



kaldi 资料归纳和总结

wbglearn (吴本谷) version 0.2

2014 年 8 月 16 日

目录

目录.....	2
0 开篇前的话.....	3
1 kaldi 的介绍.....	4
1.1 kaldi 简介.....	4
1.2 kaldi 的特色.....	4
1.3 kaldi 的声学模型.....	4
1.4 kaldi 所用到的库介绍.....	4
2 kaldi 的安装和出现错误的解决方案.....	5
2.1 安装前的配置.....	5
2.2 安装.....	5
2.2.1 Ubuntu.....	5
2.2.2 Cygwin.....	6
2.2.3 服务器或者工作站.....	6
3 kaldi 的使用.....	7
3.1 总述.....	7
3.2 yesno 的例子.....	15
3.3 timit.....	15
3.4 rm.....	15
3.5 voxforge.....	15
4 kaldi 主页上的翻译.....	16
4.1 数据准备.....	17
4.2 特征提取.....	33
4.3 声学建模代码.....	35
4.4 解码在测试部分.....	37
4.5 Karel 的神经网络训练实现.....	45
4.6 kaldi 中的关键词检索.....	49
4.7 在线识别.....	57
4.8 决策树是如何在 kaldi 中使用.....	61
4.9 HMM 拓扑结构和转移模型.....	69
4.10 kaldi tutorial.....	77
5 附录.....	85
5.1 kaldi 上搭建 TIMIT 基线系统.....	85
5.2 kaldi 里的 voxforge.....	88
5.3 在 vs 2013 中编译 kaldi.....	94
6 其他资料和资源.....	95
7 版本更新日志.....	96

0 开篇前的话

首先，先打 2 个广告。希望大家看到文档的同时，可以积极参与讨论。

第一个广告：kaldi 学习 语音深度学习群，群号是：367623211。欢迎大家的加入。

第二个广告：语音识别论坛，论坛地址：<http://www.threedweb.cn/forum-76-1.html> 论坛欢迎大家的发帖和跟帖，欢迎大家积极讨论。

其次，关注我博客（<http://blog.csdn.net/wbgxx333>）的人应该知道本文档是建立在群里大家的共同努力完成的，特别是对 kaldi 主页翻译的部分，感谢每个参与翻译的人，感谢你们的辛勤劳动和付出，希望可以对学习 kaldi 的人有用。

最后，希望大家可以积极参与讨论，让群和论坛全部活跃起来，也可以使语音识别逐渐成为大家的生活的一部分。

备注：此文档不定期更新，如果您发现有任何问题和疑问，欢迎随时在群里跟我联系或者通过邮箱（wbglearn@gmail.com 或者 354475072@qq.com）与我联系，我们将逐渐完善该文档。

2014.8.16

最近都在对 kaldi 的脚本和源码有兴趣，欢迎大家积极交流。大家的看的时候希望做下笔记，然后写下你自己的理解，希望你可以发给我，我将放在我们这里，供大家学习，希望越多的人加入进来，为未来的人铺好路。此外，大家在学习 kaldi 的过程一定要注重对代码和脚本的学习，学的深一点，不要仅仅会运行。祝大家学习愉快！

1 kaldi 的介绍

1.1 kaldi 简介

kaldi是一个开源的语音识别工具箱，是基于c++编写的，可以在windows和unix平台上编译。

1.2 kaldi 的特色

与文本无关的LVCSR系统；
基于FST的训练和解码；
最大似然训练；
各种各样的线性和映射变换；
有VTLN，SAT的脚本；

1.3 kaldi 的声学模型

支持标准的机器学习训练模型：
线性变换如：LDA HLDA,MLLT/STC；
说话人自适应:fMLLR,MLLR；
支持GMM,SGMMs,DNN

1.4 kaldi 所用到的库介绍:

- 1.OpenFst: Weighted Finite State Transducer library (加权有限状态转换器)
- 2.ATLAS/CLAPACK:标准的线性代数库
- 3.sph2pipe:由 sph 文件转成其他音频文件
- 4.firstlm:语言模型的工具箱
- 5.sctk: score benchmark (评价 ASR 基准)

2 kaldi 的安装和出现错误的解决方案

2.1 安装前的配置

安装前你需要对你的 linux 进行配置，你需要安装的软件有：

apt-get
subversion
automake
autoconf
libtool
g++
zlib
libatlas
wget

具体安装方法如下：

- (1) sudo apt-get install libtool
- (2) sudo apt-get install autoconf
- (3) sudo apt-get install wget
- (4) sudo apt-get install perl
- (5) sudo apt-get install subversion
- (6) sudo apt-get install build-essential
- (7) sudo apt-get install gfortran
- (8) sudo apt-get install libatlas-dev
- (9) sudo apt-get install libatlas-base-dev
- (10) 安装zlib

```
wget http://www.zlib.net/zlib-1.2.8.tar.gz
tar -xvzf zlib-1.2.8.tar.gz
cd zlib-1.2.8.tar.gz
./configure
make
sudo make install
```

2.2 安装

2.2.1 Unbuntu

1. svn co svn://svn.code.sf.net/p/kaldi/code/trunk kaldi-trunk
2. 先回到tool目录下，在命令行输入：make。
3. make 完后，在src目录下：

```
./configure  
make depend  
make
```

2.2.2 Cygwin

32 位的系统稳定，不会出现错误。据说 64 位的会出现 bug。
kaldi在cygwin里安装出现minsize.o错误的解决办法：
在openfst 里面执行 `configure --enable-static --disable-shared`
然后make 然后到tools 里面make .

备注：这部分有待完善，将在后面的版本中更新。

2.2.3 服务器或者工作站

实测CentOS7.0, 工作站安装。

I 安装前准备

安装前你需要对你的centos进行配置，步骤如下：

```
# yum check-update;    检查更新  
# yum install -y;    安装找到的更新，确保已安装的软件是最新版的
```

依赖软件的安装（如果已经安装可以跳过）。

- (1) `sudo yum install libtool`
- (2) `sudo yum install autoconf`
- (3) `sudo yum install wget`
- (4) `sudo yum install perl`
- (5) `sudo yum install subversion`
- (6) `sudo yum install zlib`
- (7) `sudo yum groupinstall "Development Tools"`

II 安装

(1) 静态库安装方法（默认）：

```
# cd tools; make  
# ./install_atlas.sh  
# cd ../src; ./configure; %此步骤中只能用configure， 加入--shared会报错：  
# make depend; make
```

(2) 动态库安装方法：

```
# cd tools; make
```

将 install_atlas.sh 中的 `../configure $opt --prefix=`pwd`/install || exit 1;`

改为:

```
../configure $opt --prefix=`pwd`/install --shared|| exit 1;
```

```
# ./install_atlas.sh
```

```
# cd ../src; ./configure --shared 这里如果报错, 请将--shared去掉
```

```
# make depend; make
```

(待写)

备注: 由于各个平台和你自己电脑配置的问题, 如果大家遇到什么问题, 欢迎在群里交流。

3 kaldi 的使用

3.1 总述

在跑 kaldi 里的样例时, 你需要注意三个脚本: `cmd.sh` `path.sh` `run.sh`。下面分别来说, `Cmd.sh` 脚本为:

```
# "queue.pl" uses qsub. The options to it are
# options to qsub. If you have GridEngine installed,
# change this to a queue you have access to.
# Otherwise, use "run.pl", which will run jobs locally
# (make sure your --num-jobs options are no more than
# the number of cpus on your machine.
```

```
#a) JHU cluster options
```

```
#export train_cmd="queue.pl -l arch=*64"
```

```
#export decode_cmd="queue.pl -l arch=*64,mem_free=2G,ram_free=2G"
```

```
#export mkgraph_cmd="queue.pl -l arch=*64,ram_free=4G,mem_free=4G"
```

```
#export cuda_cmd=run.pl
```

```
#b) BUT cluster options
```

```
#export train_cmd="queue.pl -q all.q@@blade -l ram_free=1200M,mem_free=1200M"
```

```
#export decode_cmd="queue.pl -q all.q@@blade -l ram_free=1700M,mem_free=1700M"
```

```
#export decodebig_cmd="queue.pl -q all.q@@blade -l ram_free=4G,mem_free=4G"
```

```
#export cuda_cmd="queue.pl -q long.q@@pco203 -l gpu=1"
#export cuda_cmd="queue.pl -q long.q@pcspeech-gpu"
#export mkgraph_cmd="queue.pl -q all.q@@servers -l ram_free=4G,mem_free=4G"
```

#c) run it locally...

```
export train_cmd=run.pl
export decode_cmd=run.pl
export cuda_cmd=run.pl
export mkgraph_cmd=run.pl
```

大家可以很清楚的看到有 3 个分类分别对应 a, b, c。a 和 b 都是集群上去运行这个样子,c 就是我们需要的。我们在虚拟机上运行的。你需要修改这个脚本。

Path.sh 的内容:

```
export KALDI_ROOT=`pwd`/../../..
export
PATH=$PWD/utils/:$KALDI_ROOT/src/bin:$KALDI_ROOT/tools/openfst/bin:$KALDI_ROOT
/tools/irstlm/bin/:$KALDI_ROOT/src/fstbin/:$KALDI_ROOT/src/gmmbin/:$KALDI_ROOT/src/
featbin/:$KALDI_ROOT/src/lm/:$KALDI_ROOT/src/sgmmbin/:$KALDI_ROOT/src/sgmm2bin/:
$KALDI_ROOT/src/fgmmbin/:$KALDI_ROOT/src/latbin/:$KALDI_ROOT/src/nnetbin:$KALDI
I_ROOT/src/nnet-cpubin/:$KALDI_ROOT/src/kwsbin:$PWD:$PATH
export LC_ALL=C
exportIRSTLM=$KALDI_ROOT/tools/irstlm
```

在这里一般只要修改 export KALDI_ROOT=`pwd`/../../..改为你安装 kaldi 的目录, 有时候不修改也可以, 大家根据实际情况。

Run.sh 里大家需要指定你的数据在什么路径下, 你只需要修改:

如:

```
#timit=/export/corpora5/LDC/LDC93S1/timit/TIMIT # @JHU
timit=/mnt/matylida2/data/TIMIT/timit # @BUT
```

修改为你的 timit 所在的路径。

其他的数据库都一样。

此外, voxforge 或者 vystadial_cz 或者 vystadial_en 这些数据库都提供下载, 没有数据库的可以利用这些来做实验。

最后, 来解释下 run.sh 脚本。我们就用 timit 里的 s5 来举例阐述:

```
#!/bin/bash
```

```
#
```

```
# Copyright 2013 Bagher BabaAli,
```

```
# 2014 Brno University of Technology (Author: Karel Vesely)
```

```
#
```



```
# TIMIT, description of the database:
# http://perso.limsi.fr/lamel/TIMIT_NISTIR4930.pdf
#
# Hon and Lee paper on TIMIT, 1988, introduces mapping to 48 training phonemes,
# then re-mapping to 39 phonemes for scoring:
# http://repository.cmu.edu/cgi/viewcontent.cgi?article=2768&context=compsci
#
```

```
./cmd.sh
[ -f path.sh ] && ./path.sh
set -e
```

这个就是之前说的你需要修改的，修改好才可以运行成功，否则会报错

```
# Acoustic model parameters
numLeavesTri1=2500
numGaussTri1=15000
numLeavesMLLT=2500
numGaussMLLT=15000
numLeavesSAT=2500
numGaussSAT=15000
numGaussUBM=400
numLeavesSGMM=7000
numGaussSGMM=9000
```

声学模型的参数，这部分我们基本默认不改，除非你自己训练自己的数据库

```
decode_nj=5
train_nj=30
```

nj 是指你需要运行 jobs 的数量，可以根据你机器的 CPU 的数量来改变。一般不超过 CPU 的数量

```
echo
```

```
=====
```

```
echo "                Data & Lexicon & Language Preparation
echo
```

```
=====
```

这部分是数据准备部分，在 kald 主页翻译中，@v 给大家翻译的足够详细，具体的可以去看那篇译文。

```
#timit=/export/corpora5/LDC/LDC93S1/timit/TIMIT # @JHU
timit=/mnt/matylda2/data/TIMIT/timit # @BUT
```

这就是你需要修改的 timit 数据路径。

```
local/timit_data_prep.sh $timit || exit 1
```

```
local/timit_prepare_dict.sh
```

```
# Caution below: we insert optional-silence with probability 0.5, which is the
# default, but this is probably not appropriate for this setup, since silence
# appears also as a word in the dictionary and is scored. We could stop this
# by using the option --sil-prob 0.0, but apparently this makes results worse.
# (-> In slite scoring the deletions of 'sil' are not scored as errors)
utils/prepare_lang.sh --position-dependent-phones false --num-sil-states 3 \
    data/local/dict "sil" data/local/lang_tmp data/lang
```

```
local/timit_format_data.sh
```

```
echo
```

这部分是提取特征部分，这里只提取了 mfcc, kald 里支持 mfcc, plp, pitch 等特征，然后做了 CMVN。

```
=====  
echo "          MFCC Feature Extration & CMVN for Training and Test set          "  
echo  
=====
```

```
# Now make MFCC features.  
mfccdir=mfcc
```

```
for x in train dev test; do  
    steps/make_mfcc.sh --cmd "$train_cmd" --nj 30 data/$x exp/make_mfcc/$x $mfccdir  
    steps/compute_cmvn_stats.sh data/$x exp/make_mfcc/$x $mfccdir  
done
```

这部分是单音素的训练和解码部分。语音识别最基础的部分。

```
echo
```

```
=====  
echo "          MonoPhone Training & Decoding          "  
echo  
=====
```

```
steps/train_mono.sh --nj "$train_nj" --cmd "$train_cmd" data/train data/lang exp/mono
```

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

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
exp/mono/graph data/dev exp/mono/decode_dev
```

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
exp/mono/graph data/test exp/mono/decode_test
```

这部分是三音素的训练和解码部分。

```
echo
```

```
=====  
echo "          tri1 : Deltas + Delta-Deltas Training & Decoding          "  
echo  
=====
```

```
steps/align_si.sh --boost-silence 1.25 --nj "$train_nj" --cmd "$train_cmd" \  
data/train data/lang exp/mono exp/mono_al
```

```
# Train tri1, which is deltas + delta-deltas, on train data.
steps/train_deltas.sh --cmd "$train_cmd" \
    $numLeavesTri1 $numGaussTri1 data/train data/lang exp/mono_ali exp/tri1
```

```
utils/mkgraph.sh data/lang_test_bg exp/tri1 exp/tri1/graph
```

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \
    exp/tri1/graph data/dev exp/tri1/decode_dev
```

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \
    exp/tri1/graph data/test exp/tri1/decode_test
```

```
echo
```

```
echo "                                tri2 : LDA + MLLT Training & Decoding"
echo
```

这部分是在三音素模型的基础上做了
LDA+MLLT 变换。

```
steps/align_si.sh --nj "$train_nj" --cmd "$train_cmd" \
    data/train data/lang exp/tri1 exp/tri1_ali
```

```
steps/train_lda_mllt.sh --cmd "$train_cmd" \
    --splice-opts "--left-context=3 --right-context=3" \
    $numLeavesMLLT $numGaussMLLT data/train data/lang exp/tri1_ali exp/tri2
```

```
utils/mkgraph.sh data/lang_test_bg exp/tri2 exp/tri2/graph
```

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \
    exp/tri2/graph data/dev exp/tri2/decode_dev
```

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" \
    exp/tri2/graph data/test exp/tri2/decode_test
```

```
echo
```

这部分是在三音素模型的基础上做了
LDA+MLLT+SAT 变换。

```
echo "                                tri3 : LDA + MLLT + SAT Training & Decoding"
echo
```

```
# Align tri2 system with train data.
steps/align_si.sh --nj "$train_nj" --cmd "$train_cmd" \
    --use-graphs true data/train data/lang exp/tri2 exp/tri2_ali
```

```
# From tri2 system, train tri3 which is LDA + MLLT + SAT.
steps/train_sat.sh --cmd "$train_cmd" \
```

```
$numLeavesSAT $numGaussSAT data/train data/lang exp/tri2_ali exp/tri3
```

```
utils/mkgraph.sh data/lang_test_bg exp/tri3 exp/tri3/graph
```

```
steps/decode_fmllr.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
exp/tri3/graph data/dev exp/tri3/decode_dev
```

```
steps/decode_fmllr.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
exp/tri3/graph data/test exp/tri3/decode_test
```

```
echo
```

这部分是在三音素模型的基础上做了
sgmm2，sgmm2 是 povey 提出的。

```
=====
```

```
echo "                                SGMM2 Training & Decoding                                "
```

```
echo
```

```
=====
```

```
steps/align_fmllr.sh --nj "$train_nj" --cmd "$train_cmd" \  
data/train data/lang exp/tri3 exp/tri3_ali
```

```
#exit 0 # From this point you can run DNN : local/run_dnn.sh
```

```
steps/train_ubm.sh --cmd "$train_cmd" \  
$numGaussUBM data/train data/lang exp/tri3_ali exp/ubm4
```

```
steps/train_sgmm2.sh --cmd "$train_cmd" $numLeavesSGMM $numGaussSGMM \  
data/train data/lang exp/tri3_ali exp/ubm4/final.ubm exp/sgmm2_4
```

```
utils/mkgraph.sh data/lang_test_bg exp/sgmm2_4 exp/sgmm2_4/graph
```

```
steps/decode_sgmm2.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
--transform-dir exp/tri3/decode_dev exp/sgmm2_4/graph data/dev \  
exp/sgmm2_4/decode_dev
```

```
steps/decode_sgmm2.sh --nj "$decode_nj" --cmd "$decode_cmd" \  
--transform-dir exp/tri3/decode_test exp/sgmm2_4/graph data/test \  
exp/sgmm2_4/decode_test
```

```
echo
```

```
=====
```

```
echo "                                MMI + SGMM2 Training & Decoding                                "
```

```
echo
```

```
=====
```

这部分是在三音素模型的基础上做了
sgmm2+MMI。

```
steps/align_sgmm2.sh --nj "$train_nj" --cmd "$train_cmd" \  
--transform-dir exp/tri3_ali --use-graphs true --use-gselect true data/train \
```

```

data/lang exp/sgmm2_4 exp/sgmm2_4_ali

steps/make_denlats_sgmm2.sh --nj "$train_nj" --sub-split "$train_nj" --cmd "$decode_cmd" \
--transform-dir exp/tri3_ali data/train data/lang exp/sgmm2_4_ali \
exp/sgmm2_4_denlats

steps/train_mmi_sgmm2.sh --cmd "$decode_cmd" \
--transform-dir exp/tri3_ali --boost 0.1 --drop-frames true \
data/train data/lang exp/sgmm2_4_ali exp/sgmm2_4_denlats \
exp/sgmm2_4_mmi_b0.1

for iter in 1 2 3 4; do
    steps/decode_sgmm2_rescore.sh --cmd "$decode_cmd" --iter $iter \
    --transform-dir exp/tri3/decode_dev data/lang_test_bg data/dev \
    exp/sgmm2_4/decode_dev exp/sgmm2_4_mmi_b0.1/decode_dev_it$iter

    steps/decode_sgmm2_rescore.sh --cmd "$decode_cmd" --iter $iter \
    --transform-dir exp/tri3/decode_test data/lang_test_bg data/test \
    exp/sgmm2_4/decode_test exp/sgmm2_4_mmi_b0.1/decode_test_it$iter
done

echo

=====
echo "
echo "
=====

# DNN hybrid system training parameters
dnn_mem_reqs="mem_free=1.0G,ram_free=0.2G"
dnn_extra_opts="--num_epochs 20 --num-epochs-extra 10 --add-layers-period 1 --shrink-interval
3"

steps/nnet2/train_tanh.sh --mix-up 5000 --initial-learning-rate 0.015 \
--final-learning-rate 0.002 --num-hidden-layers 2 \
--num-jobs-nnet "$train_nj" --cmd "$train_cmd" "${dnn_train_extra_opts[@]}" \
data/train data/lang exp/tri3_ali exp/tri4_nnet

[ ! -d exp/tri4_nnet/decode_dev ] && mkdir -p exp/tri4_nnet/decode_dev
decode_extra_opts=(--num-threads 6 --parallel-opts "-pe smp 6 -l mem_free=4G,ram_free=0.7G")
steps/nnet2/decode.sh --cmd "$decode_cmd" --nj "$decode_nj" "${decode_extra_opts[@]}" \
--transform-dir exp/tri3/decode_dev exp/tri3/graph data/dev \
exp/tri4_nnet/decode_dev | tee exp/tri4_nnet/decode_dev/decode.log

[ ! -d exp/tri4_nnet/decode_test ] && mkdir -p exp/tri4_nnet/decode_test
steps/nnet2/decode.sh --cmd "$decode_cmd" --nj "$decode_nj" "${decode_extra_opts[@]}" \
--transform-dir exp/tri3/decode_test exp/tri3/graph data/test \
exp/tri4_nnet/decode_test | tee exp/tri4_nnet/decode_test/decode.log

```

这部分是 povey 版本的 dnn 模型，不建议使用。

echo

echo " System Combination (DNN+SGMM)
echo

这部分是 dnn+sgmm 模型

for iter in 1 2 3 4; do
local/score_combine.sh --cmd "\$decode_cmd" \
data/dev data/lang_test_bg exp/tri4_nnet/decode_dev \
exp/sgmm2_4_mmi_b0.1/decode_dev_it\$iter exp/combine_2/decode_dev_it\$iter

local/score_combine.sh --cmd "\$decode_cmd" \
data/test data/lang_test_bg exp/tri4_nnet/decode_test \
exp/sgmm2_4_mmi_b0.1/decode_test_it\$iter exp/combine_2/decode_test_it\$iter
done

echo

echo " DNN Hybrid Training & Decoding (Karel's recipe)
echo

这部分是 karel 的 dnn 模型。
通用的深度学习模型框架。

local/run_dnn.sh

echo

echo " Getting Results [see RESULTS file]
echo

这部分是得到上述模型的
最后的识别结果

bash RESULTS dev
bash RESULTS test
echo

echo "Finished successfully on" `date`
echo

exit 0

此外，有人修改这个 dnn 的脚本，具体修改和详细的请看：
<http://blog.csdn.net/sytxsybm/article/details/18556403>，我会在附录5.1中给出。

3.2 yesno 的例子

1. 把 waves_yesno.zip.gz 复制到 yesno/s3 目录下，然后使用
`sudo yumzip waves_yesno.zip.gz`
`tar -xvf waves_yesno.tar`
2. 运行 ./run.sh。

3.3 timit

- 1 去 run.sh 里修改 timit 的目录，修改为你自己 timit 数据所在的目录；
- 2 根据之前的修改 cmd.sh 和 path.sh；
- 3 运行 run.sh 即可。

3.4 rm

- 1 去 run.sh 里修改 rm 的目录，修改为你自己 rm 数据所在的目录；
- 2 根据之前的修改 cmd.sh 和 path.sh；
- 3 运行 run.sh 即可。

3.5 voxforge

1. 用 getdata.sh 收集数据，数据大小大约为 11G 左右，大家保证有足够多的硬盘空间；
2. 根据之前的修改 cmd.sh 和 path.sh；
3. 你可以顺利运行 run.sh 脚本。

此外，附录中有个别人跑 voxforge 的例子，需要的人可以去附录 5.2 看一下。

4 kaldi 主页上的翻译

kaldi 主页上的翻译的事情一直是我希望做的事情，希望期待更多的人参与。这里先列出部分重点内容，欢迎随时补充：

kaldi 翻译重点：

- 1.数据准备: http://kaldi.sourceforge.net/data_prep.html @V
- 2.特征提取: <http://kaldi.sourceforge.net/feat.html> @煮8戒
- 3.声学模型: <http://kaldi.sourceforge.net/model.html> @老那
- 4.解码部分: <http://kaldi.sourceforge.net/graph.html> @karinafl
- 5.FST 部分: http://kaldi.sourceforge.net/fst_algo.html @木鱼
- 6.解码在训练部分: http://kaldi.sourceforge.net/graph_recipe_train.html @燕子飞
- 7.解码在测试部分: http://kaldi.sourceforge.net/graph_recipe_test.html
- 8.深度学习在 kaldi 上的应用 (karel 的例子): <http://kaldi.sourceforge.net/dnn1.html> @wbgleam
- 9.关键词检索: <http://kaldi.sourceforge.net/kws.html>
- 10.在线识别: http://kaldi.sourceforge.net/online_programs.html @冒顿
- 11.kaldi 手册 : <http://kaldi.sourceforge.net/tutorial.html> (这个需要5个人)

2014.8.16:

再此说明下，如果您想翻译 kaldi 主页中的任何一篇，欢迎跟我联系。因为您翻译的过程了不仅仅锻炼了您自己的英译汉的说明，而且还深刻理解了您翻译的内容。开源需要每个人的努力，欢迎您的加入。

4.1 数据准备

译者: V (shiwei@sz.pku.edu.cn) 水平有限, 如有错误请多包涵。@wbglearn校对。

介绍

在运行完示例脚本后(见[Kaldi tutorial](#)), 你可能会想用自己的数据在Kaldi上跑一下。本节主要讲述如何准备相关数据。我们假设本页的读者使用的是最新版本的示例脚本(即在脚本目录下被命名为s5的那些, 例如egs/rm/s5)。另外, 除了阅读本页所述内容外, 你还可以查看脚本目录下的那些数据准备相关的脚本。(译者: 结合起来看更易理解。) 在顶层的run.sh脚本(例如egs/rm/s5/run.sh)中, 最前面的几行命令都是和数据准备相关的, 代表数据准备的不同步骤。子目录local/下的脚本都是和数据集相关的。例如, Resource Management (RM) 数据集相应的脚本就是local/rm_data_prep.sh。对RM数据集来说, 这几行数据准备的命令为:

```
local/rm_data_prep.sh /export/corpora5/LDC/LDC93S3A/rm_comp || exit 1;
utils/prepare_lang.sh data/local/dict 'SIL' data/local/lang data/lang || exit 1;
local/rm_prepare_grammar.sh || exit 1;
```

而对于 WSJ 来说, 命令为:

```
wsj0=/export/corpora5/LDC/LDC93S6B
wsj1=/export/corpora5/LDC/LDC94S13B
local/wsj_data_prep.sh $wsj0/??-{?,??}.? $wsj1/??-{?,??}.? || exit 1;
local/wsj_prepare_dict.sh || exit 1;
utils/prepare_lang.sh data/local/dict "<SPOKEN_NOISE>" data/local/lang_tmp data/lang
|| exit 1;

local/wsj_format_data.sh || exit 1;
```

在WSJ的示例脚本中, 上述命令之后还有一些训练语言模型的命令(根据标注重新训练语言模型, 而不是使用LDC提供的), 但是上述几条命令是最重要的。数据准备阶段的输出包含两部分。一部分与“数据”相关(保存在诸如data/train/之类的目录下), 另一部分则与“语言”相关(保存在诸如data/lang/之类的目录下)。“数据”部分与数据集的录音相关, 而“语言”部分则与语言本身更相关的内容, 例如发音字典、音素集合以及其他Kaldi需要的关于音素的额外信息。如果你想用已有的识别系统和语言模型对你的数据进行解码, 那么你只需要重写“数据”部分。

数据准备-- 数据部分.

举个数据准备阶段中的关于“数据”部分例子, 请查看任何一个示例脚本目录下的“data/train”目录(假设你已经运行过一遍这些脚本了)。注意: 目录名字“data/train”本身没有什么特别的。一些被命名为其他名字的目录, 如“data/eval2000”(为一个测试集建立的), 有几乎差不多的目录结构和文件格式(说“几乎”是因为在测试集的目录下, 可能含有“sgm”和“glm”文件, 用于slite评分)。我们以Switchboard数据为例, 对应脚本在egs/swbd/s5下

```
s5# ls data/train
```

```
cmvn.scp feats.scp reco2file_and_channel segments spk2utt text utt2spk wav.scp
```

不是所有的文件都同等重要。如果要设置简单点，分段（segmentation）信息是不必要的（即一个文件里只有一段发音），你只需要自己创建“utt2spk”、“text”和“wav.scp”，“segments”和“reco2file_and_channel”是可选的，根据实际需要决定是否创建。剩下的就都交给标准脚本。

下面我们会详细描述该目录下的这些文件。首先从那些需要你手动创建的文件开始。

需要手动创建的文件

文件“text”包含每段发音的标注。

```
s5# head -3 data/train/text
sw02001-A_000098-001156 HI UM YEAH I'D LIKE TO TALK ABOUT HOW YOU
DRESS FOR WORK AND
sw02001-A_001980-002131 UM-HUM
sw02001-A_002736-002893 AND IS
```

每行的第一项是发音编号（utterance-id），可以是任意的文本字符串，但是如果在你的设置中还包含说话人信息，你应该把说话人编号（speaker-id）作为发音编号的前缀。这对于音频文件的排序非常重要。发音编号后面跟着的是每段发音的标注。你不用保证这里出现的每一个字都出现在你的词汇表中。词汇表之外的词会被映射到data/lang/oov.txt中。注意：尽管在这个特别的例子中，我们用下划线分割了发音编号中的“说话人”和“发音”部分，但是通常用破折号（“-”）会更安全一点。这是因为破折号的ASCII值更小。有人向我指出说，如果使用下划线，并且说话人编号的长度不一，在某些特殊的情况下，如果使用标准“C”语言风格对字符串进行排序，说话人编号和对应的发音编号会被排成不同的顺序。另外一个很重要的文件是wav.scp。在Switchboard例子中，

```
s5# head -3 data/train/wav.scp
sw02001-A /home/dpovey/kaldi-trunk/tools/sph2pipe_v2.5/sph2pipe -f wav -p -c 1
/export/corpora3/LDC/LDC97S62/swb1/sw02001.sph |
sw02001-B /home/dpovey/kaldi-trunk/tools/sph2pipe_v2.5/sph2pipe -f wav -p -c 2
/export/corpora3/LDC/LDC97S62/swb1/sw02001.sph |
```

这个文件的格式是：

```
<recording-id> <extended-filename>
```

其中，“extended-filename”可能是一个实际的文件名，或者就像本例中所述那样，是一段提取wav格式文件的命令。extended-filename末尾的管道符号表明，整个命令应该被解释为一个管道。等我们会解释什么是“recording-id”，但是首先，我们需要指出，如果“segments”文件不存在，“wav.scp”每一行的第一项就是发音编号。

在Switchboard设置中，我们有“segments”文件，所以下面我们就讨论一下这个文件。

```
s5# head -3 data/train/segments
sw02001-A_000098-001156 sw02001-A 0.98 11.56
sw02001-A_001980-002131 sw02001-A 19.8 21.31
```

```
sw02001-A_002736-002893 sw02001-A 27.36 28.93
```

"segments"文件的格式是:

```
<utterance-id> <recording-id> <segment-begin> <segment-end>
```

其中, segment-begin和segment-end以秒为单位。它们指明了一段发音在一段录音中的时间偏移量。“recording-id”和在“wav.scp”中使用的是同一个标识字符串。再次声明一下, 这只是一个任意的标识字符串, 你可以随便指定。文件“reco2file_and_channel”只是在你用NIST的sclite工具对结果进行评分(计算错误率)的时候使用:

```
s5# head -3 data/train/reco2file_and_channel
sw02001-A sw02001 A
sw02001-B sw02001 B
sw02005-A sw02005 A
```

格式是:

```
<recording-id> <filename> <recording-side (A or B)>
```

filename通常是.sph文件的名字, 当然需要去掉sph这个后缀; 但是也可以是任何其他你在“stm”文件中使用的标识字符串。“录音方”(recording side)则是一个电话对话中两方通话(A或者B)的概念。如果不是两方通话, 那么为保险起见最好使用“A”。如果你并没有“stm”文件, 或者你根本不知道这些都是什么东西, 那么你可能就不需要reco2file_and_channel"文件。

最后一个需要你手动创建的文件是“utt2spk”。该文件指明某一段发音是哪一个说话人发出的。

```
s5# head -3 data/train/utt2spk
sw02001-A_000098-001156 2001-A
sw02001-A_001980-002131 2001-A
sw02001-A_002736-002893 2001-A
```

格式是:

```
<utterance-id> <speaker-id>
```

注意一点, 说话人编号并不需要与说话人实际的名字完全一致——只需要大概能够猜出来就行。在这种情况下, 我们假定每一个说话方(电话对话的每一方)对应一个说话人。这可能不完全正确——有时一个说话人会把电话交给另外一个说话人, 或者同一个说话人会在不同的对话中出现——但是上述假定对我们来说也足够用了。如果你完全没有关于说话人的信息, 你可以把发音编号当做说话人编号。那么对应的文件格式就变为<utterance-id>
<utterance-id>。

在一些示例脚本中还出现了另外一个文件, 它在Kaldi的识别系统的建立过程中只是偶尔使用。在Resource Management

(RM)设置中该文件是这样的:

```
s5# head -3 ../../rm/s5/data/train/spk2gender
```

```
adg0 f
ahh0 m
ajp0 m
```

这个文件根据说话人的性别，将每个说话人编号映射为“m”或者“f”。

上述所有文件都应该被排序。如果没有排序，你在运行脚本的时候就会出现错误。在 [The Table concept](#) 中我们解释了为什么需要这样。这与（Kaldi的）I/O框架有关，归根到底是因为排序后的文件可以在一些不支持 `fseek()` 的流中，例如，含有管道的命令，提供类似于随机存取查找的功能。许多Kaldi程序都会从其他Kaldi命令中读取多个管道流，读入各种不同类型的对象，然后对不同输入做一些类似于“合并然后排序”的处理。既然要合并排序，当然需要输入是经过排序的。小心确保你的shell环境变量`LC_ALL`定义为“C”。例如，在`bash`中，你需要这样做：

```
export LC_ALL=C
```

如果你不这样做，这些文件的排序方式会与C++排序字符串的方式不一样，Kaldi就会崩溃。这一点我已经再三强调过了！

如果你的数据中包含NIST提供的测试集，其中有“stm”和“glm”文件可以用作计算WER，那么你可以直接把这些文件拷贝到数据目录下，并分别命名为“stm”和“glm”。注意，我们把评分脚本`score.sh`（可以计算WER）放到`local/`下，这意味着该脚本是与数据集相关的。不是所有的示例设置下的评分脚本都能识别`stm`和`glm`文件。能够使用这些文件的一个例子在Switchboard设置里，即`i.e.egs/swbd/s5/local/score_sclite.sh`。如果检测到你有`stm`和`glm`文件该脚本会被顶层的评分脚本`egs/swbd/s5/local/score.sh`调用。

不需要手动创建的文件

数据目录下的其他文件可以由前述你提供的文件所生成。你可以用如下的一条命令创建“`spk2utt`”文件（这是一条从`egs/rm/s5/local/rm_data_prep.sh`中摘取的命令）：

```
utils/utt2spk_to_spk2utt.pl data/train/utt2spk > data/train/spk2utt
```

这是因为 `utt2spk` 和 `spk2utt` 文件中包含的信息是一样的。`spk2utt` 文件的格式是：
`<speaker-id> <utterance-id1> <utterance-id2> ...`

下面我们讲一讲 `feats.scp` 文件。

```
s5# head -3 data/train/feats.scp
sw02001-A_000098-001156
/home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/raw_mfcc_train.1.ark:24
sw02001-A_001980-002131
/home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/raw_mfcc_train.1.ark:54975
sw02001-A_002736-002893
/home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/raw_mfcc_train.1.ark:62762
```

这个文件指向已经提取好的特征——在这个例子中我们所使用的是 MFCC。`feats.scp` 文件的格式是：

```
<utterance-id> <extended-filename-of-features>
```

每一个特征文件保存的都是 Kaldi 格式的矩阵。在这个例子中，矩阵的维度是 13（译者注：即列数；行数则和你的文件长度有关，标准情况下帧长 20ms，帧移 10ms，所以一行特征数

据对应 10ms 的音频数据。但在 Kaldi 中，实际返回的帧数可能比你预想的小一点。例如，你的音频是 5.68s，返回的通常是 565 帧而不是 568。这和 Kaldi 中处理不足以分帧的剩余数据的方式有关，目前这种方式是兼容 HTK 的。你需要一直向下传递 window size 和 frame shift 的值，才能准确算出每一帧的 timing。这种方式在实际使用中也有不少问题。)。第一行的“extended filename”，

即 /home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/raw_mfcc_train.1.ark:24，意思是，打开存档（archive）文件

/home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/raw_mfcc_train.1.ark，fseek()定位到 24（字节），然后开始读数据。

feats.scp 文件由如下命令创建：

```
steps/make_mfcc.sh --nj 20 --cmd "$train_cmd" data/train exp/make_mfcc/train $mfccdir
```

该句被顶层的“run.sh”脚本调用。命令中一些 shell 变量的定义，请查阅对应 run.sh。

\$mfccdir 是.ark 文件将被写入的目录，由用户自定义。

data/train 下最后一个未讲到的文件是“cmvn.scp”。该文件包含了倒谱均值和方差归一化的统计量，以说话人编号为索引。每个统计量集合都是一个矩阵，在本例中是 2 乘以 14 维。在我们的例子中，有：

```
s5# head -3 data/train/cmvn.scp
2001-A /home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/cmvn_train.ark:7
2001-B /home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/cmvn_train.ark:253
2005-A /home/dpovey/kaldi-trunk/egs/swbd/s5/mfcc/cmvn_train.ark:499
```

与 feats.scp 不同，这个 scp 文件是以说话人编号为索引，而不是发音编号。该文件由如下的命令创建：

```
steps/compute_cmvn_stats.sh data/train exp/make_mfcc/train $mfccdir
```

(这个例句来自 egs/swbd/s5/run.sh)。

因为数据准备阶段的错误会影响后续脚本的运行，所以我们有一个脚本来判断数据目录的文件格式是否正确。运行如下的例子：

```
utils/validate_data_dir.sh data/train
```

你可能会发现下面这个命令也很有用：

```
utils/fix_data_dir.sh data/train
```

（当然可对任何数据目录使用该命令，而不只是 data/train）。该脚本会修复排序错误，并会移除那些被指明需要特征数据或标注，但是却找不到被需要的数据的那些发音（utterances）。

数据准备-- “lang” 目录

现在我们关注一下数据准备的“lang”这个目录。

```
s5# ls data/lang
L.fst  L_disambig.fst  oov.int  oov.txt  phones  phones.txt  topo  words.txt
```

除 data/lang，可能还有其他目录拥有相似的文件格式：例如有个目录被命名为“data/lang_test”，其中包含和 data/lang 完全一样的信息，但是要多一个 G.fst 文件。该文件

是一个 FST 形式的语言模型：

```
s5# ls data/lang_test
G.fst  L.fst  L_disambig.fst  oov.int  oov.txt  phones  phones.txt  topo  words.txt
```

注意，lang_test/ 由拷贝 lang/ 目录而来，并加入了 G.fst。每个这样的目录都似乎只包含为数不多的几个文件。但事实上不止如此，因为其中 phones 是一个目录而不是文件：

```
s5# ls data/lang/phones
context_indep.csl  disambig.txt          nonsilence.txt        roots.txt
silence.txt
context_indep.int  extra_questions.int  optional_silence.csl  sets.int
word_boundary.int
context_indep.txt  extra_questions.txt  optional_silence.int  sets.txt
word_boundary.txt
disambig.csl      nonsilence.csl       optional_silence.txt  silence.csl
```

phones 目录下有许多关于音素集的信息。同一类信息可能有三种不同的格式，分别以.csl、.int 和.txt 结尾。幸运的是，作为一个 Kaldi 用户，你没有必要去一一手动创建所有这些文件，因为我们有一个脚本“utils/prepare_lang.sh”能够根据更简单的输入为你创建所有这些文件。在讲述该脚本和所谓更简单的输入之前，有必要先解释一下“lang”目录下到底有些什么内容。之后我们将解释如何轻松创建该目录。如果用户不需要理解 Kaldi 是如何工作的，而是秉着快速建立识别系统的目的，那么可以跳过下面的 [Creating the "lang" directory](#) 这一节

“lang” 目录下的内容

首先是有文件 phones.txt 和 words.txt。这些都是符号表（symbol-table）文件，符合 OpenFst 的格式定义。其中每一行首先是一个文本项，接着是一个数字项：

```
s5# head -3 data/lang/phones.txt
<eps> 0
SIL 1
SIL_B 2
s5# head -3 data/lang/words.txt
<eps> 0
!SIL 1
-'S 2
```

在 Kaldi 中，这些文件被用于在这些音素符号的文本形式和数字形式之间进行转换。大多数情况下，只有脚本 utils/int2sym.pl、utils/sym2int.pl 和 OpenFst 中的程序 fstcompile 和 fstprint 会读取这些文件。

文件 L.fst 是 FST 形式的发音字典（L，见 <http://www.cs.nyu.edu/~mohri/pub/hbka.pdf> > ["Speech Recognition with Weighted Finite-State Transducers"](#) by Mohri, Pereira and Riley, in Springer Handbook on Speech Processing and Speech Communication, 2008），

其中，输入是音素，输出是词。文件 L_disambig.fst 也是发音字典，但是还包含了为消歧而引入的符号，诸如 #1、#2 之类，以及为自环（self-loop）而引入的 #0。#0 能让消歧符号“通过”（pass through）整个语法（译者注：n 元语法，即我们的语言模型。另外前面这句话我是在不知道该怎么翻译）。更多解释见 [Disambiguation symbols](#)。但是不管明白与否，你

其实不用自己手动去引入这些符号。

文件 `data/lang/oov.txt` 仅仅只有一行：

```
s5# cat data/lang/oov.txt
<UNK>
```

在训练过程中，所有词汇表以外的词都会被映射为这个词（译者注：UNK 即 `unknown`）。“<UNK>”本身并没有特殊的地方，也不一定非要用这个词。重要的是需要保证这个词的发音只包含一个被指定为“垃圾音素”（`garbage phone`）的音素。该音素会与各种口语噪声对齐。在我们的这个特别设置中，该音素被称为<SPN>，就是“`spoken noise`”的缩写：

```
s5# grep -w UNK data/local/dict/lexicon.txt
<UNK> SPN
```

文件 `oov.int` 包含<UNK>的整数形式（从 `words.txt` 中提取的），在本设置中是 221。你可能已经注意到了，在 `Resource Management` 设置中，`oov.txt` 里有一个静音词，在 `RM` 设置中被称为“`!SIL`”。在这种情况下，我们从词汇表中任意选一个词（放入 `oov.txt`）——因为训练集中没有 `oov` 词，所以选哪个都不起作用。

文件 `data/lang/topo` 则含有如下数据：

```
s5# cat data/lang/topo
<Topology>
<TopologyEntry>
<ForPhones>
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49
50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78
79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104
105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124
125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144
145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164
165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184
185 186 187 188
</ForPhones>
<State> 0 <PdfClass> 0 <Transition> 0 0.75 <Transition> 1 0.25 </State>
<State> 1 <PdfClass> 1 <Transition> 1 0.75 <Transition> 2 0.25 </State>
<State> 2 <PdfClass> 2 <Transition> 2 0.75 <Transition> 3 0.25 </State>
<State> 3 </State>
</TopologyEntry>
<TopologyEntry>
<ForPhones>
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
</ForPhones>
<State> 0 <PdfClass> 0 <Transition> 0 0.25 <Transition> 1 0.25 <Transition> 2 0.25
<Transition> 3 0.25 </State>
<State> 1 <PdfClass> 1 <Transition> 1 0.25 <Transition> 2 0.25 <Transition> 3 0.25
<Transition> 4 0.25 </State>
```

```

<State> 2 <PdfClass> 2 <Transition> 1 0.25 <Transition> 2 0.25 <Transition> 3 0.25
<Transition> 4 0.25 </State>
<State> 3 <PdfClass> 3 <Transition> 1 0.25 <Transition> 2 0.25 <Transition> 3 0.25
<Transition> 4 0.25 </State>
<State> 4 <PdfClass> 4 <Transition> 4 0.75 <Transition> 5 0.25 </State>
<State> 5 </State>
</TopologyEntry>
</Topology>

```

这个文件指明了我们所用 HMM 模型的拓扑结构。在这个例子中，一个“真正”的音素内含 3 个发射状态，呈标准的三状态从左到右拓扑结构——即“Bakis”模型。（发射状态即能“发射”特征矢量的状态，与之对应的就是那些“假”的仅用于连接其他状态的非发射状态）。音素 1 到 20 是各种静音和噪声。之所以会有这么多，是因为对词中的不同位置的同一音素进行了区分（word-position-dependency）。这种情况下，实际上这里的静音和噪声音素大多数都用不上。不考虑在词位话，应该只有 5 个代表静音和噪声的音素。所谓静音音素（silence phones）有更复杂的拓扑结构。每个静音音素都有一个起始发射状态和一个结束发射状态，中间还有另外三个发射状态。你不用手动创建 data/lang/topo。

data/lang/phones/下有一系列的文件，指明了音素集合的各种信息。这些文件大多数有三个不同版本：一个“.txt”形式，如：

```

s5# head -3 data/lang/phones/context_indep.txt
SIL
SIL_B
SIL_E

```

一个“.int”形式，如：

```

s5# head -3 data/lang/phones/context_indep.int
1
2
3

```

以及一个“.csl”形式，内含一个冒号分割的列表：

```

s5# cat data/lang/phones/context_indep.csl
1:2:3:4:5:6:7:8:9:10:11:12:13:14:15:16:17:18:19:20

```

三种形式的文件包含的是相同的信息，所以我们只关注人们更易阅读的“.txt”形式。文件“context_indep.txt”包含一个音素列表，用于建立文本无关的模型。也就是说，对这些音素，我们不会建立需要参考左右音素的上下文决策树。实际上，我们建立的是更小的决策树，只参考中心音素和 HMM 状态。这依赖于“roots.txt”，下面将会介绍到。关于决策树的更深入讨论见 [How decision trees are used in Kaldi](#)。

文件 context_indep.txt 包含所有音素，包括那些所谓“假”音素：例如静音（SIL），口语噪声（SPN），非口语噪声（NSN）和笑声（LAU）：

```

# cat data/lang/phones/context_indep.txt
SIL
SIL_B
SIL_E

```



```
SIL_I
SIL_S
SPN
SPN_B
SPN_E
SPN_I
SPN_S
NSN
NSN_B
NSN_E
NSN_I
NSN_S
LAU
LAU_B
LAU_E
LAU_I
LAU_S
```

因为考虑了词位，这些音素都有许多变体。不是所有的变体都会被实际使用。这里，`SIL` 代表静音词，会被插入到发音字典中（是一个词而不是一个词的一部分，可选的）；`SIL_B` 则代表一个静音音素，应该出现在一个词的开端（这种情况应该永不出现）；`SIL_I` 代表词内静音（也很少存在）；`SIL_E` 代表词末静音（不应该存在）；而 `SIL_S` 则表示一种被视为“单词”（singleton word）的静音，意指这个音素只对应一个词——当你的发音字典中有“静音词”且标注中有明确的静音段时会有用。

`silence.txt` 和 `onsilence.txt` 分别包含静音音素列表和非静音音素列表。这两个集合是互斥的，且如果合并在一起，应该是音素的总集。在本例中，`silence.txt` 与 `context_indep.txt` 的内容完全一致。我们说“非静音”音素，是指我们将要估计各种线性变换的音素。所谓线性变换是指全局变换，如 LDA 和 MLLT，以及说话人自适应变换，如 EMLLR。根据之前的实验，我们相信，加入静音对这些变换没有影响。我们的经验是，把噪声和发声噪声都列为“静音”音素，而其他传统的音素则是“非静音”音素。在 Kaldi 中我们还没有通过实验找到一个最佳的方法来这样做。

```
s5# head -3 data/lang/phones/silence.txt
SIL
SIL_B
SIL_E
s5# head -3 data/lang/phones/nonsilence.txt
IY_B
IY_E
IY_I
```

`disambig.txt` 包含一个“消歧符号”列表 (见 [Disambiguation symbols](#)):

```
s5# head -3 data/lang/phones/disambig.txt
```

```
#0
#1
#2
```

这些符号会出现在 `phones.txt` 中，被当做音素使用。

`optional_silence.txt` 只含有一个音素。该音素可在需要的时候出现在词之间：

```
s5# cat data/lang/phones/optional_silence.txt
SIL
```

（可选静音列表中的）音素出现在词之间的机制是，在发音字典的 FST 中，可选地让该音素出现在每个词的词尾（以及每段发音的段首）。该音素必须在 `phones/` 中指明而不是仅仅出现在 `L.fst` 中。这个原因比较复杂，这里就不讲了。

文件 `sets.txt` 包含一系列的音素集，在聚类音素时被分组（被当做同一个音素），以便建立文本相关问题集（在 Kaldi 中，建立决策树时使用自动生成的问题集，而不是具有语言语义的问题集）。本设置中，`sets.txt` 将每个音素的所有不同词位的变体组合为一行：

```
s5# head -3 data/lang/phones/sets.txt
SIL SIL_B SIL_E SIL_I SIL_S
SPN SPN_B SPN_E SPN_I SPN_S
NSN NSN_B NSN_E NSN_I NSN_S
```

文件 `extra_questions.txt` 包含那些自动产生的问题集之外的一些问题：

```
s5# cat data/lang/phones/extra_questions.txt
IY_B B_B D_B F_B G_B K_B SH_B L_B M_B N_B OW_B AA_B TH_B P_B OY_B
R_B UH_B AE_B S_B T_B AH_B V_B W_B Y_B Z_B CH_B AO_B DH_B UW_B
ZH_B EH_B AW_B AX_B EL_B AY_B EN_B HH_B ER_B IH_B JH_B EY_B NG_B
IY_E B_E D_E F_E G_E K_E SH_E L_E M_E N_E OW_E AA_E TH_E P_E OY_E
R_E UH_E AE_E S_E T_E AH_E V_E W_E Y_E Z_E CH_E AO_E DH_E UW_E
ZH_E EH_E AW_E AX_E EL_E AY_E EN_E HH_E ER_E IH_E JH_E EY_E NG_E
IY_I B_I D_I F_I G_I K_I SH_I L_I M_I N_I OW_I AA_I TH_I P_I OY_I R_I UH_I
AE_I S_I T_I AH_I V_I W_I Y_I Z_I CH_I AO_I DH_I UW_I ZH_I EH_I AW_I AX_I
EL_I AY_I EN_I HH_I ER_I IH_I JH_I EY_I NG_I
IY_S B_S D_S F_S G_S K_S SH_S L_S M_S N_S OW_S AA_S TH_S P_S OY_S R_S
UH_S AE_S S_S T_S AH_S V_S W_S Y_S Z_S CH_S AO_S DH_S UW_S ZH_S
EH_S AW_S AX_S EL_S AY_S EN_S HH_S ER_S IH_S JH_S EY_S NG_S
SIL SPN NSN LAU
SIL_B SPN_B NSN_B LAU_B
SIL_E SPN_E NSN_E LAU_E
SIL_I SPN_I NSN_I LAU_I
SIL_S SPN_S NSN_S LAU_S
```

你可以看到，所谓一个问题就是一组音素。前 4 个问题是关于普通音素的词位信息，后面五个则是关于“静音音素”的。“静音”音素也可能没有像 `_B` 这样的后缀，比如 `SIL`。这些可被作为发音字典中可选的静音词的表示，即不会出现在某个词中，而是单独成词。在具有

语调和语气的设置中，extra_questions.txt 可以包含与之相关的问题集。

word_boundary.txt 解释了这些音素与词位的关联情况：

```
s5# head data/lang/phones/word_boundary.txt
SIL nonword
SIL_B begin
SIL_E end
SIL_I internal
SIL_S singleton
SPN nonword
SPN_B begin
```

这和音素中的后缀（_B 等等）是相同的信息，但我们并不想给音素的文本形式附加这样的强制限制——记住一件事，Kaldi 的可执行程序从不使用音素的文本形式，而是整数形式。

（译者注：即 Kaldi 内部使用的都是音素的整数标号来表示和传递音素）。所以我们使用文件 word_boundary.txt 来指明各音素与词位间的对应关系。建立这种对应关系的原因是因为我们需要这些信息从音素网格中恢复词的边界（例如，lattice-align-words 需要读取 word_boundary.txt 的整数版本 word_boundary.int）。找出词的边界是有用的，其中之一是用作 NIST 的 sclite 评分，该工具需要词的时间标记。还有其他的后续处理需要这些信息。

roots.txt 文件包含如何建立音素上下文决策树的信息：

```
head data/lang/phones/roots.txt
shared split SIL SIL_B SIL_E SIL_I SIL_S
shared split SPN SPN_B SPN_E SPN_I SPN_S
shared split NSN NSN_B NSN_E NSN_I NSN_S
shared split LAU LAU_B LAU_E LAU_I LAU_S
...
shared split B_B B_E B_I B_S
```

暂时你可以忽略”shared “和” split “——这些与我们建立决策树时的具体选项有关(更多信息见 [How decision trees are used in Kaldi](#))。像 SIL SIL_B SIL_E SIL_I SIL_S 这样，几个音素出现在同一行的意义是，在决策树中它们都有同一个“共享根”(shared root)，因此状态可在这些音素间共享。对于带语气语调的系统，通常所有与语气和语调相关的音素变体都会出现在同一行。此外，一个 HMM 中的 3 个状态（对静音来说有 5 个状态）共享一个根，且决策树的建立过程需要知道状态（的共享情况）。HMM 状态间共享决策树根节点，这就是 roots 文件中“shared”代表的意思。

建立"lang"目录

data/lang/目录下有很多不同的文件，所以我们提供了一个脚本为你创建这个目录，你只需要提供一些相对简单的输入信息：

```
utils/prepare_lang.sh data/local/dict "<UNK>" data/local/lang data/lang
```

这里，输入目录是 data/local/dict/，<UNK>需要在字典中，是标注中所有 OOV 词的映射词（映射情况会写入 data/lang/oov.txt 中）。data/local/lang/ 只是脚本使用的一个临时目录，data/lang/才是输出文件将会写入的地方。

作为数据准备者，你需要做的事就是创建 data/local/dict/这个目录。该目录包含以下信息：

```
s5# ls data/local/dict
extra_questions.txt  lexicon.txt  nonsilence_phones.txt  optional_silence.txt
silence_phones.txt
```

（实际上还有一些文件我们没有列出来，但那都是在创建目录时所遗留下的临时文件，可以忽略）。下面的这些命令可以让你知道这些文件中大概都有些什么：

```
s5# head -3 data/local/dict/nonsilence_phones.txt
IY
B
D
s5# cat data/local/dict/silence_phones.txt
SIL
SPN
NSN
LAU
s5# cat data/local/dict/extra_questions.txt
s5# head -5 data/local/dict/lexicon.txt
!SIL SIL
-'S S
-'S Z
-'T K UH D EN T
-1K W AH N K EY
```

正如你看到的，本设置（Switchboard）中，这个目录下的内容都非常简单。我们只是分别列出了“真正”的音素和“静音”音素，一个叫 `extra_questions.txt` 的空文件，以及一个有如下格式的 `lexicon.txt`：

```
<word> <phone1> <phone2> ...
```

注意：`lexicon.txt` 中，如果一个词有不同发音，则会在不同行中出现多次。如果你想使用发音概率，你需要建立 `exiconp.txt` 而不是 `lexicon.txt`。`lexiconp.txt` 中第二域就是概率值。注意，一个通常的作法是，对发音概率进行归一化，使最大的那个概率值为 1，而不是使同一个词的所有发音概率加起来等于 1。这样可能会得到更好的结果。如果想在顶层脚本中找一个与发音概率相关的脚本，请在 `egs/wsj/s5/run.sh` 目录下搜索 `pp`。

需要注意的是，在这些输入中，没有词位信息，即没有像 `_B` 和 `_E` 这样的后缀。这是因为脚本 `prepare_lang.sh` 会添加这些后缀。

从空的 `extra_questions.txt` 件中你会发现，可能还有些潜在的功能我们没有利用。这其中就包括重音和语调标记。对具有不同重音和语调的同一音素，你可能会想用不同的标记去表示。为展示如何这样做，我们看看在另外一个设置 `egs/wsj/s5` 的这些文件。结果如下：

```
s5# cat data/local/dict/silence_phones.txt
SIL
SPN
NSN
s5# head data/local/dict/nonsilence_phones.txt
```

```

S
UW UW0 UW1 UW2
T
N
K
Y
Z
AO AO0 AO1 AO2
AY AY0 AY1 AY2
SH
s5# head -6 data/local/dict/lexicon.txt
!SIL SIL
<SPOKEN_NOISE> SPN
<UNK> SPN
<NOISE> NSN
!EXCLAMATION-POINT EH2 K S K L AH0 M EY1 SH AH0 N P OY2 N T
"CLOSE-QUOTE K L OW1 Z K W OW1 T
s5# cat data/local/dict/extra_questions.txt
SIL SPN NSN
S UW T N K Y Z AO AY SH W NG EY B CH OY JH D ZH G UH F V ER AA IH M
DH L AH P OW AW HH AE R TH IY EH
UW1 AO1 AY1 EY1 OY1 UH1 ER1 AA1 IH1 AH1 OW1 AW1 AE1 IY1 EH1
UW0 AO0 AY0 EY0 OY0 UH0 ER0 AA0 IH0 AH0 OW0 AW0 AE0 IY0 EH0
UW2 AO2 AY2 EY2 OY2 UH2 ER2 AA2 IH2 AH2 OW2 AW2 AE2 IY2 EH2
s5#

```

你可能已经注意到了，`nonsilence_phones.txt` 中的某些行，一行中有多个音素。这些是同一元音的与重音相关的不同表示。注意，在 CMU 版的字典中，每个音素有 4 种表示：例如，`UW UW0 UW1 UW2`。基于某些原因，其中一种表示没有数字后缀。行中音素的顺序没有关系。通常，我们建议将每个“真实音素”的不同形式都组织在单独的一行中。我们使用 CMU 字典中的重音标记。文件 `extra_questions.txt` 中只有一个问题包含所有的“静音”音素（实际上这是不必要的，只是脚本 `prepare_lang.sh` 会添加这么一个问题），以及一个涉及不同重音标记的问题。

这些问题对利用重音标记信息来说是必要的，因为在 `nonsilence_phones.txt` 中每个音素的不同重音表示都在同一行，这确保了他们在 `data/lang/phones/roots.txt` 和 `data/lang/phones/sets.txt` 也属同一行，这又反过来确保了它们共享同一个（决策）树根，并且不会有决策问题弄混它们。因此，我们需要提供一个特别的问题，能为决策树的建立过程提供一种区分音素的方法。注意：我们在 `sets.txt` 和 `roots.txt` 中将音素分组放在一起的原因是，这些同一音素的不同重音变体可能缺乏足够的去稳健地估计一个单独的决策树，或者是产生问题集时需要的聚类信息。像这样把它们组合在一起，我们可以确保当数据不足以对它们分别估计决策树时，这些变体能在决策树的建立过程中“聚集在一起”（stay together）。

写到这里我们需要提一点，脚本 `utils/prepare_lang.sh` 支持很多选项。下面是该脚本的用法，可让你们了解这些选项都有哪些：

```
usage: utils/prepare_lang.sh <dict-src-dir> <oov-dict-entry> <tmp-dir> <lang-dir>
e.g.: utils/prepare_lang.sh data/local/dict <SPOKEN_NOISE> data/local/lang data/lang
options:
    --num-sil-states <number of states>                # default: 5, #states in silence
models.
    --num-nonsil-states <number of states>              # default: 3, #states in
non-silence models.
    --position-dependent-phones (true|false)            # default: true; if true, use _B, _E,
_S & _I
                                                         # markers on phones to
indicate word-internal positions.
    --share-silence-phones (true|false)                 # default: false; if true, share pdfs
of
                                                         # all non-silence
phones.
    --sil-prob <probability of silence>                 # default: 0.5 [must have 0 <
silprob < 1]
```

一个可能的重要选项是`--share-silence-phones`。该选项默认是 `false`。如果该选项被设为 `true`，所有静音音素——如静音、发声噪声、噪声和笑声——的概率密度函数（PDF，高斯混合模型）都会共享，只有模型中的转移概率不同。现在还不清楚为什么这样做有用，但我们发现这对 IARPA 的 BABEL 项目中的广东话数据集非常有效。该数据集非常乱，其中有很长的未标注的部分，我们试着将其与一个特别标记的音素对齐。我们怀疑训练数据可能没能成功正确对齐，而且基于某些不明原因，将上述选项设置为 `true` 则改变了结果。

另外一个可能的重要选项是“`--sil-prob`”。通常，对这些选项我们所作的实验都不多，所以对具体如何设置也不能给出非常详细的建议。

创建语言模型或者语法文件

前面的关于如何创建 `lang/` 目录的教程没有涉及如何产生 `G.fst` 文件。该文件是语言模型——或者称为语法——的有限状态转换器格式的表示，我们解码时需要它。实际上，在一些设置中，为做不同的测试，我们可能会有许多“`lang`”目录。这些目录中有不同的语言模型和字典。以华尔街日报（WSJ）的设置为例：

```
s5# echo data/lang*
data/lang data/lang_test_bd_fg data/lang_test_bd_tg data/lang_test_bd_tgpr
data/lang_test_bg \
    data/lang_test_bg_5k data/lang_test_tg data/lang_test_tg_5k data/lang_test_tgpr
data/lang_test_tgpr_5k
```

根据我们使用的语言模型的不同——是统计语言模型还是别的种类的语法形式——生成 `G.fst` 的步骤会不同。在 RM 设置中，使用的是二元语法，只允许某些词对。我们将总概率值 1 分配给所有向外的弧，以确保每个语法状态的概率和为 1。在 `local/rm_data_prep.sh` 中有这样一句代码：


```
local/make_rm_lm.pl $RMROOT/rm1_audio1/rm1/doc/wp_gram.txt > $tmpdir/G.txt ||
exit 1;
```

脚本 local/make_rm_lm.pl 会建立一个 FST 格式的语法文件（文本格式，不是二进制格式）。该文件包含如下形式的行：

```
s5# head data/local/tmp/G.txt
0 1 ADD ADD 5.19849703126583
0 2 AJAX+S AJAX+S 5.19849703126583
0 3 APALACHICOLA+S APALACHICOLA+S 5.19849703126583
```

到 www.openfst.org 上查阅更多关于 OpenFst 的信息（他们有一个很详细的教程）。脚本 local/rm_prepare_grammar.sh 会将文本格式的语法文件转换为二进制文件 G.fst。所用命令如下：

```
fstcompile --isymbols=data/lang/words.txt --osymbols=data/lang/words.txt
--keep_isymbols=false \
--keep_osymbols=false $tmpdir/G.txt > data/lang/G.fst
```

如果你要建立自己的语法文件，你也应做类似的事。

注意：这种过程只适用于一类语法：用上述方法你不能创建上下文无关的语法，因为这类语法不能被表示为 OpenFst 格式。在 WFST 框架下还是有办法这么做（见 Mike Riley 最近关于 push down transducers 的研究工作），但是在 Kaldi 中我们还没实现这些功能。

在 WSJ 设置中，我们使用了一个统计语言模型。脚本 local/wsj_format_data.sh 将 WSJ 数据库提供的 ARPA 格式的语言模型文件转换为 OpenFst 格式的。脚本中关键的命令如下：

```
gunzip -c $lmdir/lm_${lm_suffix}.arpa.gz | \
utils/find_arpa_oovs.pl $test/words.txt > $tmpdir/oovs_${lm_suffix}.txt
...
gunzip -c $lmdir/lm_${lm_suffix}.arpa.gz | \
grep -v '<s> <s>' | \
grep -v '</s> <s>' | \
grep -v '</s> </s>' | \
arpa2fst - | fstprint | \
utils/remove_oovs.pl $tmpdir/oovs_${lm_suffix}.txt | \
utils/eps2disambig.pl | utils/s2eps.pl | fstcompile --isymbols=$test/words.txt \
--osymbols=$test/words.txt --keep_isymbols=false --keep_osymbols=false | \
fstmrmspace > $test/G.fst
```

这里，变量 \$test 的值形如 data/lang_test_tg。最重要的一条命令是 arpa2fst，这是一个 Kaldi 程序。该程序将 ARPA 格式的语言模型转换为一个加权有限状态转换器（实际上是一个接收器）。grep 命令移除语言模型中“不可用”的 N 元语法。程序 remove_oovs.pl 移除包含集外词的 N 元语法（如果不移除会引起 fstcompile 崩溃）。eps2disambig.pl 将回退弧上的 <eps>（ ϵ ）符号转换为一个特殊的符号 #0，以保证语法文件是确定的（determinizable），见 [Disambiguation symbols](#)。如果你不知道“回退弧”是什么，你可以参考关于回退 N 元语法的文献，例如 Goodman 的“A bit of progress in language modeling”，以及我们前面引用的 Mohri 的论文。命令 s2eps.pl 将句首和句末符号 <s> 和 </s> 转换为 epsilon（<eps>），意即“没有符号”。fstcompile 是一个 OpenFst

命令，可将文本形式的 FST 转换为二进制形式的。 `fstrmepsilon` 也是一个 OpenFst 命令，可将 FST 中由 `<s>` 和 `</s>` 替换而来的少量的 `<eps>` 符号移除掉。

版权所有 翻版必究

4.2 特征提取

本翻译原文 <http://kaldi.sourceforge.net/feat.html>，由@煮八戒翻译，@wbglearn 校对和修改。

简介

我们做特征提取和波形读取的这部分代码，其目的是为了得到标准的 MFCC(译注：梅尔倒谱系数)和 PLP(译注：感知线性预测系数)特征，设置合理的默认值但留了一部分用户最有可能想调整的选项（如梅尔滤波器的个数，最小和最大截止频率等等）。这部分代码只读取 wav 文件里的 pcm(译注：脉冲编码调制)数据。这类文件通常带.wav 或.pcm 后缀（虽然有时.pcm 后缀会用于 sph 文件；这种情况下必须转换该文件）。假如源数据不是 wav 类文件，则用户可自由选择命令行工具来转换，而我们提供的 sph2pipe 工具已能满足一般的情况。

命令行工具 `compute-mfcc-feats` 和 `compute-plp-feats` 计算特征；同其它 Kaldi 工具一样，不带参数地运行它们会给出一个选项列表。例子脚本里显示了这些工具的用法。

计算 MFCC 特征

这里我们介绍如何使用命令行工具 `compute-mfcc-feats` 计算 MFCC 参数。该程序需要两个命令行参数：`rspecifier` 是用来读.wav 数据（以发音为索引）和 `wspecifier` 是用来写特征（以发音为索引）；参见 [The Table concept](#) 和 [Specifying Table formats: wspeifiers and rspeifiers](#) 获取更多关于这些术语的解释。典型的用法是，将数据写入一个大的“archive”文件，也写到一个“scp”文件以便随机存取；参见 [Writing an archive and a script file simultaneously](#) 解释。程序没有添加增量功能（如需添加，参见 `add-deltas`）。它接收选项 `-channel` 来选择通道（如 `-channel=0`, `-channel=1`），该选项在读取立体声数据时很有用。

计算 MFCC 特征由 `Mfcc` 类型的对象完成，它有 `Compute()` 函数可以根据波形计算特征。

完整的 MFCC 计算如下所示：

- 计算出一个文件中帧的数目（通常帧长 25ms 帧移 10ms）。
- 对每一帧：
 - 提取数据，可选做 **dithering**（注：直译为抖动，在这里可以理解为类似归一化的预处理），预加重和去除直流偏移，还可以和加窗函数相乘（此处支持多种选项，如汉明窗）
 - 计算该点能量（假如用对数能量则没有 C0）
 - 做 FFT（译注：快速傅里叶变换）并计算功率谱
 - 计算每个梅尔滤波器的能量；如 23 个部分重叠的三角滤波器，其中心在梅尔频域等间距
 - 计算对数能量并作余弦变换，根据要求保留系数（如 13 个）
 - 选做倒谱变换；它仅仅是比例变换，确保系数在合理范围。

上下截止频率根据三角滤波器界定，由选项`-low-freq` 和`-high-freq` 控制，通常分别设置为 0Hz 和奈奎斯特频率附近，如对 16kHz 采样的语音设置为`-low-freq=20` 和 `-high-freq=7800`。

Kaldi 的特征和 HTK 的特征在很多方面不同，但是几乎所有这些不同归结于有不同的默认值。用选项`-htk-compat=true` 并正确设置参数，能得到同 HTK 非常接近的特征。一个可能重要的选项是我们不支持能量最大归一化。这是因为我们希望能把无状态方式应用到归一化方法，且希望从原理上计算一帧帧特征仍能给出相同结果。但是程序 `compute-mfcc-feats` 里有`-subtract-mean` 选项来提取特征的均值。对每个语音做此操作；每个说话人可以有不同的方式来提取特征均值。（如在脚本里搜“`cmvn`”，表示倒谱均值和方差归一化）。

计算 PLP 特征

计算 PLP 特征的算法与 MFCC 的算法前期是一样的。稍后我们也许会在此部分增加些内容，但目前参见 Hynek Hermansky 《语音的感知线性预测 (PLP) 分析》，*Journal of the Acoustical Society of America*, vol. 87, no. 4, pages 1738–1752 (1990).

特征级声道长度归一化 (VTLN)

程序 `compute-mfcc-feats` 和 `compute-plp-feats` 接收一个 VTLN 弯折因子选项。在目前的脚本中，这仅用作线性版的 VTLN 的初始化线性转换的一种方法。VTLN 通过移动三角频率箱的中心频率的位置来实现。移动频率箱的弯折函数是一个在频域空间分段线性的函数。为理解它，记住以下数量关系：

$$0 \leq \text{low-freq} \leq \text{vtln-low} < \text{vtln-high} < \text{high-freq} \leq \text{nyquist}$$

此处，`low-freq` 和 `high-freq` 分别是用于标准 MFCC 或 PLP 计算的最低和最高频率（忽略更低和更高的频率）。`vtln-low` 和 `vtln-high` 是用于 VTLN 的截止频率，它们的功能是确保所有梅尔滤波器有合适的宽度。

我们实现的 VTLN 弯折函数是一个分段线性函数，三个部分映射区间`[low-freq, high-freq]`至`[low-freq, high-freq]`。记弯折函数为 $W(f)$ ， f 是频率。中段映射 f 到 f/scale ，`scale` 是 VTLN 弯折因子（通常范围为 0.8 到 1.2）。 x 轴上低段和中段的连接点是满足 $\min(f, W(f)) = \text{vtln-low}$ 的 f 点。 x 轴上中段和高段的连接点是满足 $\max(f, W(f)) = \text{vtln-high}$ 的 f 点。要求低段和高段的斜率和偏移是连续的且 $W(\text{low-freq}) = \text{low-freq}$ ， $W(\text{high-freq}) = \text{high-freq}$ 。这个弯折函数和 HTK 的不同；HTK 的版本中，“`vtln-low`”和“`vtln-high`”的数量关系是 x 轴上可以不连续的点，这意味着变量“`vtln-high`”必须基于弯折因子的可能范围的先验知识谨慎选择（否则梅尔滤波器可能为空）。

一个合理的设置如下（以 16kHz 采样的语音为例）；注意这反映的是我们理解的合理值，并非任何非常细致的调试实验的结果。

low-freq	vtln-low	vtln-high	high-freq	nyquist
40	60	7200	7800	8000

4.3 声学建模代码

（翻译：那兴宇）

简介

我们先来简单说说 Kaldi 中用于建模的代码的思路，以及为什么这样设计。我们的目标是使 Kaldi 支持传统模型，也就是对角阵高斯混合模型（GMM）和子空间高斯混合模型（SGMM），同时也要便于扩展到新的模型。在上一轮代码设计过程中，我们设计了一个虚基类，从其中衍生出 GMM 类和 SGMM 类，然后编写了一个命令行工具用于操作这两种模型。但是根据我们的经验，基类并不像事先想象的那样能派上大用场，因为这两种模型之间的差异太大（例如，它们支持的自适应算法不同）。为了构建我们想象中具有“普适性”的代码，使它能够处理不同模型类别，我们不得不不断的扩展基类。最终，我们的命令行工具已经很难修改和维护了。

在重设计代码的时候，我们决定使用更“现代化”的软件工程方法，不为了增强普适性而过于强调类的层级结构，而是专注于创建简单的、可复用的组件。例如，我们的解码器代码是很通用的，因为它的需求很少，它只要我们从简单的基类 [DecodableInterface](#) 中 创建一个衍生的实例。这个实例操作起来类似于由一个句子的声学似然度组成的矩阵（参见 [Kaldi 工具中的解码器](#)）。同时，每个命令行工具更简单易用，例如 gmm-align 用于给定对角 GMM 获得一句话的状态级对齐结果。总的想法是，要实现某种新技术只要编写一个新的命令程序，而不是增加已有的命令程序的复杂度。

对角 GMM

[DiagGmm](#) 类表示一个对角协方差矩阵的高斯混合模型。一组 [DiagGmm](#) 的对象，由从 0 开始的“pdf-ids”索引组合在一起，组成声学模型。这个声学模型是由 [AmDiagGmm](#) 类实现的。尽管这个声学模型类的接口比单个 GMM 丰富，但你可以简单地把 [AmDiagGmm](#) 看作是 [DiagGmm](#) 的矢量。把声学模型表示为独立模型的集合，每个模型用于表示一个 pdf，我们可以想象，这不是对所有模型都适用的构建方法。例如，SGMM 就不能这样表示。而且，如果我们想实现不同状态之间 GMM 子成分的捆绑，就不能把每个 pdf 单独表示了。

（译者注：pdf，科技文献中写作 p.d.f.，是概率密度函数的缩写。）

独立 GMM

理论上，[DiagGmm](#) 类是一个简单的、被动的对象，负责存储高斯混合模型的参数，并且有一个用于计算似然度的成员函数。它完全不知道自己会被如何使用，它只提供获取其成员的接口。它不获取累计量（Accumulation），不更新参数（对于参数更新，参见 [MLEstimateDiagGmm](#) 类）。[DiagGmm](#) 类存储参数的方式是：方差的倒数，以及均值乘以方差倒数的结果。这意味着，通过简单的内积操作就可以计算似然度。与 HTK 的 gconst 不同，Kaldi 中的 gconst 是独立于均值的。

由于以这种方式调整高斯参数很不方便，我们提供了 [DiagGmmNormal](#) 类，用于以简单直观的方式存储参数。同时，我们提供了 [DiagGmm](#) 和 [DiagGmmNormal](#) 实例互相转换的函

数。大部分的参数更新代码作用于 [DiagGmmNormal](#) 类。

基于 GMM 的声学模型

[AmDiagGmm](#) 类 表示一组 [DiagGmm](#) 对象，由 [pdf-id](#) 索引。该类不表示 HMM-GMM 模型，仅仅是一组 GMM 的集合。将其与 HMM 组合在一起由其他代码负责，主要是负责拓扑结构和转移概率部分的代码以及负责整合解码图的代码（参见 [HMM 拓扑与转移建模](#)）。在这里需要指出的是，在声学模型中，我们没有只存储一个 [AmDiagGmm](#) 对象，而是同时保存了一个 [TransitionModel](#) 对象和一个 [AmDiagGmm](#) 对象。这是为了避免在存储器上写入太多的独立文件。这样做的原因是，高斯模型和转移概率的更新往往是同时进行的。这样做的初衷是，在创建其他类型的模型时，我们可以将一个 [TransitionModel](#) 和一个目标模型的对象保存在同一个文件中。这样，那些只需要读取转移模型的程序（如创建解码图）就可以直接读取模型文件，而不需要知道其中声学模型 的类型。

[AmDiagGmm](#) 类 是一个相对简单的对象，并不负责模型估计（参见 [AccumAmDiagGmm](#)）和变换矩阵估计等工作。在 Kaldi 中，有其他的代码来负责这些工作，参见 [Kaldi 中的特征空间与模型空间变换](#)。

满协方差 GMM

对于满协方差矩阵的 GMM，我们创建了 [FullGmm](#) 类，其作用与 [DiagGmm](#) 相似，区别只在与协方差矩阵的形式。这个类主要是用来在 SGMM 建模中训练满协方差的广义背景模型(UBM)。唯一用于满方差 GMM 的命令行工具是用于训练全局混合模型的，也就是 UBM。我们没有编写 [AmDiagGmm](#) 类 的满方差版本以及对应的命令行工具。不过要实现这个功能也不是难事。

子空间高斯混合模型（SGMM）

子空间高斯混合模型（SGMM）由 [AmSgmm](#) 类 实现。这个类从本质上实现了“The Subspace Gaussian Mixture Model – a Structured Model for Speech Recognition”（by D. Povey, Lukas Burget et. al, Computer Speech and Language, 2011.）这篇论文中的方法。[AmSgmm](#) 类 表示一个完整的 pdf 的集合。Kaldi 中没有用于表示一个 SGMM 的类（如用于 GMM 的 [DiagGmm](#)）。SGMM 参数的估计是由 [MleAmSgmmAccs](#) 类 和 [MleAmSgmmUpdater](#) 类 完成的。

对于如何训练一个基于 SGMM 系统的示例脚本，参见 [egs/rm/s1/steps/train_ubma.sh](#), [egs/rm/s1/steps/train_sgmma.sh](#) 和 [egs/rm/s1/steps/decode_sgmma.sh](#)。

4.4 解码在测试部分

Firstly, we cannot hope to introduce finite state transducers and how they are used in speech recognition.

首先，我们不详细介绍有限状态转换以及它在语音识别中的应用。

For that, see "[Speech Recognition with Weighted Finite-State Transducers](#)" by Mohri, Pereira and Riley (in Springer Handbook on Speech Processing and Speech Communication, 2008). The general approach we use is as described there, but some of the details, particularly with respect to how we handle disambiguation symbols and how we deal with weight-pushing, differ.

需要了解的，可以看"[Speech Recognition with Weighted Finite-State Transducers](#)" by Mohri, Pereira and Riley (in Springer Handbook on Speech Processing and Speech Communication, 2008). 我们要用的主要方法那里面有介绍，但是一些细节，尤其是像怎样控制混淆符，怎样处理 weight-pushing 可能与那篇论文中的不一样。

Overview of graph creation

The overall picture for decoding-graph creation is that we are constructing the graph $HCLG = H \circ C \circ L \circ G$. Here

整个解码网络都是围绕 $HCLG = H \circ C \circ L \circ G$ 的图来建立的。

- G is an acceptor (i.e. its input and output symbols are the same) that encodes the grammar or language model.
- G 是用来编码语言模型的接收器
- L is the lexicon; its output symbols are words and its input symbols are phones.
- L 是发声词典；输出是词，输入是音素。
- C represents the context-dependency: its output symbols are phones and its input symbols represent context-dependent phones, i.e. windows of N phones; see Phonetic context windows.
- C 表示上下文关系，输出是音素，输入是上下文相关的音素例如 N 个音素组成的窗。具体看 Phonetic context windows 里面的介绍。
- H contains the HMM definitions; its output symbols represent context-dependent phones and its input symbols are transitions-ids, which encode the pdf-id and other information (see Integer identifiers used by TransitionModel)
- H 包含 HMM 的定义；输出符表示上下文相关的音素，输入是 transitions-ids(转移 id)，transitions-ids 用来编码 pdf-id 或者其他信息(自转或向后转)，具体看(Integer identifiers used by TransitionModel)

This is the standard recipe. However, there are a lot of details to be filled in. We want to ensure that the output is determinized and minimized, and in order for HCLG to be

determinizable we have to insert disambiguation symbols. For details on the disambiguation symbols, see below Disambiguation symbols.

以上是标准方法，然而还有好多细节没有说明。我们想确保输出是确定化的和最小化的，为了让 HCLG 确定化，我们插入了混淆符。就混淆符更多的简介，看下面的 Disambiguation symbols。

We also want to ensure the HCLG is stochastic, as far as possible; in the conventional recipes, this is done (if at all) with the "push-weights" operation. Our approach to ensuring stochasticity is different, and is based on ensuring that no graph creation step "takes away" stochasticity; see Preserving stochasticity and testing it for details.

我们也想让 HCLG 尽可能的随机。传统的方法是通过"push-weights"操作来达到这种效果。我们确保随机性方法与传统的不同，是基于确保没有能够消除随机性的图构建步骤；具体看 Preserving stochasticity and testing it for details。

If we were to summarize our approach on one line (and one line can't capture all the details, obviously), the line would probably be as follows, where `asl` = "add-self-loops" and `rds` = "remove-disambiguation-symbols", and `H'` is `H` without the self-loops:

`HCLG = asl(min(rds(det(H' o min(det(C o min(det(L o G))))))))`

如果我们用一行公式来总结我们的方法（很明显一行不能覆盖所有的细节），那一行应该是下面这样，`asl` = "add-self-loops"，`rds` = "remove-disambiguation-symbols"，`H'` 是 `H` 不带自环的：

`HCLG = asl(min(rds(det(H' o min(det(C o min(det(L o G))))))))`

Weight-pushing is not part of the recipe; instead we aim to ensure that no stage of graph creation will stop the result from being stochastic, as long as `G` was stochastic. Of course, `G` will typically not be quite stochastic, because of the way Arpa language models with backoff are represented in FSTs, but at least our approach ensures that the non-stochasticity "stays put" and does not get worse than it was at the start; this approach avoids the danger of the "push-weights" operation failing or making things worse.

Weight-pushing 没有被我们采用。相反，我们的目标是只要 `G` 是随机的就要确保没有消除结果随机性的图的构建过程。当然，`G` 不会是绝对随机的，因为带补偿的 Arpa 语言模型被用 FSTs 表示，但是至少我们的方法确保不随机的部分不会增加，不会比开始变的更糟糕；这种方法避免了"push-weights"操作失败或者是变更坏的危险。

Disambiguation symbols

混淆符

Disambiguation symbols are the symbols `#1`, `#2`, `#3` and so on that are inserted at the end of phoneme sequences in the lexicon. When a phoneme sequence is a prefix of another phoneme sequence in the lexicon, or appears in more than one word, it needs to have one of these symbols added after it. These symbols are needed to ensure that the product `L o G` is determinizable. We also insert disambiguation symbols in two

more places. We have a symbol #0 on the backoff arcs in the language model G; this ensures that G is determinizable after removing epsilons (since our determinization algorithm does remove epsilons). We also have a symbol #-1 in place of epsilons that appear on the left of context FST C, at the beginning of the utterance (before we start outputting symbols). This is necessary to fix a rather subtle problem that happens when we have words with an empty phonetic representation (e.g. the beginning and end of sentence symbols and-).

混淆符是在被插在词典中音素序列末尾的类似 #1, #2, #3 的符号。在词典中当一个音素序列是另一个音素序列的前缀, 或者是这个音素序列出现在多个词中, 就需要把这些符号加在它的后面。这些符号用来确保 $L \circ G$ 输出时确定化的。我们也插入混淆符在两个其它的地方。我们把 #0 加在语言模型 G 的补偿弧上, 确保 G 是确定化的当删除 epsilons 后 (确定化的方法删除 epsilons)。我们也加入 #-1 在出现在 context FST C 左边的 epsilons 的地方, 这种情况在句子的开始。这对于解决当有的词是用空音素 ($\langle s \rangle$) 表示的问题是必要的。

We give the outline of how we would formally prove that the intermediate stages of graph compilation (e.g. LG, CLG, HCLG) are determinizable; this is important in ensuring that our recipe never fails. By determinizable, we mean determinizable after epsilon removal. The general setup is: first, we stipulate that G must be determinizable. This is why we need the #0 symbols (G is actually deterministic, hence determinizable). Then we want L to be such that for any determinizable G, $L \circ G$ is determinizable. [The same goes for C, with $L \circ G$ on the right instead of G]. There are a lot of details of the theory still to be fleshed out, but I believe it is sufficient for L to have two properties:

下面是关于怎样证明图编译的中间过程 (e.g. LG, CLG, HCLG) 是确定化的一个概述。这对于确保我们的方法永远不失败是很重要的。我们这里讲的确定化是删除 epsilon 后的确定化。主要步骤是: 1. 保证 G 必须是确定化的, 这就是为什么需要 #0 (这里 G 确实是确定化的)。然后我们想让 L 也是这样对于任何确定化的 G, 那么 $L \circ G$ 也是确定化的。[The same goes for C, with $L \circ G$ on the right instead of G]. 这里还有很多理论细节需要充实, 但是我认为对于 L 有以下两点属性就够了:

- must be functional
- 必须是函数形式的

equivalently: any input-sequence on L must induce a unique output-sequence

相当于: 对于任何的输入序列在 L 中必须有唯一的输出序列

- equivalently: for any linear acceptor A, $A \circ L$ is a linear transducer or empty.
- 相当于: 对于任何线性接受器 A, $A \circ L$ 是线性转换或是空。
- L has the twins property, i.e. there are no two states reachable with the same input-symbol sequence, that each have a cycle with the same input sequence but different weight or output sequence.

- L 具有双胞胎属性, i.e. there are no two states reachable with the same input-symbol sequence, that each have a cycle with the same input sequence but different weight or output sequence.

The same applies of course to the C transducer. We believe that the transducers as our scripts and programs currently create them have these properties.

这对 C transducer 同样适用。我们认为我们的脚本和代码创建的转换器具有这些属性。

The ContextFst object

The ContextFst object (C) is a dynamically created FST object that represents a transducer from context-dependent phones to context-independent phones. The purpose of this object is to avoid us having to enumerate all possible phones in context, which could be difficult when the context width (N) or the number of phones is larger than normal. The constructor ContextFst::ContextFst requires the context-width (N) and central-position (P) which are further explained in Phonetic context windows; their normal values corresponding to a triphone system are 3 and 1, respectively. It also requires the integer id of a special symbol, the "subsequential symbol" (referred to as '\$' in the reference given above), which the FST outputs N-P-1 times after all phones are seen (this ensures that the context FST is output-deterministic). In addition to this it requires lists of the integer id's of the phones and the disambiguation symbols. The vocabulary on the output side of the ContextFst consists of the set of phones and disambiguation symbols plus the subsequential symbol. The vocabulary on the input side is dynamically generated and (apart from epsilon, which is not used), corresponds to phones in context, disambiguation symbols, and a special symbol that we write as #-1 that takes the place of epsilon in the "traditional recipe", but which we treat as any other disambiguation symbol (it is necessary to ensure determinizability in the sense we prefer, i.e. with epsilon removal). The subsequential symbol '\$' has nothing corresponding to it on the input side (this is also the case in the traditional recipe). The symbol id's corresponding to the disambiguation symbols on the input side are not necessarily the same integer identifiers as used on the output side for the corresponding symbols. The ContextFst object has a function ILabelInfo() which returns a reference to an object of type std::vector < which enables the user to work out the "meaning" of each symbol on the input side. The correct interpretation of this object is described further in The ilabel_info object.

ContextFst 类的对象 (C) 是一个动态建立的 FST 对象用来表示一个从上下文相关的音素到上下文独立的音素的转换器。这个对象是用来避免我们罗列所有可能的音素在上下文中, 当上下文窗长 (N) 或音素数量比平常大很多时罗列所有音素的方法将会很困难。ContextFst::ContextFst 转换器需要上下文窗长 context-width (N) 和 中间音素位置 central-position (P) 这些将会在 Phonetic context windows 中有进一步的解释。对于一个三音素系统他们的常用值是 3 和 1,。特殊符号的整数 id 也是需要的, "subsequential symbol" (上文中提到的 \$), 当所有音素都被处

理完后 FST 输出 N-P-1 次\$(用来确保 context FST 的输出是确定化的)。除此之外也要整数 id 的音素和混淆符列表。ContextFst 输出端的词表包括音素, 和加上 subsequential symbol 的混淆符 disambiguation symbols。输入端的词表是自动生成的(除了 epsilon, 它未被使用), 相当于内容有上下文相关的音素, 混淆符, 以及一些特殊符号在传统方法中用 #-1 代替 epsilon, 然而这里我们把他们看成别的混淆符(这是用来确保确定化的, i.e. 删除 epsilon)。

subsequential symbol '\$'和输入端没有太多联系(对传统方法来说也一样)。输入端混淆符的 id 和输出端的相关符号的整数 id 不绝对一样。ContextFst 类对象有一个 ILabelInfo()的函数, 是用来对 `std::vector<std::vector<int32>>` 型对象返回引用, 使用户能求出输入端每个符号的 "meaning"。相关的解释可以查找 [The ilabel_info object](#)。

There is a special "Matcher" object called ContextMatcher which is intended to be used in composition algorithms involving the ContextFst (a Matcher is something that OpenFst's composition algorithm uses for arc lookup; for more explanation, see the OpenFst documentation;). The ContextMatcher makes the use of the ContextFst object more efficient, by avoiding the allocation of more states than is necessary (the issue is that with the normal matcher, every time we want any arc out of a state, we would have to allocate the destination states of all other arcs out of that state). There is an associated function, ComposeContextFst(), which performs FST composition in the case where the left hand argument to composition is of type ContextFst; it uses this matcher. There is also a function ComposeContext(), which is similar but creates the ContextFst object itself.

这是 ContextMatcher 类的特殊对象 "Matcher", 它是在组合算法中被使用 (Matcher 是组合算法用来做弧查找, 更多细节请看 the OpenFst documentation)。ContextMatcher 让 ContextFst 的对象使用更高效通过避免分配更多不必要的状态((the issue is that with the normal matcher, every time we want any arc out of a state, we would have to allocate the destination states of all other arcs out of that state)。这是一个相关函数, ComposeContextFst(), 是用来执行 FST 组合操作的当组合的左边参数是 ContextFst 型时。也有一个函数很类似 ComposeContext(), 但是用来创建 ContextFst 对象。

Avoiding weight pushing

We deal with the whole question of weight pushing in a slightly different way from the traditional recipe. Weight pushing in the log semiring can be helpful in speeding up search (weight pushing is an FST operation that ensures that the arc probabilities out of each state "sum to one" in the appropriate sense). However, in some cases weight pushing can have bad effects. The issue is that statistical language models, when represented as FSTs, generally "add up to more than one" because some words are counted twice (directly, and via backoff arcs).

我们处理权重前推的整个问题上和传统方法上有微小的差别。权重前推对于 log 半环在加速搜索是有效的(权重前推是一个 FST 操作确保每个状态出弧概率的和在合适的情况下相加为 1)。然

而在一些情况下 权重前推有负面影响。这个问题来自用 FSTs 表示统计语言模型，通常相加的结果大于 1，因为一些词被记了两次（直接的或通过回退弧）。

We decided to avoid ever pushing weights, so instead we handle the whole issue in a different way. Firstly, a definition: we call a "stochastic" FST one whose weights sum to one, and the reader can assume that we are talking about the log semiring here, not the tropical one, so that "sum to one" really means sum, and not "take the max". We ensure that each stage of graph creation has the property that if the previous stage was stochastic, then the next stage will be stochastic. That is: if G is stochastic, then LG will be stochastic; if LG is stochastic, then $\det(LG)$ will be stochastic; if $\det(LG)$ is stochastic, then $\min(\det(LG))$ will be stochastic, and so on and so forth. This means that each individual operation must, in the appropriate sense, "preserve stochasticity". Now, it would be quite possible to do this in a very trivial but not-very-useful way: for instance, we could just try the push-weights algorithm and if it seems to be failing because, say, the original G fst summed up to more than one, then we throw up our hands in horror and announce failure. This would not be very helpful.

我们决定避免可能的权重前推，所以使用了不同的方法来处理问题。1.定义"stochastic" FST，它的权重和为 1，读者可以假设这里只考虑 log 半环，而不是 tropical，所以"sum to one"就是加和的意思，不是取最大。要确保图创建的每一过程都具有那种属性：如果前一个过程是随机的，后一个过程也将是随机的。也就是说，如果 G 是随机的， LG 也是随机的；如果 LG 是随机的，那么 $\det(LG)$ 也是随机的；如果 $\det(LG)$ 是随机的，那么 $\min(\det(LG))$ 也是随机的，等等。意思就是，每个独立的操作必须在一定情况下保持随机性。现在通过一种琐碎但是很有用的方法可以实现：例如：we could just try the push-weights algorithm and if it seems to be failing because, say, the original G fst summed up to more than one, then we throw up our hands in horror and announce failure. This would not be very helpful.

We want to preserve stochasticity in a stronger sense, which is to say: first measure, for G , the min and max over all its states, of the sum of the (arc probabilities plus final-prob) out of those states. This is what our program "fstisstochastic" does. If G is stochastic, both of these numbers would be one (you would actually see zeros from the program, because actually we operate in log space; this is what "log semiring" means). We want to preserve stochasticity in the following sense: that this min and max never "get worse"; that is, they never get farther away from one. In fact, this is what naturally happens when we have algorithms that preserve stochasticity in a "local" way. There are various algorithms that we need to preserve stochasticity, including:

我们想保持随机性在某方面，也就是说首先，对于 G ，所有状态中最小和最大的，以及这些状态的出弧概率与最后的概率的和。这是我们程序"fstisstochastic"能实现的。如果 G 是随机的，所有的这些数字就都是 1（你也从程序中看到是 0，因为我们在 log 空间中执行的操作；这里是 log 半环）。我们想保持随机性在一下方面：当最小或最大从来不变的最坏；换句话说，他们从来不

会离 1 很远。事实上，这很自然会发生，当我们的算法用局部的方法维护随机性。还有好多算法需要保持随机性，包括：

- Minimization 最小化操作
- Determinization 确定化操作
- Epsilon removal 删除空
- Composition 组合操作(with particular FSTs on the left) There are also one or two minor algorithms that need to preserve stochasticity, like adding a subsequential-symbol loop. Minimization naturally preserves stochasticity, as long as we don't do any weight pushing as part of it (we use our program "fstminimizeencoded" which does minimization without weight pushing). Determinization preserves stochasticity as long as we do it in the same semiring that we want to preserve stochasticity in (this means the log semiring; this is why we use our program fstdeterminizestar with the option --determinize-in-log=true). Regarding epsilon removal: firstly, we have our own version of epsilon removal "RemoveEpsLocal()" (fstrmepslocal), which doesn't guarantee to remove all epsilons but does guarantee to never "blow up". This algorithm is unusual among FST algorithms in that, to to what we need it to do and preserve stochasticity, it needs to "keep track of" two semirings at the same time. That is, if it is to preserve equivalence in the tropical semiring and stochasticity in the log semiring, which is what we need in practice, it actually has to "know about" both semirings simultaneously. This seems to be an edge case where the "semiring" concept somewhat breaks down. Composition on the left with the lexicon L, the context FST C and the H transducer (which encodes the HMM structure) all have to preserve stochasticity. Let's discuss this the abstract: we ask, when composing $A \circ B$, what are sufficient properties that A must have so that $A \circ B$ will be stochastic whenever B is stochastic? We believe these properties are:
 - For any symbol sequence that can appear on the input of B, the inverse of A must accept that sequence (i.e. it must be possible for A to output that sequence), and:
 - For any such symbol sequence (say, S), if we compose A with an unweighted linear FST with S on its input, the result will be stochastic.

These properties are true of C, L and H, at least if everything is properly normalized (i.e. if the lexicon weights sum to one for any given word, and if the HMMs in H are properly normalized and we don't use a probability scale). However, in practice in our graph creation recipes we use a probability scale on the transition probabilities in the HMMs (similar to the acoustic scale). This means that the very last stages of graph creation typically don't preserve stochasticity. Also, if we are using a statistical language model,

G will typically not be stochastic in the first place. What we do in this case is we measure at the start how much it "deviates from stochasticity" (using the program `fstisstochastic`), and during subsequent graph creation stages (except for the very last one) we verify that the non-stochasticity does not "get worse" than it was at the beginning.

至少如果每个都能被恰当的归一的话(例如, 如果对任何给定的词, 词典权重之和都为 1, 如果 H 中的 HMMs 状态被恰当的归一, 而没用概率尺度表示) 这些属性对 C,L,H 也都是真实存在的。然而我们在建图的实际应用中用概率尺度表示 HMMs 的转移概率(类似声学尺度)。这就是说图构建的最后的过程不能保持随机性。如果我们用了一个随机的语言模型, G 首先不会是随机的。

还没翻译完, 有好多理论的东西没搞懂, 我再弄弄, 你先看下!

版权所有 翻版必究

4.5 Karel 的深度神经网络训练实现

Karel's DNN training implementation

综述

Karel Vesely 的 DNN 实现需要以下的技术：

- 基于 RBMS(受限波尔兹曼机)，每一层进行预训练
- 每一帧进行交叉熵训练
- 用格子框架通过 sMBR 准则（状态的最小贝叶斯风险），对序列的区分性训练

这个系统主要是建立在 LDA-MLLT-fMLLR 特征上（从辅助的 GMM 模型得到的）。整个 DNN 训练是在一个用 CUDA 的单 GPU 上运行的，但是 cuda 矩阵库是为了没有 GPU 的机器上运行的，但是似乎慢 10 倍。（备注：这个例子只支持单 GPU 或者单 CPU，而 povey 的例子支持多 GPU 或者多 CPU）

我们用 CUDA (4.2, 5.0, 5.5) 发布版本来运行和测试我们例子。

标准数据库(rm/wsj/swbd)的脚本在：`egs/{rm,wsj}/s5/local/run_dnn.sh` `egs/swbd/s5b/local/run_dnn.sh`。

DNN 训练阶段

Pre-training（预训练）

每一层的 RBM 预训练是参考 <http://www.cs.toronto.edu/~hinton/absps/guideTR.pdf>。由于 DNN 的输入是高斯特征，我们的第一个 RBM 用的是 Gaussian-Bernoulli 单元，这层学习率小，和更多的训练次数。接下来的 RBMs 模型都用 Bernoulli-Bernoulli 单元。我们是用马尔科夫链蒙特卡洛采样 (CD-1)组成的对比散度算法。在这些实验里，调整学习率是非常重要的。然而，为了更好的调整，我们需要做很多次的 fine-tuning 实验。在这个样例中我们使用 L2 正则化来增加 RBM 训练的混合率，在前 50h 的数据上动量项(momentum)是从 0.5 到 0.9 是逐步线性增加的。为了有效的保持学习速率不变，我们需要通过一个因子(1-mmt)来减小学习率。

在原始的 TIMIT 的实验里，我们有大约 200h 的数据，相对应我们的迭代次数是 50。因此我们在 switchboard 数据库上每层训练一次。根据我们跟 Toronto 里的人的讨论，我们拥有的数据越多，RBM 的预训练就越不重要。但是能让我们更好的初始化深度网络，使得反向传播算法能有一个好的初始化位置（点）。

在 RBM 训练中，我们在句子级别和帧级别上分别置乱来模仿从训练数据分布里提取样本，每一个 mini-batches 更新一次。当我们训练 Gaussian-Bernoulli 单元，RBM 对权重扩张非常敏感，特别是有很大的学习率和许多隐含单元。为了避免这样，我们在 min-batches 里比较训练数据和重建数据的变化。如果差别很大，权重将重新调整，学习率将暂时的减小。

Frame-level cross-entropy training

在这个阶段，我们把每一帧分成三音素状态来训练 DNN。这个是通过 mini-batch 随机梯度下降算法来进行，默认的学习率为 0.008, minibatch 的大小为 256;我们不使用动量项或者正则化。

学习率在最初的几次迭代中是保持不变的，当神经网络不再提高，我们在每次训练时将学习率减半，直到它再次停止提高。

就像在 RBMs 里，我们用句子级别和帧级别的置乱来模仿从训练数据分布中日趋样本，在训练工具里就做了帧级别的置乱，在 GPU 存储空间中。

一般所有的训练工具需要从 `-feature-transform` 接受参数，这个是进行临近特征变换的神经网络，这个通常在 GPU 里完成。在这个脚本里，我们用来把数据分块和归一化数据，但是通常是由像瓶颈特征提取这样的更多的复杂变换。

Sequence-discriminative training (sMBR)

这个阶段，神经网络被训练用来正确的对整个句子分类，这个是比较帧级别的训练更加贴近普通的 ASR 的目标。sMBR 的目标就是最大化我们期望的正确率，但需要从参考的 transcriptions 的状态标签获得。我们使用 lattice framework 来表示 competing hypothesis。

这个训练是通过对 per-utterance 进行随机梯度下降更新。我们使用更低的学习速率 $1e-5$ 不变跑 3-5 次。第一次训练后，重新生成 lattices 收敛会更快。

我们支持 MMI, BMMI, MPE and sMBR 训练。所有的方法使用非常相似，只有 sMBR 仅仅好一点点。在 sMBR 优化中，我们排除了静音帧。更多的细节 http://www.danielpovey.com/files/2013_interspeech_dnn.pdf。

代码

DNN 的代码在 `src/nnet`，工具在 `src/nnetbin`。因为所有的计算都通过 GPU，所有的代码都建立在 `src/cudamatrix` 基础上。

神经网络的表示

神经网络 `Nnet` 通过 `Component` 向量操作的。一个 `Component` 与神经网络的一层不是很准确的对应。对于每一层分别有 `Components` 对应映射变换和非线性变换。`Nnet` 最重要的思想是：

- `Nnet::Propagate`: 从输入传到输出
- `Nnet::Backpropagate`: 将目标函数的导数通过网络后向传播
- `Nnet::Feedforward`: 从输入传播到输出，用 2 个 flipping buffers 节省内存存储
- `Nnet::SetTrainOptions`: 设置训练的综合参数，学习率，动量项，L1,L2 代价。通过 `const` 引用，Components and buffers 是可以接受的。

请记住 DNN 是可以扩展的，我们定义了二个接口：

- [Component](#)：神经网络一层的接口，这个不需要包含可训练的参数（非线性性和权重是分开 Components）。
- [UpdatableComponent](#)：通过添加支持的可训练的参数扩展的 [Component](#)。
[Component](#) and [UpdatableComponent](#) 最重要的虚拟方法有：
- [Component::PropagateFnc](#)：在前向传播中计算数据变换
- [Component::BackpropagateFnc](#)：在后向传播中转化导数
- [UpdatableComponent::Update](#)：从导数和输入中计算梯度，然后更新

[Component](#) 的实现通过：

- [Sigmoid](#)：计算逻辑 sigmoid 变换和他的导数
- [Softmax](#)：计算 softmax 变换，假设导数已经是 wrt. 激活函数了 (BackpropagateFnc 复制原来的导数)

[UpdatableComponent](#) 的实现通过：

- [AffineTransform](#)：含有一个线性变换和一个偏置项
- [Rbm](#)：与 Restricted Boltzmann Machine 对应，包含它的参数和单元类型(当添加到 DBN，就转化 [AffineTransform](#) 和 [Sigmoid](#))

目标函数的表示

每一帧的目标函数是通过 classes [Mse](#) (均方误差), [Xent](#) (交叉熵) 来想实现的。他们的导数是通过 methods [Mse::Eval](#), [Xent::Eval](#) and [Xent::EvalVec](#) 计算的，每个 epoch 的目标值的统计性是累积的。

Frame-level shuffling

The frame-level shuffling 是通过 Cache 和 CacheTgtMat 完成的。在 Cache 里我们假设 targets 编码成向量的索引，在 CacheTgtMat 目标中是矩阵。

Extending network by a new component

用一个新的成分来扩展 NN 框架是非常简单的，所有的可以通过以下做到：

- 1) 定义新的入口：[Component::ComponentType](#)
- 2) 在表 [Component::kMarkerMap](#) 定义新的行
- 3) 添加动态的构造函数的调用到像工厂函数 [Component::Read](#)
- 4) 实现界面 [Component::](#) 或者 [UpdatableComponent::](#)

Binary tools

这些训练工作都在 src/nnetbin，最重要的工具是：

- nnet-train-xent-hardlab-frmsbuff.cc: 帧级别的训练，对后验概率和目标之前的交叉熵做最小化，允许在时间轴上做拼接，帧级别的置乱是在特征转换后完成，进行一次训练。学习率减半是在 `bash` 脚本上完成的。
- nnet-forward.cc: 这个工具通过神经网络传播输入特征，他可以去除运行中的 `softmax` 和基于三音素状态的数目来通过先验划分。这个工具常常用来特征提取中的解码部分，如果你运行在一个没有 GPU 的机器，一个 CPU 就被代替。
- rbm-train-cd1-frmsbuff.cc: 训练 RBM，可以通过逐步线性增加动量项来计算迭代的时候
- nnet-train-mmi-sequential.cc: 可以用来 MMI/bMMI 训练
- nnet-train-mpe-sequential.cc: 可以用来 MPE / sMBR 训练

版权所有 翻版必究

4.6 kaldi 中的关键词检索

(@刘诗涵翻译)

Keyword Search in Kaldi

Introduction

本部分主要对 kaldi 中关键字搜索模块的内容做相关介绍。我们会简要介绍在论文 "[Lattice indexing for spoken term detection](#)", [D. Can, M. Saraclar, Audio, Speech, and Language Processing](#) 中提出的关键字搜索算法, 但是这里会着重介绍搜索算法的具体实现过程以及我们对算法做的一些延伸。

Lattice Indexing

在此我们只是简单的介绍一下算法, 如果了解更多的细节, 你可以去查看相关网格索引 (lattice indexing) 的文章。

Related Semiring

索引算法是基于一种特殊设计的半环 (semiring), 此种半环可以同时存储时间和权重 (置信度) 的信息。强烈建议读者去仔细阅读理解半环的详细定义, 但是以防读者忘记相关概念, 我们在此对其进行简单回顾:

定义 1: 幺半群 (注 monoid: 幺半群是指一个带有可结合二元运算和单位元的代数结构) 是一个三元组 $(k, \otimes, \bar{1})$, 其中 \otimes 是在数据集 k 上的一个封闭的二元结合运算, $\bar{1}$ 是 \otimes 中的单位元。如果二元运算 \otimes 在数据集 k 是可交换的那么此幺半群也是可交换的。

定义 2: 半环 (semiring: 在抽象代数中, 半环是类似于环但没有加法逆元的代数结构) 是一个五元组 $(k, \oplus, \otimes, \bar{0}, \bar{1})$, 并且符合如下条件:

1. $(k, \otimes, \bar{1})$ 是一个幺半群;
2. $(k, \oplus, \bar{0})$ 是一个可交换幺半群;
3. \otimes 对 \oplus 符合分配率, 即

$$\forall a, b, c \in k, a \otimes (b \oplus c) = a \otimes b \oplus a \otimes c \text{ 且 } (a \oplus b) \otimes c = a \otimes c \oplus b \otimes c;$$

$$\forall a \in k, \bar{0} \otimes a = a \otimes \bar{0} = \bar{0}$$

下面介绍在算法中使用到的 5 种半环 (semiring):

对数半环 (Log semiring): 对数半环 (Log semiring) 的定义如下:

$$L = (R \cup \{-\infty, +\infty\}, \oplus_{\log}, +, +\infty, 0),$$

其中 $\forall a, b \in \mathbb{R} \cup \{-\infty, +\infty\}$;

$a \oplus_{\log} b = -\log(e^{-a} + e^{-b})$, 而 $e^{-\infty} = 0$, $-\log(0) = \infty$ 。

热带半环 (Tropical semiring): 此种半环的格式为

$$T = (\mathbb{R} \cup \{-\infty, +\infty\}, \min, +, +\infty, 0),$$

其中操作符 \min (得到最小值) 是用来做 tropical 加法 (tropical addition) 的。

北极半环 (Arctic semiring): 此种半环的格式为

$$T = (\mathbb{R} \cup \{-\infty, +\infty\}, \max, +, +\infty, 0),$$

其中操作符 \max (得到最大值) 是用来做 arctic 加法 (Arctic addition)。因为其操作正好与热带半环 (tropical semiring) 相反, 因此得名北极半环 (arctic semiring)。

$\forall a_1, a_2 \in k_A, b_1, b_2 \in k_B$, 则 $(a_1, b_1) \oplus_{\times} (a_2, b_2) = (a_1 \oplus_{k_A} a_2, b_1 \oplus_{k_B} b_2)$

产品半环 (Product semiring): 此种半环的格式为:

$$A \times B = (k_A \times k_B, \oplus_{\times}, \otimes_{\times}, \bar{0}_{k_A} \times \bar{0}_{k_B}, \bar{1}_{k_A} \times \bar{1}_{k_B}),$$

其中 $A = (k_A, \oplus_{k_A}, \otimes_{k_A}, \bar{0}_{k_A}, \bar{1}_{k_A})$ 和 $B = (k_B, \oplus_{k_B}, \otimes_{k_B}, \bar{0}_{k_B}, \bar{1}_{k_B})$ 是偏序半环; \oplus_{\times} 和 \otimes_{\times} 式能量算子符号 (component-wise operators), 例如:

词典半环 (Lexicographic semiring): 词典半环的定义为:

$$A * B = (k_A \times k_B, \oplus_*, \otimes_*, \bar{0}_{k_A} \times \bar{0}_{k_B}, \bar{1}_{k_A} \times \bar{1}_{k_B}),$$

其中 $A = (k_A, \oplus_{k_A}, \otimes_{k_A}, \bar{0}_{k_A}, \bar{1}_{k_A})$ 和 $B = (k_B, \oplus_{k_B}, \otimes_{k_B}, \bar{0}_{k_B}, \bar{1}_{k_B})$ 是偏序半环; \oplus_* 是字典算

符优先符号, 例如: $\forall a_1, a_2 \in k_A, b_1, b_2 \in k_B$, 则

$$(a_1, b_1) \oplus (a_2, b_2) = \begin{cases} (a_1, b_1 \oplus_{k_B} b_2), & \text{if } a_1 = a_2 \\ (a_1, b_1), & \text{if } a_1 = (a_1 \oplus_{k_A} a_2) \neq a_2 \\ (a_2, b_2), & \text{if } a_1 \neq (a_1 \oplus_{k_A} a_2) = a_2 \end{cases}$$

Algorithm

索引算法是把由 ASR 过程中生成的网格 (Lattice) 转变为加权系数转换器 (weighted factor transducer), 这样就可以通过结合结果转换器 (resulting transducer) 与经过合理定义的关键词转换器 (keywords transducers) (也包括关键句子, 但是为了简单在接下来的文中仅使用关键字), 就可以得到关键字的时间信息和表示置信度的分数 (scores indicating the confidence). resulting factor transducer 也称作索引, 有如下性质: 每一条路径代表原网格中的部分路径, 而路径上的权重信息则表示相关时间信息和得分/置信度。需要强调的是最后索引上的权重的回推到整条路径, 独立的每条边的权重没有意义。

从原始网格中建立如上的所以大概分为以下五个步骤:

权重推进和聚类 (Weight Pushing and Clustering): 这两步有时候可以理解为预处理。在原始网格中的权重信息通常对应于由语言模型和声学模型所分配的联

合概率，而权重推进步骤把这些权重信息转化成所需要的基于网格的后验概率。聚类的目的是把连接相同的字以及相同时间段内的边合并。聚类是通过“标签”实现的：如果有几条边属于相同的类（连接相同的字以及有时间上的重叠），那么就给这些边相同的输出“标签”（输入标签仅仅是简单的单词）。当在转换器（transducer）上面进行确定化处理时才会对不同类之间进行真正的合并。

因子生成 (Factor Generation) : 正像我们前面所提及的，在最后索引中的每条路径都是原网格的部分路径，这是通过因子生成（factor generation）这步来实现的。对于每一个网格，我们新增加两个状态，起始状态和终止状态。网格中的每个状态，新增加一条从起始状态到此状态的路径，以及此状态到终止状态的路径。在此步中还对半环做了修改，从对数半环 L 变化到半环 $L \times T \times T'$ ，其中 T 是热带半环（tropical semiring）， T' 是北极半环（arctic semiring）， \times 表示产品半环（product semiring）。其中使用对数半环 L 来表示置信度得分，热带半环 T 表示部分路径的起始时间，以及北极半环 T' 表示部分路径的结束时间。需要注意在新增加的起始状态和终止状态之间添加路径的时候，必须保证推进权重（“pushed” weight）中的时间信息代表此部分路径的起始和结束时间。

因子合并 (Factor merging) : 此步骤完成在聚类步骤中提到的对聚类进行合并，是通过确定化编码处理来实现的。在完成合并后，就没有北极半环（arctic semiring）并且产品半环（product semiring）也变为词典半环（lexicographic semiring）。即，把 $L \times T \times T'$ 变为 $T * T * T$ ，同时去掉在聚类步骤中生成的聚类标签。

因子消歧 (Factor Disambiguation) : 在结果传感器（resulting transducer）中的每条连接最后状态的边均对应着在原始网格中不同的独立的部分路径，通过消歧因子使其能在后续步骤中保持相互独立。

优化 (optimization) : 此步中包括空符号移除（epsilon removal），确定化和最小化等的优化操作。

通过上述的几个步骤，就可以把每个网格转化为小的索引，通过结合不同的小的索引变为一个大的索引。进一步的优化可以在最后完成的大的索引上进行。

Implementation Details

因为 kaldi 是基于 openfst 的工具箱，这就使得其实现变得相对简单。直接从 Kaldi:CompactLattice 开始，在其中收集时间信息和置信度得分信息，然后编译使其生成因子传感器。通常情况下使用在索引文章中所描述的算法，但是会需

要做一些本地自适应和修改。在此部分中我们列出此些修改。

半环定义 (semiring definition)

我们对在 `openfst` 中现有的大部分半环进行了模块化的处理,除了北极半环(`arctic semiring`), 是因为北极半环没有可以直接使用的半环。模块化的半环如下所示:

```
// The T*T*T semiring
typedef fst::LexicographicWeight<TropicalWeight, TropicalWeight> StdLStdWeight;
typedef fst::LexicographicWeight<TropicalWeight, StdLStdWeight> StdLStdLStdWeight;
typedef fst::ArcTpl<StdLStdLStdWeight> StdLStdLStdArc;

// The LxTxT' semiring
typedef fst::ProductWeight<TropicalWeight, ArcticWeight> StdXStdprimeWeight;
typedef fst::ProductWeight<LogWeight, StdXStdprimeWeight> LogXStdXStdprimeWeight;
typedef fst::ArcTpl<LogXStdXStdprimeWeight> LogXStdXStdprimeArc;

// Rename the weight and arc types to make them look more "friendly".
typedef StdLStdLStdWeight KwsLexicographicWeight;
typedef StdLStdLStdArc KwsLexicographicArc;
typedef fst::VectorFst<KwsLexicographicArc> KwsLexicographicFst;
typedef LogXStdXStdprimeWeight KwsProductWeight;
typedef LogXStdXStdprimeArc KwsProductArc;
typedef fst::VectorFst<KwsProductArc> KwsProductFst;
```

其中 `fst::ArcticWeight` 是在 lat/arctic-weight.h 中定义的。

权重推进 (Weight Pushing)

在原始的算法中,权重推进是在预处理步骤完成的,但是我们发现如果在因子生成 (factor generation) 步骤中完成权重推进会更高效。计算先验概率 `alphas`、后验概率 `betas`, 通过手动完成权重推进。我们把此步向后移,这样就可以使得在因子生成 (factor generation) 步骤中计算最短路径的时候可以重用 `alphas` 和 `betas` 的信息。对于一条给定的弧,假定 `c` 为它的 `cost`, 通过如下步骤修改 `cost`:

```
non-final:
c <-- c - beta[destination-state of arc] + beta[start-state of arc]
final:
c <-- c + beta[start-state of arc]
```

在完成修改之后把 `betas` 置为 0, 通过如下公式来修改 `alphas`:

```
alpha[s] <--- alpha[s] + beta[s] - beta[initial-state]
```

现在“好像”是通过 pushed FST 计算出 alphas 和 betas。

静音消除 (Long Silence Removal)

在因子生成 (factor generation) 步骤中形成的传感器 (transducer) 可以使得我们可以搜索原网格中的任意一条部分路径，但是在这其中只有部分路径是有意义的。因此可以在因子传感器 (factor transducer) 的顶层使用一些滤波技术。在我们实现的过程中使用基于词间静音长度的滤波器。如果两个词间的静音长度太长（在 babel 项目中此时间设为 0.5 秒），那么这两个单词之间也许是不连续的，我们可以在索引中去掉此路径。如果两个单词之间的静音时间很长，则把指向下个单词的弧转而指向立即状态，然后在整个传感器上运行 Connect ()，此步会去掉没有意义的路径。

消歧符号 (Disambiguation Symbol)

在 Dogan 和 Murat's 的文章中，在完成索引上的最后优化后，去掉输入标签域的消歧符号，使得在搜索时间内完成与关键字 FST 的结合。但是这样使得在搜索的结果上很难做进一步的优化（通过结合索引与关键字 FST 形成结果 FST），因为来自相同句子的不同实例可能会结合。在我们实现的过程中并没有完全去掉消歧符号。我们把消歧符号与句子的 ID 编码在一起，从最后边中去掉原始的消歧符号，并且用编码后的标签去替代最后边的输出标签。这样我们就可以在结果 FST 中进行进一步优化（例如确定化和最小化），因为有时候结果 FST (resulting FST) 会非常大。

Running KWS in Kaldi

在 kaldi 上运行 KWS

在 kaldi 中运行 KWS 时有两个重要的文件：lattice-to-kws-index 和 kws-search。从他们各自的名字中我们也可以分辨其各自的作用，lattice-to-kws-index 是读入网格信息输出索引，kws-search 则是在索引上进行关键字的搜索。

lattice-to-kws-index 可以通过下面的命令来运行：

```
lattice-to-kws-index ark:utter_id "ark:gzlat.1.gz|" ark:1.index
```

其中 lat.1.gz 是 kaldi 中网格的 ARK 格式文件（注：.ark 格式可以是 text 或 binary 格式，（你可以写为 text 格式，命令行要加 ‘t’，binary 是默认的格式）文件里面数据的格式是：key（如句子的 id）空格后接数据。摘自博客 my9084 的博客 <http://blog.sina.com.cn/u/1145070023>），1.index 包含在 achieve 中每个网格的一个索引。Utter_id 是对每个句子的 ID 建立的符号表，如下所示：

```
10713_A_20111024_220917_000665 1
10713_A_20111024_220917_001321 2
10713_A_20111024_220917_002097 3
10713_A_20111024_220917_002550 4
10713_A_20111024_220917_003311 5
10713_A_20111024_220917_003903 6
10713_A_20111024_220917_004616 7
```

```
10713_A_20111024_220917_004976 8
10713_A_20111024_220917_005656 9
10713_A_20111024_220917_006296 10
.....
```

在检索之前，为了得到更多准确的时间信息，网格通常需要“对齐”，有个时候对某些特定的词赋予“惩罚”系数（**penalty**）。在检索之后，会有一个合并的过程这样使得在原 **archive** 中的所有网格可以形成单个索引。一些实例脚本如下：

```
lattice-add-penalty "ark:gzip -cdf lat.1.gz|" ark:- |\
  lattice-align-words $word_boundary $model ark:- ark:- |\
  lattice-scale --acoustic-scale=$acwt --lm-scale=$lmwt ark:- ark:- |\
  lattice-to-kws-index ark:utter_id ark:- ark:- |\
kws-index-union ark:- "ark:gzip -c > index.1.gz"
```

为了在已编译好的索引上搜索关键字，关键字的编译为 FST 格式，如下：

```
transcripts-to-fsts ark:keywords.int ark:keywords.fsts
```

其中 **keywords.int** 文件由一系列形如<关键字 id，关键字内容>所组成，如下：

```
KW101-0001 2669 4878
KW101-0002 3419
KW101-0003 6792
KW101-0004 867 9757
KW101-0005 2055 8016
KW101-0007 5959 5450 16796
KW101-0008 1560 867 9818
KW101-0009 4213
KW101-0010 4305
.....
```

在上述文件中关键字已经映射到其对应的符号上，其文本格式如下所示：

```
KW101-0001 有 晒
KW101-0002 化疗
KW101-0003 屋企人
KW101-0004 个 时间
KW101-0005 保障 成本
KW101-0006 彭鱼
KW101-0007 好 多 野
KW101-0008 今 个 星期日
KW101-0009 吃饭
KW101-0010 后卫
```

在编译完关键字的 FST 后，搜索仅需要一条命令：

```
kws-search "ark:gz| -cdf index.1.gz|" ark:keywords.fsts \
"ark,t:|int2sym.pl -f 2 utter_id > result.1"
```

我们把搜索的结果存储在 result.1 文件中，此文件中的每一行都包括五个部分，对应单独的一条搜索结果。下图所示的是在 result.1 文件中的前几行内容，其中每行的第一部分表示关键字的 ID，第二部分为搜索句子的 ID，第三部分是相对于所搜索句子起始位置的关键字出现的起始帧，第四部分为相对的关键字的结束帧，最后一部分为搜索结果“否定”的对数概率（"negated" log probability）：

```
KW101-0001-A 10713_A_20111024_220917_026395 163 194 2.876953
KW101-0001-A 10713_A_20111024_220917_008532 154 187 5.734375
KW101-0003-A 10733_A_20111021_141006_001440 448 510 1.737305
KW101-0010-A 10713_A_20111024_220917_028639 471 522 7.064453
KW101-0010-A 10733_A_20111021_141006_043478 224 275 8.200195
KW101-0012-A 10713_A_20111024_220917_026395 255 314 7.344727
KW101-0016-A 10733_A_20111021_141006_028914 738 800 0.1992188
KW101-0016-A 10733_A_20111021_141006_028914 738 803 4.556641
KW101-0016-A 10713_A_20111024_220917_020531 224 258 4.624023
KW101-0016-A 10733_A_20111021_141006_020971 233 273 8.793945
```

所有以上所提及的步骤都可以通过现有脚本来实现，脚本的位置在 egs/wsj/s5/run.sh。

Proxy Keywords

在 Kaldi 中的关键字模块中，引入代理关键字来处理不在词汇表内（out Of vocabulary, OOV）的关键字的问题。这里的“代理”指的是与原 OOV 关键字发音较接近的字。代理关键字可以通过下述公式产生：

$$K' = \text{project}(\text{ShortestPath}(K \circ L_2 \circ E' \circ (L_1)^{-1}))$$

其中 K 是原关键字， L_2 是由类似于 Sequitur 的 G2P 工具生成的 OOV 词典， E' 是包含音素模糊（phone confusion）的距离编辑传感器（edit Distance transducer）， L_1 是原来的词典部分关键词可能不在此词典中。假设 OOV 词典 L_2 可以有标准的工具生成，对于每个因素的模糊因素都可以收集到，代理关键字的生成可以通过下列步骤完成：

首先，要生成距离编辑传感器（edit Distance transducer）。假设我们已经得到了 phone_confusion，可以通过运行 egs/babel/s5 中的如下命令来实现：

```
local/build_edit_distance_fst.pl \
--confusion-matrix phone_confusion phones.txt - | \
fstcompile --isymbols=phones.txt --osymbols=phones.txt - Edit.fst
```

文件 phones.txt 是包含在词典中要使用到的因素的文件。Phone_confusion 文件包含每对因素组合因素模糊（phone confusion）的“否定”对数概率（"negated" log

probability)，如下：

```
o oj 7.18083119904456
v tS 4.35670882668959
u ow 5.20766299827991
dZ S 4.91998092582813
h dZ 6.21460809842219
z A 4.18965474202643
iw U 3.93182563272433
b uj 6.40025744530882
j N 3.56751859710967
D z 3.85014760171006
```

这两个词典都需要编译为 FST 格式

```
cat oov.lex | utils/make_lexicon_fst.pl - \|
    fstcompile --isymbols=phones.txt --osymbols=words.txt - \|
    fstinvert | fstarcsort --sort_type=olabel > L2.fst
cat original.lex | utils/make_lexicon_fst.pl - \|
    fstcompile --isymbols=phones.txt --osymbols=words.txt - \|
    fstarcsort --sort_type=ilabel > L1'.fst
fstcompose L2.fst Edit.fst \|
    fstarcsort --sort_type=olabel > L2xE.fst
```

最后可以通过下述命令调用 `generate-proxy-keywords` 来生成代理关键字。

```
generate-proxy-keywords --verbose=1 \
    --cost-threshold=$beam --nBest=$nbest \
    L2xE.fst L1'.fst ark:oov_keywords.int ark:proxy.fsts
```

所有上述的步骤都被整合在一个脚本里，这个脚本就是：

`egs/babel/s5/local/generate_proxy_keywords.sh`

在生成代理关键字后，可以当作普通的关键在原索引上进行搜索。但是需要注意的是对返回的搜索结果可能会需要不同的处理（例如：它们可能需要不同的阈值）以及进行调优。

4.7 在线识别

(@冒顿翻译)

在kaldi的工具集里有好几个程序可以用于在线识别。这些程序都位在src/onlinebin文件夹里，他们是由src/online文件夹里的文件编译而成(你现在可以用make ext 命令进行编译).这些程序大多还需要tools文件夹中的portaudio 库文件支持， portaudio 库文件可以使用tools文件夹中的相应脚本文件下载安装。

这些程序罗列如下：

online-gmm-decode-faster: 从麦克风读取语音，并将识别结果输出到控制台
online-wav-gmm-decode-faster: 读取wav文件列表中的语音，并将识别结果以指定格式输出。

online-server-gmm-decode-faster: 从UDP连接数据中获取语音MFCC向量，并将识别结果打印到控制台。

online-net-client : 从麦克风录音，并将它转换成特征向量，并通过UDP连接发送给online-server-gmm-decode-faster

online-audio-server-decode-faster : 从tcp连接中读取原始语音数据，并且将识别结果返回给客户端

online-audio-client : 读出 wav 文件列表，并将他们通过 tcp 发送到online-audio-server-decode-faster,得到返回的识别结果后，将他存成指定的文件格式

代码中还有一个java版的online-audio-client,他包含更多的功能，并且有一个界面。另外，还有一个与GStreamer 1.0兼容的插件，他可以对输入的语音进行识别，并输出的文字结果。这个插件基于onlinefasterDecoder,也可以做为在线识别程序

在线语音服务器。

online-server-gmm-decode-faster 和online-audio-server-decode-faster的主要区别是各自的输入不同，前者接受特征向量做为输入，而后者则接收原始音频。后者的好处是他可以处理任意客户端直接传过来的数据：不管他是互联网上的计算机还是一部移动设备。最主要的是客户端不需要知道训练模型使用了何种特征集，只要设定好预先定义的采样率和位深度，他就可以跟服务器进行通信。使用特征向量做为输入的服务器的好处是，在客户端和服务器之间传输的数据比音频服务器小得多，而这只需使用简单的音频代码就可以实现。

在online-audio-client和online-audio-server-decode-faster之间的通信包括两个步骤：第一步客户端将原始音频数据包发送给服务器，第二步服务器将识别结果发回给客户端。这两步可以同步进行，这意味着解码器不用等音频数据发送完毕，就可以在线输出结果，这给未来新的应用提供了很多可以扩展的功能。

音频数据

音频数据格式必须是采样率16KHz, 16-bit 位深, 单声道, 线性PCM编码的硬编码数据。

通信协议数据分为块和前缀, 每一个块包含4字节指示此块的长度。这个长度值也是一个long型little-endian值, 可以是正也可以是负 (因为是16-bit采样)。最后一个长度为0的包被当作音频流的结束, 这将迫使解码器导出所有剩余的结果, 并且结束识别过程

识别结果

服务器返回两种类型的结果。时间对齐结果和部分结果。

解码器识别出每一个词都会立即发送一个部分结果, 而当解码器识别到一次发音结束时则会发送一个时间对齐结果 (这可能是也可能不是一次停顿或者句子结束)。

每一个部分结果包会有以字符串: "PARTIAL:" 开头, 后面跟一个词, 每一个词发送一次不同的部分结果包。时间对齐结果包会以字符串: "RESULT:" 开头。后面是以逗号为分格的key=value参数列表, 这些参数列表包含一些有用信息。

NUM: 后面词的个数

FORMAT: 返回结果的格式, 目前只支持WSEC格式 (词-开始-结束-置信度), 未来会允许修改

RECO-DUR: 识别语音所使用的时间。(以秒计的浮点型值)

INPUT-DUR: 输入语音的时间长度 (以秒计的浮点型值)

你可以用识别时间除以输入时间得到解码器的实时识别速度。当服务器将识别结果全部返回后会发送一个"RESULT:DONE", 这种情况下服务器会等待输入或是断开。

在数据头部后面包含有NUM行以FORMAT格式组成的词行, 现在的WSEC格式, 只是简单以逗号分隔的四个部分: 词, 起始时间, 终止时间 (以秒计的浮点数), 置信度 (0-1之间浮点数) 要记住这些词只是在词典中的编码, 因此客户端必须执行转换操作。online-audio-client 在生成webvtt文件时不进行任何字符转换, 因此你需要用iconv将结果文件转换成UTF8

服务器返回的结果的例子如下:

```
RESULT:NUM=3,FORMAT=WSEC,RECO_DUR=1.7,INPUT_DUR=3.22
one,0.4,1.2,1
two,1.4,1.9,1
three,2.2,3.4,0.4
RESULT:DONE
```

使用举例：

命令行启动服务器：

```
online-audio-server-decode-faster --verbose=1 --rt-min=0.5 --rt-max=3.0  
--max-active=6000 --beam=72.0 --acoustic-scale=0.0769 final.mdl graph/HCLG.fst  
graph/words.txt '1:2:3:4:5' 5010 graph/word_boundary_phones.int final.mat
```

各参数含义如下：

final.mdl :声音模型文件

HCLG.fst :完全的fst

words.txt:发音词典（将词的id号映射到对应的字符上）

'1:2:3:4:5' : sil 的id号

5010 : 服务器的端口号

word_boundary_phones.int : 词对齐时使用的音素边界信息

final.mat : 特征的LDA矩阵

命令行启动客户端

```
online-audio-client --htk --vtt localhost 5010 scp:test.scp
```

各参数含义如下

?-htk 结果存成htk 标注文件

?-vtt 结果存成webvtt文件

?localhost 服务器地址

?5010 服务器端口号

?scp:test.scp Wav文件的列表文件

命令行方式启动java客户端

```
java -jar online-audio-client.jar
```

或者直接在图形界面上双击jar文件

命令行方式运行识别wav文件

```
online-wav-gmm-decode-faster --verbose=1 --rt-min=0.8 --rt-max=0.85\  
--max-active=4000 --beam=12.0 --acoustic-scale=0.0769 \  
scp:$decode_dir/input.scp $ac_model/model $ac_model/HCLG.fst \  
$ac_model/words.txt '1:2:3:4:5' ark,t:$decode_dir/trans.txt \  
ark,t:$decode_dir/ali.txt $trans_matrix;;
```

各参数含义如下：

GStreamer 插件

Kaldi 工具集为GStreamer 多媒体流框架提供了一个插件（1.0或兼容版）。此插件可以做为过滤器，接收原始音频做为输入并将识别结果做为输出。这个插件的主要好处是他使得kaldi 在线语音识别功能对所有支持GStreamer1.0的编程语言都可用(包括python,ruby,java,java等等)。这也使得可以把kaldi 在线解码器集成到支持GStreamer 通信标准的应用程序中

安装

GStreamer 插件的源码位于src/gst-plugin 目录。要完成编译，kaldi工具集的其它部分必须使用共享库进行编译。因此在配置的时候必须使用--shared 参数，同时需要编译在线扩展(make ext).确保支撑GStreamer 1.0开发头文件的包已经安装在你的系统中。在Debian Jessie系统版本中，需要包libgstreamer1.0-dev。在Debian Wheezy 版本中，GStreamer 1.0 可以从backports源中下载到。需要安装gstreamer1.0-plugins-good 和 gstreamer1.0-tools 这两个包。演示程序还需要PulseAudio Gstreamer插件，包名：gstreamer1.0-pulseaudio

最后，在src/gst-plugin目录中运行make depend和make。这样就会生成一个文件src/gst-plugin/libgstkaldi.so，他包含了GStreamer 插件。为确保GStreamer 可以找到kaldi 插件，必须把src/gst-plugin目录加到他的插件搜索目录。因此需要把这个目录加入环境变量中

```
export GST_PLUGIN_PATH=$KALDI_ROOT/src/gst-plugin
```

当然，你需要把\$KALDI_ROOT修改成你自己文件系统中kaldi所在 的根目录。

4.8 决策树是如何在 kald 中使用

How decision trees are used in Kaldi

介绍

这部分将介绍音素决策树在 kald 中是如何建立和使用的，以及是如何将训练和图建立相交互的。对于决策树的构建的代码，可以看 [Decision tree internals](#)；对于建立图解码的更多细节，可以看 [Decoding graph construction in Kaldi](#)。

最基本的实现方法就是自顶向下贪婪的分裂，这里我们又很多的方法来分裂我们的数据，比如我们可以用左音素、右音素、中间音素，我们所在的状态等等。我们实现的方法与标准的方法很相似，具体的你可以去看 Young, Odell 和 Woodland 的论文"Tree-based State Tying for High Accuracy Acoustic Modeling"。在这个方法里，我们通过局部最优问题来分裂数据，比如假设我们通过一个单高斯来分裂我们需要建模的数据，通常我们会用最大似然函数的增加来作为分裂的依据。与标准实现方法不同的有：对怎么计算树的根节点来添加 flexibility；对 hmm 状态和中间音素决策的能力；但是在 kald 脚本默认的是，决策是对数据自顶向下的自动实现二叉树聚类，意思就是我们不需要提供手工产生的决策。不考虑树的根节点的构造：它可能用在一个单共享组中的所有音素的所有统计量来进行分裂(包括中心音素和HMM状态的问题)，或者用一个单一的音素，或者 HMM 状态里的音素，来作为分裂的树的根节点，或者音素组作为树的根节点。对于在标准的脚本里如何构建，你可以看 [Data preparation](#)。在实际中，我们让每一个树的根节点对应一个真正的音素，意思就是我们重组了所有 all word-position-dependent, 每一个音素的 tone-dependent 或者 stress-dependent 组成一个组来成为树的根节点。

这页的剩下部分大多数介绍代码层的一些细节。

上下文相关音素窗 (Phonetic context windows)

这里我们将解释在我们的代码里如何描述上下文相关音素。一个特定的树含有两个值，他们分别描述上下文音素窗的宽度和中心位置。下表是总结这些值：

Name in code	Name in command-line arguments	Value (triphone)	Value (monophone)
N	-context-width=?	3	1
P	-central-position=?	1	0

N 代表上下文相关音素窗的宽度，P 表示指定中心音素。正常地 P 是窗的中心（因此名字叫中心位置）；举个例子，当 N=3，我们将有 P=1，但是你可以自由的选择0到 N-1中的任何值；比如，P=2 和 N=3 表示有2个音素在左边的上下文，而右边没有上下文。在代码中，当我们讨论中心音素时，我们的意思是第 P 个音素，也许是或者不是上下文相关音素窗的中心音素。

一个整型的 vector 表示一个典型的三音素上下文窗也许就是：

```
// probably not valid C++
```

```
vector<int32> ctx_window = { 12, 15, 21 };
```

假设 $N=3$ 和 $P=1$, 这个就代表音素15有一个右边的上下文21和左边的上下文12。这种方式我们处理尾部时就用0 (表示不是一个有效的音素, 因为在 OpenFst 里的 `epsilon` 表示没有符号), 所以举个例子:

```
vector<int32> ctx_window = { 12, 15, 0 };
```

表示音素15 有一个左上下文和没有右上下文, 因为是一句话的结尾处。尤其在一个句子的结尾处, 这种用0的方式也许有一点出乎意料, 因为最后一个音素事实上是后续符号"\$" (看 [Making the context transducer](#)), 但是在决策树代码里为了方便, 我们不把后续符号放在这些上下文窗中, 我们直接给其赋0。注意如果我们要有 $N=3$ 和 $P=2$, 上述的上下文窗将无效, 因为第 P 个元素是0, 它不是一个真正的音素; 当然, 如果我们用一个树的 $N=1$, 所有的窗都将无效, 因为他们是错误的大小。在单因素的情况下, 我们有一个窗像:

```
vector<int32> ctx_window = { 15 };
```

所以单音素系统里是上下文相关系统的一个特殊情况, 窗的大小 $N=1$ 和一个不做任何事情树。

树建立的过程 (The tree building process)

在这部分我们将介绍在 `kaldi` 中的树建立的过程。

即使一个单因素系统有一个决策树, 但是也比较简单。看函数 [MonophoneContextDependency\(\)](#) 和 [MonophoneContextDependencyShared\(\)](#), 他们将返回这个简单的树。这个在命令行里叫 `gmm-init-mono`; 他们主要的输入是 [HmmTopology](#) 类和它们的输出是树, 通常作为类 [ContextDependency](#) 写到一个名字叫树的文件(`tree`), 和它们的模型文件(模型文件含有一个 [TransitionModel](#) 类和一个 [AmDiagGmm](#) 类)。如果程序 `gmm-init-mono` 接受一个叫 `-shared-phones` 的选项, 它将在音素特定集中共享 pdfs; 否则将所有的音素分离。

在训练一个以 `flat start` 开始的单音素系统, 我们训练树, 采用单音素对齐和使用函数 [AccumulateTreeStats\(\)](#) (称为 `acc-tree-stats`) 来累积统计量。这个程序不仅仅在在单音素对齐中使用; 也可以对上下文相关音素来对齐, 所以我们可以根据比如三音素对齐来建立树。为树建立的统计量将写到 `disk`, 作为类型 [BuildTreeStatsType](#) (看 [Statistics for building the tree](#))。函数 [AccumulateTreeStats\(\)](#) 采用值 N 和 P , 像我们在之前的部分解释那样; 命令行将默认的各自地设定它们为3和1, 但是这个可以用 `-context-width` 和 `-central-position` 选项来覆盖。程序 `acc-tree-stats` 采用上下文音素 (e.g. `silence`) 的列表, 但是如果是上下文音素, 他们是不需要的。它仅仅是一个减少统计量的一个机制。对于上下文相关音素, 程序将积累相对应的统计量 the corresponding statistics without the keys corresponding to the left and right phones defined (c.f. [Event maps](#))。

当统计量已经积累, 我们将使用程序 `build-tree` 来建立树。输出是一棵树。程序 `build-tree` 需要三个东西:

- 统计量(BuildTreeStatsType 类的)
- 问题的构建([Questions](#) 类)
- 根文件(看下面)

统计量是通过程序 `acc-tree-stats` 得到的；问题构建类是通过问题构建程序得到的，这需要再音素问题集的一个 `topology` 表中 (在我们的脚本里，他们通常从程序 `cluster-phones` 的树建立统计量自动获得)。根文件指定音素集，这些音素集是在决策树聚类处理中共享根节点，对于每一个音素需要设定2个东西：

- "shared" or "not-shared" 意思是对于每一个 [pdf-classes](#) 是否有单独的根节点(例如：典型的情况，HMM 的状态)，或者根节点是否共享。如果我们打算去分裂(下面有个"split" 选项)，我们执行时，根节点是共享的。
- "split" 或者 "not-split" 意思决策树分裂是否应该通过根节点的问题来决定(对于 `silence`，我们一般不 `split`)。

要小心，因为符号有些棘手。在根文件里的行"shared" 是表示我们是否在一个单一的树的根节点上共享的所有的三个 HMM 状态。但是我们经常共享所有音素的根节点，这些音素在一个根文件的单一的一行上。这个不是通过这些字符串来构建的，因为如果你不想共享它们，你会把它们分布在根文件的分离的行上。

下面是根文件的一个例子；这里假设音素1是 `silence` 和所有的其他有分离的根节点。

```
not-shared not-split 1
shared split 2
shared split 3
...
shared split 28
```

当我们有像 `position and stress-dependent` 音素时，在同一行上有许多音素是非常有用的。这种情况下，每一个"real" 音素将对应整数音素 `ids` 的一个集合。在这种情况下，我们对一个特定标注的音素的所有情况将共享根节点（译者注：意思应该就是一个音素可能有不同的标注版本，我们用的时候就相当于一个，具体的可以看下面的根文件）。下面是 `wsj` 数据库的一个根文件，在 `egs/wsj/s5` 脚本里(这个在 `text` 里，不是整数形式；它将经过 `kaldi` 转为整数形式)：

```
not-shared not-split SIL SIL B SIL E SIL I SIL S SPN SPN_B SPN_E SPN_I
SPN_S NSN NSN_B NSN_E NSN_I NSN_S
shared split AA B AA E AA I AA S AA0 B AA0 E AA0 I AA0 S AA1 B AA1 E
AA1 I AA1 S AA2 B AA2 E AA2 I AA2 S
shared split AE B AE E AE I AE S AE0 B AE0 E AE0 I AE0 S AE1 B AE1 E
AE1 I AE1 S AE2 B AE2 E AE2 I AE2 S
shared split AH B AH E AH I AH S AH0 B AH0 E AH0 I AH0 S AH1 B AH1 E
AH1 I AH1 S AH2 B AH2 E AH2 I AH2 S
shared split AO B AO E AO I AO S AO0 B AO0 E AO0 I AO0 S AO1 B AO1 E
AO1 I AO1 S AO2 B AO2 E AO2 I AO2 S
shared split AW B AW E AW I AW S AW0 B AW0 E AW0 I AW0 S AW1 B
AW1 E AW1 I AW1 S AW2 B AW2 E AW2 I AW2 S
shared split AY B AY E AY I AY S AY0 B AY0 E AY0 I AY0 S AY1 B AY1 E
AY1 I AY1 S AY2 B AY2 E AY2 I AY2 S
shared split B_B B_E B_I B_S
```



```
shared split CH B CH E CH I CH S
shared split D_B D_E D_I D_S
```

当创建根文件时，你应该确保每一行至少有一个音素。举例来说，在这种情况下，如果音素 AY 应该至少在 stress 和 word-position 有一些组成，我们就认为是可以的。

在这里例子里，我们有许多 word-position-dependent 静音的变异体等等。在这里例子里，他们将共享他们的 pdf，因为他们是在同一行和是 "not-split"，但是他们有不同的转移参数。事实上，静音的大多数变异体从来不被用作词间出现的静音；这些都为了未来做准备，以防止在未来有人做了一些奇怪的变化。

我们对初始阶段的高斯混合是用之前(比如：单音素)的建立来做对齐；对齐通过程序 [convert-ali](#) 来把一个树转换成其他的。

PDF 标识符 (identifiers)

PDF 标识符(pdf-id)是一个数字，从0开始，被用作概率分布函数(p.d.f.)的一个索引。在系统中的每一个 p.d.f. 有自己的 pdf-id，和他们是邻近的(典型的的就是在一个 LVCSR 有上千上万个)。当树一开始建立的时候，他们就被初始化了。这有可能取决于树是如何建立的，对于每一个 pdf-id，都有一个音素与之对应。

上下文相关类 (Context dependency objects)

类 [ContextDependencyInterface](#) 对于一个与图建立相相互的特定书来说，是一个虚拟的基类。这种交互仅仅需要4个函数：

- [ContextWidth\(\)](#) 返回树所需要的 N(context-width)的值。
- [CentralPosition\(\)](#) 返回树所需要的 P(central-position)值。
- [NumPdfs\(\)](#) 返回由树定义的 pdfs 的数量，他们的值是从0到 NumPdfs()-1。
- [Compute\(\)](#) 是计算一个特定的上下文 (和 [pdf-class](#))的 [pdf-id](#) 的函数。

函数 [ContextDependencyInterface::Compute\(\)](#) [Compute\(\)](#) 如下声明：

```
class ContextDependencyInterface {
...
virtual bool Compute(const std::vector<int32> &phoneseq, int32 pdf_class,
int32*pdf_id) const;
}
```

如果可以计算这个上下文和 [pdf-class](#) 的 pdf-id 就返回 true。如果是 false，就说明许多种类的错误或者不匹配。用这个函数的一个例子是：

```
ContextDependencyInterface *ctx_dep = ... ;
vector<int32> ctx_window = { 12, 15, 21 }; // not valid C++
int32 pdf_class = 1; // probably central state of 3-state HMM.
int32 pdf_id;
if(!ctx_dep->Compute(ctx_window, pdf_class, &pdf_id))
KALDI_ERR<< "Something went wrong!"
else
KALDI_LOG<< "Got pdf-id, it is " << pdf_id;
```

从类 [ContextDependencyInterface](#) 继承的唯一一个类就是类 [ContextDependency](#),

将轻微的丰富这种交互；唯一的一个重要的添加就是函数 [GetPdfInfo](#)，它将在类 [TransitionModel](#) 中计算一个特定的 pdf 对应某一个音素(通过枚举所有的上下文，这个函数可以模拟特定的 [ContextDependencyInterface](#) 的交互)。

对象 [ContextDependency](#) 事实上是对象 [EventMap](#) 的一点点的改动；可以看 [Decision tree internals](#)。我们想要尽最大可能地隐藏树的真正的实现来使以后我们需要重组代码的时候更加的简单。

决策树的一个例子(An example of a decision tree)

决策树文件的格式不是以可读性为第一的标准创建的，而是由于大家的需要，我们将尝试解释如何让翻译这个文件。可以看下面的例子，是一个在 wsj 数据库略的一个三音素的决策树。它由一个名字叫 [ContextDependency](#) 对象开始的，然后 N (the context-width) 等于3，P (the "central position" of the context window) 等于1，比如这个音素上下文位置的中心，我们是从0开始编号的。文件的剩下部分包含一个单一的对象 [EventMap](#)。[EventMap](#) 是一个多态类型，它包含指向其他 [EventMap](#) 对象的指针。更多的细节可以看 [Event maps](#)；它是一个决策树或者决策树集合的表示，它把 key-value 对的集合(比如：left-phone=5, central-phone=10, right-phone=11, pdf-class=2)映射成一个 pdf-id (比如158)。简单的说，它有三个类型：[SplitEventMap](#) (像决策树里的一个分裂)，[ConstantEventMap](#) (像决策树里的叶子，只包含一个数字，代表一个 pdf-id)，和 [TableEventMap](#) (像含有其他的 EventMaps 表中查找)。[SplitEventMap](#) 和 [TableEventMap](#) 都有一个他们查询的"key"，在这种情况下可以为0, 1 or 2，与之对应的是 left, central or right context，或者-1代表"pdf-class"的标识。通常 pdf-class 的值与 HMM 状态的索引值一样，比如0, 1 or 2。尝试不要因为这样而迷惑：key 的值是-1，但是 value 是0, 1 or 2，和他们在上下文窗里的音素的 keys 的 0, 1 or 2 没有任何联系。[SplitEventMap](#) 有一组值，这组值将触发树的"yes"分支。下面是解释树文件格式的一种 quasi-BNF 标识。

```
EventMap := ConstantEventMap | SplitEventMap | TableEventMap | "NULL"
ConstantEventMap := "CE" <numeric pdf-id>
SplitEventMap := "SE" <key-to-split-on> "[" yes-value-list "]" "{" EventMap EventMap "}"
TableEventMap := "TE" <key-to-split-on> <table-size> "(" EventMapList ")"
```

在下面的这个例子里，树的顶层 [EventMap](#) 是一个分裂的 key 为1，表示中心音素 [SplitEventMap](#) (SE)。在方括号里的是 phone-ids 的邻近值。就像发生的那样，他们不是代表一个问题，而仅仅是在音素上的分裂的方式，以至于我们可以得到真正的每个音素的决策树。关键就是这个树是通过"shared roots"来建立的，所以这里有许多 phone-ids，对应相同音素的不同版本的 word-position-and-stress-marked，他们共享树节点。我们不可能在树的顶层使用 [TableEventMap](#) (TE)，或者我们不得不重复决策树很多次 (因为 [EventMap](#) 是一个纯树，不是一个普通的图，它没有任何的机制被指向"shared")。"SE" label 的接下来的几个例子也是这个"quasi-tree"的一部分，它一开始就用中心音素来分类 (当我们继续往下看这个文件，我们将更加对这个树深入；注意到括号"{" 是打开的而不是闭合的)。然后我们的字符串"TE -1 5 (CE 0 CE 1 CE 2 CE 3 CE 4)"，表示在 pdf-class "-1"(effectively, the HMM-position)用 [TableEventMap](#) 来分裂，和通过4返回值0。这个值代表对静音和噪声音素 SIL, NSN and SPN 的5个

pdf-ids; 在我们的建立中, 这个 pdfs 是在这三个非语音音素(对于每一个非语音音素只有转移矩阵是特定的)是共享的。注意: 对于这些音素, 我们有5个状态而不是三个状态的 HMM, 因此有5个不同的 pdf-ids。接下来是"SE -1 [0]"; 和这些可以被认为在树中是第一个"real" 问题。当中心音素的值是5到19时, 我们可以从上面提供的看出 SE 问题, 这个是音素 AA 的不同版本; 问题就是 pdf-class (key -1)的值是否为0 (i.e. the leftmost HMM-state)。假如这个答案是"yes", 接下来的问题就是"SE 2 [220 221 222 223]", 这就问音素的右边是不是音素"M"的各种形式(一个相对不直观的问题被问, 因为他们是 leftmost HMM-state); 如果是 yes, 我们将问"SE 0 [104 105 106 107... 286 287]" 音素的右边的一个问题, 如果是 yes, 然后 pdf-id 就是5 ("CE 5") 和如果是 no,就是696 ("CE 696")。

```
s3# copy-tree --binary=false exp/tri1/tree - 2>/dev/null | head -100
ContextDependency 3 1 ToPdf SE 1 [ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73
74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139
140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158
159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178
179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197
198 199 200 201 202 203 204 205 206 207 208 209 210 211 ]
{ SE 1 [ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82
83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106
107 108 109 110 111 ]
{ SE 1 [ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55
56 57 58 59 60 61 62 63 64 65 66 67 ]
{ SE 1 [ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 ]
{ SE 1 [ 1 2 3 ]
{ TE -1 5 ( CE 0 CE 1 CE 2 CE 3 CE 4 )
SE -1 [ 0 ]
{ SE 2 [ 220 221 222 223 ]
{ SE 0 [ 104 105 106 107 112 113 114 115 172 173 174 175 208 209 210 211 212 213
214 215 264 265 266 267 280 281 282 283 284 285 286 287 ]
{ CE 5 CE 696 }
SE 2 [ 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 36 37 38 39 40 41 42 43 44 45 46 47
48 49 50 51 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 248
249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 268 269 270 271
288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 ]
```

下面是一个简单的例子: 从 rm 数据库的一个单音素 tree。顶层 [EventMap](#) 是一个 [TableEventMap](#) ("TE 0 49 ...")。key "0" 表示的就是中心音素, 因为 context width (N) 为1。在这个表中项的数量是 49 (这种情况下, 音素的数量加1)。在表(索引为0)中的第一个 [EventMap](#) 是 NULL, 因为这里没有索引为0的音素。接下来是一个有三个元素的 [TableEventMap](#), 与第一个音素的三个 HMM 状态 (technically, pdf-classes) 想对应: "TE -1 3 (CE 0 CE 1 CE 2)"。

```
s3# copy-tree --binary=false exp/mono/tree - 2>/dev/null | head -5
ContextDependency 1 0 ToPdf TE 0 49 ( NULL TE -1 3 ( CE 0 CE 1 CE 2 )
TE -1 3 ( CE 3 CE 4 CE 5 )
```

```
TE -1 3 ( CE 6 CE 7 CE 8 )
TE -1 3 ( CE 9 CE 10 CE 11 )
TE -1 3 ( CE 12 CE 13 CE 14 )
```

The `ilabel_info` object

CLG 图(看 [Decoding graph construction in Kaldi](#)) has symbols on its input side that represent context-dependent phones (as well as disambiguation symbols and possibly epsilon symbols)。在途中，这些被表示成整数型的标签。我们在代码中用一个对象，在文件名我们通常称为 `ilabel_info`。`ilabel_info` 对象跟 [ContextFst](#) 对象有很强的联系，可以看 [The ContextFst object](#)。像 `kaldi` 中许多其他的类型一样，`ilabel_info` 是一个通用(STL)的类型，但是我们可以使用一致的变量名，以至于可以被识别出来。就是下面的类型：

```
std::vector<std::vector<int32>> ilabel_info;
```

它是一个 `vector`，是通过 FST 输入标签来索引的，它可以给每一个输入标签相对应的音素上下文窗(看上面的 [Phonetic context windows](#))。举例来说，假设符号 1500 是音素 30，有一个 12 的右边的上下文和 4 的左边的上下文，我们将有：

```
// not valid C++
ilabel_info[1500] == { 4, 30, 12 };
In the monophone case, we would have things like:
ilabel_info[30] == { 28 };
```

这是对处理歧义符号的一个特殊处理(看 [Disambiguation symbols](#) 或者 Springer Handbook paper referenced above for an explanation of what these are)。如果一个 `ilabel_info` entry 对应一个歧义符号，我们把它放在歧义符号的符号表的负部分 (note that this is not the same as the number of the printed form of the disambiguation symbol as in #0, #1, #2 etc., 这些数字对应符号表文件的歧义符号，这个在我们现在的脚本中叫 `phones_disambig.txt`)。举例来说，

```
ilabel_info[5] == { -42 };
```

意思就是在 HCLG 中的符号数字 5 对应一个整数 id 为 42 的歧义符号。我们不是为了脚本的方便，所以解释 `ilabel_info` 对象的程序对于给定的歧义符号列表不是很需要，在单音素的情况下，是为了可以从真正的音素中区分它们。这里是二个其他的特殊的例子，我们有：

```
ilabel_info[0] == { }; // epsilon
ilabel_info[1] == { 0 }; // disambig symbol #-1;
// we use symbol 1, but don't consider this hardwired.
```

第一个表示正常的 `epsilon` 符号，我们将给它一个空的 `vector` 作为它的 `ilabel_info` entry。这个符号不会出现 CLG 的左边。第二个是一个特殊的歧义符号，它的印刷形式为 "#-1"。在正常的脚本里，我们使用它，这里的 `epsilon`s 被用作 C transducer 的输入；它是确保在空音素表示的词里对 CLG 确定性。

程序 [fstmakecontextsyms](#) 是可以创建一个与 `ilabel_info` 对象的打印形式相对应的符号表；这主要用来调试和诊断。

就像你看到的那样，`ilabel_info` 对象不是很愉快，因为它涉及到歧义符号，但是与它密切交互的代码不是很多：仅仅 [fst::ContextFst](#) 类(和一些相关的东西；可

以看 ["Classes and functions related to context expansion"](#))，程序 [fstmakecontextsyms.cc](#)，和一些在 [Classes and functions for creating FSTs from HMMs](#) 列出来的函数)。ContextDependency 对象，特别地，仅仅可以看到代表音素上下文窗的长度 N 的有效的序列。

版权所有 翻版必究

4.9 HMM 拓扑结构和转移模型

HMM topology and transition modeling

介绍

在这里我们将介绍在 Kaldi 中如何表示 HMM topologies 和我们如何让建模和训练 HMM 转移概率的。我们将简要的说下它是如何跟决策树联系的；决策树你可以在 [How decision trees are used in Kaldi](#) 和 [Decision tree internals](#) 这些地方看到更详细的；对于这个里面的类和函数，你可以看 [Classes and functions related to HMM topology and transition modeling](#)。

HMM topologies

[HmmTopology](#) 类是为用户指定到 HMM 音素的拓扑结构。在平常的样例里，脚本创建一个 [HmmTopology](#) 对象的文本格式的文件，然后给命令行使用。为了说明这个对象包含什么，接下来的就是一个 [HmmTopology](#) 对象的文本格式的文件(the 3-state Bakis model):

```
<Topology>
<TopologyEntry>
  <ForPhones> 1 2 3 4 5 6 7 8 </ForPhones>
  <State> 0 <PdfClass> 0
  <Transition> 0 0.5
  <Transition> 1 0.5
</State>
  <State> 1 <PdfClass> 1
  <Transition> 1 0.5
  <Transition> 2 0.5
</State>
  <State> 2 <PdfClass> 2
  <Transition> 2 0.5
  <Transition> 3 0.5
</State>
  <State> 3
</State>
</TopologyEntry>
</Topology>
```

这里在这个特定的 [HmmTopology](#) 对象里有一个 TopologyEntry，它覆盖音素1到8(所以在这个例子里有8个音素和他们共享相同的 topology)。他们有3个发射态；每一个有个自环和一个转移到下一个状态的。这里也有一个第四个非发射态，状态3 (对于它没有<PdfClass> entry) 没有任何的转移。这是一个这些 topology entries 的标准特征；Kaldi 把第一个状态(state zero) 看做开始状态。最后一个状态，一般都是非发射状态和没有任何的转移，有最终的概率的那一个。在一个 HMM 中，你可以把转移概率看做最后一个状态的最终概率。在这个特定的例子里所有的发射状态他们有不同的 pdfs (因为 PdfClass 数目是不同的)。我们可以强制<PdfClass>数目一样。在对象 [HmmTopology](#) 里概率是用来初始化训练的；这个训练概率对于上下文相关的 HMM 是特定的和存储在 [TransitionModel](#) 对象。[TransitionModel](#) 用一个类名来存储 [HmmTopology](#) 对象，但是要注意，在初始化 [TransitionModel](#) 后，一般不使用在 [HmmTopology](#)

对象的转移概率。这里有个特例，对于那些不是最终的非发射状态(比如：他们有转移但是没有<PdfClass> entry)，Kaldi 不训练这些转移概率，而是用 [HmmTopology](#) 对象里给的概率。对于不支持非发射状态的转移概率的训练，简化了我们的训练机制，和由于非转移状态有转移是不正常的，我们就觉得这些没有很大的损失。

Pdf-classes

pdf-class 是一个与 [HmmTopology](#) 对象有关的概念。[HmmTopology](#) 对象为每个音素指定了一个 prototype HMM。"prototype HMM"的每一个状态数有一个 "pdf_class"变量。如果两个状态有相同的 pdf_class 变量，他们将共享相同的概率分布函数(p.d.f.)，如果他们在相同的音素上下文。这是因为决策树代码不能够直接“看到”HMM 状态，它仅仅可以看到 pdf-class。在正常的情况下，pdf-class 和 HMM 状态的索引(e.g. 0, 1 or 2)一样，但是 pdf classes 为用户提供一种强制共享的方式。如果你想丰富转移模型但是想离开声学模型，这个是非常有用的，要不然就变成一样的。pdf-class 的一个功能就是指定非发射状态。如果一些 HMM 状态的 pdf-class 设定常量 [kNoPdf](#) = -1，然后 HMM 状态是非发射的(它与 pdf 没有联系)。这个可以通过删除<PdfClass>标签和相联系的数目来简化对象的文本格式。

一个特定的 prototype HMM 的 pdf-classes 的设置期望是从0开始和邻近的(e.g. 0, 1, 2)。这些是为了图建立代码的方便和没有导致任何的损失。

Transition models (the TransitionModel object)

[TransitionModel](#) 对象存储转移概率和关于 HMM topologies 的信息 (它包括一个 [HmmTopology](#) 对象)。图建立代码根据 [TransitionModel](#) 对象来获得 topology 和转移概率(它也需要一个 [ContextDependencyInterface](#) 对象来获得与特定音素上下文的 pdf-ids)。

How we model transition probabilities in Kaldi

The decision that underlies a lot of the transition-modeling code is as follows: 我们很据下面的4件事情决定一个上下文相关的 HMM 状态的转移概率(你可以把他们看做4-tuple):

- 音素(HMM 里的)
- 源 HMM-state (我们解释成 [HmmTopology](#) 对象, i.e. normally 0, 1 or 2)
- [pdf-id](#) (i.e. 与状态相关的 pdf 索引)
- 在 [HmmTopology](#) 对象的转移的索引

这4项可以看做在 [HmmTopology](#) 对象里的目标 HMM 状态的编码。根绝这些事情来求转移概率的原因是，这是最细粒度 (the most fine-grained way) 的方式，我们在不增加编译解码图的大小来对转移建模。对于训练转移概率也是非常方便的。事实上，用传统的方法来对转移尽可能精确的建模可能没有任何区别，和在单音素层面上的共享转移的 HTK 方式可能更加有效。

TransitionModel 对象当被初始化时,就被设置成许多整数的映射,也被其他部分的代码用作这些映射。除了上面提到的数量,也有转移标识符(transition-ids),转移索引(这个与转移标识符不是一样的),和转移状态。之所以要介绍这些标识符和相联系的映射,是因为我们可以用一个完整的基本 FST 的训练脚本。最自然的基本 FST 的设置是输入标签是 pdf-ids。然而,考虑到我们的树建立的方法,它不是总可能从 pdf-id 唯一的映射到一个音素,所以那样很难从输入的标签序列映射到音素序列,和这不是很方便的很多原因;一般很难仅仅用 FST 的信息来训练转移概率。基于这个原因,我们把转移标识符 transition-ids 作为 FST 的输入标签,和把这些映射到 pdf-id,也可以映射到音素,也可以映射到一个 prototype HMM (在 **HmmTopology** 对象里)具体的转移概率。

Integer identifiers used by TransitionModel

以下类型的标识符都用于 **TransitionModel** 接口,它们所有的都用 int32类型。注意其中的一些量是从1开始索引的,也有有的是从0开始索引的。我们尽可能的避免从1索引,因为与 C++数组索引不兼容,但是因为 OpenFst 把0看做特定的情况(代表特殊符号 epsilon),我们决定把那些频繁在 FST 中用作输入符号的从1开始索引,我们就允许从1开始。更重要的是,transition-ids 是从1开始的。因为我们没有想到作为 FST 的标签, pdf-ids 出现很频繁,和因为我们经常使用他们作为 C++数组索引,我们决定使他们从0开始,但当 pdf-ids 作为 FST 的输入标签,我们就给 pdf-ids 加1。当我们阅读 **TransitionModel** 代码时,我们应该对索引从1开始的量,如果是这种情况我们就减去1,如果不是就不用减去。我们记录这些声明的成员变量。在任何情况,这些代码不是公开接口的,因为那会引起很多困惑。在 **TransitionModel** 中用的许多整数量如下:》

- 音素(从1开始): 标识符的类型贯穿整个工具箱;用 OpenFst 符号表可以把 pdf-ids 转换成一个音素。这类的 id 不一定是邻近的(工具箱允许跳过音素索引)。
- Hmm 状态(从0开始): 这是 **HmmTopology::TopologyEntry** 类型的索引,一般取值为{0, 1, 2}里的一个。
- Pdf 或者 pdf-id (从0开始): 这是 p.d.f.的索引,由决策树的聚类来初始化(看 **PDF identifiers**)。在一个系统里通常由几千个 pdf-ids。
- 转移状态,或者 trans_state (从0开始): 这是由 **TransitionModel** 自己定义的。每个可能的三(phone, hmm-state, pdf) 映射到一个自己的转移状态。转移状态可以认为是 HMM 状态最细化的分割尺度,转移概率都是分别评估的。
- 转移索引或者 trans_index (从0开始): 这是 **HmmTopology::HmmState** 类型中的转移数组的索引,这个索引记录了转移出某个转移状态的次数。
- 转移标识符或者 trans_id (从0开始): 其中的每一个都对应转移模型中的唯一的一个转移概率。用从(transition-state, transition-index) 到 transition-id 的映射,反之亦然。

在转移模型代码中也参考如下的概念:

- A triple 意思就是 a triple (phone, hmm-state, pdf) 从转移状态中映射。
- A pair 意思就是 a pair (transition-state, transition-index) 从转移 id 中映射。

Training the transition model

转移模型的训练步骤是非常简单的，我们创建的 FST(训练和测试中) 将 transition-ids 作为他们的输入标签。在训练阶段，我们做维特比解码，生成输入标签的序列，就是每一个 transition-ids (每个特征向量一个，意思也就是对应每帧输出有限个 transition-id 的序列)的序列。训练转移概率时我们积累的统计量就是数每个 transition-id 在训练阶段出现了多少次(代码本身用浮点数来数，但在正常的训练时我们仅仅用整数)。函数 Transition::Update() 对每一个 transition-state 做最大似然更新。这个明显是非常有用的。这里也有一些小的和概率下线有关的苦难，如果一个特定的 transition-state 不知道，我们怎么做？更多的细节，请看代码。

Alignments in Kaldi

这里我们将介绍对齐的概念。说到对齐，我们意思就是<int32>类型的向量，含有 transition-ids 的序列，(c.f. [Integer identifiers used by TransitionModel](#)) 它的长度跟需要对齐的那句话一样长。transition-ids 序列可以通过对输入标签序列解码获得。对齐通常用在训练阶段的维特比训练和在测试时间的自适应阶段。因为 transition-ids 编码了音素的信息，所以从对齐中计算出音素的序列是可行的(c.f. [SplitToPhones\(\)](#) and [ali-to-phones.cc](#))。

我们经常需要处理以句子 id 做索引的一系列对齐。为了方便，我们用表来读和写“对齐”，更多的信息可以看 [I/O with alignments](#)。

函数 [ConvertAlignment\(\)](#) (c.f. 命令行 [convert-ali](#)) 把对齐从一个转移模型转换成其他的。这种情况就是你使用一个转移模型(由特定的决策树来创建)来做对齐，并希望将其转换为不同的树的其他的转移模型。最好的就是从最原始的音素映射到新的音素集；这个特征通常不是一定要的，但是我们使用它来处理一些由减少的音素集的简化的模型。

在对齐的程序里通常使用一个后缀"-ali"。

State-level posteriors

状态级的后验概率是“对齐”这个概念的扩展，区别于每帧只有一个 transition-ids，这里每帧可以有任意多的 transition-ids，每个 transition-ids 都有权重，存储在下面的结构体里：

```
typedef std::vector<std::vector<std::pair<int32, BaseFloat>>> Posterior;
```

如果我们有个 Posterior 类型的对象"post"，post.size()就等于一句话的帧数，和 post[i] 是一串 pair (用向量存储)，每一个 pair 有一个 (transition-id, posterior)构成。

在当前的程序里，有两个方式创建后验概率：

- 用程序 `ali-to-post` 来讲对齐转换为后验,但是生成的 `Posterior` 对象很细小,每帧都有一个单元后验的 `transition-id`
- 用程序 `weight-silence-post` 来修改后验,通常用于减小静音的权重。

未来,当加入 `lattice generation`,我们将会从 `lattice` 生成后验。

与"-ali"相比,读入的后验程序没有后缀。这是为了简洁;读取状态层的后验要考虑对齐信息的这类程序的默认值。

Gaussian-level posteriors

一个句子的高斯级别的后验概率用以下的形式存储:

```
typedef std::vector<std::vector<std::pair<int32, Vector<BaseFloat>>>> GauPost;
```

这个和状态级的后验概率结构体非常相似,除了原来是一个浮点数,现在是一个浮点数向量(代表状态的后验),每个值就是状态里的每个高斯分量。向量的大小跟 pdf 里的高斯分量的数目一样, pdf 对应于转移标识符 `transition-id`,也就是每个 `pair` 的第一个元素。

程序 `post-to-gpost` 把后验概率的结构体转换为高斯后验的结构体,用到模型和特征来计算高斯级别的后验概率。当我们需要用不同的模型或特征计算高斯级别的后验时,这个就有用了。读取高斯后验程序有一个后缀"-gpost"。

Functions for converting HMMs to FSTs

把 HMMs 转换为 FSTs 的整个函数和类的列表可以在 [here](#) 这里找到。

GetHTransducer()

最重要的函数就是 `GetHTransducer()`, 声明如下:

```
fst::VectorFst<fst::StdArc>*
GetHTransducer(const std::vector<std::vector<int32>> &ilabel_info,
const ContextDependencyInterface &ctx_dep,
const TransitionModel &trans_model,
const HTransducerConfig &config,
std::vector<int32> *disambig_syms_left);
```

如果没有介绍 `ilabel_info` 对象, `ContextDependencyInterface` 接口, 和 `fst` 在语音识别里的一些应用基础, 这个函数的很多方面大家会很难理解。这个函数返回一个 FST, 输入标签是 `transition-ids`, 输出标签是代表上下文相关的音素(他们有对象 `ilabel_info` 的索引)。FST 返回一个既有初始状态又有最终状态, 所有的转移都有输出标签(CLG 合理利用)。每一个转移都是一个3状态的 HMM 结构, 和循环返回最初状态。FST 返回 `GetHTransducer()` 仅仅对表示 `ilabel_info` 的上下文相关音素有效, 我们称为特定。这是非常有用的, 因为对于宽的上下文系统, 有很多数量的上下文, 大多是没有被使用。`ilabel_info` 对象可以从 `ContextFst` (代表 C)对象获得, 在合并它和一些东西, 和它含有已经使用过的上下文。我们提供相同的 `ilabel_info` 对象给 `GetHTransducer()` 来获得覆盖我们需要的一个 H 转换器。

注意 [GetHTransducer\(\)](#) 函数不包括自环。这必须通过函数 [AddSelfLoops\(\)](#) 来添加；当所有的图解码优化后，仅仅添加自环是很方便的。

The HTransducerConfig configuration class

The [HTransducerConfig](#) configuration class 控制了 GetHTransducer 行为。

- 变量 `trans_prob_scale` 是一个转移概率尺度。当转移概率包含在图里，他们就包含了这个尺度。命令行就是 `-transition-scale`。对尺度的合理使用和讨论，可以看 [Scaling of transition and acoustic probabilities](#)。
- 一个变量 [reverse](#) 和二个其他变量是可以修改的，如果 "reverse" 选项是 true。"reverse" 选项是为后向解码创建一个 time-reversed FSTs。为了使这个有用，我们需要添加功能到 Arpa 语言模型里，我们将在以后去做。

The function GetHmmAsFst()

函数 [GetHmmAsFst\(\)](#) 需要一个音素上下文窗和返回一个用 transition-ids 作为符号的对应的有限状态接受器。这个在 [GetHTransducer\(\)](#) 使用。函数 [GetHmmAsFstSimple\(\)](#) 有很少的选择被提供，是为了表示在原理上是怎么工作的。

AddSelfLoops()

The [AddSelfLoops\(\)](#) function adds self-loops to a graph that has been created without self-loops. A typical setup is to create the H transducer without self-loops, compose with CLG, do determinization and minimization, and then add the self-loops. This enables more efficient determinization and minimization. The [AddSelfLoops\(\)](#) function has the option to reorder the transitions; see below [Reordering transitions](#) for more details on this. It also takes a transition-probability scale, "self_loop_scale", which does not have to be the same as the normal transition-probability scale; for more on this, see below [Scaling of transition and acoustic probabilities](#).

Adding transition probabilities to FSTs

The [AddTransitionProbs\(\)](#) function adds transition probabilities to an FST. The reason this is useful is so that graphs can be created without transition probabilities on them (i.e. without the component of the weights that arises from the HMM transitions), and these can be added in later; this makes it possible to use the same graph on different iterations of training the model, and keep the transition-probabilities in the graph up to date. Creating the graph without transition-probabilities is accomplished by using a zero value for `trans_prob_scale` (command-line option: `-transition-scale`). In training time, our scripts tend to store

the graphs on disk without the transition probabilities, and then each time we realign we add in the currently valid transition probabilities.

Reordering transitions

The [AddSelfLoops\(\)](#) function takes a boolean option "reorder" which tells it to reorder transition-probabilities so the self-loop comes after the transition out of the state. Where applicable this becomes a boolean command-line option, e.g. you can do `--reorder=true` to enable reordering during graph creation. This option makes the "simple" and "faster" decoders more efficient (see [Decoders used in the Kaldi toolkit](#)), although it is not compatible with the "kaldi" decoder.

The idea of reordering is that we switch the order of the self-loop arcs with all the other arcs that come out of a state, so the self-loop is located at the destination state of each of the other arcs. For this to work, we have to ensure that the FST has certain properties, namely that all the arcs into a particular state must induce the same self-loop (also, a state with a self-loop cannot have input arcs with epsilon inputs, or be the start state). The [AddSelfLoops\(\)](#) function modifies the graphs to ensure that they have this property. A similar property is required even if the "reorder" option is set to false. The graphs created with the "reorder" option are exactly equivalent to the non-reordered graphs in terms of the acoustic and transition-model probabilities you get when decoding an utterance. The transition-ids on the resulting alignment are in a different order, but this does not matter given the ways that we make use of these alignments.

Scaling of transition and acoustic probabilities

There are three types of scaling that can be applied in Kaldi:

Name in code	Name in command-line arguments	Example value (train)	Example value (test)
acoustic_scale	<code>--acoustic-scale=?</code>	0.1	0.08333
self_loop_scale	<code>--self-loop-scale=?</code>	0.1	0.1
transition_scale	<code>--transition-scale=?</code>	1.0	1.0

You may notice that there is no language model scale on this list; everything is scaled relative to the language model. Also we don't support a word insertion penalty, in general (although the "kaldi" decoder does support this). The idea is that the language model represents "real" probabilities so it makes sense to scale everything else relative to them. The scales during training time are the scales we use in decoding to get Viterbi alignments. In general, we use a figure of 0.1 whenever a parameter is not too critical and is expected to be small. The acoustic scale used during test is quite critical and is typically tuned to the task. We now explain what these three scales do:

- The acoustic scale is the scale applied to the acoustics (i.e. to the log-likelihood of a frame given an acoustic state).
- The transition scale is the scale on the transition probabilities, but this only applies to HMM states that have multiple transitions out of them; it applies to the relative weight between such transitions. It does not have any effect for typical topologies.
- The self-loop scale is the scale that we apply to the self-loops. More specifically, when we add the self-loop, let the probability mass given to the self-loop be p and the mass given to the rest be $(1-p)$. We add a self-loop with log-probability $\text{self_loop_scale} * \log(p)$, and add $(\text{self_loop_scale} * \log(1-p))$ to all the other log transition probabilities out of that state. In typical topologies, the self-loop scale is the only scale that matters.

The reason we feel it might make sense to apply a different probability scale to the self-loops versus the normal transition scale is we think they could be dictating the probabilities of events at different timescales. A slightly more subtle argument is the following. All the transition probabilities can be regarded as "real" probabilities (comparable to LM probabilities), because the problem of correlation between acoustic probabilities does not occur for transitions. However, a problem arises because we use the Viterbi algorithm in testing (and in our case, in training too). The transition probabilities would only represent real probabilities when summed over, as in the forward-backward algorithm. We expect this to be more of an issue for the self-loops than for probabilities that dictate the weight to give entirely different paths through the HMM, as in the latter case the acoustic distributions will often be quite disjoint, and the difference between forward-backward and Viterbi will be small.

4.10 kaldi tutorial

以下是我看 kaldi 教程记的些笔记,希望能对你有所帮助(你可以把这个文档当成 kaldi tutorial 的简要翻译) 命令行我都加了下划线。数据准备

这部分基本略过了, 比较简单。

从 data/lang 说起。 data/lang 是由 prepare_lang.sh 生成的。

首先生成的是 words.txt 和 phones.txt, 这是 openfst 格式的 symbol tables(后来我就喜欢直接写英文了, 与其费心思翻译为合适的中文, 不如直接用原词来得原汁原味), 它们是字符串(字符串是指标注文本的单位,) 到整数的映射, 下面这个是日语识别中的部分 words.txt:

```
<eps> 0
a 1
b 2
ch 3
d 4
e 5
f 6
g 7
h 8
i 9
j 10
k 11
m 12
n 13
```

再看 phones 文件夹下以 .csf 为后缀的文件, 分别有 disambig.csf nonsilence.csf optional_silence.csf silence.csf, 它们的内容都是以冒号作为分割。他们是后续命令行中偶尔会用到的选项。

看 phones.txt, 在 data/lang/下。这个是 phone symbol table, 也包含了标准 FST 脚本中的“消歧符号”, 这些符号就是 1, 2, 3,...。我们在此加了个 0 符号, 代替了语言模型中的 epsilon transitions。

L.fst 文件是编译后的 FST 格式的词典 (lexicon)。想看里面有什么信息, 输入: (from s5/)

```
fstprint --isymbols=data/lang/phones.txt --osymbols=data/lang/words.txt data/lang/L.fst | head
```

然后下面是对应的输出:

```
0 1 <eps> <eps> 0.693147182
0 1 sil <eps> 0.693147182
1 2 a a 0.693147182
1 1 a a 0.693147182
1 1 b b 0.693147182
1 2 b b 0.693147182
1 1 ch ch 0.693147182
1 2 ch ch 0.693147182
1 1 d d 0.693147182
1 2 d d 0.693147182
```

G.fst 是个描述这种语言语法结构的 FST, 下面是这个 G.fst 的内容:

```
0 2 21 21 0.000315946876
0 1 29 0 8.03088665
1 29 25 25 12.1202555
1 28 6 6 8.34519005
1 27 15 15 6.83906889
1 26 28 28 5.35383272
1 25 19 19 3.59180236
1 24 16 16 5.82719803
1 23 26 26 4.43024254
1 22 17 17 4.63756752
```

我觉得这应该就是对应词、或 phone 出现次数统计后算的概率。

提取特征

这一段是提取训练特征，首先跑对应的命令，然后来看生成的文件。

看看 `exp/make_mfcc/train/make_mfcc.1.log`，首先给出的跑过的命令行，(kaldi 总是在 log 最上面显示命令行，)

`split_scp.pl` 就是把 `scp` 分成几个小的 `scp` (`.ark` 和 `.scp` 是 kaldi 中两种记录数据的格式，`.ark` 是数据 (二进制文件)，`scp` 是记录对应 `ark` 的路径)

`.ark` 文件一般都是很大的 (因为他们里面是真正的数据)，可以通过下面这条命令来看：
`copy-feats ark:mfcc/raw_mfcc_train.1.ark ark,t:- | head`

以下是对应的输出：

```
NF089001 [
53.54222 -31.82449 -9.899872 -0.02364012 -5.681367 2.072489 -19.41396 -15.6856 14.83652
25.04876 -11.34208 1.64803 9.309975
49.06616 -28.41237 1.188962 -0.5514585 -14.60496 -6.065259 -12.19813 -17.75549 -9.185356
4.032361 -9.320414 1.339788 12.23572
48.10678 -24.78042 8.86155 -4.958602 -4.843619 1.443337 -8.813286 0.4328361 -3.807028
0.8784758 9.743609 7.107668 9.02508
63.17915 -21.53388 -22.33113 5.595533 -12.11316 -4.990936 -14.4953 -10.58425 2.666025
-0.3021607 -11.49867 -1.502062 3.861568
70.48519 -19.16981 -25.84126 10.23085 -15.72831 -5.344745 -22.62867 -12.71542 0.8277165
-4.167449 -19.62204 -5.533485 2.644755
52.99891 -16.45959 0.7519462 -4.386663 3.804989 -1.37611 -24.83507 5.490471 -3.33739
-8.404724 -17.6997 -0.2677126 5.236793
54.01795 -19.39126 -3.082492 -1.624617 -8.421985 -11.15252 -18.0968 -11.92423 -6.684193
-11.88862 -8.570399 -3.803415 5.675081
55.33753 -18.44497 -9.369541 -7.717715 -8.041488 -11.45842 -19.81938 -12.43418 -1.97697
-4.627994 -7.774594 4.451687 7.557387
55.72844 -20.32559 -12.32121 -9.614379 -2.77022 -8.572324 -14.91047 -6.382179 -7.155323
-7.767553 -17.01464 1.11917 -2.572359
```

同名的 `archive (.ark)` 和 `script(.scp)` 文件代表的同一部分数据，注意，这些命令行都有前缀 `"scp:"` 或 `"ark:"`，kaldi 不会自己判断这到底是个 `script` 还是 `archive` 形式，这需要我们加前缀告诉 kaldi 这是什么格式的文件。对于 `code` 而言，这两种格式对它来说都是一样的。

这两种格式都是‘表 (table)’的概念。一个‘表’就是一组有序的事物，前面是识别字符串 (如句子的 id)，一个‘表’不是一个 `c++` 的对象，因为对应不同的需求 (写入、迭代、随机读入) 我们分别有 `c++` 对象来读入数据。

`.scp` 格式是 `text-only` 的格式，每行是个 `key` (一般是句子的标识符 (id)) 后接空格，接这个句子特征数据的路径。

`.ark` 格式可以是 `text` 或 `binary` 格式，(你可以写为 `text` 格式，命令行要加 `'t'`，`binary` 是默认的格式) 文件里面数据的格式是：`key` (如句子的 id) 空格后接数据。

下面是关于 `script` 和 `archive` 的几点说明：

用于说明如何读表的字符串叫做“`rspecifier`”，如 `ark:gunzip -c my/dir/foo.ark.gz`

用于说明如何写入表的字符串叫做“`wspecifier`”，如 `ark,t:foo.ark`

`archive` 可以合并成大的 `archive`，仍然有效

`code` 可以读入这两种格式通过顺序读入或随机读入。用户级的 `code` 只知道它是在迭代还是查找，并不知道它接触的数据格式是什么 (是 `script` 还是 `archive`)

通过随机读入 (`random access`) 来读取 `archive` 数据，内存使用效率较低，要想高效的随机读入 `archive`，那么生成时就生成对应的 `ark` 和 `scp` 文件，读入时通过 `scp` 文件读入。

这部分更多信息在官网 (google 输入 kaldi) Kaldi I/O mechanisms 中。

训练单音子系统 (`monophone`)

输入：

```
gmm-copy --binary=false exp/mono/0.mdl - | less
```

下面是对应的输出：

```
<TransitionModel>
<Topology>
```

```
<TopologyEntry>
<ForPhones>
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
</ForPhones>
<State> 0 <PdfClass> 0 <Transition> 0 0.75 <Transition> 1 0.25 </State>
<State> 1 <PdfClass> 1 <Transition> 1 0.75 <Transition> 2 0.25 </State>
<State> 2 <PdfClass> 2 <Transition> 2 0.75 <Transition> 3 0.25 </State>
<State> 3 </State>
</TopologyEntry>
<TopologyEntry>
<ForPhones>
1
</ForPhones>
<State> 0 <PdfClass> 0 <Transition> 0 0.5 <Transition> 1 0.5 </State>
<State> 1 <PdfClass> 1 <Transition> 1 0.5 <Transition> 2 0.5 </State>
<State> 2 <PdfClass> 2 <Transition> 2 0.75 <Transition> 3 0.25 </State>
<State> 3 </State>
</TopologyEntry>
</Topology>
<Triples> 84
1 0 0
1 1 1
1 2 2
2 0 3
2 1 4
2 2 5
3 0 6
3 1 7
3 2 8
4 0 9
4 1 10
4 2 11
5 0 12
5 1 13
5 2 14
6 0 15
6 1 16
6 2 17
7 0 18
7 1 19
7 2 20
8 0 21
8 1 22
8 2 23
9 0 24
9 1 25
9 2 26
10 0 27
10 1 28
10 2 29
11 0 30
11 1 31
11 2 32
12 0 33
12 1 34
12 2 35
13 0 36
13 1 37
13 2 38
14 0 39
14 1 40
14 2 41
15 0 42
```

15 1 43
15 2 44
16 0 45
16 1 46
16 2 47
17 0 48
17 1 49
17 2 50
18 0 51
18 1 52
18 2 53
19 0 54
19 1 55
19 2 56
20 0 57
20 1 58
20 2 59
21 0 60
21 1 61
21 2 62
22 0 63
22 1 64
22 2 65
23 0 66
23 1 67
23 2 68
24 0 69
24 1 70
24 2 71
25 0 72
25 1 73
25 2 74
26 0 75
26 1 76
26 2 77
27 0 78
27 1 79
27 2 80
28 0 81
28 1 82
28 2 83

[illegible]


```

-0.2876821 -1.386294 -0.2876821 -1.386294 -0.2876821 -1.386294 -0.2876821 -1.386294 ]
</LogProbs>
</TransitionModel>
<DIMENSION> 39 <NUMPDFS> 84 <DiagGMM>
<GCONSTS> [ -94.49178 ]
<WEIGHTS> [ 1 ]
<MEANS_INVVARS> [
-0.005838452 -0.0107621 0.007483369 0.002269829 0.01010145 0.001220717 -0.002948278
0.004102771 -0.009732663 0.005548568 -0.00846673 0.003018271 0.002561719 -0.001072273
-0.0003676935 0.0009567018 0.0004904701 0.001004559 0.0006702438 0.002065411
0.001736847 -0.0004884294 -0.0001839283 0.000573744 -6.096664e-06 0.0008038587
0.000548786 0.0005939789 -0.001607142 -0.0008620437 0.0002163016 -0.0002253224
0.0009042169 0.0007718542 0.0001247094 -0.0003084296 -0.001637235 0.0004870822
0.002509772 ]
<INV_VARS> [
0.002302196 0.004163655 0.005391662 0.002574274 0.003217114 0.002863562 0.005842444
0.004294651 0.003149447 0.005013018 0.004209091 0.00450796 0.006656926 0.06917789
0.07472826 0.08657464 0.05778877 0.05123304 0.05053843 0.08058306 0.07275081 0.0605317
0.07780149 0.07870268 0.08504011 0.1153897 0.4811986 0.4609138 0.5185907 0.3938675
0.3216299 0.3013106 0.4305585 0.4167711 0.3496516 0.4075994 0.4405422 0.4886224
0.6218032 ]
</DiagGMM>
<DiagGMM>
<GCONSTS> [ -94.49178 ]
<WEIGHTS> [ 1 ]
<MEANS_INVVARS> [
-0.005838452 -0.0107621 0.007483369 0.002269829 0.01010145 0.001220717 -0.002948278
0.004102771 -0.009732663 0.005548568 -0.00846673 0.003018271 0.002561719 -0.001072273
-0.0003676935 0.0009567018 0.0004904701 0.001004559 0.0006702438 0.002065411
0.001736847 -0.0004884294 -0.0001839283 0.000573744 -6.096664e-06 0.0008038587
0.000548786 0.0005939789 -0.001607142 -0.0008620437 0.0002163016 -0.0002253224
0.0009042169 0.0007718542 0.0001247094 -0.0003084296 -0.001637235 0.0004870822
0.002509772 ]
<INV_VARS> [
0.002302196 0.004163655 0.005391662 0.002574274 0.003217114 0.002863562 0.005842444
0.004294651 0.003149447 0.005013018 0.004209091 0.00450796 0.006656926 0.06917789
0.07472826 0.08657464 0.05778877 0.05123304 0.05053843 0.08058306 0.07275081 0.0605317
0.07780149 0.07870268 0.08504011 0.1153897 0.4811986 0.4609138 0.5185907 0.3938675
0.3216299 0.3013106 0.4305585 0.4167711 0.3496516 0.4075994 0.4405422 0.4886224
0.6218032 ]
</DiagGMM>
<DiagGMM>
<GCONSTS> [ -94.49178 ]
<WEIGHTS> [ 1 ]

```

首先给出的 topo 的信息, 有一个 phone 和其他 phone 的 topology 不同, 通过对比 phones.txt, 可知这个不同的 phone 是 sil (代表 silence), topo 文件的惯例是第一个状态是初始状态 (概率为 1), 最后一个状态是结束状态 (概率为 1) 在这里面, 明显, -1 是初始状态, 0, 1, 2 是 HMM 中间的转移状态, 3 是结束状态。

.mdl 文件的惯例是包含两部分信息: 一部分的类型是 Transition Model (转换模型), 包含拓扑信息 (topo), 作为 HMMtopology 的一个成员变量, 另一部分是相关模型类型 (叫 AmGmm), 这种类型的文件不是‘表’, 写入是 binary or text, 取决于命令行选项 --binary=true or --binary=false, ‘表’就是指 script 和 archive。

看上面数据, 会发现, 0.mdl 是初始化模型, 所以参数都是初始化的, 这个模型训练 40 次, 所以 40.mdl 中的概率是最终的参数。

以上信息更多见官网 HMM topology and transition modeling。

再提重要的一点: 在 kald 中 p.d.f.'s 使用数字标识符表示的, 从 0 开始 (这些数字我们叫做 pdf-ids), 在 HTK 中他们没有名字。mdl 文件没有足够的信息能在 context-dependent phones 和 pdf-ids 间建立映射, 为看这个, 看 tree 文件, 输入:
copy-tree --binary=false exp/mono/tree - | less

以下是输出:

3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 6

这是对训练数据做的维特比 alignment(对齐)。每句话是一行,再看看上面 exp/mono/tree 文件, p.d.f. id 数值最大的是 83, 而这里的数值远大于这个, 是因为, alignment 文件用的不是

p.d.f. id, 这里用的是更细分的标识符 (identifier), 称作“transition-id”, 这些 id 将 phone 和它们在拓扑原型结构中的转移概率也编码进来了。若想知道“transition-id”是什么, 输入:

show-transitions data/lang/phones.txt exp/mono/0.mdl

以下是对应的输出:

Transition-state 1: phone = sil hmm-state = 0 pdf = 0
Transition-id = 1 p = 0.5 [self-loop]
Transition-id = 2 p = 0.5 [0 -> 1]
Transition-state 2: phone = sil hmm-state = 1 pdf = 1
Transition-id = 3 p = 0.5 [self-loop]
Transition-id = 4 p = 0.5 [1 -> 2]
Transition-state 3: phone = sil hmm-state = 2 pdf = 2
Transition-id = 5 p = 0.75 [self-loop]
Transition-id = 6 p = 0.25 [2 -> 3]
Transition-state 4: phone = a hmm-state = 0 pdf = 3
Transition-id = 7 p = 0.75 [self-loop]
Transition-id = 8 p = 0.25 [0 -> 1]
Transition-state 5: phone = a hmm-state = 1 pdf = 4
Transition-id = 9 p = 0.75 [self-loop]
Transition-id = 10 p = 0.25 [1 -> 2]
Transition-state 6: phone = a hmm-state = 2 pdf = 5
Transition-id = 11 p = 0.75 [self-loop]
Transition-id = 12 p = 0.25 [2 -> 3]
Transition-state 7: phone = b hmm-state = 0 pdf = 6
Transition-id = 13 p = 0.75 [self-loop]
Transition-id = 14 p = 0.25 [0 -> 1]
Transition-state 8: phone = b hmm-state = 1 pdf = 7
Transition-id = 15 p = 0.75 [self-loop]
Transition-id = 16 p = 0.25 [1 -> 2]
Transition-state 9: phone = b hmm-state = 2 pdf = 8
Transition-id = 17 p = 0.75 [self-loop]
Transition-id = 18 p = 0.25 [2 -> 3]
Transition-state 10: phone = ch hmm-state = 0 pdf = 9
Transition-id = 19 p = 0.75 [self-loop]
Transition-id = 20 p = 0.25 [0 -> 1]
Transition-state 11: phone = ch hmm-state = 1 pdf = 10
Transition-id = 21 p = 0.75 [self-loop]
Transition-id = 22 p = 0.25 [1 -> 2]
Transition-state 12: phone = ch hmm-state = 2 pdf = 11
Transition-id = 23 p = 0.75 [self-loop]
Transition-id = 24 p = 0.25 [2 -> 3]

.....

Transition-state 82: phone = z hmm-state = 0 pdf = 81
Transition-id = 163 p = 0.75 [self-loop]
Transition-id = 164 p = 0.25 [0 -> 1]
Transition-state 83: phone = z hmm-state = 1 pdf = 82
Transition-id = 165 p = 0.75 [self-loop]
Transition-id = 166 p = 0.25 [1 -> 2]
Transition-state 84: phone = z hmm-state = 2 pdf = 83
Transition-id = 167 p = 0.75 [self-loop]
Transition-id = 168 p = 0.25 [2 -> 3]

显然, 上面这个是训练前的初始状态

为了增加可读性, 输入:

show-alignments data/lang/phones.txt exp/mono/40.mdl exp/mono/40.occs | less
(.occs 文件是指 occupation counts)

以下是对应的输出:

Transition-state 1: phone = sil hmm-state = 0 pdf = 0
 Transition-id = 1 p = 0.934807 count of pdf = 1.13866e+06 [self-loop]
 Transition-id = 2 p = 0.0651934 count of pdf = 1.13866e+06 [0 -> 1]
 Transition-state 2: phone = sil hmm-state = 1 pdf = 1
 Transition-id = 3 p = 0.889584 count of pdf = 672302 [self-loop]
 Transition-id = 4 p = 0.110416 count of pdf = 672302 [1 -> 2]
 Transition-state 3: phone = sil hmm-state = 2 pdf = 2
 Transition-id = 5 p = 0.7137 count of pdf = 259284 [self-loop]
 Transition-id = 6 p = 0.2863 count of pdf = 259284 [2 -> 3]
 Transition-state 4: phone = a hmm-state = 0 pdf = 3
 Transition-id = 7 p = 0.713307 count of pdf = 390711 [self-loop]
 Transition-id = 8 p = 0.286693 count of pdf = 390711 [0 -> 1]
 Transition-state 5: phone = a hmm-state = 1 pdf = 4
 Transition-id = 9 p = 0.594051 count of pdf = 275931 [self-loop]
 Transition-id = 10 p = 0.405949 count of pdf = 275931 [1 -> 2]
 Transition-state 6: phone = a hmm-state = 2 pdf = 5
 Transition-id = 11 p = 0.594987 count of pdf = 276569 [self-loop]
 Transition-id = 12 p = 0.405013 count of pdf = 276569 [2 -> 3]
 Transition-state 7: phone = b hmm-state = 0 pdf = 6
 Transition-id = 13 p = 0.590539 count of pdf = 19660 [self-loop]
 Transition-id = 14 p = 0.409461 count of pdf = 19660 [0 -> 1]
 Transition-state 8: phone = b hmm-state = 1 pdf = 7
 Transition-id = 15 p = 0.417553 count of pdf = 13821 [self-loop]

这个用的是 40.mdl，得到的概率都是模型最后用的转移概率
 要想了解更多关于 HMM 拓扑结构，转移标识符 (transition-ids)，转移模型之类的，看官网
 HMM topology and transition modeling 这部分。

接下来看看训练是如何进行的，输入：
 grep Overall exp/mono/log/acc.{?.? ,?.??,???,??,??}.log

以下是输出的最后一部分：

exp/mono/log/acc.35.10.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115)
 exp/mono/log/acc.37.12.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -99.1242 over 595815 frames.
 exp/mono/log/acc.38.10.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -99.0045 over 666753 frames.
 exp/mono/log/acc.38.11.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -95.769 over 793715 frames.
 exp/mono/log/acc.38.12.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -99.0953 over 595815 frames.
 exp/mono/log/acc.39.10.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -98.9901 over 666753 frames.
 exp/mono/log/acc.39.11.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -95.7472 over 793715 frames.
 exp/mono/log/acc.39.12.log:LOG (gmm-acc-stats-ali:main():gmm-acc-stats-ali.cc:115) Overall
 avg like per frame (Gaussian only) = -99.0786 over 595815 frames.

你可以看到每次迭代的声学似然概率。

就写到这儿吧。

hope it is helpful (well, it is to me), hope you like it.

5 附录

5.1 kaldi 上搭建 TIMIT 基线系统

因工作需要，要搭建几个基于 kaldi 的语音识别系统，也正在学习 kaldi，把过程和经验同大家分享，希望互相帮助共同学习，欢迎与我联系，若需转载请注明出处。

注：kaldi 自带的示例提供了实验流程，但是其 monophone 的 HMM-GMM 基线是 39 phones 的，而后的 DNN 基线是 triphone 的，和我们目前需要的 61 monophone 不一致，故需要重新搭建，本文档的基线系统包括 61 monophone 的 HMM-GMM 系统和 61 monophone, 183 states 的 DNN 基线，包括 fMLLR 等改进技术可以参见其它相关示例程序加入。

一. Kaldi 下搭建 TIMIT 数据集的 HMM-GMM 基线系统

1. 复制 egs/timit/s5 到同级目录下并改名为 61monophone，方便后面直接在脚本里面改动，可以在 run.sh 里面修改或者新建一个 61monophone_run.sh。

2.

`./cmd.sh`

本机运行选择 run it locally... 注释掉其余部分

`./path.sh`

3. Data & Lexicon & Language Preparation

`timit=/home/sfxue/data/TIMIT`

`local/timit_data_prep.sh $ timit`

将会新产生两个目录/data/local/data 和/data/local/nist_lm 后者用不到不去管它，这个脚本主要是对数据集的结构、数据集完整性进行检查然后生成一些列表和文件，示例本身是要将 61 phone 的标注列表转成 39 phone 的，而我们需要的是 61 phone 的，所以要对其中的那条转换命令进行修改，修改后的命令为：

`cat ${x}.trans | $local/timit_norm_trans.pl -i - -m $conf/phones.61-61-61.map -to 61 | sort > $x.txt || exit 1;`

phones.61-61-61.map 文件是 phones.60-48-39.map 修改得来，我们的目的只是将 h#转换成 sil，而 q 保留，同时 timit_norm_trans.pl 也需要稍作修改来让命令可以执行。打开/data/local/data 里面的文件看看，应该可以理解它们的功能。

`local/timit_prepare_dict.sh`

示例中的这段注释说的很明显了，打开这几个文件看看就行了：

The parts of the output of this that will be needed are

[in data/local/dict/]

lexicon.txt

extra_questions.txt

nonsilence_phones.txt

optional_silence.txt

silence_phones.txt

```
utils/prepare_lang.sh --position-dependent-phones false --num-sil-states 3
data/local/dict "sil" data/local/lang_tmp data/lang
local/timit_format_data.sh
```

4. MFCC Feature Extraction & CMVN

```
# Now make MFCC features.
mfccdir=mfcc
use_pitch=false
use_ffv=false
for x in train dev test; do
steps/make_mfcc.sh --cmd "$train_cmd" --nj 10 data/$x exp/make_mfcc/$x $mfccdir
    || exit 1;
steps/compute_cmvn_stats.sh data/$x exp/make_mfcc/$x $mfccdir
done
```

生成 mfcc 特征的命令是 `compute-mfcc-feats`，默认 channel 是 13，特征在此处没有做差分，后面在每个实验中可以选择差分操作。`--nj 10` 是并行任务数，后面三个参数分别是数据目录，`exp` 目录和特征存储目录，默认压缩存储所以占用空间不大。CMVN 是句子级的。

4. Monophone Training

```
steps/train_mono.sh --nj "$train_nj" --cmd "$train_cmd" data/train data/lang
exp/mono || exit 1;
utils/mkgraph.sh --mono data/lang_test_bg exp/mono exp/mono/graph || exit 1;
```

该步进行 MLE 的训练，首先要弄清楚模型结构，可以用 `gmm-copy -binary=false exp/mono/0.mdl 0.model.txt` 来讲模型文件转成可阅读的 txt 文件，浏览里面的结构可以看到和 HTK 的差异，具体可以看网站上的指南文档。

Kaldi 的 MLE 训练中每个状态的混合高斯数目是不一样的，在 `train_mono.sh` 中有几个比较关键的参数设定：

```
num_iters= # Number of iterations of training
max_iter_inc= # Last iter to increase #Gauss on. (最后涨高斯的一次迭代)
totgauss= # Target #Gaussians. (训练最终希望达到的高斯总数)
realign_iters= (re-alignment 分别在哪些迭代进行)
```

6.decode

```
steps/decode.sh --nj "$decode_nj" --cmd "$decode_cmd" exp/mono/graph data/test
exp/mono/decode_test || exit 1;
```

解码本身没有什么需要修改，只是在 score 的时候需要注意我们要将 61 monophone 映射到 39 monophone 进行。在 `sore.sh` 文件里修改两处：

```
phonemap="conf/decodephones.61-48-39.map"
```

```
cat $data/text | sed 's:::g' | sed 's:::g' | local/decodetimit_norm_trans.pl -i -m
conf/decodephones.61-48-39.map -to 39 | sort > $dir/scoring/test_filt.txt
```

`decodephones.61-48-39.map` 是修改自 `phones.60-48-39.map`，只是将 `h#` 换成 `sil`。

`decodetimit_norm_trans.pl` 修改自 `timit_norm_trans.pl` 只是保留了丢掉 `q` 的操作。

PER 识别结果 28.11。

二. Kaldi 下搭建 TIMIT 数据集的 DNN 基线

流程主要包括三个部分 force alignment, pre-train 和 fine-tune。

1. Force Alignment

```
traindata_dir=data/train
language_dir=data/lang
src_dir=exp/mono
ali_dir=exp/mono_ali
#steps/align_si.sh --nj 10 --cmd $train_cmd $traindata_dir $language_dir $src_dir
$ali_dir
```

2. pre-train

```
# Pre-train the DBN
dir=exp/pretrain-dbn
(tail --pid=$$ -F $dir/_pretrain_dbn.log 2>/dev/null)&
$cuda_cmd $dir/_pretrain_dbn.log \
steps/pretrain_dbn.sh --nn_depth 6 --delta_order 2 --hid-dim 1024 --rbm-iter 15
data/train $dir
```

kaldi 的 pre-train 使用一个初始学习速率，然后随着迭代递减，同时 momentum 从 0.5 涨到 0.9，我目前使用默认的 B-B 0.4，G-B 0.01 参数，效果还可以。迭代次数 G-B 是 B-B 的两倍，也就是 30,15,15,.....。--delta_order 2 表示使用二阶差分，kaldi 提供的脚本里有个 bug，不能够使用差分选项，否则 CMVN 时会出错，找到出错的地方仿照 train_net.sh 修改即可。

3. fine-tune

```
dir=data/train
# split the data : 90% train 10% cross-validation (held-out)
utils/subset_data_dir_tr_cv.sh $dir ${dir}_tr90 ${dir}_cv10

# Train the MLP
dir=exp/pretrain-dbn_dnn
ali=exp/mono_ali
feature_transform=exp/pretrain-dbn/final.feature_transform
dbn=exp/pretrain-dbn/6.dbn
#with pretrain
(tail --pid=$$ -F $dir/_train_nnet.log 2>/dev/null)&
$cuda_cmd $dir/_train_nnet.log \
steps/train_nnet.sh --feature-transform $feature_transform --delta_order 2 --dbn
$dbn --hid-layers 0 --learn-rate 0.008 \
data/train_tr90 data/train_cv10 data/lang $ali $ali $dir || exit 1;
## decode (reuse HCLG graph)
steps/decode_nnet.sh --nj 10 --cmd $decode_cmd --config conf/decode_dnn.config
--acwt 0.1 \
exp/mono/graph data/test $dir/decode || exit 1;
```

--delta_order 2 表示使用二阶差分，--dbn \$dbn 表示使用之前的 dbn 初始化 DNN 除最后一层的网络，--hid-layers 0 要特别注意指的是增加一个隐层数为 0，

也即输入输出共两层的网络，也就是我们一般加在 dbn 之后的随机层，如果没有之前的--dbn \$dbn 选项的话要训练一个 6 隐层的网络这里也需要设成 6，个人感觉 kald 在这个参数的设计上不够明智，很容易让人出错。--learn-rate 0.008 表示初始学习速率而 0.008。

最终的识别结果：22.95

5.2 kald 里的 voxforge

Some weeks ago there was a question on the Kaldi's mailing list about the possibility of creating a Kaldi recipe using VoxForge's data. For those not familiar with it, VoxForge is a project, which has the goal of collecting speech data for various languages, that can be used for training acoustic models for automatic speech recognition. The project is founded and maintained, to the best of my knowledge, by Ken MacLean and thrives thanks to the generous contributions of great number of volunteers, who record sample utterances using the Java applet available on the website, or submit pre-recorded data. As far as I know this is the largest body of free(in both of the usual senses of the word) speech data, readily available for acoustic model training. It seemed like a good idea to develop a Kaldi recipe, that can be used by people who want to try the toolkit, but don't have access to the commercial corpora. My previous recipe, based on freely available features for a subset of RM data can be also used for that purpose, but it has somewhat limited functionality. This post describes the data preparation steps, specific to VoxForge's data.

Prerequisites

As usual the following instructions assume you are using Linux operating system. The scripts try to install some of the external tools needed for their work, but also assume you have some tools and libraries already installed. These are things that should be either already available on your system or can be easily installed through its package manager. I will try to enumerate the main dependencies when describing the respective part of the recipe making use of them.

If you don't have Kaldi installed you need to download and install it by following the steps in the [documentation](#). If you already have it installed you need to make sure you have a recent version, which includes this recipe. It can be found in `egs/voxforge/s5` under Kaldi's installation directory. From this point on all paths in this blog entry will be relative to this directory, unless otherwise noted.

Downloading the data

Before doing anything else you should set `DATA_ROOT` variable in `./path.sh` to point to a directory residing on a drive, which has enough free space to store several tens of gigabytes of data (about 25GB should be enough in the default recipe config at the time of writing).

You can use the `./getdata.sh` to download VoxForge's data. As many other parts of the

recipe it hasn't been extensively tested, but hopefully will work. It assumes you have `wget` installed on your system and downloads the 16KHz versions of the data archives to `$ {DATA_ROOT}/tgz` and extracts them to `$ {DATA_ROOT}/extracted`. If you want to free several gigabytes by deleting the archives after the extraction finishes, you can add a `"--deltgz true"` parameter:

`view sourceprint?`

```
./getdata.sh --deltgz true
```

Configuration and starting the recipe

It's usually recommended to run Kaldi recipes by hand by copy/pasting the commands in the respective `run.sh` scripts. If you do this be sure to source `path.sh` first in order to set the paths and `LC_ALL=C` variable. If this environment variable is not set you may run into sometimes hard to diagnose problems, due to the different sorted orders used in other locales. Or if you are like me and prefer to just start `/bin/bash run.sh` to execute all steps automatically, you can do this too but you may want to modify several variables first. If you happen to have a cluster of machines with Sun Grid Engine installed you may want to modify the train and decode commands in `cmd.sh` to match your config. A related parameter in `run.sh` is `njobs`, which defines the maximum number of parallel processes to be executed. The current default is 2 which is perhaps too low even for a relative new home desktop machine.

There are several parameters toward the start of `run.sh` script, that I will explain when discussing the parts of the recipe they affect.

The recipe is structured according to the currently recommended("s5") Kaldi script style.

The main sub-directories used are:

- `local/` - hosts scripts that are specific to each recipe. This is mostly data normalization stuff, which is taking the information from the particular speech database and is transforming it into the `files/data` structures that the subsequent steps expect. The work of using new data with Kaldi, including this recipe, involves writing and modifying scripts that fall into this category.
- `conf/` - small configuration files, specifying things like e.g. feature extraction parameters and decoding beams.
- `steps/` - scripts implementing various acoustic model training methods, mostly through calling Kaldi's binary tools and the scripts in `"utils/"`
- `utils/` - scripts performing small, low-level task, like adding disambiguation symbols to lexicons, converting between symbolic and integer representation of words etc.
- `data/` - This is where various metadata, produced at the recipe run time is stored. Most of it is result from the work of the scripts in `local/`
- `exp/` - The most important output of the recipe goes there. This includes acoustic models and recognition results.

- tools/ - A non-standard directory, specific to this recipe, that I am using to put various external tools into (more about these later).

Note: Keep in mind that steps/ and utils/ are shared between all "s5" scripts and are just symlinks pointing to egs/wsj/s5. That means you should be careful if you make changes there, as this may affect the other "s5" recipes.

Data subset selection

The recipe has the option to train and test on a subset of the VoxForge's data. For (almost) every submission directory there is a etc/README file, with meta-information like pronunciation dialect and gender of the speaker. For English there are many pronunciation dialects, but from an (admittedly very limited) sample of the data I've got the impression that some of the recordings with dialect set to say "European English" and "Indian English" sound distinctively non-native. So you have the option to select a subset of the dialects if you like. In my limited experience it seems, that the speech tagged with "American English" has relative lower percentage of non-native speech intermixed. The default for the recipe is currently set to produce an acoustic model with good coverage over (presumably) native English speakers:

view sourceprint?

```
dialects="((American)|(British)|(Australia)|(Zealand))"
```

The local/voxforge_select.sh script creates symbolic links to the matching submission subdirectories in \$ {DATA_ROOT}/selected and the subsequent steps work just on this data.

If you want to select all of VoxForge's English speech you should perhaps set this to:

view sourceprint?

```
dialects="English"
```

I don't have experience with the non-English audio at VoxForge, so I can't comment how the scripts should be modified to work with it.

Mapping anonymous speakers to unique IDs

VoxForge allows for anonymous speaker registration and speech submission. This is not ideal from Kaldi's scripts viewpoint, because they perform various speaker-dependent transforms. The "anonymous" speech is recorded under different environment/channel conditions (microphones used, background noise etc), by speakers that may be both males and females and have different accents. So instead of lumping all this data together I decided to give each speaker unique identity. The script local/voxforge_fix_data.sh renames all "anonymous" speakers to "anonDDDD"(D is a decimal digit), based on the submission date. This is not entirely precise of course because it may give two different IDs to the same speaker who made recordings on two different dates, and also give the same ID to two or more different "anonymous" speakers who happened to submit speech on the same date.

Train/test set splitting and normalizing the metadata

The next steps is to split the data into train and test sets and to produce the relevant

transcription and speaker-dependent information. These steps are performed by a rather convoluted and not particularly efficient script called `local/voxforge_data_prep.sh`. The number of speakers to be assigned to the test set is defined in `run.sh`. The actual test set speakers are chosen at random. This is probably not an ideal arrangement, because the speakers will be different each time `voxforge_data_prep.sh` is started and probably the WER for the test set will be slightly different too. I think this is not very important in this case, however because VoxForge still doesn't have predefined high-quality train and test sets and test time language model. It has something at [here](#), but there are speakers, for which there are utterances in both sets and I wanted the sets to be disjoint.

The script assumes that you have `FLAC` codec installed on your machine, because some of the files are encoded in this lossless compression format. One solution is to convert the files beforehand into `WAV` format, but the recipe is using a nifty feature of Kaldi called `extended filenames` to convert the audio on-demand. For example `data/local/train_wav.scp`, which contains a list of files to be used for feature extraction, there are lines like:

```
benkay_20090111-ar-ar-09 flac -c -d --silent  
/media/secondary/voxforge/selected/benkay-20090111-ar/flac/ar-09.flac |
```

This means in effect, that when `ar-09.flac` needs to be converted into MFCC features `flac` is invoked first to decode the file, and the decoded `.wav-format` stream is passed to `compute-mfcc-feats` using Unix pipes.

It seems there are missing or not properly formatted transcripts for some speakers and these are ignored by the data preparation script. You can see these errors in `exp/data_prep/make_trans.log` file.

Building the language model

I decided to use `MITLM` toolkit to estimate the test-time language model. `IRSTLM` is installed by default under `$ KALDI_ROOT/tools` by Kaldi installation scripts, but I had a mixed experience with this toolkit before and decided to try a different tool this time. A script called `local/voxforge_prepare_lm.sh` installs `MITLM` in `tools/mitlm-svn` and then trains a language model on the train set. The installation of `MITLM` assumes you have `svn`, `GNU autotools`, `C++` and `Fortran` compilers, as well as `Boost C++` libraries installed. The order of the LM is given by a variable named `lm_order` in `run.sh`.

Preparing the dictionary

The script, used for this task is called `local/voxforge_prepare_dict.sh`. It downloads the CMU's pronunciation dictionary first, and prepares a list of the words that are found in the train set, but not in `cmudict`. Pronunciations for these words are automatically generated using `Sequitur G2P`, which is installed under `tools/g2p`. The installation assumes you have `NumPy`, `SWIG` and `C++` compiler on your system. Because the training of `Sequitur` models takes a lot of time this script is downloading and using a pre-built model trained on `cmudict` instead.

Decoding

These were the most important steps specific to this recipe. Most of the rest is just borrowed from WSJ and RM scripts.

The decoding results for some of the steps are as follows:

`view sourceprint?`

`exp/mono/decode/wer_10`

`%WER 64.96 [5718 / 8803, 320 ins, 1615 del, 3783 sub]`

`%SER 96.29 [935 / 971]`

`exp/tri2a/decode/wer_13`

`%WER 44.63 [3929 / 8803, 412 ins, 824 del, 2693 sub]`

`%SER 87.95 [854 / 971]`

`exp/tri2b_mmi/decode_it4/wer_12`

`%WER 38.94 [3428 / 8803, 401 ins, 753 del, 2274 sub]`

`%SER 81.77 [794 / 971]`

These are obtained using a monophone(mono), a maximum likelihood trained triphone(tri2a) and a discriminatively trained(tri2b_mmi) triphone acoustic models. Now, I know the results are not very inspiring and may look even somewhat disheartening, but keep in mind they were produced using a very poor language model - a bigram estimated just on the quite small corpus of training set transcripts. Another factor contributing to the relatively poor results is that a typically about 2-3% of the words(depending on the random test set selection) are found just in test set, and thus considered out-of-vocabulary. Not only they don't have any chance to be recognized themselves, but are also reducing the chance for the words surrounding them to be correctly decoded.

One thing that I haven't had the time to try yet is to change the number of states and Gaussians in the acoustic model. As I've mentioned the training and decoding commands in `run.sh` were copied from RM recipe. My guess is that if the states and PDFs are increased somewhat that will lower the WER by 2-3% at least. This may be done by tweaking the relevant lines in `run.sh`, e.g.

`view sourceprint?`

`# train tri2a [delta+delta-deltas]`

`steps/train_deltas.sh --cmd "$train_cmd" 1800 9000 \`

`data/train data/lang exp/tri1_ali exp/tri2a || exit 1;`

In the above 1800 is the number of PDFs and 9000 is the total number of Gaussians in the system.

There is also something else worth mentioning here. As you may have noticed already I am using relatively small number of speakers(40) for testing. This is mostly because the set of prompts used in audio submission applet is limited and so there are many utterances duplicated across speakers. Because the test utterances are excluded when training the LM, we don't want too many test set speakers, because this would also

mean less text for training language model and thus even worse performance.
Just to check everything is OK with the acoustic model I ran some tests using language models trained on both the train and the test set (WARNING: this is considered to be "cheating" and a very bad practice - please don't do this at home!). The results were about 3% WER with a trigram and about 17% with a bigram model with a discriminatevely trained AM.

版权所有 翻版必究

5.3 在 vs 2013 中编译 kaldi

@王立军在 cnblog 上发表。

在 visual studio 2013 中编译 kaldi 的主要困难的地方在于正确编译 kaldi 的依赖库，主要是：openfst, ATLAS, pthreadVC2，其中又属 ATLAS 最难编译，因此，为求整个编译逻辑清晰简洁，在此将忽略 kaldi 所有依赖库的编译，如果有时间，我会直接提供其所依赖的库。

编译环境：

- 1.操作系统：windows8 （推荐）
- 2.编译工具：visual studio2013
- 3.其他：要安装完整的 CygWin 环境

编译步骤：

- 1.新建解决方案：打开 visual studio 2013，新建一个 Win32项目，不妨将解决方案名称名为 kalid_vs13。可以先简单写个 HelloWorld 程序，用来测试项目是否成功建立
- 2.新建项目：在解决方案 kalid_vs13 中用 DLL 项目模板添加一个新的项目 kalid-lib
- 3.添加源文件：把 kald-trunk/src/下的*.cc 文件加入项目 kalid-lib 中，注意这里不要把一些*bin 目录里的测试文件以及其他的一些文件名中含有关键字 test 的*.cc 文件加入 kalid-lib 中
- 4.配置 kalid-lib 的相关属性，包括：1) 配置头文件的引用目录，2) 预处理器定义的宏(HAVE_ATLAS)，3) 依赖库的引用(openfst.lib,atlas.lib,pthreadVC2.lib)
- 5.其他配置：因为 nnet1 和 nnet2 目录下有相同的文件名：nnet-nnet.cc 和 nnet-componet.cc，所以需要配置他们各自的“输出文件”中的“对象文件名”，比如可以将 nnet1 中的 nnet-nnet.cc 的对象文件名设为\$(IntDir)nnet1\，而将 nnet2 中的 nnet-nnet.cc 的对象文件名设为\$(IntDir)nnet2\，其他重复文件类推

备注：

1. 要想顺利编译，还需要修改一些 kaldi 源文件，有时间我会向官方提交一些 patch，这样大家就能顺利编译 kaldi 了
2. 以上的编译环境是我自己的，其他的环境是否能行我就不确定了
3. 这篇随笔主要叙述 kaldi 的整体编译思路，暂时不会关注编译细节！

6 其他资料和资源

- 1.kaldi 主页: <http://kaldi.sourceforge.net/>
- 2.Povey 主页: <http://www.danielpovey.com/>
- 3.Kaldi+pdnn: <http://www.cs.cmu.edu/~ymiao/kaldipdnn.html>
- 4.@马源的有关 kaldi 的介绍: <http://blog.csdn.net/sytxsybm/article/category/1867123>
- 5.@神牛的 csdn 博客里关于 kaldi 的内容:
<http://blog.csdn.net/lifeitengup/article/category/2095113>
以及新的博客: <http://ftli.farbox.com/>
- 6.@天大的同学的博客: <http://blog.csdn.net/jojzhangju/article/category/2121691>
- 7.wbglearn 的博客 (我的博客): <http://blog.csdn.net/wbgxx333>

欢迎大家在这里添加资料和资源, 如果你觉得有哪些比较好的, 欢迎和我联系, 我将逐渐的完善这部分。

7 版本更新日志

2014.7.14	V0.1	kaldi 全部资料第一版发布
2014.8.16	V0.2	更新：1.添加新的翻译：决策树在 kaldi 上的应用 2.kaldi 在 vs 上的调试建议

版权所有 翻版必究