

文章编号: 1004-4353(2021)03-0273-06

基于语音识别的朝鲜语语音检索方法

徐博文, 金小峰*

(延边大学 工学院, 吉林 延吉 133002)

摘要: 针对基于语音识别的语音检索方法对语言模型的强依赖问题,通过改进声学模型学习框架提出了一种新的朝鲜语语音检索方法. 该方法首先修改 KoSpeech 框架的网络模型,通过训练得到了朝鲜语的声学模型;其次通过语音文档分割方法构建了语音文档索引库;最后利用编辑距离匹配的方法实现了语音检索. 实验结果表明,改进的朝鲜语声学模型学习框架降低了语音检索方法对语言模型的依赖和大规模数据集的要求. 当 k 取 9 时, top- k 评价方法的检索均值平均精度达到 86.74%, 召回率达到 95.25%, 该结果表明本文提出的方法是有效的,具有一定的实际应用价值.

关键词: 语音检索; 语音识别; 声学模型; 语音切分

中图分类号: TP391.42

文献标识码: A

Korean speech retrieval method based on speech recognition

XU Bowen, JIN Xiaofeng*

(College of Engineering, Yanbian University, Yanji 133002, China)

Abstract: Aiming the issue that recognition based speech retrieval method relies heavily on language model, a novel Korean speech retrieval method based on improved acoustic model learning framework is proposed. First, Korean acoustic model is trained by modified KoSpeech framework network model. Second, speech documents index library is constructed by speech document segmentation method. Finally, Levenshtein distance matching method is used to implementation speech retrieval. Experiments result show that proposed improved model of Korean acoustic reduces the dependency of language model and the requirement of largescale dataset for retrieval method. For the top- k evaluation method, mAP and recall rate reach best to 86.74% and 95.25% respectively when $k = 9$, so it is firmly demonstrated that the proposed method is effective and has certain practical application value.

Keywords: speech retrieval; speech recognition; acousticmodel; speech segmentation

0 引言

语音检索是指在语音文档中查找与检索语音相关的语音片段及其定位信息的方法^[1]. 目前,语音检索大多采用的是分步策略的方法,即首先通过上游的语音识别技术得到语音的转写结果,然后再经过下游的检索得到最终的结果. 2006 年,

Burget 等^[2]在语音识别的基础上基于多字查询算法和对三音素进行索引提出了一种语音识别词格网络的检索方法,该方法的检索效果显著优于单音素网络. 2011 年,李伟^[3]提出了一种基于内容的汉语语音检索方法,该方法可有效提高检索效率. 金惠琴^[4]利用特征级融合和 PCA 降维的方

收稿日期: 2021-04-07

基金项目: 吉林省教育厅“十三五”科学技术项目(JJKH20191126KJ);延边大学外国语言文学世界一流学科建设项目(18YLPY14)

* 通信作者: 金小峰(1970—),男,硕士,教授,研究方向为语音信息处理、计算机视觉及机器人技术.

法设计了一种维吾尔语关键词检索系统,该系统可有效提高维吾尔语的检索速度和重音检测率. Liu 等^[5]提出了一种利用区分性局部空间-时间描述符对中文语音关键词进行检索的方法,该方法可有效提高语音检索中的抗噪能力. 王朝松等^[6]提出了一种侧重于关键词的深度神经网络声学建模方法,该方法利用非均匀的最小分类错误准则来调整深度神经网络声学建模中的参数,并利用 AdaBoost 算法来动态调整声学建模中的关键词权重,从而提高了关键词检索的性能. 李鹏等^[7]提出了一种将不同语音识别系统的词图进行相交融合的关键词检索方法,该方法能综合利用各词图的得分信息来减小冗余,进而可有效提高关键词的检索效率. Chen 等^[8]提出了一种将语音识别和填充模型相融合的方法,该方法可提高关键词的检出性能. Zhuang 等^[9]利用 LSTM-CTC (long short term memory-connectionist temporal classification) 提出了一种基于深度学习的非限制词表关键词的检索方法,该检索方法具有词典无关的优点. Dhananjay 等^[10]提出了一种基于音素子空间特征增强的关键词检索方法,实验结果表明该方法优于传统 DNN 后验概率的方法. 2021 年, Huang 等^[11]在编码器-解码器网络中引入了多头注意机制和软三重损失函数,该方法可有效提高检索性能. 本文借鉴上述研究中的分步策略,利用改进的朝鲜语语音识别框架 KoSpeech^[12]学习来得到朝鲜语声学模型,并在此基础上提出了一种基于语音识别的朝鲜语语音检索方法.

1 改进的朝鲜语声学模型学习框架

2021 年, Kim 等^[12]提出了一种专门针对朝鲜语语音识别的 KoSpeech 框架,该框架的核心编解码器为 LAS (listen, attend and spell) 模型^[13]. LAS 模型由 1 个声学模型编码器和 1 个基于注意力机制的字符解码器组成,如图 1 所示. 由于 LAS 模型包含注意机制,因此该模型可以通过学习直接得到语音和文本之间的映射关系. 且当训练数据充足时,该模型还可以实现声学模型与语言模型的联合学习.

LAS 框架有 2 个核心模块,即 Listener 模块

和 AttendAndSpeller 模块. 其中, Listener 模块由多层双向长短时记忆网络 (Bi-directional long short-term memory, BLSTM) 堆叠而成. 声学特征编码序列 $\mathbf{x} = (x_1, x_2, \dots, x_T)$ 经由 Listener 模块输出时,其被转换为更为高级的编码形式 $\mathbf{h} (\mathbf{h} = (h_1, h_2, \dots, h_U))$, 其中 $U < T$, 该编码过程可表示为:

$$\mathbf{h} = \text{Listen}(\mathbf{x}). \quad (1)$$

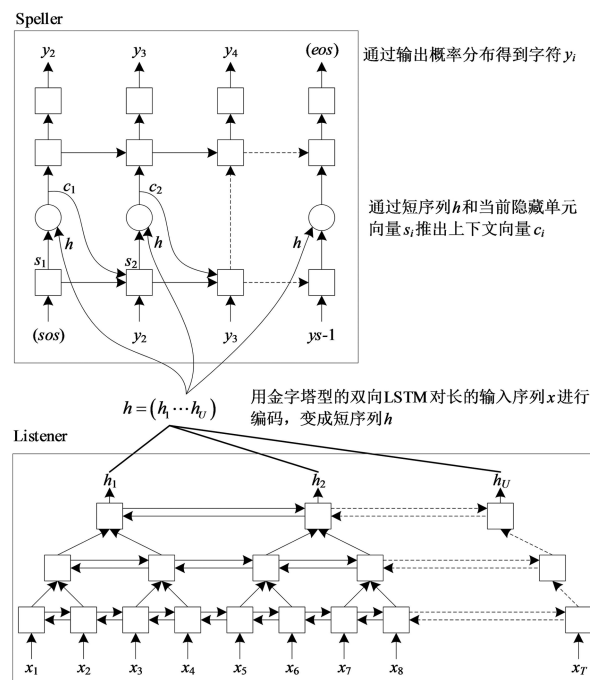


图 1 LAS 模型的结构

AttendAndSpell 是一个基于注意力机制的 LSTM 解码器函数,为 Speller 的核心操作,它的主要作用是完成声学特征编码 \mathbf{x} 与对应文本字符序列 \mathbf{y} 的映射. 设 $\mathbf{y} = (\langle \text{sos} \rangle, y_1, \dots, y_s, \langle \text{eos} \rangle)$, 则 AttendAndSpell 可以通过 (\mathbf{h}, \mathbf{y}) 计算得到 \mathbf{y} 在输入 \mathbf{x} 时的概率分布 $P(\mathbf{y} | \mathbf{x})$, 即:

$$P(\mathbf{y} | \mathbf{x}) = \text{AttendAndSpell}(\mathbf{h}, \mathbf{y}). \quad (2)$$

由式 (2) 可知, AttendAndSpell 会根据先前产生的所有字符预测下一个输出字符的概率分布,即根据上一时间步的输出字符 y_{i-1} 、解码器状态向量 \mathbf{s}_{i-1} 和上下文向量 \mathbf{c}_{i-1} 来推理当前时间步的向量 \mathbf{s}_i . 该过程可表示为:

$$\mathbf{s}_i = \text{RNN}(\mathbf{s}_{i-1}, y_{i-1}, \mathbf{c}_{i-1}), \quad (3)$$

其中 RNN 在 LAS 网络内是一个两层的 LSTM, $\mathbf{c}_i (\mathbf{c}_i = \text{AttentionContext}(\mathbf{s}_i, \mathbf{h}))$ 由注意力机制得到. 由 \mathbf{s}_i 和 \mathbf{c}_i 可推出当前时间步的输出字符 y_i

的概率分布. 设 CharacterDistribution 为多层感知机的 softmax 输出, 第 i 个时间步的字符输出概率的计算公式为: $P(y_i | \mathbf{x}, y_{<i}) = \text{Character-Distribution}(s_i, c_i)$, 其中 $y_{<i}$ 为之前所有的输出字符. 因此 AttendAndSpell 在解码时可将输入的语音声学特征预测为最有可能的字符序列, 即:

$$\hat{\mathbf{y}} = \arg \max_y \log P(\mathbf{y} | \mathbf{x}). \quad (4)$$

由于 LSTM 性能与 GRU 性能相近^[14], 因此

本文对 KoSpeech 框架做如下改进: 用图 2(b) 中的单向门控循环单元 (GRU) 替换图 2(a) 中的 Bidirectional-LSTM 和 Unidirectional-LSTM. 由于 GRU 比 LSTM 少 1 个门, 因此其能够降低运算量, 进而提高学习的速度. 改进后的 KoSpeech 框架可通过学习得到的朝鲜语声学模型将语音文档和检索语音的语音信号转写为文本并输出.

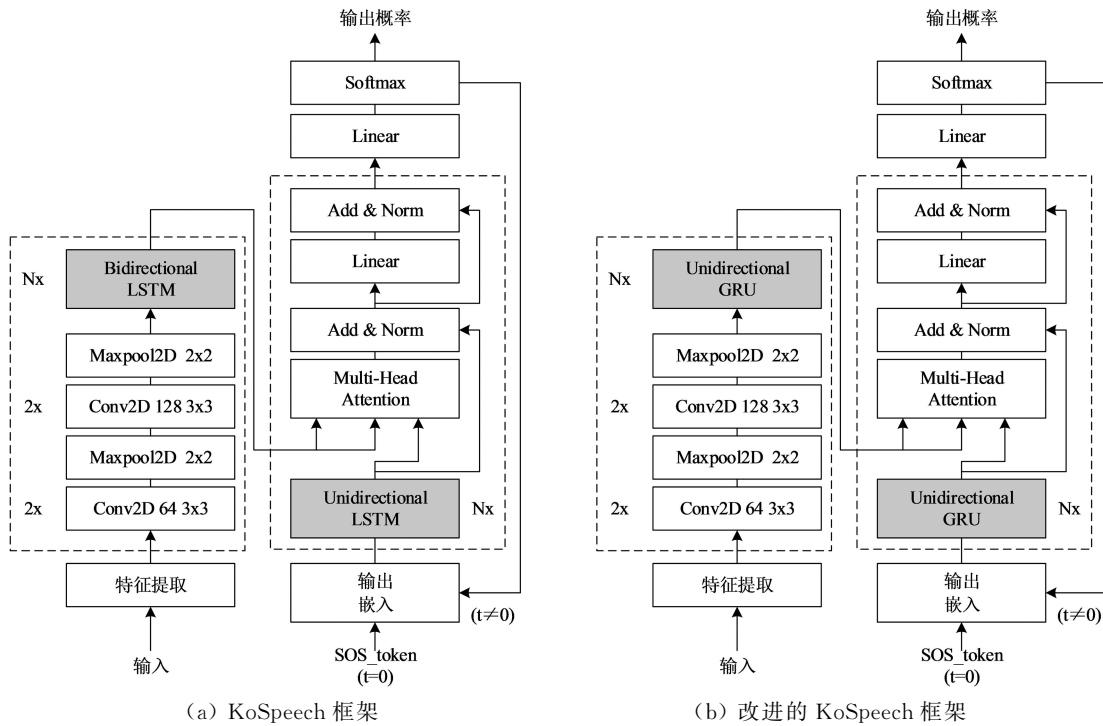


图2 改进前后的 KoSpeech 框架

2 语音文档的分割方法

语音检索需要建立语音文档索引库, 其主要目的是为了便于集中管理语音文档和提高检索速度. 另外, 为提高语音文档转写文本的准确性和语音检索的时间定位精度, 本文提出了一种语音文档语义分割方法, 即将语音文档分割为一系列具有相对语义完整性的子语音文档. 语音文档的分割方法是依据人们的说话习惯和文语语料数据集中语音数据的平均时长, 将 5 s 作为每段子语音的初始分割时长, 然后再结合语音信号的能量谱和阈值进一步精确定位分割点. 具体的分割步骤如下:

输入: 语音文档

输出: 子语音文档分割信息

初始化: 初始分割语音段的时长 T ($T=5$ s),

初始分割起止位置 ($pstart = pend = 0$).

Step1 读取一个语音文档数据 W , 计算静音段的平均能量阈值 e 、能量谱 E 和语音文档的总时长 L .

Step2 while($pend \leq L$):

$pend \leftarrow pend + T$

if $E(pend) \geq e$ then

 向后搜索, 直至 $E(pend + \Delta t) \leq e$,

$pend \leftarrow pend + \Delta t$

将 $W[pstart, pend]$ 作为分割片段, 转写分割片段, 记录该片段的起止时间

$pstart \leftarrow pend$

Step3 把 W 子语音文档分割信息记录到索引库中.

Step4 如果处理完所有语音文档,转 Step5,否则转 Step1.

Step5 输出所有语音文档的分割信息.

图 3 是语音文档采用上述分割方法处理后的结果,图中第 1 行是该语音文档的完整波形图,第 2 行至第 7 行是分割的子语音文档的波形.由图 3 可以看出,在保证分割语义完整性的前提下,各子语音文档的时长存在差异.

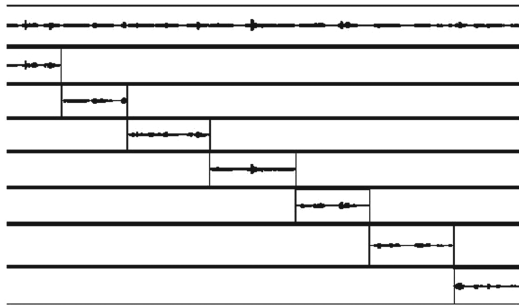


图 3 语音文档切分结果

语音文档索引库包含的主要信息有语音文档名称、子语音文档中的相对完整的起止时间以及子语音文档的转写文本信息.为了验证本文提出

的语音文档分割方法的有效性,对 277 个语音文档进行了分割实验.分割共得到了 1 978 个子语音文档,其中最长为 7.5 s、最短为 0.32 s、平均为 4.3 s.该分割结果与人工分割结果接近(见表 1),由此说明本文提出的语音文档分割方法是有效的.

表 1 语音文档分割实验结果

分割方法	语音文档数	子语音文档分割信息			
		分割数	平均/s	最长/s	最短/s
本文方法	277	1 978	4.3	7.5	0.32
人工方法	277	1 970	4.4	7.3	0.50

3 朝鲜语语音检索方法

本文提出的朝鲜语语音检索方法采用分步策略:第 1 步,通过改进的 KoSpeech 框架学习得到朝鲜语的声学模型,以此实现语音文档和检索语音等语音信号的文本转写输出;第 2 步,将语音检索任务转化为文本检索任务,即在语音文档索引库的转写文本中匹配和定位检索语音的转写文本.本文提出的语音检索方法的处理流程见图 4.

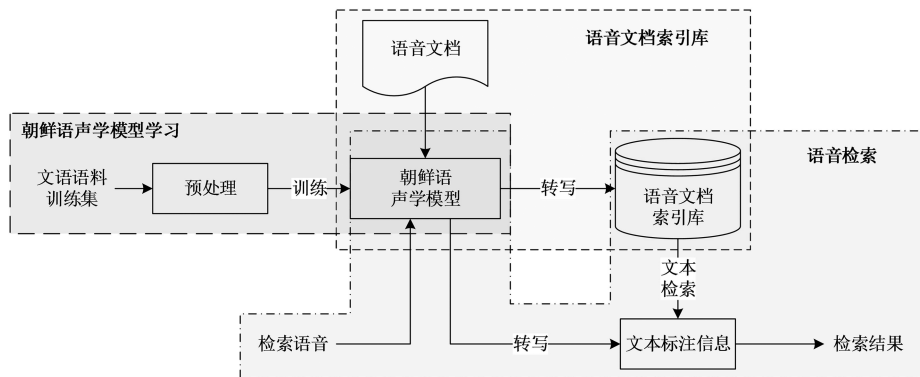


图 4 朝鲜语语音检索的处理流程

本文采用的文本检索方法是基于编辑距离(levenshtein distance)的文本相似度度量方法,即对检索语音转写文本与索引库转写文本的匹配度进行打分,并以 top- k 评价检索的准确率.计算两个字符串 a 和 b 的编辑距离的公式为:

$$\text{lev}_{a,b}(i,j) = \begin{cases} \max(i,j), & \min(i,j) = 0; \\ \min \begin{cases} \text{lev}_{a,b}(i-1,j) + 1 \\ \text{lev}_{a,b}(i,j-1) + 1 \\ \text{lev}_{a,b}(i-1,j-1) + \mathbf{1}_{(a_i \neq b_j)} \end{cases}, & \text{其他.} \end{cases}$$

其中, $\text{lev}_{a,b}(i,j)$ 表示的是 a 中前 i 个字符与 b 中前 j 个字符之间的距离.当 $\min(i,j) = 0$ 时,意味着 i 和 j 中有一个为 0, $\text{lev}_{a,b}(i,j) = \max(i,j)$,即编辑距离为 i 和 j 中的最大值.当 $\min(i,j) \neq 0$ 时,编辑距离为删除操作($\text{lev}_{a,b}(i-1,j) + 1$)、插入操作($\text{lev}_{a,b}(i,j-1) + 1$)和替换操作($\text{lev}_{a,b}(i-1,j-1) + \mathbf{1}_{(a_i \neq b_j)}$)下的最小值. $\mathbf{1}_{(a_i \neq b_j)}$ 是指示函数,当 $a_i = b_j$ 时取 0,当 $a_i \neq b_j$ 时取 1.

若将本文检索语音的转写文本字符串设定为 a ,将索引库中的每个子语音文档所对应的转写

文本字符串设定为 b ,则 a 和 b 的匹配度得分可由式(5)计算获得.得分越高表示两者越相似.按照得分值大小降序排序即可生成候选的检索结果.

$$score = \left(1 - \frac{\text{lev}_{a,b}(i,j)}{i+j}\right) \times 100\%. \quad (5)$$

4 结果与分析

实验数据采用公开数据集 AI-hub 中的朝鲜语语料,每个样本由一个语义完整的语音文件和一个对应的人工标注文本文件组成,所有的语音数据时长为 1 000 h.由于受计算机硬件资源的限制,本文滤除了文本字符数超过 100 的样本,并在滤除后的文本中随机选出 1.1 万条文语料用于声学模型训练,其中训练集 8 000 条、验证集 2 000 条、测试集 1 000 条.样本的音频格式为 PCM,采样频率为 16 000 Hz,音频特征为 80 维 fbank,帧长度为 20 ms,帧移为 10 ms,窗口使用汉明窗.

4.1 网络超参数调优实验

在改进的 KoSpeech 架构的超参数调优实验中,分别对批处理大小、迭代次数和隐层单元数进行了调优,评价指标采用字符错误率(character error rate, CER).CER 越低表示语音识别方法的性能越好.具体调优实验过程如下:

1) 批处理大小参数调优.预设隐层单元数为 256,分别取批处理大小 4、8、16.当迭代次数为 40、批处理大小为 16 时 CER 收敛(稳定在 0.613),因此最佳批处理大小为 16.

2) 隐层单元数参数调优.批处理大小取 16,隐层单元数分别取 128 和 256.当迭代次数为 20 次、隐层单元数为 256 时 CER 收敛(稳定在 0.801),因此最佳隐层单元数为 256.

3) 迭代次数参数调优.批处理大小和隐层单元数分别取 16 和 256.当迭代次数为 300 次时 CER 收敛(稳定在 0.225),因此最佳迭代次数为 300.

超参数调优实验的具体结果见表 2.由以上结果可知,当批处理大小、隐层单元数以及迭代次数分别取 16、256、300 时模型的性能最优.

表 2 超参数调优实验结果

超参数	参数值	CER
批处理大小	4	0.623
	8	0.620
	16	0.613
隐藏层神经元个数	128	0.900
	256	0.801
迭代次数	20	0.801
	40	0.613
	150	0.318
	300	0.225

4.2 改进的 KoSpeech 框架的性能验证

为了验证改进的 KoSpeech 架构的性能,本文将改进的 KoSpeech 架构与文献[12]中的 KoSpeech 原型架构进行对比实验,结果见表 3.由表 3 可以看出,改进的 KoSpeech 架构的 CER 指标与 KoSpeech 原型架构基本接近,表明二者的语音转写性能相近.另外,改进的 KoSpeech 架构的网络参数规模和迭代平均耗时显著低于 KoSpeech 原型架构,表明改进的 KoSpeech 架构的学习速度优于 KoSpeech 原型架构.

表 3 改进前后的 KoSpeech 架构的性能

架构方法	网络参数规模/MB	迭代平均耗时/s	CER
KoSpeech 原型架构 ^[12]	2.56	212.6	0.133 2
本文改进的架构	1.92	160.3	0.133 6

4.3 朝鲜语语音检索实验

检索实验采用另外准备的 451 条检索语音,评价指标使用基于 top- k 的召回率(recall)和均值平均精度(mean average precision, mAP).计算 mAP 时,首先利用式(6)计算不同 k 值对应的平均精度值(average precision, AP),然后再利用式(7)求出 mAP 值.

$$AP(q) = \frac{1}{N} \sum_{k=1}^N \text{rel}(k) / k, \text{rel}(k) \in \{0, 1\}, \quad (6)$$

$$mAP = \frac{1}{Q} \sum_{q=1}^Q AP(q), \quad (7)$$

其中 Q 表示查询结果的个数, $AP(q)$ 表示第 q 个查询精度, N 表示检索数据库的语音个数, $\text{rel}(k)$ 表示检索的语音是否和查询语音相关(1 为

相关, 0 为不相关)。

表 4 为 k 取不同值($k=1, 2, 3, \dots, 10$) 时的实验结果, mAP 和 recall 随 k 值的变化见图 5。从表 4 和图 5 可以看出:随着 k 值的增加, 召回率显著提高, 并且在 $k=9$ 时召回率达到最大值(95.25%); mAP 在 $k \geq 2$ 时呈现小幅上升随后趋于稳定, 且在 $k=9$ 时达到最大值(86.74%)。mAP 虽然总体上低于召回率, 但是较高的召回率对于语音检索任务而言比 mAP 更具实用意义。

表 4 本文方法的检索实验结果 %

top- k	mAP	recall	top- k	mAP	recall
1	81.75	81.75	6	86.59	94.00
2	85.13	88.50	7	86.62	94.25
3	86.00	91.00	8	86.72	95.00
4	86.15	91.75	9	86.74	95.25
5	86.55	93.75	10	86.74	95.25

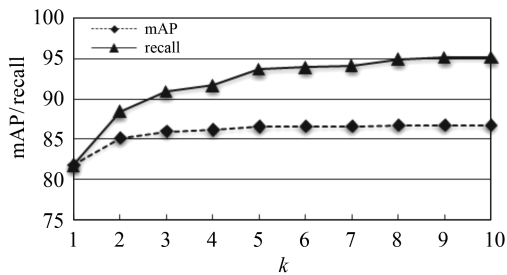


图 5 mAP 和 recall 随 k 值的变化

5 结论

研究表明, 本文以构建朝鲜语声学模型为目标而改进的 KoSpeech 框架可以降低基于语音识别的语音检索方法对数据集规模和语言模型的依赖, 进而可以减少模型参数规模, 提高训练速度。本文提出的语音文档的分割方法能够有效地分割出具有相对完整语义的子语音文档, 有助于提高语音文档转写文本的准确性和语音检索的时间定位精度。当 $k=9$ 时, 本文方法语音检索的召回率和均值平均精度分别达到了 95.25% 和 86.74%, 该结果表明本文提出的语音检索方法是有效的, 可应用在朝鲜语的语音检索中。在今后的研究中, 我们将尝试构建音素级的朝鲜语声学模型, 以此进一步提高语音转写的准确率。

参考文献:

[1] AKIBA T, NISHIZAKI H, NANJO H, et al.

Overview of the NTCIR-12 SpokenQuery&Doc-2 task[C]//Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: NII, 2016:167-179.

[2] BURGET L, CERNOCK J, FAPSO M, et al. Indexing and search methods for spoken document[J]. Lect Notes Comput Sci, 2006, 4188(1):351-358.

[3] 李伟. 基于内容的汉语语音检索技术与系统实现[D]. 北京:清华大学, 2011.

[4] 金惠琴. 基于维吾尔语语音关键词检索的研究[D]. 乌鲁木齐:新疆大学, 2013.

[5] LIU H, FAN T, WU P. Audio-visual keyword spotting for mandarin based on discriminative local spatial-temporal descriptors[C]//2014 22nd International Conference on Pattern Recognition. Stockholm: IEEE, 2014:785-790.

[6] 王朝松, 韩纪庆, 郑铁然. 基于非均匀 MCE 准则的 DNN 关键词检测系统中声学模型的训练[J]. 智能计算机与应用, 2015, 5(5):15-17.

[7] 李鹏, 屈丹. 采用词图相交融合的语音关键词检测方法[J]. 信号处理, 2015(6):702-709.

[8] CHEN I F, NI C, LIM B P, et al. A keyword-aware language modeling approach to spoken keyword search[J]. J Signal Process Syst, 2016, 82(2): 197-206.

[9] ZHUANG Y, CHANG X, QIAN Y, et al. Unrestricted vocabulary keyword spotting using LSTM-CTC[C]//Proceedings of 2016 Interspeech. San Francisco: ISCA, 2016:938-942.

[10] DHANANJAY R, AFSANEH A, HRVÉ B. Phonetic subspace features for improved query by example spoken term detection[J]. Speech Commun, 2018, 103:27-36.

[11] HUANG J, GHARBIEH W, SHIM H S, et al. Query-by-example keyword spotting system using multi-head attention and softtriple loss[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021:6858-6862.

[12] KIM S, BAE S, WON C. Open-source toolkit for end-to-end Korean speech recognition[J]. Software Impacts, 2021, 7:100054.

[13] CHAN W, JAITLEY 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). Shanghai: IEEE, 2016:4960-4964.

[14] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv:1412.3555v1, 2014.