CSLT TECHNICAL REPORT-2016/12/21 [Thursday 5th January, 2017]

基于Kaldi的哈萨克语语音识别系统

Ying Shi

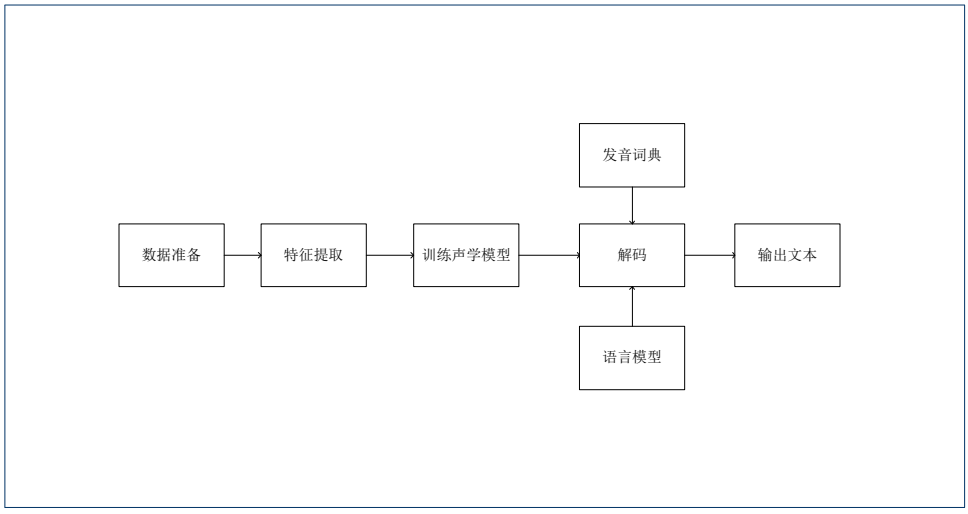
Correspondence:
shiying@cslt.riit.tsinghua.edu.cn
Full list of author information is available at the end of the article

Abstract

近年来深度学习技术 [1]在人工智能领域的得到广泛应用，语音识别作为人工智能的一个分支也得到了前所未有的发展，越来越多的科研机构及公司投身到了语音识别领域。尽管如此，哈萨克语作为我国众多小语种中较为重要的一种，依然没有一套成型的可供科学研究及工程应用的语音识别系统。本文将基于当前最为著名的语音识别工具Kaldi，从数据准备阶段开始介绍如何从无到有构建哈萨克语语音识别系统。本文的受众对象为初入语音识别领域的读者，所以在这篇文章将没有过多的专业知识的介绍，旨在降低语音识别及Kaldi 的入门门槛，让没有太多基础的读者能够快速使用自己的数据构建语音识别系统。

Keywords: 深度学习; 语音识别; Kaldi

1 简介



上图展示了基于Kaldi的语音识别的流程，如图所示除了最后的输出文本外基于Kaldi 的语音识别系统，总共包括6步，其中绝大多数步骤融合在Kaldi的脚本中（特征提取，训练声学模型，解码）。所以用哈萨克语的输入及输出搭建语音识别系

统，我们唯一要做的就是数据准备（其中发音词典和语言模型被包含在这一步中）。数据准备完成后，剩下的工作就是数据放在特定的目录下然后运行相应的脚本并等待结果。而这对于没有很多基础知识的读者来说并不是太难的事情。在后文中我们将先介绍哈语的特性再基于上图逐步介绍我们是如何构建哈萨克语语音识别体统的。其中我们将详细介绍数据准备部分。

2 哈语的特点

包含很多词根和词缀(待补充)

3 数据准备

在这个章节中我们将分目录介绍需要准备哪些数据，以及准备这些数据时需要注意的一些事项。另外在准备音频文件时我们建议读者将音频文件设计为如下形式。

/work3/shiying/kazak/kazak-spdb-src/lvcsr/data/train/F0101001.wav

这是我们准备的哈语音频文件的形式，其中第一个字母代表说话人的性别，F代表女性，M代表男性，这个字母与紧接着的4位数字0101 和共同构成了说话人编号，最后的三位数字001代表这位说话人所说的第一句话，这个文件名除后缀以外的部分共同构成了某句话的ID，这种命名方式最大程度上包含了我们后边需要用到信息，在一定程度上降低了数据准备所需要的工作量。

3.1 data/{train,test,dev}

在这个子章节中我们将向读者介绍data/{train,test,dev}这三个目录下所包含的内容以及各自的意义，实际上这三个目录下的文件的名称是一样的，只不过内容和各自的用途不一样，train目录下的文件用于训练声学模型，test集下的文件用于测试，dev是一个较为特殊的目录在Kaldi 的先前版本（nnet1）中被用到，我们所做的哈萨克语识别是基于Kaldi nnet3 版本的，所以在这篇中没有过多的关于dev集的介绍，感兴趣的读者可以去Kaldi 官网查看官方文档。我们以training集为例来介绍这四个文件应该包括的内容。

在data/train目录下包含以下的几个文件

wav.scp	text	utt2spk	spk2gender	spk2utt

其中text文件的内容如下：

F0101_001	psyhykaleq masElEneN jastanwGa bEtalwenan ...
F0102_002	bEyjyNtyanjynhebEy vux ozEn aterawe jwjyaN ...

在这个文件中包含了两列，其中第一列代表某句话的ID，可以看到与之前提到的命名方式相同下划线前的部分代表说话人的ID下划线后边的部分代表该说话人所说的第几句话。第二列则代表这句话的内容。对于出现在第二列的内容，最好不要包括标点符号，因为第二列的每个元素都要在后文我们将提到的Lexicon中有对应的发音，为了省去不必要的麻烦，我们去掉了所有的标点符号，同时我们也建议初学者这样做。另外我们得到的最原始的哈语语料并不是英文字母形式，而是阿拉伯文的形式，从阿拉伯字母到英文的字母的转换我们使用了新疆大学米吉提老师提供的CodeMap。用法如下：

在chars目录下
./program/codetransform.pl orgfilename newname

其中orgfilename代表需要转换的文件，转换后的文件将被写在newname中。

wav.scp

F0101_001	/work3/shiying/kazak-spdb-src/lvcsr/train/F0101001.wav
F0101_002	/work3/shiying/kazak-spdb-src/lvcsr/train/F0101002.wav

在这个文件中，第一个元素与text中第一个文件相同，第二个文件则代表这句话的音频文件的具体位置，这里最好使用绝对路径。在wav.scp 中每个音频必须在text中有唯一的一条语句与其对应。

utt2spk

F0101_001	F0101
F0102_002	F0102

utt2spk表示句子和说话人的对应关系，其中每一行的第一个元素代表句子的ID第二个元素代表说话人的ID。

spk2utt

F0101	F0101_001 F0101_002 F0101_003...
F0102	F0102_001 F0102_002 F0102_003...

与utt2spk相同spk2utt代表说话人与句子的对应关系，每一行的第一个元素代表说话人的ID，另一个元素则代表句子的ID。在Kaldi的众多脚本中使用kaldi/egs/thchs30/utils/{utt2spk_to_spk2utt.pl ,spk2utt_to_utt2spk.pl} 这两个脚本可以实现utt2spk 与spk2utt这两个文件之间的相互转换。所以读者只需要手动创建其中的一个文件，另一个可以通过脚本生成。

spk2gender

F0101	f
M0102	m

很显然这个文件包括的是说话人以及他们的性别信息。所以第一个元素代表说话人的ID另一个元素则代表该说话人的性别，其中f代表女性m代表男性（性别需要用小写）。以上所述的四个文件都需要进行排序，值得注意的是，Kaldi源码是使用C++语言实现的，读者最好在自己的环境变量里加上“export LC_ALL=C”，否则在排序时可能发生排序原则与C++不符合，导致在后边的脚本中出现错误。

3.2 data/dict/{lexicon.txt,nonsilence_phones.txt,silence_phones.txt,optional_silence.txt}

lexicon.txt

#	SIL
$< SPOKEN\_NOISE >$	SIL
SIL	SIL
Anen	A n e n
ENbEge	E N b E g e

lexicon.txt是dict目录下较为重要的一个文件，这个文件其实是一个字典，但与普通字典不同的是，这个字典包含的是完整的字以及这个字对应的发音，所以在lexicon.txt中第一列代表字，第二列则代表这个字的发音序列，在哈语中每个字母都有自己的发音，所以在经过阿拉伯字母到英文字母的转换后，每个字的发音序列就是组成这个字的字母的序列。在这个文件中每个字都必须对应至少一个发音，可以存在多音字。理论上lexicon.txt 应该包含某种语言的所有字和所有的发音，但是由于哈语语料的限制，我们只使用了出现在哈语语料中的所有的字。在这个文件的前三行，是三个较为特殊的字符。#代表可能出现在语料中的某种字符，$< SPOKEN_NOISE >$代表噪音，SIL代表静音，我们将这三个特殊字符都对应到了SIL（silence）静音上。

nonsilence_phones.txt

A
E
G
H
N
.
.
.

这个文件中必须包含所有非静音的音素。

silence_phones.txt

SIL

这个文件仅包含一个元素，就是静音的音素的符号。在哈语中我们使用SIL。

optional_silence.phones.txt

SIL

这个文件中包含被设为静音的音素，在哈语中也是只有SIL。

3.3 语言模型

数据准备的最后一步就是制作语言模型，通过使用一个强大的制作语言模型的工具srilm这个步骤会变得异常的简单。在这里我们要简略介绍一下n-gram [2]模型，n-gram 模型是一种基于统计的模型，即一个句子中的一个词出现的概率只与前N-1个词有关，与其他词无关，根据概率的知识很容易得到一句话出现的概率就是所有词出现概率的乘积，所以1-gram模型就是基本的词频统计，2-gram模型考虑词对，3-gram模型则考虑三元组。

制作语言模型只需要两个文件，第一个是语料，第二个则是词表。用于训练语言模型的语料，原则上要求足够大能够涵盖该语言在各个领域的表达，但是这么全面的语言模型用在较为简单的语音识别的任务上往往是非常浪费的，所以读者可以为自己的语音识别任务准备一个较小的语言模型，但是这个语言模型最好能与test集的语言处于相近或相同的领域，否则会对语音识别的效果有一定的影响，同时我们需要能够保证用于训练语言模型的语料每一行都具有实际意义，对于哈语我们的做法是按照句号问好感叹号换行，保证每一行是一个完整的句子，对于语料中出现的数字也有一些特别的要求，一般语料中的数字都是阿拉伯数字，对于这些阿拉伯数字最好的做法就是将数字转换成相应语言的特定格式，例如："2016"转换成汉语的"二零一六"，但是对于哈语将所有的阿拉伯数字转化成哈语形式的数字是非常困难的，所以我们的做法是将数字用空格隔开，即"2016"转换成"2 0 1 6"，再将0-9这10个阿拉伯数字加入到前边提到的lexicon.txt 中即可。训练语言模型所使用的词表就是组成语料的词表，但是我们往往不需要用全部的词表的去做语言模型，读者可以根据自己的需要对词表按词频进行裁剪，对于哈萨克语我们按词频从高到低排列取了前10 万的词。

我们所做的哈萨克语的语音识别所使用的语言模型是3-gram模型，制作语言模型的脚本为run_lm.sh（这个脚本由工程师张志勇提供）。

run_lm.sh corous order name vocab

其中corpus代表用于训练语言模型的语料，order代表使用几gram模型，name是制作出来的语言模型的命名，vocab是词表。这样制作的语言模型对于语音识别任务来说总是存在一些冗余，往往需要进行裁剪，幸运的是srilm提供了裁剪工具。如下：

```
ngram -prune parameter -lm LMname -write-lm newLMname
```

prune代表裁剪语言模型所用的参数，例如对于哈语语音模型的裁剪时的参数，我们使用了2e-7，代表词频概率取对数后小于$2 * 10^{-7}$ 的全部去掉。LMname代表需要裁剪的语言模型，newLMname代表裁剪后的语言模型的命名。语言模型的准备工作到这里就基本全部结束了，剩下的工作就是就运行Kaldi中的脚本然后等待结果。

4 特征提取

在接下来的章节中，我们将对Kaldi提供的一些脚本进行简略的说明，以Kaldi/egs/thchs30/s5/run.sh 为例，事实上我们所做的哈萨克语的语音识别也是以这个脚步为参照做的。

语音识别所用的机器学习的算法是监督学习，经过上一步的数据准备，我们已经将监督学习所需要的输入以及输出准备完成了，但是Kaldi 对于用于训练模型的输入与输出有自己的格式要求，特征提取其实就是将输入转化为Kaldi所要求的格式，对于输入数据，Kaldi 提供了两种不同的特征提取的方式：MFCC以及FBANK。其中MFCC多用于训练GMM [3]模型，而FBANK 多用于DNN模型，这两种特征提取所用的脚本分别为“kaldi/egs/thchs30/s5/steps/make_mfcc.sh”和“kaldi/egs/thchs30/s5/make_fbank.sh”。这两个脚本所需要传入的参数都是一样的，所以我们以make_mfcc.sh为例介绍这两个脚本所需要的参数以及他们所生成的文件的含义。

make_mfcc.sh

```
steps/make_mfcc.sh –nj $n –cmd ”$train_cmd” data/mfcc/train
exp/make_mfcc/ mfcc/train
```

每个参数的含义：

- -nj	这个参数代表所使用的job的个数，例如我们将这个参数设置为8，那么Kaldi 则会将输入数据分为8份然后投递到8个节点去运行
- -cmd	这个参数代表使用哪种q函数
data/mfcc/train	输入数据所在的目录
exp/make_mfcc/	脚本的log 文件将被写在这个目录下
mfcc/train	特征提取的结果将被写在这个目录下

这个脚本最终的产物是feats.scp它的内容为:

F0101_001	/work3/shiying/kazak/mfcc/train/raw_mfcc_train.1.ark:10

Kaldi会通过特征提取将音频文件转换成矩阵格式，每一句在一个矩阵中，矩阵的每一行对应音频文件的一帧(一般为25ms)。feats.scp与前边数据准备时所用的wav.scp相对应，第一列是某句话的ID，第二列是这句话在Kaldi标准格式的矩阵中所在的起始位置。以上述的列子来说"/work3/shiying/kazak/mfcc/train/raw_mfcc_train.1.ark:10"，代表F0101_001这句话在raw_train_1.ark中起始位置为第十个字节。

提取提取的第二步时计算cmvn

steps/compute_cmvn_stats.sh data/mfcc/train exp/mfcc_cmvn/train mfcc/train

每个参数的含义:

data/mfcc/train	输入数据所在的目录
exp/make_cmvn/	log文件将被写在这个目录下
mfcc/train	计算cmvn的结果被写在这个目录下

通过计算cmvn所得到的文件与feats.scp的格式是相同的:

F0101	/work3/shiying/kazak/mfcc/train/cmvn_train.ark:6

其中第一列元素代表说话人的ID，第二列元素也是一个矩阵，这个矩阵包含了说话人的统计倒谱均值和做过normalization的方差。

与输入音频文件相同输出也需要通过脚本转换成Kaldi的标准格式。

utils/prepare_lang.sh - -position_dependent_phones false data/dict "$< SPOKEN\_NOISE >$" data/local/lang data/lang

每个参数的含义为:

- -position_dependent_phones false	是否使用词位信息，对于哈萨克语的识别我将它设为false
data/dict	这个脚本所需要的输入数据所在的目录
$< SPOKEN\_NOISE >$	oov词汇将被映射到这个符号上
data/local/lang	临时输出的目录
data/lang	最后的结果将被写在这个目录下。

这个脚本主要围绕数据准备阶段所准备的lexicon相关的文件做一些处理。在data/lang目录下会生成很多文件其中比较重要的两个是L.fst和L_disambig.fst这两个其实是lexicon的fst（finite state transducers 有限状态转换机）版本。对于初学者没有必要深究这其中的实现方式。

另外一个脚本主要是对语言模型的处理:

```
utils/format_lm.sh data/lang data/graph/kazak.3.lm.gz
kazak/lm/lexicon.txt data/graph/lang
```

每个参数的含义为:

data/lang	L.fst 所在的目录
data/graph/kazak.3.lm.gz	所准备的语言模型（注意语言模型需要被压缩为.gz格式）
kazak/lm/lexicon.txt	lexicon.txt
data/graph/lang	最后的输出将被写在这个目录下

这个脚本的主要产物为G.fst，与前文提到的L.fst 相同它是语言模型的fst版本。

5 模型训练以及解码

Kaldi对于模型的训练以及解码都遵循同样的流程即：训练模型-解码测试-数据标注- 使用标准数据训练下一个模型。在这一章节中我们以训练monophone为例讲解每个脚本的作用以及参数的意义。

5.1 训练模型

训练monophone所使用的脚本为：

```
steps/train_mono.sh - -boost-silence 1.25 - -nj $n - -cmd "$train_cmd"
data/mfcc/train  data/lang exp/mono
```

每个参数的含义为:

- -boost-silence 1.25	提升静音似然度的因子。
- -cmd "$train_cmd"	与前文所介绍的cmd意义相同
- -nj $n	与前文所介绍的nj意义相同
data/mfcc/train	输入数据的目录
data/lang	L.fst 所在的目录
exp/mono	最终的模型将被写在这个目录下

模型训练结束后，在exp/mono目录下我们会得到一个名为fina.mdl（gmm 模型）的模型。这个文件是最终GMM模型的二进制拓扑结构，我们可以使用“copy-transition-model –binary=false final.mdl final.txt”将这个二进制文件转换成txt格式以便查看其中的内容。

5.2 解码

在解码的时候，我们需要使用前边的结果准备一个新名为：HCLG.fst的文件,这个文件是由H C L G通过特定方式合成的。其中L与G分别代表前边提到的L.fst和G.fst。H 则代表HMM相关的内容，C代表音素级别的上下文，如果读者对这个文件的内容感兴趣可以通过查看Kaldi的官方文档以获取关于这个文件的细节描述性 [4]。

用于生成HCLG.fst的脚本为

```
utils/mkgraph.sh - - mono data/graph/lang exp/mono mono/graph
```

每个参数的意义为:

- - mono	是否使用mono生成解码图
data/graph/lang	G.fst与L.fst所在的目录
exp/mono	tree所在的目录
mono/graph	HCLG.fst 将被写在这个目录下

decode的脚本为:

```
steps/decode.sh - - cmd ”$decode_cmd” mono/graph data/mfcc/test
exp/mono/decode
```

每个参数的意义为：

- - cmd	与前文介绍的cmd 的意义相同
mono/graph	HCLG.fst 所在的目录
data/mfcc/test	测试数据所在的目录
exp/mono/decode	最后的结果将被写在这个目录下

解码结束后，在exp/mono/decode目录下将会生成很多类似wer_10.0.0这样的文件，这些文件代表识别的错误率（wer word error rate），我们可以使用“grep wer* | kaldi/egs/thchs30/utils/best_wer.sh” 查看最低的错误率可以达到多少。

5.3 标注

标注的任务是使用当前的模型对数据进行标注，得到每一帧对应的音素（实际上Kaldi是将数据标注为帧-transition id,并不是音素为了便于初学者理解，我们将它写为帧-音素），标注的脚本为

steps/align_si.sh - -boost-silence 1.25 –nj $n - -cmd ”$train_cmd”
data/mfcc/train data/lang exp/mono exp/mono_align

每个参数的意义为（前文介绍过的参数没有在这里重复列出）:

data/mfcc/train	这个脚本所需要的输入数据所在的目录
data/lang	L.fst所在的目录
exp/mono	final.mdl 所在的目录
exp/mono_align	标注结果将被写在这个目录下

标注结束后在exp/mono_align目录下会生成很多类似ali.1.gz的文件。对于这些标注文件Kaldi 提供一些系列工具查看每一帧对应的音素和pdf等 [4]：

ali-to-phones final.mdl ark:1.ali ark,t:file	查看每一帧对应的音素序号
ali-to-pdf final.mdl ark:1.ali ark,t:file	查看每一帧对应的pdf
copy-int-vector ark:1.ali ark,t:file	查看每一帧对应的transition-id

5.4 TDNN

在这一章节中，我们将以TDNN（time delay neural network,kaldi/egs/wsj/s5/local/nnet3/run_tdnn.sh）为模板像读者介绍Kaldi 如何实现一个DNN 模型。

steps/nnet3/train_tdnn.sh - - parameter data/mfcc/train data/lang
exp/tri4b_ali exp/nnet3/tdnn

每个参数的意义为：

- -stage	训练状态数
- -num-epochs	epochs的数量，iteration将会由这个参数生成
- - splice-indexes	TDNN的结构由这个参数生成
- -feat-type	feat的种类
- -initial-effective-lrate	起始时的学习率
- -final-effective-lrage	结束时的学习率
- -cmd	与前文介绍的cmd意义相同
- -pnorm-input-dim	pnorm 的输入维度（pnorm是一种激活函数，Kaldi还提供了其他种类的激活函数比如：Tanh，Sigmoid，Relu等）
- -pnorm-output-dim	pnorm 的输出维度
data/train	训练数据所在的目录
data/lang	L.fst所在的目录
exp/tri4b_ali/	标注数据所在的目录
exp/nnet3/tdnn	最终模型将被写在这个目录下

对于初学者来说参数列表中的参数基本不需要自己手动修改，直接运行这个脚本然后等待结果即可，但是很多时候我们的程序会因为一些非脚本内部逻辑错误而中断，比如说硬件死机等等，这时候通过修改- -stage这个参数可以迅速的从程序中断的状态恢复运行。Kaldi在训练模型的时候每训练一轮就会在存放结果的目录下保存一个当前的模型，一般被命名为iteration.mdl。iteration 就代表当前模型训练到了多少轮，所以在我们的程序因为一些因素中断后，我们可以将- -stage的设为iteration的值，程序就可以从中断的状态继续训练，而不是从头开始，可以节约很多时间成本。

6 总结

本文所介绍的哈语的语音识别的模型，只是最初步的语音识别模型，在数据处理方面其实还存在一些不是很合理地方，未来还会为提高哈语识别率做其他改进工作，另外这篇文章的受众群体为刚开始接触语音识别和Kaldi 的读者，对于语音识别和Kaldi更细节的知识与技术，还需要读者在熟悉Kaldi 的流程后多做一些探索与实践。

7 推荐读物

http://kaldi-asr.org/doc/
https://www.inf.ed.ac.uk/teaching/courses/asr/
http://deeplearning.net/reading-list/
Automatic Speech Recognition A Deep Learning Approcach By Dong Yu and Li Deng
Deep Learning by Yoshua Bengio,Ian Goodfellow,Aaron Courville

References

1. Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1798-1828.
2. https://en.wikipedia.org/wiki/N-gram
3. https://www.inf.ed.ac.uk/teaching/courses/asr/2015-16/asr03-hmmgmm-handout-nup.pdf
4. http://kaldi-asr.org/doc/
5. Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]//Proceedings of INTERSPEECH. ISCA, 2015: 2440-2444.