一．Kaldi 训练 HMM/GMM 的过程

1. mono-phone model

（1）先生成模型的拓扑结构(topo)，指定每个音素（建模单元）的 HMM 状态数及其 id 等; 生成训练图 training graph;

（2）根据 topo 来生成(gmm-init-mono)初始模型 0.mdl 和 tree。（可选：使用少量的训练集来初始化 0.mdl）；

（3）将训练语音帧往 HMM 状态上做平均化对齐(align-equal-compiled)，并统计对齐后的统计量(gmm-acc-stats-ali)，得到 0.*.acc； (*指的是 job id)

（4）根据统计结果，在 0.mdl 基础上更新模型参数(gmm-est)，得到新的模型 1.mdl;（可选：输出状态的 occupancy(1.occs))

（5）使用 x.mdl 对训练语音做对齐(gmm-align-compiled)，得到对其结果 ali.*.gz; (这是迭代里的第一步，x.mdl 指的是当前模型，第一次进入时 x.mdl 就是 1.mdl)

（6）统计对齐后的统计量(gmm-acc-stats-ali)，得到 x.*.acc;

（7）根据统计结果，在 x.mdl 基础上更新模型参数(gmm-est)，得到新的模型 x+1.mdl;（可选：输出状态的 occupancy(x+1.occs)）

（8）判断是否已经到达最大迭代步骤，如果达到了就训练结束，否则返回(6)继续迭代。

流程图见图 1

２．model tree 等输出成文本形式（可读格式）

gmm 模型： gmm-copy

dnn 模型: nnet-copy

tree： copy-tree

draw-tree (invoke it and learn how to print specific format(ps or pdf))

matrix： copy-matrix

*.acc: while invoking acc-tree-stats、gmm-acc-stats-ali command, add --binary=false

training graph:compile-train-graphs 的输出形式为 ark,t 这样会得到文本格式的训练图。但是这个图中的输入和输出不是 symbol 而是它们对应的 int 值。

所以，可以直接 fstcompile（不加 isymbols 和 osymbols 选项），生成对应的 fst，然后再 fstdraw（isymbols 很可能是 transition-ids，可以把 transition-ids 和自己映射，osymbols 是 words.txt）等输出 ps或者 pdf。

G.fst,L.fst,LG.fst can be created in normal way.

C.fst: (in data/graph/lang/tmp)

(1)subseq_sym=`tail -1 ../phones/disambig.int | awk '{print $1+1;}'`

(2)fstmakecontextfst ../phones.txt $subseq_sym ilabels_3_1 > C.fst

(can not draw C.fst to ps or pdf format)

(3)print C.fst:

A. add a phone(arbitrary one) to phones.txt:

B. fstmakecontextsyms ../phones.txt ilabels_3_1 > context_syms.txt

C. fstprint --isymbols=context_syms.txt --osymbols=phones.txt C.fst C.txt

(4)print CLG.fst

fstprint --isymbols=context_syms.txt --osymbols=../words.txt

CLG_3_1.fst CLG_3_1.txt

(5)print ilabels_3_1:

fstmakecontextsyms phones.txt ilabels_3_1 > ilabels_3_1.txt

In fact, we don't create C.fst independently because in this way it will contain all the possible triphones despite never appear in the LG.fst or even is illegal in the language, instead use fstcomposecontext to craete C.fst dynamically and compose it with LG.fst.

H.fst and HCLG.fst can be created according to mkgraph.sh

3.Other utils

(1) rspecifier 读压缩文件： ark:gunzip -c $dir/ali.JOB.gz |

wspecifier 写压缩文件： ark,t:|gzip -c >$dir/ali.JOB.gz

Questions:

1. 为什么训练图是那个样子？（sil 状态之间转移很复杂，如何得到的？）

Just a sequence of the HMM models of the phones got from the transcription.

2. GMM 混合数在训练时逐步递增，具体如何实现的？
3. 训练时，对训练语音的帧做对齐，对齐具体如何做的？
4. 解码如何做的？
5. WFST 细节：消歧符等
6. Why SIL is output in the final decoding result in dear project and not in other ones?
7. There is no SIL between the phones in L.fst?

Yes, no SIL between phones consisting a word.

如果说话人的语音内容是"ii~ SIL iao1"，那么他说的可能是“伊奥”（等两个词），或者“腰”（等一个词），就存在歧义。把这个语音识别为两个词是相对合理的。如果"腰“的词典构成中包含了"ii SIL iao1"（也就是说在两个音素之间加了 SIL)，那么这个语音也可能被识别为”腰"；如果不包含，就不会被识别为“腰”。至于构成一个词的因素之间要不要加 SIL，就要视应用场景而定。比如普通话识别中，认为不会有把一个词分开读的可能，那么就不要加 SIL；如果在十个数字上做识别，不会有这种歧义发生，那么可以允许把词分开读，加入 SIL 之后，能提高系统的鲁棒性。

8. What is the decision tree and what is the purpose of making a decision tree?

How to accumulating tree stats, cluster phones (to make questions)?

In dear/s2/tri1/tree, pdf-id 492 repeats.(Unreasonable!)

How is the tree built?(details)

9. How are the parameters of final HMM/GMM models got?(e.g. #states, #pdf, #transition, #GM)
10. What is ilable_3_1 (in data/graph/lang/tmp)?
11. While training HMM/GMM, why initial equal alignment can be effective? Output the alignments from different iteration of training.

图 1．单音素模型训练流程图