词匹配和去重

在得到语音识别结果中各个单词的时间起止点和置信度后,在识别结果中进行关键词检索。以往的研究显示,对于关键词检索任务而言,仅仅在解码得到的首选假设(即 N-最佳假设中解码得分最高的单词序列)中进行关键词匹配,得到的检索结果是不够充分的,会遗漏潜在的关键词结果[13]。为了发掘更多潜在的关键词结果,我们在 N-最佳假设中进行关键词匹配。由于 N-最佳假设之间有一定的重复单词,对于语音中同一个位置出现的关键词,在 N-最佳假设中可能会进行多次的匹配,因此,我们利用时间起止点和置信度对匹配结果进行去重。

首先,将最终结果列表置为空。

其次,从 N-最佳假设中分数排名第 N 的假设开始到排名第 1 的假设为止,倒序遍历各个假设,在各个假设中匹配关键词,并将匹配关键词结果加入最终的结果列表。

最后,若最终结果列表中存在与待加入匹配结果有重叠的时间范围且关键词相同的匹配结果,那么将上述两个匹配结果中置信度分数较低的一个从最终结果列表中排除,将置信度较高的一个加入最终结果列表。

4 实验配置及结果

4.1 实验语种

我们的实验在越南语上进行。越南语是越南的官方语言,也是我国京族的民族语言,在系属分类上属于南亚语系-越芒语族-越语支,在结构分类上属于孤立语。越南书面语使用基于拉丁字母的一套越南文字母进行书写,越南称之为"国语字",属拼音文字。越南语单词中的各个音节之间,在书写时以空格隔开。字母或字母组合有固定的发音,每个音节都可以按照一套规则从拼写方法映射到实际发音。端到端越南语语音识别因为能够采用字母、子词等与单词边界无关的单元进行建模,所以可以摆脱对单词级别词典等语言学专家知识的依赖。另外,相比于存在大量同音词、多音词的汉语、英语而言,越南语的发音和文字拼写之间有更明显且更确定的对应关系,不规则现象较少,这一点也使得越南语适合使用端到端的建模方法,而且方便本文提出的关键词检索框架在推理阶段进行拼写及到音素序列的映射,无需考虑集外词。

4.2 实验数据及预处理

本文实验中使用的语音数据为越南语自然对话数据,其中训练集约 100 h,开发集和测试集各约 5 h。我们使用 Kaldi[14] 进行语音预处理和特征提取,采用 40 维高分辨率梅尔频率倒谱系数以及 3 维声调特征作为语音特征[15],并在训练集上进行全局的倒谱均值方差归一化。

本文使用的越南语关键词列表中包含 200 个关键词,每个关键词的长度均在 1 到 4 个音节之间。我们使用全部训练集人工转录文本,用字节对编码(Byte Pair Encoding, BPE)算法[16]生成 1 000 个文本建模单元,作为端到端语音识别的输出单元。BPE 模型的训练和文本的子词切分基于 Sentencepiece 库实现。

4.3 关键词检索系统配置

4.3.1 基于 DNN-HMM 语音识别的基线系统

我们使用 Kaldi 平台[14]进行基于 DNN-HMM 语音识别的关键词检索实验作为基线,采用 Kaldi 标准流程训练了基于双向长短期记忆网络[17-18](Bidirectional Long Short-Term Memory, BiLSTM)-HMM 的三音子建模声学模型,对基于决策树聚类产生的 6 964 个绑定状态进行建模。BiLSTM 包含 3 个隐层,隐含节点数为 1 024。我们采用了一个包含约 10 000 个音节的越南语词表,并借助由训练集转录文本统计的四元文法语言模型进行解码,在解码得到的词图中进行关键词检索。

4.3.2 基于端到端语音识别的系统

我们在 ESPnet 平台[19]上开展基于端到端语音识别和帧级别对齐的关键词检索框架的相关实验。具体实验配置如下:音素分类器与语音识别共享底层的 9 层 Transformer 编码器,另外各自独占较高的 3 层编码器。语音识别部分的解码器为 6 层。各个编码器与解码器层中多头注意力的维度为 320,头数为 4,前馈神经网络维度为 2 048。音素分类器的建模单元为越南语的 24 个辅音、11 个不带调元音以及静音,共 36 个音素标签。模型训练时使用带 Noam 学习率衰减的 Adam 优化器进行优化,并使用 dropout(概率为 0.1)、标签平滑(系数为 0.1)、训练热身(25 000 步)和梯度裁剪(阈值为 5)训练,多任务学习损失插值系数 α 和 β 分别设置为 0.3 和 0.1,模型训练 25 轮。使用一个三音子建模的高斯混合模型-HMM 语音识别系统得到音素分类器训练所需的训练集语音逐帧音素标注。

在推理阶段,使用 CTC 权重为 0.5 的 CTC/注意力联合解码,在解码得到的 2-最佳假设中进行关键词检索,以达到最佳检索性能。我们使用的混合置信度中,解码器置信度与帧级别对齐置信度的插值系数分别为 0.7 和 0.3。关于如何选取这一组系数,将在 4.4.2 节进行详细介绍。

4.3.3 消融实验

为了进一步研究各个因素为本文提出的关键词解码框架带来的性能提升,我们在 4.3.2 节的系统基础上,通过操作以下 3 组实验条件进行了消融实验:

(1)使用基于 DNN-HMM 语音识别的基线系统声学模型中的 BiLSTM 结构相同的神经网络作为编码器,两层的单向长短期记忆网络作为解码器,代替 Transformer 网络结构。

(2)使用越南语字母和空格代替由 BPE 算法生成的子词,作为语音识别输出基本单元。

(3)去掉音素分类器分支,单独训练语音识别前端部分。利用 CTC 尖峰时间点获取单词起止时间点,采用单词中各个标签的解码器输出后验概率得分的均值作为单词置信度。解码和关键词匹配去重算法保持不变。

4.4 实验结果与分析

4.4.1 各系统的性能比较

各个关键词检索系统的性能如表 1 所列。我们使用全部检出的关键词(不考虑置信度)打分得到的 F1 值[20](KWS F1 Score)作为系统性能的综合评价指标。在打分过程中,判断某个检出关键词是否命中的条件是:该检出关键词与一个参

考答案单词相同，且检出关键词的时间中点处于该答案的时间起始点与终止点之间。除 F1 值外，表 1 还列出了各个系统的关键词检索精度(KWS Precision)、召回率(KWS Recall)和语音识别词错误率(ASR WER)作为参考。其中，Hybrid BiLSTM-HMM 为 4.3.1 节所述的基线系统；E2E Transformer BPE TA 为 4.3.2 节所述的本文提出的基于端到端语音识别的系统；中间的 5 行为 4.3.3 节所述的消融实验系统。在各消融实验系统中，Transformer 或 BiLSTM 表示神经网络结构；BPE 或 char 表示使用由 BPE 算法生成的子词或单个字母作为语音识别输出单元；包含 TA 的系统表示使用 3.2 节描述的帧级别对齐算法生成关键词时间起止点和置信度；不包含 TA 的系统则表示使用基于 CTC 的时间起止点和解码器输出的置信度。

表 1 各个关键词检索系统的性能

Table 1 Efficiencies of various KWS systems

(单位：%)

KWS system	KWS F1 Score	KWS Precision	KWS Recall	ASR WER
Hybrid BiLSTM-HMM	72.0	74.3	70.0	49.2
E2E BiLSTM,BPE	67.2	70.4	64.3	48.3
E2E Transformer char	70.6	71.7	69.5	47.1
E2E Transformer BPE	71.3	72.2	70.4	44.0
E2E BiLSTM BPE TA	74.3	77.9	71.0	48.0
E2E Transformer char TA	76.4	78.1	74.7	47.1
E2E Transformer BPE TA	**77.6**	**78.9**	**76.4**	**43.9**

实验结果显示，本文提出的 E2E Transformer BPE TA 系统在各个系统中展示出了最佳性能，比基线系统 Hybrid BiLSTM-HMM 的关键词检索 $F1$ 值提高了 7.8%，且精度和召回率均有提升。E2E BiLSTM BPE TA 系统相比 Hybrid BiLSTM-HMM 系统，其 $F1$ 值也有 3.2%的提升，这证明了在神经网络结构相似的条件下，采用 BPE 子词建模单元的端到端 CTC/注意力架构相比使用音节字典的传统 DNN-HMM 架构，在越南语关键词检索任务上有一定的优势。

消融实验证明了本文提出的音素分类器和帧级别对齐算法在基于端到端语音识别的关键词检索中的作用。表 1 的结果表明，3 个使用帧级别对齐的系统，相比各自对应的不使用帧级别对齐的系统，在语音识别词错误率基本相同的基础上，其 $F1$ 值分别提升了 8.2%～10.6%，3 个系统的 $F1$ 值平均提升了 9.2%。此外，我们直接计算了在测试集上的音素分类器准确率，为 86.9%。因此，音素分类器和帧级别对齐在不给语音识别性能造成负面影响的同时，能够为检出关键词提供更准确的时间起止点，提高系统的检索性能。

表 1 的实验结果还展示了用 BPE 算法生成的子词单元进行建模的优势。采用 BPE 子词作为语音识别建模单元相比使用单个越南语字母建模，$F1$ 值有 1.6%的提升。

4.4.2 置信度可靠性比较

上文使用关键词检索系统的全部检索结果进行了性能评价。除此之外，我们还使用置信度阈值对关键词检索结果进行过滤，保留分数高于阈值的部分检索结果并进行打分。通过设定若干组置信度阈值，能够调节关键词检索系统的精度和召回率，绘制出精度-召回率曲线[21](P-R 曲线)。对于 E2E Transformer BPE TA 系统的检索结果，我们使用 3 种不同的关键词置信度，即解码器置信度、对齐置信度以及两者线性插值得到的混合置信度，分别绘制了 P-R 曲线，如图 2 所示。曲线下面积越大代表置信度越可靠。在实验中，插值系数为 0.7 和 0.3 的混合置信度展示出了最佳的 P-R 曲线。为了使曲线图更加清晰，我们省略了其他插值系数的混合置信度的 P-R 曲线。观察图 2 可以看出，解码器置信度在高置信度阈值区域(即曲线左上角)有一处明显的凹陷，这表示在这一置信度范围内解码器置信度可靠性较差，难以通过置信度过滤进一步提高关键词检索精度。通过与对齐置信度进行插值发现，混合置信度有效弥补了这一缺陷，且在其他置信度区域保持了与解码器置信度一致的 P-R 曲线走势，证明了本文提出的方法对整体的关键词置信度可靠性的提升有很大帮助。

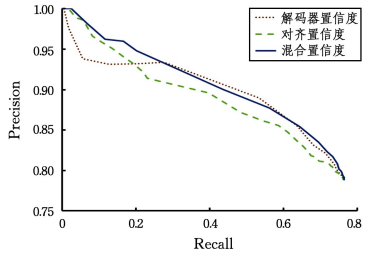


图 2 不同置信度的 P-R 曲线

Fig. 2 P-R curves for different confidence scores

结束语 本文提出了一个基于端到端语音识别的关键词检索框架，利用音素分类器和帧级别对齐解决了端到端语音识别自身无法为关键词检索任务提供准确的时间起止点和置信度的问题。本文使用一个与语音识别模型联合训练的音素分类器输出逐帧音素概率，采用一个基于动态规划的对齐算法，利用音素概率为语音识别结果中的各个单词计算时间起止点和置信度，在 N-最佳假设中对关键词进行匹配。本文在一个越南语自由交谈数据集上进行了实验，结果表明，本文提出的关键词检索系统相比基于传统 DNN-HMM 语音识别的关键词检索系统，具有更高的关键词检索 F1 值，而且能够提供可靠的关键词置信度用于检索结果的精度-召回率调节。未来我们会将本文工作扩展到更多的语种和方言上，并探索采用不依赖于额外音素分类器的方法来进行基于端到端语音识别的关键词检索。

参 考 文 献

[1] SHAO J,ZHAO Q,ZHANG P,et al. A fast fuzzy keyword spotting algorithm based on syllable confusion network[C]// Eighth Annual Conference of the International Speech Communication Association. 2007.

[2] ZHANG P,SHAO J,HAN J,et al. Keyword spotting based on phoneme confusion matrix[C]// Proc. of ISCSLP. 2006:408-419.

[3] AUDHKHASI K,ROSENBERG A,SETHY A,et al. End-to-end ASR-free keyword search from speech[J]. IEEE Journal of Selected Topics in Signal Processing,2017,11(8):1351-1359.

[4] MYER S,TOMAR V S. Efficient keyword spotting using time delay neural networks[C]// Proc. Interspeech 2018. 2018:1264-1268.

[5] KINGSBURY B, CUI J, CUI X, et al. A high-performance Cantonese keyword search system[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 8277-8281.

[6] CHOROWSKI J, BAHDANAU D, SERDYUK D, et al. Attention-based models for speech recognition[C]// Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. 2015: 577-585.

[7] CHAN W, JAITLY N, LE Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016: 4960-4964.

[8] GRAVES A, FERNÁNDEZ S, GOMEZ F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]// Proceedings of the 23rd International Conference on Machine Learning. 2006: 369-376.

[9] LI J, YE G, DAS A, et al. Advancing acoustic-to-word CTC model[C]// 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 5794-5798.

[10] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[12] NAKATANI T. Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration[C]// Proc. Interspeech 2019. 2019: 1408-1412.

[13] SARACLAR M, SPROAT R. Lattice-based search for spoken utterance retrieval[C]// Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. 2004: 129-136.

[14] POVEY D, GHOSHAL A, BOULIANNE G, et al. The Kaldi speech recognition toolkit[C]// IEEE 2011 workshop on automatic speech recognition and understanding. IEEE Signal Processing Society, 2011 (CONF).

[15] ZHENG C J, WANG C L, JIA N. Survey of Acoustic Feature Extraction in Speech Tasks[J]. Computer Science, 2020, 47(5): 110-119.

[16] GAGE P. A new algorithm for data compression[J]. C Users Journal, 1994, 12(2): 23-38.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

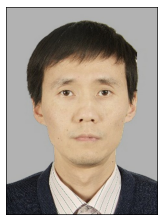
[18] ZHANG S, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification[C]// Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. 2015: 73-78.

[19] WATANABE S, HORI T, KARITA S, et al. Espnet: End-to-end speech processing toolkit[C]// Interspeech. 2018: 2207-2211.

[20] NANCY C. MUC-4 evaluation metrics[C]// Conference on Message Understanding. Association for Computational Linguistics, 1992.

[21] RAGHAVAN V, BOLLMANN P, JUNG G S. A critical investigation of recall and precision as measures of retrieval system performance[J]. ACM Transactions on Information Systems (TOIS), 1989, 7(3): 205-229.

YANG Run-yan, born in 1996, postgraduate. His main research interests include multi-lingual automatic speech recognition and keyword search.



LIU Jian, born in 1971, Ph. D, professor, master's supervisor. His main research interests include continuous automatic speech recognition, embedded speech recognition, and music retrieval.

(责任编辑:李亚辉)