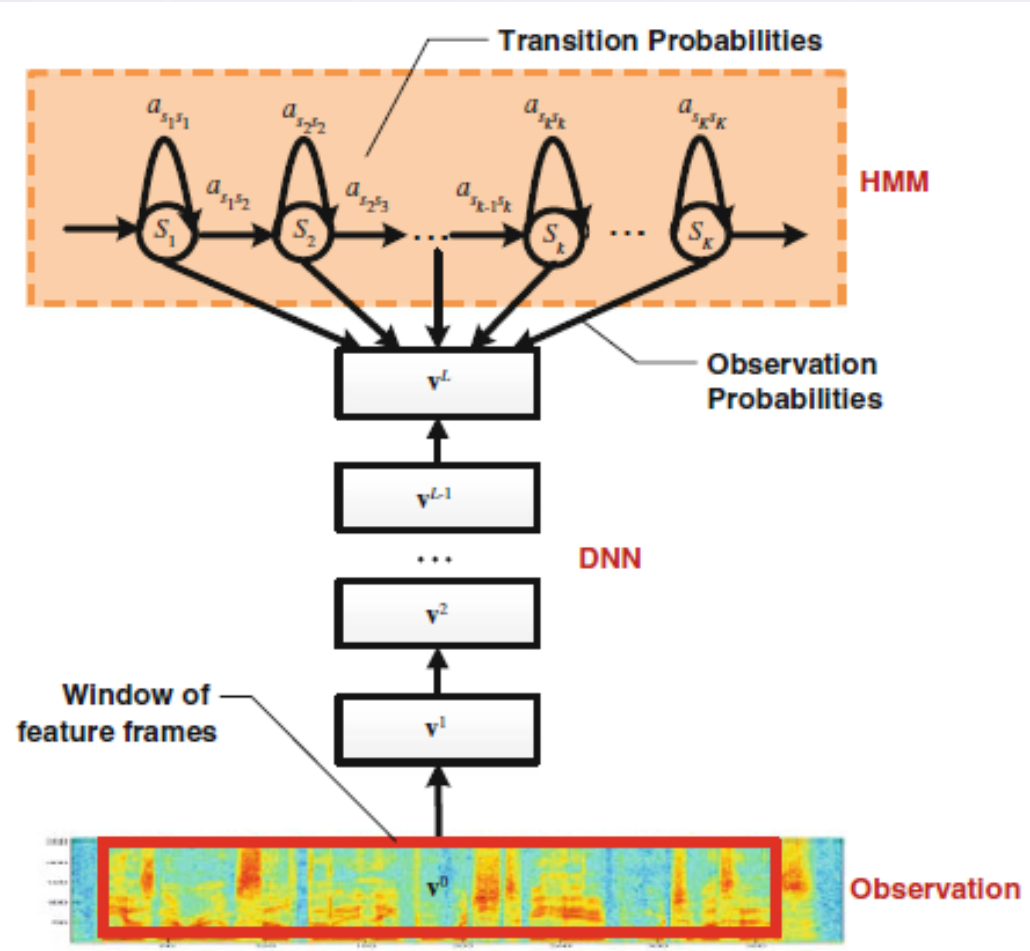
语音识别介绍

胡浩基
浙江大学信息与电子工程学院
haoji_hu@zju.edu.cn

隐马尔科夫模型

（略，见参考资料和板书）

结合深度网络模型的语音识别



DNN-HMM模型框架

结合深度网络模型的语音识别

DNN-HMM 与 GMM-HMM对比：

假设输入语音为$\{x_1, x_2, \dots, x_T\}$，且HMM有N个状态。

1. GMM-HMM (Gaussian Mixture Models-Hidden Markov Models) 是用GMM来模拟概率密度函数 $p(x_t|s_i)$, 其中$i = 1,2, \dots, N$。

2. DNN-HMM (Deep Neural Networks-Hidden Markov Models) 是用DNN来模拟概率密度函数 $p(s_i|x_t)$，其中$i = 1,2, \dots, N$

结合深度网络模型的语音识别

DNN-HMM 理论推导

$$p(\mathbf{x}_t|q_t = s) = p(q_t = s|\mathbf{x}_t)p(\mathbf{x}_t)/p(s)$$

由于$p(x_t)$对不同的HMM模型都不变，在识别过程中可以忽略，因此我们可以简化如下：

$$\bar{p}(\mathbf{x}_t|q_t) = p(q_t = s|\mathbf{x}_t)/p(s)$$

结合深度网络模型的语音识别

DNN-HMM 理论推导 – 识别流程

在识别中，某段语音属于某个“单词”w 是这样判断的：

$$\hat{w} = \arg\max_w p(w|\mathbf{x}) = \arg\max_w p(\mathbf{x}|w)\,p(w)/p(\mathbf{x})$$
$$= \arg\max_w p(\mathbf{x}|w)\,p(w),$$

其中$p(w)$表示某个“单词”w 出现的先验概率，可以通过统计获得。而

$$p(\mathbf{x}|w) = \sum_q p(\mathbf{x}|q,w)\,p(q|w)$$
$$\approx \max\pi(q_0)\prod_{t=1}^{T} a_{q_{t-1}q_t}\prod_{t=0}^{T} p(q_t|\mathbf{x}_t)/p(q_t)$$

结合深度网络模型的语音识别

DNN-HMM 理论推导 – 识别流程

在识别中，某段语音属于某个“单词”w 是这样判断的：

$$\hat{w} = \arg\max_w p(w|\mathbf{x}) = \arg\max_w p(\mathbf{x}|w)p(w)/p(\mathbf{x})$$
$$= \arg\max_w p(\mathbf{x}|w)p(w),$$

其中$p(w)$表示某个“单词”w 出现的先验概率，可以通过统计获得。而

$$p(\mathbf{x}|w) = \sum_q p(\mathbf{x}|q,w)p(q|w)$$
$$\approx \max\pi(q_0)\prod_{t=1}^{T} a_{q_{t-1}q_t}\prod_{t=0}^{T} p(q_t|\mathbf{x}_t)/p(q_t)$$

最终决策过程：

$$\hat{w} = \arg\max_w \left[\log p(\mathbf{x}|w) + \lambda \log p(w)\right]$$

结合深度网络模型的语音识别

DNN-HMM 的训练流程

$$p(\mathbf{x}|w) = \sum_{q} p(\mathbf{x}|q, w)\, p(q|w)$$

$$\approx \max \pi(q_0) \prod_{t=1}^{T} a_{q_{t-1}q_t} \prod_{t=0}^{T} p(q_t|\mathbf{x}_t)/p(q_t)$$

问题：如何获得$\pi(q_0), a_{q_{t-1}q_t}, p(q_t)$和$p(q_t|x_t)$?

回答：首先训练一个GMM-HMM模型，由GMM-HMM模型获得$\pi(q_0)$, $a_{q_{t-1}q_t}$，。 通过GMM-HMM预测每个x_t的标签q_t，统计获得$p(q_t)$。最后用深度网络获得$p(q_t|x_t)$。

Dong Yu and Li Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer, 2014. (Chapter 6)

结合深度网络模型的语音识别

Algorithm 6.1 Main steps involved in training CD-DNN-HMMs

procedure TRAINCD- DNN- HMM($\mathbb{S}$) ▷ $\mathbb{S}$ is the training set
  $hmm0 \leftarrow$ TrainCD-GMM-HMM($\mathbb{S}$); ▷ $hmm0$ is used in the GMM system
  $stateAlignment \leftarrow$ ForcedAlignmentWithGMMHMM($\mathbb{S}$, $hmm0$);
  $stateToSenoneIDMap \leftarrow$ GenerateStateTosenoneIDMap($hmm0$);
  $featureSenoneIDPairs \leftarrow$GenerateDNNTrainingSet($stateToSenoneIDMap$, $stateAlignment$);
  $ptdnn \leftarrow$ PretrainDNN($\mathbb{S}$); ▷ Optional
  $hmm \leftarrow$ ConvertGMMHMMToDNNHMM($hmm0$, $stateToSenoneIDMap$);
    ▷ hmm is used in the DNN system
  $prior \leftarrow$ EstimatePriorProbability($featureSenoneIDPairs$)
  $dnn \leftarrow$ Backpropagate($ptdnn$, $featureSenoneIDPairs$);
  Return $dnnhmm = \{dnn, hmm, prior\}$
end procedure

算法流程

结合深度网络模型的语音识别

实验结果 （9层神经网络，用自编码器初始化）

Table 6.3 Sentence error rate (SER) on the voice search development set when the context-independent monophone state labels and context-dependent triphone senone labels are used

Model	Monophone state (%)	Senone (761) (%)
CD-GMM-HMM (MPE)	–	34.5
DNN-HMM (3× 2 K)	35.8	30.4

(Summarized from Dahl et al. [7])

Table 6.4 Word error rate (WER) on Hub5'00-SWB using 309 h training set and ML alignment label

Model	Monophone state (%)	Senone (9304) (%)
CD-GMM-HMM (BMMI)	–	23.6
DNN-HMM (7× 2 K)	34.9	17.1

When the context-independent monophone state labels and context-dependent triphone senone labels are used. (Summarized from Seide et al. [25])

大词汇量连续语音识别（LVCSR）

大词汇量连续语音识别（Large-scale Vocabulary Continuous Speech Recognition, LVCSR）

问题1. 每一个HMM模型所表达的“单词”是什么？

问题2. 在识别流程中如何对测试声音文件做时间轴的划分，使每一个分段（SEGMENT）对应一个“单词”？

问题3. 如何搜索最佳的“单词”组合？

问题4. 如何构造语言模型 (Language Model)?

大词汇量连续语音识别（LVCSR）

每一个HMM模型所表达的“单词”是什么？

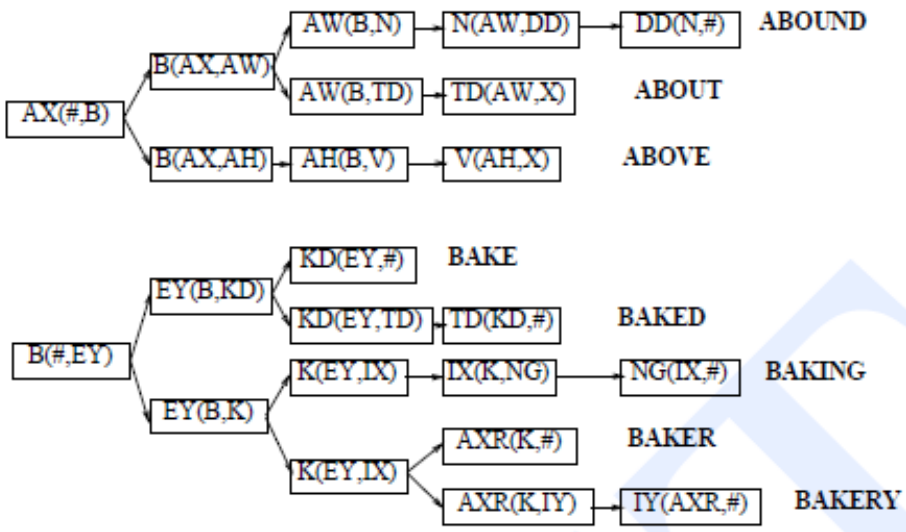


Figure 10.10 A tree-structured lexicon from the Sphinx-II recognizer (after Ravishankar (1996)). Each node corresponds to a particular triphone in a slightly modified version of the ARPAbet; thus EY(B,KD) means the phone EY preceded by a B and followed by the closure of a K.

三连音（Triphone）

英语中有效的Triphone个数大致在55000左右

（过多，需要简化！）

大词汇量连续语音识别（LVCSR）

每一个HMM模型所表达的“单词”是什么？

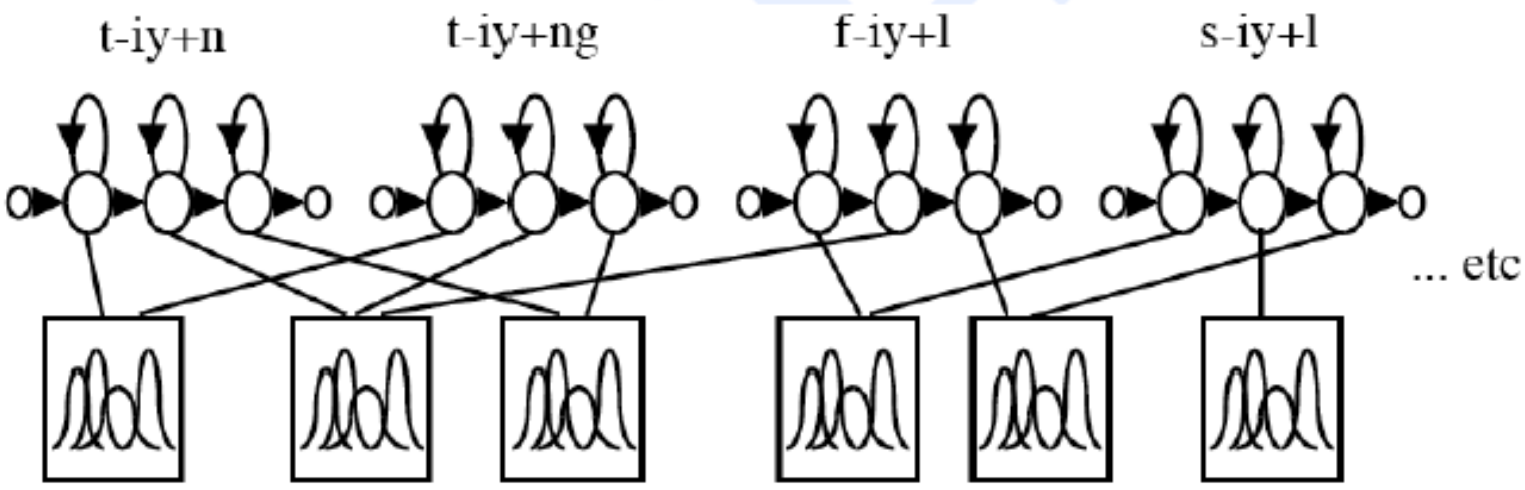


Figure 10.13 PLACEHOLDER FIGURE. Four triphones showing the result of clustering. Notice that the initial subphone of [t-iy+n] and [t-iy+ng] is tied together, i.e. shares the same Gaussian mixture acoustic model. From Young et al. (1994).

多个Triphone 合并（Tying）

大词汇量连续语音识别（LVCSR）

每一个HMM模型所表达的“单词”是什么？

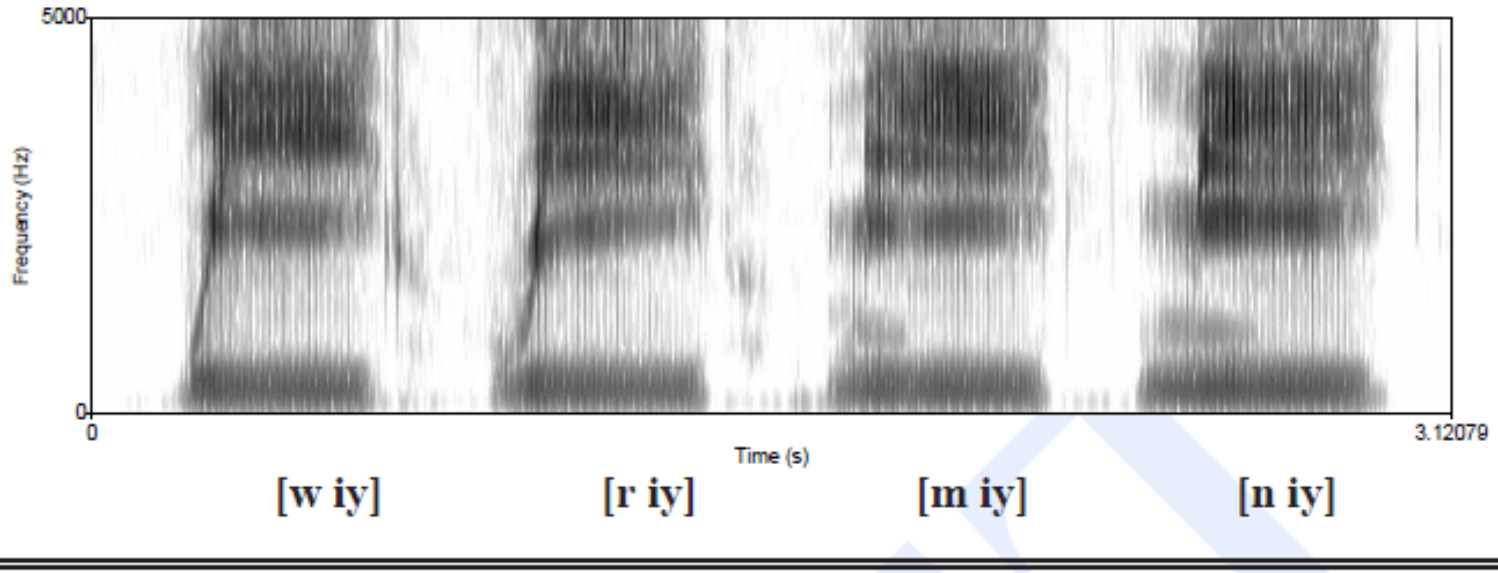


Figure 10.12 The words *we*, *re*, *me*, and *knee*. The glides [w] and [r] have similar effects on the beginning of the vowel [iy], as do the two nasals [n] and [m].

大词汇量连续语音识别（LVCSR）

每一个HMM模型所表达的“单词”是什么？

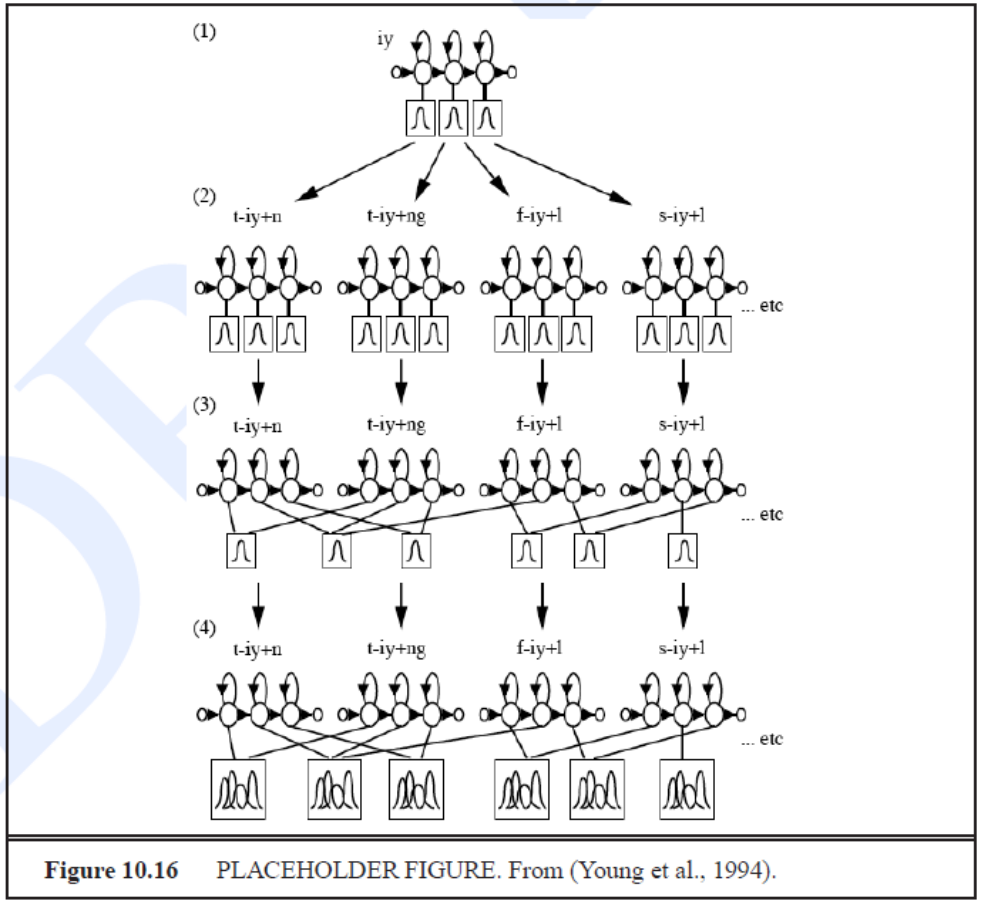


Figure 10.16 PLACEHOLDER FIGURE. From (Young et al., 1994).

多个Triphone 联合训练（Tying）

大词汇量连续语音识别（LVCSR）

汉语中Triphone个数：音节内270多个，音节间3800多个，这是包含声调后的结果 。这就意味着，汉语构造声学模型比英语更容易。

大词汇量连续语音识别（LVCSR）

如何对声音文件做时间轴的划分并搜索最佳“单词”组合？

这是一个搜索问题，搜索就是在由语句构成的空间中，寻找最优句子的过程，也就是利用已掌握的声学只是、语音学只是、语言模型及语法语义知识等，在状态（指词组、词、HMM的状态）空间中找到最优的状态序列。

搜索方法有很多种，这里归纳如下：

（1）VITERBI搜索 （有多种形式）

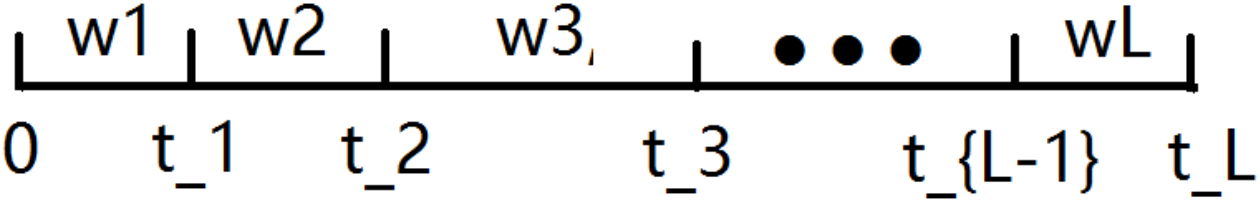
（2）A*搜索

（3）随机搜索

大词汇量连续语音识别（LVCSR）

如何对声音文件做时间轴的划分并搜索最佳“单词”组合？

语音：



待求变量：L, 所有t, 所有w。

这是一个搜索问题，方法有很多种，这里归纳如下：

（1）VITERBI搜索 （有多种形式）

（2）A*搜索

（3）随机搜索

大词汇量连续语音识别（LVCSR）

VITERBI搜索的一种： Two-Level Dynamic Programming

假设一个单词w的起始时刻为b, 终止时刻为e, 所有候选单词个数为V。 定义$\widetilde{D(b,e)}$为从b开始到e终止的最佳单词匹配的距离：

$$\widetilde{D(b,e)}=min_{1\le v\le V}\ dist(v,b,e)$$

其中 $dist(v,b,e)$ 可以通过v的HMM求得。

定义 $D_l(e)$ 为终止于e且总共有l个单词的最佳匹配距离，那么根据VITERBI算法，有：

$$D_l(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_{l-1}(b-1)\right]$$

大词汇量连续语音识别（LVCSR）

VITERBI搜索的一种： Two-Level Dynamic Programming

步骤1：$D_0(0)=0$，$D_l(0)=0$, 其中$1 \le l \le l_{max}$

步骤2： $D_1(e)=\widetilde{D(1,e)}$, 其中$2 \le e \le M$

步骤3： For $l=2\ to\ l_{max}$ do

$$D_2(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_1(e)\right],\ 3 \le e \le M$$

$$D_3(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_2(e)\right],\ 4 \le e \le M$$

……

$$D_l(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_{l-1}(e)\right],\ l+1 \le e \le M$$

步骤4： 最终决策： $D^*=min_{1\le l\le l_{max}}[D_l(M)]$, 通过每一步的记录获得单词。

大词汇量连续语音识别（LVCSR）

VITERBI搜索的一种： Two-Level Dynamic Programming

步骤1：$D_0(0)=0$，$D_l(0)=0$, 其中$1 \le l \le l_{max}$

步骤2： $D_1(e)=\widetilde{D(1,e)}$, 其中$2 \le e \le M$

步骤3： For $l=2\ to\ l_{max}$ do

$$D_2(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_1(e)\right],\ 3 \le e \le M$$

$$D_3(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_2(e)\right],\ 4 \le e \le M$$

……

$$D_l(e)=min_{1\le b<e}\left[\widetilde{D(b,e)}+D_{l-1}(e)\right],\ l+1 \le e \le M$$

步骤4： 最终决策： $D^*=min_{1\le l\le l_{max}}[D_l(M)]$, 通过每一步的记录获得单词。

大词汇量连续语音识别（LVCSR）

如何构造语言模型？

定义（N-gram）：一个单词出现的概率，只与它前面的N个单词相关。

$$P(w1, w2, w3, \dots, wn)=P(w1)P(w2|w1)P(w3|w1w2)P(w4|w1w2w3)\dots P(wn|w1w2\dots wn-1)$$

（1）在1-gram模型下

$$P(w1, w2, w3, \dots, wn)=P(w1)P(w2|w1)P(w3|w1w2)P(w4|w1w2w3)\dots P(wn|w1w2\dots wn-1)$$
$$\approx P(w1)P(w2|w1)P(w3|w2)P(w4|w3)\dots P(wn|wn-1)$$

（2）在2-gram模型下：

$$P(w1, w2, w3, \dots, wn)=P(w1)P(w2|w1)P(w3|w1w2)P(w4|w1w2w3)\dots P(wn|w1w2\dots wn-1)$$
$$\approx P(w1)P(w2|w1)P(w3|w1w2)P(w4|w2w3)\dots P(wn|wn-2wn-1)$$

（3）在3-gram模型下：

$$P(w1, w2, w3, \dots, wn)=P(w1)P(w2|w1)P(w3|w1w2)P(w4|w1w2w3)\dots P(wn|w1w2\dots wn-1)$$
$$\approx P(w1)P(w2|w1)P(w3|w1w2)P(w4|w1w2w3)\dots P(wn|wn-3wn-2wn-1)$$

大词汇量连续语音识别（LVCSR）

如何构造语言模型？

在N-Gram 中， N越大，模型越复杂，对训练样本需求越多。当然，样本足够情况下，N越大，训练后效果会更好。因此需要选一个合适的N来平衡准确度与样本数量要求。

一般来说，英语N=3, 汉语N=4。

参考资料

参考书和论文

1. L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice-Hall International , 1997.
2. Dong Yu and Li Deng, Automatic Speech Recognition: A Deep Learning Approach, Springer, 2014.
3. D. Jurafsky and J. H. Martin, Speech and Language Processing: An introduction to natural language processing, 2006.
4. L. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceedings of the IEEE, 1-30, 1989.

工具包

1. HTK Speech Recognition Toolkit, http://htk.eng.cam.ac.uk/
2. CMU Sphinx， https://cmusphinx.github.io/
3. Kaldi, https://github.com/kaldi-asr/kaldi

Thank you and comments are welcomed