基于动态贝叶斯网络的语音识别及音素切分研究*

孙阿利[1]，蒋冬梅[1]，吕国云[1]，Hichem Sahli[2]，Werner Verhelst[2]
(1. 西北工业大学 计算机学院，西安 710072；2. 比利时布鲁塞尔自由大学 电子与信息工程系，比利时)

摘 要：研究了一种基于动态贝叶斯网络(dynamic bayesian networks，DBN)的语音识别建模方法，利用 GMTK(graphical model tool kits)工具构建音素级音频流 DBN 语音训练和识别模型，同时与传统的基于隐马尔可夫的语音识别结果进行比较，并给出词与音素的切分结果。实验表明，在各种信噪比测试条件下，基于 DBN 的语音识别结果与基于 HMM 的语音识别结果相当，并表现出一定的抗噪性，音素的切分结果也比较准确。

关键词：动态贝叶斯网络；图模型；图模型工具包
中图分类号：TP391.42 **文献标志码**：A **文章编号**：1001-3695(2007)10-0104-03

Research on DBN-based continuous speech recognition and phoneme segment

SUN A-li[1]，JIANG Dong-mei[1]，LV Guo-yun[1]，Hichem Sahli[2]，Werner Verhelst[2]
(1. School of Computer Science，Northwestern Polytechnical University，Xi'an 710072，China；2. Dept. of Electronics & Information Engineering，Vrije Universiteit Brussel，Belgium)

Abstract: This paper described a dynamic Bayesian network (DBN) based technique on continuous speech recognition. The word recognition accuracies and phoneme segment accuracies of the DBN based system (implemented using the graphical model tool kit) were compared with those from classical HMM. Results show that under various SNRs, DBN based system and HMM based system has similarity performance for speech recognition and phoneme segment, especially in much lower SNR circumstance, DBN get even much better performance than HMM.

Key words: DBN; GM(graphical models); GMTK

随着语音技术的走向应用，实际环境对语音识别的声学噪声鲁棒性要求越来越高。仅仅依靠音频特征进行语音识别，已经不能满足低信噪比的环境应用。由于语音中的视觉特征具有抗噪性强的特点，近年来，将视觉特征与听觉特征相结合进行听视觉语音识别，已经成为提高语音识别系统鲁棒性的一条新途径，并且获得了较高的识别率[1]。对于听视觉语音识别的建模方法，传统采用 product HMM[2] 和多流 HMM[3]。

利用多流 HMM 模型可以表示听视觉之间的关系，然而只能对听视觉异步关系进行音素级的建模。研究实验证明，对于连续语音识别，协同发音现象非常普遍，使得听视觉间的异步关系已经超过音素边界。另外，HMM 在结构上只允许一个时间片具有一个状态，严重限制了对细节的描述。同时，product HMM 也带来了状态空间过大、计算量增加等问题。针对这些问题，对于听视觉语音识别急需寻找一种新的反映这种异步关系的建模方法。近年来，基于 DBN 的单流或多流语音模型应用于连续语音识别[4~6]，并取得了较高的识别结果。Zhang Yi-min 等人[7] 利用 DBN 的建模优势，提出了一种多流 DBN 模型(multi-stream DBN，MSDBN)。该模型使用 DBN 对各种声学特征进行同步和异步建模。实验证明该方法比传统的基于 HMM 的融合方法带来更高的识别率。然而现有的 DBN 模型结构中，并没有针对音素级切分结果的比较。为此，本文利用 GMTK，构建了音素级的单流 DBN 模型训练和识别模型；同时还给出了识别率统计结果及词、音素切分结果，并与手工切分、HTK 切分结果进行比较。

1 基于动态贝叶斯网络的语音模型

1.1 图模型与贝叶斯网络

图模型[8] 是一种将概率论与图论相结合的抽象统计模型。透过它，可进一步深入地观察和研究随机过程中一些极为重要的特性，以及这些随机过程式图表示的物理现象。除了强大的表示能力之外，GM 还提供了一套高效的概率计算和决策算法。

图模型通常分为无向图模型(undirected GM)、有向图模型(directed GM)以及链路图(chain graph)。贝叶斯网络(Bayesian network)是一种有向无环图网络。它采用一种直接的方式表示变量集合，以及变量之间联合概率分布的因式分解。这种联合概率因式分解可以表示为

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n) = \prod_i P(X_i = x_i \mid \prod_i = \pi_i)$$

其中：$\prod_i (i = 1, 2, \cdots, n)$ 表示变量 $X_i (i = 1, 2, \cdots, n)$ 直接前驱。

1.2 动态贝叶斯网络

语音信号是随时间变化的随机过程，如果将 BN 应用于语

收稿日期：2006-07-04；修返日期：2006-10-18 基金项目：西北工业大学基金资助项目(04XD0102)；中国科技部与比利时弗拉芒大区科技合作资助项目(国科外函[2004] 487)

作者简介：孙阿利(1979-)，女，硕士研究生，主要研究方向为数字语音处理(sun_ally@hotmail.com)；蒋冬梅(1973-)，女，副教授，主要研究方向为音频信号处理、语音处理、听视觉融合的语音识别和说话人头部动画；吕国云(1975-)，男，山西万荣人，博士研究生，研究方向为模式识别等；Hichem Sahli，教授，主要研究方向为图像处理、听视觉语音处理、视频编解码；Werner Verhelst，教授，主要研究方向为语音合成、听视觉语音处理等.

音建模中, 需要将 BN 与时间联系起来。动态贝叶斯网络[9]是在时间上对 BN 的扩展, 非常适合对时间序列进行建模。DBN 在有限时间内, 将变量之间的因果关系用联合概率关系的形式表示出来, 并继承了 GM 和 BN 强大的表示能力。它是继 HMM 之后, 建立更为复杂的语音模型的新选择。

1.3　用于连接词语音识别的 DBN 模型构建

以华盛顿大学的 Bilmes 为代表的研究者利用动态贝叶斯网络, 构建了用于语音识别的 GMTK[4, 5], 大大简化了语音训练及识别的模型结构, 提高了识别的运算速度。本文采用 GMTK 工具包来构建需要的音素级的模型结构。GMTK 中使用脚本语言, 定义了语音模型的基本结构, 对各个节点变量的类型以及范围作了详细定义; 同时定义了节点之间的概率转移关系, 并用条件概率关系表(CPT) 和决策树(decision trees, DTs) 进行描述。

1.3.1　GMTK 动态模型结构

GMTK 中动态模型的基本结构如图 1 所示。

针对图 1 中的结构图, 将中间 frame1 的结构进行扩展, 得到图 2 所示的结构, 从而显示地描述更长的时间序列。

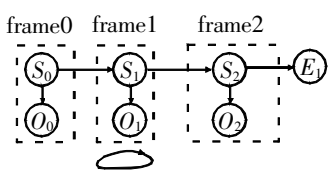


图 1　动态模型的基本结构

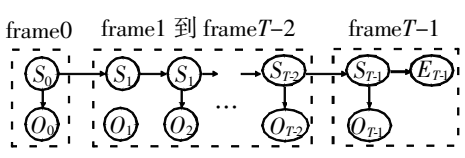


图 2　扩展后的结构图

1.3.2　音素级模型结构描述

在音视频语音识别的研究中, 可以分别利用音频特征和视频特征进行语音识别, 因而根据 GMTK 模型结构的基本框架, 构建同时适用于音频特征和视频特征语音识别的单流 DBN 模型。模型结构如图 3、4 所示。

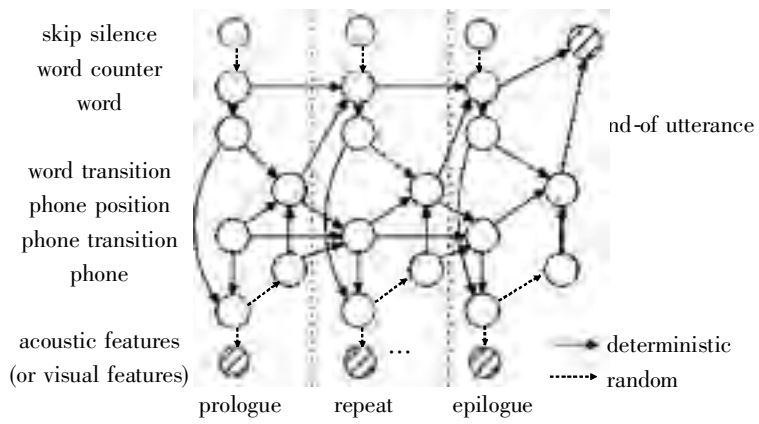


图 3　用于连接词语音识别的 DBN 训练模型

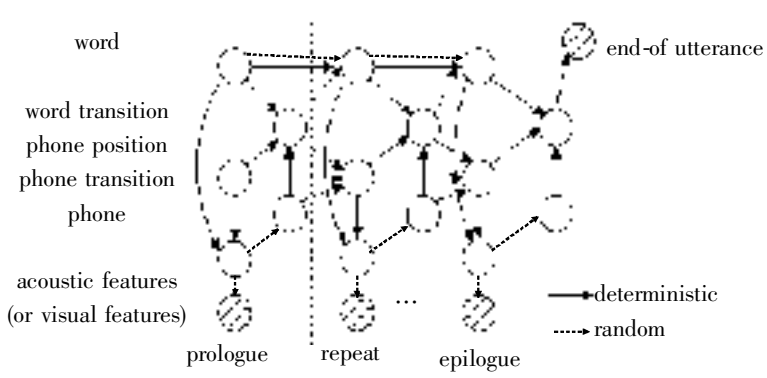


图 4　用于连接词语音识别的 DBN 识别模型

该模型被分为三个部分: prologue、repeat、epilogue。对 repeat 块进行扩展, 使得结构能够显示地表示语音的发音过程。

图 3 和图 4 中, 实线箭头表示确定性的条件概率, 虚线箭头条表示随机概率。充分利用这种确定性概率关系, 可以大大提高 DBN 在推理过程中的计算效率。在识别模型结构图中, 可以看到从节点 word transition 到 word 之间有一条虚线有向边连接, 将节点 WT 称做转换节点(switching parent)[4]。CPT 的选择是根据该节点的取值来决定的。

模型中节点的具体含义为:

Skip silence(SS) ——随机变量, 相当于静音或者语音之间的停顿。

Word counter(WC) ——表示句子中词所在的位置(只有训练模型中有该节点)。

Word(W) ——当前词。

Word transition(WT) ——如果值为 1, 表示词发生转移。

Phone position(PP) ——表示音素在某个词中的位置。

Phone transition(PT) ——如果值为 1, 表示音素发生转移。

Phone(P) ——词中具体的音素。

Observation(O) ——音频或视频的观测向量。

图 1 中的联合概率关系表示为

$$p(W_{1:T}, WT_{1:T}, PP_{1:T}, PT_{1:T}, P_{1:T}, O_{1:T}) = \prod_t p(O_t|p_t) \times p(PT_t|P_t)\, p(P_t|+PP_t, W_t) \times p(WT_t|W_t, PP_t, W_t) \times p(PP_{t-1}|WT_{t-1}, PP_{t-1}, PT_{t-1}) \times p(W_t|W_{t-1}, WT_{t-1})$$

其中: 下标 t 表示当前时间; $t-1$ 表示前一时间片。

1.3.3　节点条件概率关系描述

采用条件概率分布(conditional probability distributions, CPD) 来描述各个节点之间的依赖关系。

a) Word counter(WC)

Word counter 用来标志当前词在句子中的位置, 其父节点为 word counter(-1)、word transition(-1)、skip sil(0)。当词没有发生转移时, 即 $WT_{t-1}=0$, $WC_t = WC_{t-1}$; 当词发生转移时, 即 $WT_t=1$, 词的位置计数超过句子的边界时, $WC_t = WC_{t-1}$; 当词没有超过边界时, 观察是否有停顿 skip sil, 如果没有停顿出现, 则计数加 1, 即 $WC_t = WC_{t-1}+1$; 如果有停顿, 则观察前一个词是否为真正意义上的词, 如果是, 则计数加 2, 转移到下一个词中; 否则, 计数加 1, 转移到下一个真正的词上。概率关系表达式为

$$p(WC_t = i | WC_{t-1} = j, WT_{t-1} = k, SS = l) = \begin{cases} 1 & i = j \text{ and } k = 0 \\ 1 & i = j \text{ and bound}(w, j) \text{ and } k = 1 \\ 1 & i = j+1 \text{ and } \sim\text{bound}(w, j) \text{ and } l = 0 \text{ and } k = l \\ 1 & i = j+2 \text{ and } \sim\text{bound}(w, j) \text{ rdalword}(w) \text{ and } l = 1 \text{ and } k = 1 \\ 1 & i = j+1 \text{ and } \sim\text{bound}(w, j) \; l = 1 \text{ and } \sim\text{realword and } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

b) Word(W)

当词没有发生转移时, 即 $WT_{t-1}=0$, $W_t = W_{t-1}$, 该条件概率是确定的; 当词发生转移时, 即 $WT_{t-1}=1$, 遵从二元文法(bigram) 的概率分布。概率关系表达式为

$$p(W_t = i | W_{t-1} = j, WT_t = k) = \begin{cases} 1 & \text{if } i = j \text{ and } k = 0 \\ \text{bigram} & \text{if } k = 1 \\ 0 & \text{otherwise} \end{cases}$$

c) Word transition(WT)

当音素在词中的位置 PP 到达当前词所在的最后一个音素位置时, 且有音素发生转移, 即 $PT_t=1$ 时, $PP_t = k$, $W_t = w$ (其中词 w 有 k 个音素), $WT_t=1$。概率关系表达式为

$$p(WT_t = i | W_{t-1} = w, PP_t = k, PT_i = j) = \begin{cases} 1 & \text{if } i = 1 \text{ and } j = 1 \text{ and lastphone}(k, w) \\ 1 & \text{if } i = 0 \text{ and } j = 1 \text{ and lastphone}(k, w) \\ 1 & i = 0 \text{ and } j = 0 \\ 0 & \text{otherwise} \end{cases}$$

其中: lastphone(k, w) = true 表示当且仅当 k 是词 w 的最后一个音素时。

d) Phone position(PP)

如果从前一时间 $t-1$ 到时间 t, 没有音素发生转移, 即 $PT_{t-1}=0$, 则有下列概率关系:

$$p(PP_i = i \mid PP_{t-1} = i, WT_{t-i} = j, PT_{t-1} = 0) = 1$$

如果音素发生转移, 且不是一个词的最后一个位置, 则有

$$PP_t = PP_{t-1} + 1$$

如果音素发生转移, 且音素的位置处于一个词的最后一个音素位置, 则有

$$p(PP_t = 0 \mid PP_{t-1} = i, WT_{t-1} = 1, PT_{t-1} = 1) = 1$$

总的概率关系为

$$p(PP_t = i \mid PP_{t-1} = j, WT_{t-1} = k, PT_{t-1} = l) = \begin{cases} 1 & i = j \text{ and } l = 0 \\ 1 & i = j+1 \text{ and } k = 0 \text{ and } l = 1 \\ 1 & i = 0 \text{ and } k = 1 \text{ and } l = 1 \\ 0 & \text{otherwise} \end{cases}$$

其中: bound(w, j) 表示当前词计数没有超过句子的边界; real-word(w) 表示当前词是否为真正意义上的词。

e) Phone transition (PT)

该节点的概率关系表达式为

$$P(PT_t = j \mid P_t = i) = \begin{cases} \alpha_{ii} & \text{if } j = 0 \\ 1 - \alpha_{ii} & \text{if } j = 1 \end{cases}$$

其中: α_{ii} 表示停留在音素 i 的概率; $1-\alpha_{ii}$ 表示从音素 i 转移到音素 $i+1$ 的概率。

f) Phone(P)

Phone 用来表示词中的某个音素。

$$P(P_t = i \mid W_t = w, PP_t = j) = \begin{cases} 1 & \text{if } i = \text{phone}(w, j) \\ 0 & \text{otherwise} \end{cases}$$

其中: phone(w, j) 表示词 w 的第 j 个音素。

2 模型脚本的更改及实现

为了能够输出音素, 同时切分出音素对应的时间, 可将原有模型中状态上层的结构不进行改动, 而只是将整词状态级更换为真正意义上的音素级, 同时构造音素之间的转移概率以及音素到观测向量产生的概率。

在 GMTK 中, 对于模型结构、节点之间的概率关系以及参数的初始化设置, 都是用脚本语言的方式描述的, 因而根据 GMTK 的文件描述格式, 对其中涉及到的脚本文件进行修改。在这里, 本文构建了连接词中 zero 到 oh 这 11 个单词的音素对应表(表 1) 。

表 1 连接词音素对应表

词	音素组成	词	音素组成
zero	z ih1 r ow1 l	six	s ih1 k s
one	w ah1 n	seven	s eh1 v ax n
two	t uw1	eight	ey1 t
three	th r iy1	nine	n ay1 n
four	f ao1 r	oh	ow1
five	f ay1 v		

定义了 37 个音素, 相应地定义了 37 个混合高斯, 每个高斯分量个数为 1, 均值和方差分别为 42 维。对 phone position 节点到 phone 之间的概率关系进行定义, 针对每个词输出相应的音素。

对模型中的 chunk 块进行扩展, 将 prologue、repeat、epilogue 这三个部分分别进行三角化(triangulate) 处理、简化模型结构; 然后又为每个部分建立决策树(junction trees) ; 最后连接成一个完整的三角化决策树。一旦模型结构经过三角化处理后, 就可以运用 EM 算法对模型参数进行重估。

3 实验结果与分析

3.1 实验设置

采用 Aurura 3.0 数据库中的连接词作为实验数据, 训练样本为 100 句, 测试样本为 30 句, 同时对原始语音数据加入高斯白噪声, 可将形成各种信噪比的语音数据作为测试样本。特征采用 13 维的 mfcc 特征, 选用窗长为 25 ms 的 Hamming 窗。同时计算 13 维的一阶差分和二阶差分特征向量, 再加上一维的能量特征向量, 构成 42 维特征向量。高斯混合模型采用 37 个高斯, 每个状态对应一个高斯, 一个高斯由两个高斯混合分量组成。

3.2 实验数据分析

针对信噪比分别对 0 ~40 db 的数据进行识别, 并与 HTK 的识别结果进行比较。词一级的识别结果如表 2 所示。

表 2 连接词识别结果比较

模型	clean	40 db	30 db	20 db	15 db	10 db	0 db
HTK	99.06	99.06	99.06	94.34	81.13	58.49	30.19
GMTK	99.06	99.06	98.11	87.74	84.91	69.81	35.85

从表 2 中可以看出, 基于 DBN 的模型识别结果明显高于基于 HMM 的模型识别结果; 尤其在低信噪比的环境中, GMTK 的识别率要比 HTK 的词识别率高出很多。

针对音素的切分, 选取一个实验句子, 原始语音为 five six。原始语音波形图如图 5 所示, 并给出使用 HTK 和 GMTK 音素的切分结果。

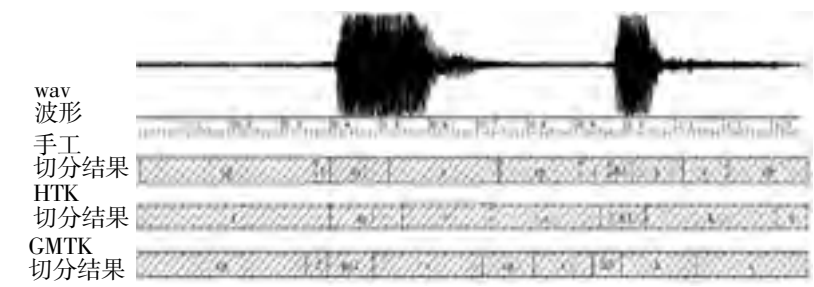


图 5 音素切分结果比较图

从图 5 中可以看出, 词边界的切分结果基本接近, 音素的切分结果也基本在可以接受的范围之内。

4 结束语

本文利用动态贝叶斯网络构建了针对连接词的语音识别和训练模型, 通过对不同信噪比的语音数据进行识别发现, 基于 DBN 的模型显示出较强的识别结果, 尤其是在低信噪比的环境中, 其识别结果高于基于 HMM 模型结构, 从而体现出 DBN 具有较强的噪声鲁棒性。同时, 通过对音素切分结果进行比较, GMTK 能够实现词以及音素的切分功能, 并且切分结果具有有效性, 其词的切分结果与 HTK 的词切分结果基本接近。当然, 音素的切分结果还不是很理想的, 今后还需要对模型中的参数进行调整, 以使音素切分结果更加准确; 同时也为今后针对基于 DBN 的音/视频模型, 研究音/视频之间的异步关系奠定了基础。

参考文献:

[1] OTAMIANOS G, NETI C, GRAVIER G, *et al*. Recent advances in the automatic recognition of audiovisual speech[J]. IEEE, 2003, 91(9): 1306-1326.

(下转第 127 页)

的平均长度为 23.11, 如图 5(a) 所示。标记域存储过的概率为 0.002, 如图 6(a) 所示。而相对频率相等且都为 255, 其他条件都相同的情况下, 所占 ls 平均长度为 24.95, 较之 23.11 要稍大一些, 如图 5(b) 所示。标记域被存储的概率为 0.001, 如图 6(b) 所示。从图 6(b) 还可以看到, 当跳数为 21、路由器 R 的上游路由器数量为 3 时, 所占 ls 的平均长度超过了 30 bit, 所以标记域必须存储到路由存储器中。因此在 32 跳中, 标记域被存储在路由器的概率为 2.34%。

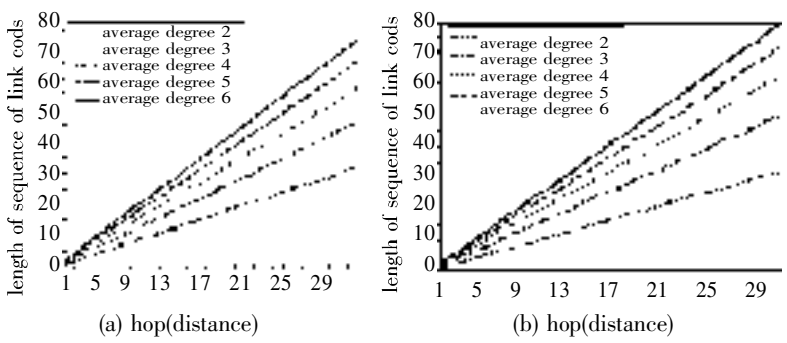
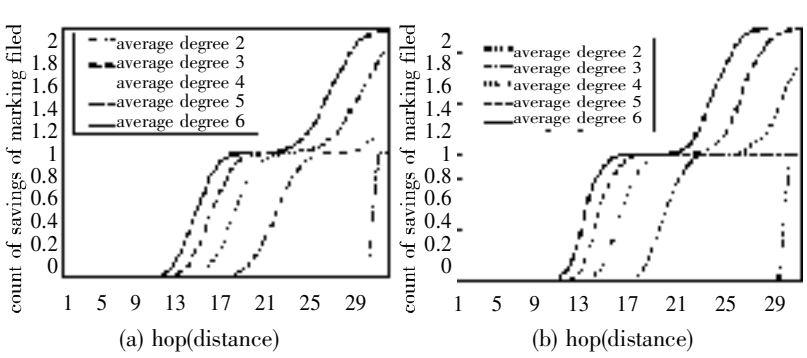


图 5　不同(相同)频率下跳数与占用 ls 的关系



从图 7 链接数量与 Huffman 编码之间的关系可以看出, 频率差值不等情况下所需的平均 Huffman 编码长度比频率差值相等情况下小, 随着频率差值的增加这种差距越明显。

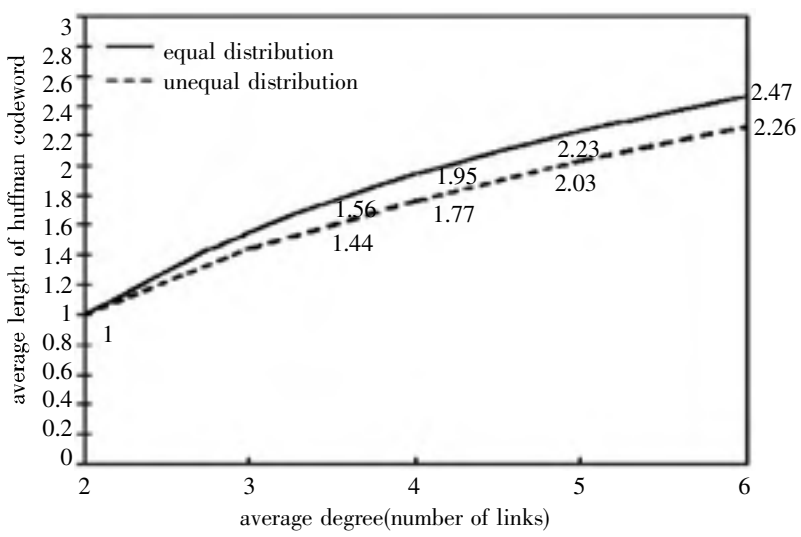


图 7　链接数量与 Huffman 编码的关系

4　结束语

与目前已有的很多方案一样, 该方案也可以实现攻击时和攻击后的追踪。目前的概率包标记方案[1,4] 需要成百上千个攻击包才可重构攻击路径, 而该方案一个很大的特点是只需一个包(无论是攻击包还是合法的包) 就能重构 DoS 和 DDoS 攻击路径, 因为不同的攻击路径有不同的 ls, 而不同的序列流能够解密不同的攻击路径; 与 ingress filter、logging、packet marking等方案相比, 该方案需要的存储空间更少这也是本方案最成功的地方。存储标记域增加了路由器的开销是该方案与其他方案相比存在的不足之处, 不过从实验结果来看此开销还是很小的。进一步的研究方向是将该方案运用到 IPv6 中去, 并且尽量将路由器的开销减少到最低程度。

参考文献:

[1] AVAGE S, WETHERALL D, KARLIN A. Practical network support for IP traceback[C] //Proc of the ACM SIGCOMM Conference. New York: ACM Press, 2000: 295-306.

[2] FERGUSON P, SENIE D. RFC 2827, Network ingress filtering: defeating denial of service attacks which employ IP source address spoofing [S] . 2000.

[3] BABA T. Tracing network attack to their sources [J] . IEEE Internet Computing, 2002, 6(2) : 20-26.

[4] SONE D X, ADRIAN P. Advanced and authenticated marking schemes for IP traceback[C] //Proc of IEEE INFOCOM'01. Anchorage, Alaska: IEEE Press, 2001: 878-886.

[5] PAPK K, LEE H. On the effectiveness of probabilistic packet marking for IP traceback under denial of service attack[C] //Proc of IEEE INFOCOM. Anchorge: IEEE, 2001: 338-347.

[6] 严蔚敏, 吴伟民. 数据结构: C 语言版[M] . 北京: 清华大学出版社, 1997: 144-146.

[7] STONE R. Center track: an IP overlay network for tracking DoS floods [C] //Proc of the 9th Usenix Security Symp. Usenix Assoc: [s. n.] , 2000: 199-212.

[8] STRAYER T. SPIE-IPv6: single IPv6 packet traceback, local computer networks[C] //Proc of the 29th Aunual IEEE International Conference. Boston: IEEE Standards Office, 2004: 118-125.

[9] BURCH H, CHESWICK B. Tracing anonymous packets to their approximate source[C] //Proc of the 14th Conf Systems Administration. [S. l.] : Usenix Assoc, 2000: 319-327.

(上接第 106 页)

[2] UKUNDH N, SREEIVAS T V. Product-HMM: a novel class of HMMs for sub-sequence modeling[EB/OL] . (2003-01-09) . [2006-06-04] . http: //www. isca-speech. org/orchive/wslp-117. html.

[3] HAGEN A, MORRIS A C. Recent advances in the multi-stream HMM/ANN hybrid approach to noise robust ASR [J] . Computer Speech & Language, 2005, 19(1) : 3-30.

[4] BILMES J. GMTK: the graphical models toolkit[EB/OL] . [2006-06-04] . http: //ssli. ee. washington. edu/ ~bilmes/gmtk/doc. pdf.

[5] BILMES J A, CHRIS B. Graphical model architectures for speech recognition[J] . IEEE Signal Processing, 2005, 22(5) : 89-100.

[6] KEVIN P M. Dynamic Bayesian networks: representation, inference and learning[D] . Berkeley: University of California, 2002.

[7] ZHANG Yi-min, DIAO Qian, *et al*. DBN based multi-stream models for speech[C] //Proc of IEEE Int Conference on Acoustics, Speech, and Signal Processing. 2003: 836-839.

[8] ZWEIG G, RUSSELL S. Speech recognition with dynamic Bayesian networks[C] //Proc of the 15th Nat Conf Artificial Intelligence and 10th Innovative Applications of Artificial Intelligence Conf(AAAI-'98) . 1998: 173-180.

[9] RUASSELL S, NOORVIG P. 人工智能: 一种现代方法. [M] . 中文版. 北京: 人民邮电出版社, 2004: 430-437.