唇读中的 HLM 模型及其文字流解析*)

王 丹 姚鸿勋 万玉奇 洪晓鹏
(哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

摘 要 由于唇动序列和语言序列是一对多的映射,计算机自动唇读识别仅使用 HMM 是远远不够的。以 HMM 为基础,结合语言先验知识,建立了新的唇动识别模型——HLM (HMM and Bigram Language Model)。HLM 突破了单纯采用 HMM 计算声学后验概率进行识别的传统框架,将 HMM 和语言背景知识紧密联系起来,依据语言模型对语言背景知识进行统计,在识别阶段融合声学后验概率和语言学先验概率进行判决。实验结果表明,HLM 可使单音识别率提高 7.3%,句子识别率提高 19.5%。另外,采用语言模型对文字流进行解析,而不再是盲目文字匹配,单一视觉流的解析精确率达 70.5%。

关键词 唇读,识别模型,HLM,HMM

Lipreading HLM and Text Flow Analysis

WANG Dan YAO Hong-xun WAN Yu-qi HONG Xiao-peng
(School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract Since lip movement sequence and language sequence are one-to-many mapping, it is far from sufficiency to use only HMM for lip-reading recognition. Proposed a novel recognition model HLM(HMM and Bigram Language Model), which is based on HMM, and combined with prior knowledge of language. In contrary to the traditional framework, which adopts pure acoustic HMM posterior probability calculation for recognition, HLM combines closely language background knowledge and HMM. It carries on background knowledge of the language statistics according to language model. Acoustic posterior probability and linguistics prior probability are fused for judgments in the recognition stage. Experimental results demonstrated that applying HLM, syllable accuracy can increase by 7.3%, and sentence accuracy can increas by 19.5%. In addition, exploited language model for text flow analysis, rather than blindly text matching. In single video channel the accuracy can be up to 70.5%.

Keywords Lipreading, Recognition model, HLM, HMM

1 引言

计算机自动唇读能够在噪声环境和多话者环境下,识别话者的说话内容,或根据说话方式识别话者身份,因此成为人机交互和生物特征识别的重要方式之一。本文针对单一视觉通道的唇读问题,重点突破结合语言背景的识别和文字流解析。

传统的唇读系统从视频信息中提取特征,然后利用 HMM,DTW 等分类器进行识别[1,2,5]。这种框架下,系统已经取得了不错的识别率,但因为口型序列和拼音的一对多关系而得不到唯一解[3,4]。利用语言的相关特征,建立正确的模型,是得到唯一解的可能方法,因此唇读研究一定要结合语言模型。同时,语言模型也能够为文字流解析提供启发式信息。

语言模型是语言客观事实的数学模拟[6],可分为传统的文法模型[7,8]和基于统计的语言模型。基于大规模真实语料的统计语言模型,随着计算语言学的兴起得到了广泛应用。文献[9]将统计语言模型和复杂特征集、词汇主义一同列为自然语言处理技术的三大里程碑。其实用价值已经在语音识别、机器翻译、中文键盘输入等领域得到了充分的证实,将它应用到唇读中很可能带来性能上的提高。

统计语言模型将自然语言看作一个随机序列,文本中的每个语言单位(字、词、句子、文本)都是具有一定分布的随机变量。建模过程采用大量的文本资料,统计各个词出现的概率及其相互关联的条件概率。该模型的典型代表有:N-Gram 模型、N-Pos 模型、决策树模型和最大熵模型等。N-Pos 模型是统计词类的概率分布,其精度不如 N-Gram 模型;决策树模型是一种更加通用的模型,但树的构造时空消耗非常大[10]。目前应用较广泛的还有 HMM,最大熵模型(Maximum entropy model)等。最大熵模型[6]的主要优点是可以将所有不同概率分布以约束的形式提出,然后求解满足约束条件的解,但计算量很大。N-Gram 是最经典的统计语言模型,也是实际系统中应用最多的,技术相对成熟。本文采用 N-Gram 模型。

本文以 HMM 为基础,结合 Bigram 统计语言模型(可以扩展到其他语言模型),建立了 HLM(HMM and Bigram Language Model)。HLM 利用口型序列训练 HMM 各项参数,同时利用语料库训练 Bigram,识别在 HMM 和 Bigram 的共同作用下进行。其实质是利用语言的先验知识建立模型,辅助

*)黑龙江省自然科学基金项目(E2005-29),哈尔滨工业大学"新世纪人才支持计划"(NCET-05-03 34)。王 丹 硕士生,主要研究领域为图像处理、模式识别及自然人机交互;姚鸿勋 博士,教授,博士生导师,主要研究领域为图像处理、模式识别、多媒体技术及自然人机交互;万玉奇 硕士,主要研究领域为图像处理、模式识别;洪晓鹏 博士生,主要研究领域为图像处理、模式识别。

口型序列识别以得到最大概率的可能解。实验表明，在不同的特征提取方法下，HLM 较 HMM 的拼音识别率提高 7.3%，句子识别率提高 19.5%。语言模型背景下单一视觉通道的文字流解析率达 70.5%。

2 唇读中语言模型的应用

2.1 N-Gram 语言模型

N-Gram 模型假设语言是一个 Markov 过程，每个预测变量只与和它相邻的前 $N-1$ 个元素有关，而与其他任何元素不相关，整句的概率就是各个词出现概率的乘积。设 w_i 是句子 W 中的任意一个语言单位(字或词)，则：

$$P(W)\approx\prod_{i=1}^{n}P(w_i\mid w_{i-N+1}w_{i-N+2}\cdots w_{i-1}) \tag{1}$$

$$P(w_i\mid w_{i-N+1},w_{i-N+2},\cdots,w_{i-1})=\frac{F(w_{i-N+1}w_{i-N+2}w_{i-1}w_i)}{F(w_{i-N+1}w_{i-N+2}w_{i-1})} \tag{2}$$

其中，$F(W)$表示串 W 在训练语料中出现的次数。当 $N=1$，2，3 时，分别称为 Unigram，Bigram，Trigram 模型。由于计算量大，N 很少取 4 或者更大的数，本文系统采用 Bigram 模型。

2.2 HLM——HMM 与语言模型的联合概率模型：$P(A|W)P(W)$

目前唇读中应用最多的识别器是 HMM，它基于模式转移和类独立性的统计假设[5]。如图 1 所示，传统的系统框架先对每个音的观察值序列 a 进行特征提取，用特征矢量训练每个音 w 的 HMM。测试时同样先进行特征提取，将特征矢量送入识别器，计算当前口型序列在各个音的 HMM 下出现的条件概率 $P(a|w)$，选取概率最大的音 w^* 作为识别结果。

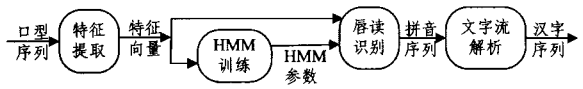

图 1 传统的唇读流程图

此框架有效利用了口型序列的时间特性，在手工参与预处理的情况下，取得了很好的识别率[11]。但是由于口型序列与拼音是一对多的关系，单纯通过特征提取方法和识别算法的改进，无法对具有相似甚至相同口型的不同音进行区分，必须借助语言的先验知识完成识别。比如，拼音中口型相似的/qi/，/ji/，/yi/三个音，口型序列仅由视频信息很难区分。而如果在一定的语境下，比如“chong zheng qi gu”(重整旗鼓)，利用上下文信息，就可能避免误识为：“chong zheng ji gu”或“chong zheng yi gu”，从而提高识别率。

事实上，如果输入的口型序列为 A，正确识别结果为拼音串 W^*，那么识别任务就是口型序列为 A 时后验概率最大的拼音串，即

$$W^*=\arg\max_W P(W|A)=\arg\max_W\frac{P(A|W)P(W)}{P(A)}=\arg\max_W P(A|W)P(W) \tag{3}$$

$P(A|W)$对应 HMM 下的概率，$P(W)$则对应语言模型下的概率，该模型即为 HLM(HMM and Bigram Language Model)，它恰好是 HMM 和语言模型的乘积模型。

训练过程除了要训练 HMM 各项参数外，还要统计语料库中的先验知识。由于考虑的是 Bigram 模型，因此只关注相邻词对应拼音的相关性。对于待识别的口型序列，当前的唇读系统还不能做到理想的连续词语或句子识别，而是一一识别单个音的序列，然后串结成句子，因此语言模型的应用与其它领域有所不同。

设句子的拼音串为 $W^*=w_1^*w_2^*\cdots w_n^*$，$w_i^*$ $(i=1,2,\cdots,n)$为单个音，待识别的口型序列 $A=a_1a_2\cdots a_n$，a_i $(i=1,2,\cdots,n)$为单个音对应的口型序列。文中的分析基于这样的假设：

(1)口型序列已被精确分割，即 w_i 与 a_j $(j\neq i)$不相关；

(2)每个拼音对应的口型序列的出现是相互独立的；

(3)拼音被自左向右逐个识别；

(4)拼音一旦被识别，即认为识别正确，后面的音可根据与它的相关性进行识别；

(5)当前音的概率分布只与和它相邻的前一个音有关。

那么，对于一个句子的每一个拼音 w_i^* $(i=1,2,\cdots,n)$的识别过程考虑两种情况：

①$i=1$，即待识别的拼音位于句首：

$$w_1^*=\arg\max_{w_1}P(a_1|w_1)P(w_1) \tag{4}$$

$P(a_1|w_1)$对应某个音的 HMM 下的概率，$P(w_1)$则应该利用语言中的先验知识，这里使用当前拼音出现在句首的概率是合理的。

②若 $i>1$，即拼音 w_i 为非句首元素。

根据假设(3)－(5)，与之相邻的前一个音 w_{i-1} 已经识别正确，我们可以根据两者的相关性为识别增加辅助信息。先计算给定口型序列在各个音的 HMM 下出现的概率，而后计算各自在 w_{i-1}^* 后出现的概率，最后联合两个概率信息给出识别结果。具体推导过程如下：

$$\begin{aligned}
&w_{i-1}^*w_i^*\\
&=\arg\max_{w_{i-1}w_i}P(a_{i-1}a_i|w_{i-1}w_i)P(w_{i-1}w_i) && \text{由公式(3)}\\
&=\arg\max_{w_{i-1}w_i}P(a_{i-1}|w_{i-1}w_i)P(a_i|w_{i-1}w_i)P(w_{i-1}w_i)\\
& && \text{由假设(2)}\\
&=\arg\max_{w_{i-1}w_i}P(a_{i-1}|w_{i-1})P(a_i|w_i)P(w_{i-1}w_i) && \text{由假设(1)}\\
&=\arg\max_{w_{i-1}w_i}P(a_{i-1}|w_{i-1})P(a_i|w_i)P(w_{i-1})P(w_i|w_{i-1})\\
& && \text{由公式(1)}
\end{aligned}$$

事实上，由假设(3)，(4)知，上式中的 $w_{i-1}=w_{i-1}^*$，是常量，因此有

$$w_i^*=\arg\max_{w_i}P(a_i|w_i)P(w_i|w_{i-1}^*) \tag{5}$$

其中，$i=2,3,4,\cdots,n$。

由步骤①、②，整句话的拼音被一一识别出来。每个音的识别都利用了语言模型的辅助信息，使识别更加准确。

2.3 文字流解析

唇读的最终目标是将口型序列解析为文字流，而汉语的一大特点就是存在大量的同音字。根据国家标准 GB2312-80，国家规定的一级和二级常用汉字总共有 6763 个，而拼音总共只有大约 400 个，平均每个拼音对应 17 个汉字，个别音节(如 yi)对应的汉字达到 100 多个，因此无辅助信息的盲目解析准确率极低。而借助语言模型则能形成“启发式”文字流解析，带来性能的提高。

唇读文字流解析的数学模型可描述为找对应问题，$X=x_1,x_2,\cdots,x_n$，$Y=y_1,y_2,\cdots,y_n$，给 Y 求 X。如果 X 和 Y 之间不是一一对应关系，则一般不存在唯一“正确”的对应。统计的方法就是找近似正确的即最可能的对应 $\hat{X}$，使其在 Y 条件下的出现概率最大，即

$$\hat{X}=\arg\max_X P(X|Y)=\arg\max_X\frac{P(Y|X)P(X)}{P(Y)}=\arg\max_X P(Y|X)P(X)\approx\arg\max_X P(X) \tag{6}$$

这里，X 表示识别的汉字串，Y 表示拼音串。在不考虑多音字的情况下，$P(Y|X)=1$，因此只需关注 $P(X)$，利用 Bigram 统计语言模型给出的定义：

$$P(X)=P(x_1,x_2,\cdots,x_n)$$
$$=P(x_1)P(x_2|x_1)\cdots P(x_n|x_{n-1}) \quad (7)$$

后处理类似于上一小节的方法。应用语言模型不可避免地涉及到数据稀疏问题(data sparseness)，即由于语言模型的训练语料有限，测试集中许多词间搭配都未出现的现象[12]，可以采用一定的数据平滑方法来解决。目前所采用的数据平滑方法有加性平滑、Jelinek-Meccer 平滑、Katz 平滑和 Kneser-Ney平滑等[13,14]。

3 实验设计

实验针对特定人的唇读识别。整个系统包括特征提取、识别模型的训练、口型序列的拼音识别和文字流解析。改进后的系统流程如图 2。为了验证语言模型在唇读中的作用，我们对不同的特征提取方法下语言模型应用前后的识别率进行了对比。Gabor+PCA 和 DCT+PCA[11] 是目前比较有效的特征提取方法。Gabor 小波变换由于能够抓住图像局部区域内多个方向的空间频率和局部性结构特征，因而在特征提取中得到广泛应用。对整个唇区做 Gabor 变换，将导致维度过高，因此实验采用对唇区特殊点进行变换的方法，根据唇区的几何形状特征，选取 25 个特殊点提取 Gabor 特征。DCT (Discrete Cosine Transform)具有可分离和快速计算的特性，并被证明对特征提取是有效的。实验中采用分块 DCT 的方法，即将原始的 32×16 唇区分成 8 个 8×8 的块，分别进行 DCT，最后将 8 个块的特征联合作为整个唇区的特征。两种特征提取方法得到的特征矢量均采用 PCA(Principal Component Analysis)进行降维，以去除特征之间的相关性。

唇动过程和说话内容是双随机序列，采用 HMM 能够比较理想地描述唇动过程，实验中采用了半连续的 HMM——SCHMM，状态数 $N=6$，混合项数 $M=8$。

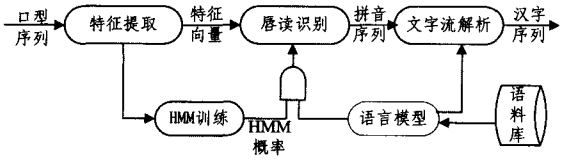

图 2 HLM 和文字流解析流程图

系统的具体实现步骤如下。

步骤 1 通过统计语料库建立拼音序列的统计语言模型，本实验中为 Bigram；

步骤 2 通过唇读训练建立 96 个音的 HMM 匹配模型；

步骤 3 对口型序列进行识别，将 HMM 识别概率进行归一化；

步骤 4 使用式(3)将 HMM 识别概率与统计语言模型概率联合，选择最大者输出。

4 实验结果及分析

实验采用的数据库是 HIT Bi-CAV Database，采集了 10 个人(5 男 5 女)简单背景下的正面人脸图像，图像采集速率为 25 帧/秒，存储为 256×256 大小的 24 位真彩图像。该库由 200 个句子组成，涵盖了汉语拼音中的 96 个音节，所有句子录制三遍，前两遍用来训练，最后一遍用来识别，每个音对应 5～25 幅的图片序列。所有实验都针对特定人，给出了在音的识别中应用 HLM 和在文字流解析阶段中应用语言模型的结果。

4.1 HLM

图 3 是采用 Gabor+PCA 方法提取不同维数的特征时，应用语言模型前后的识别率比较。

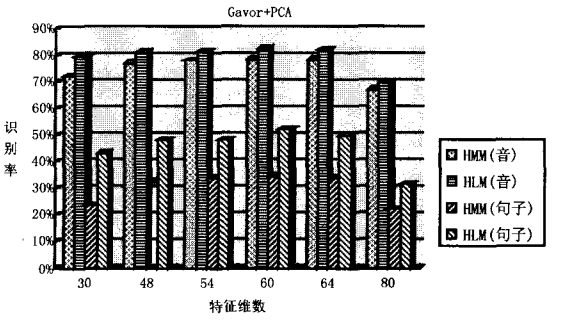

图 3 Gabor+PCA 方法下的性能比较

可以看出，结合语言模型后的 HLM 方法比单纯使用 HMM 识别性能有很大提高。对于单个音的识别，识别率最多提高 7.3%；句子识别率最多提高 19.5%。

HLM 方法首先要计算口型序列在所有 HMM 下的概率，最后只保留概率较大的前 m 个 HMM，然后联合语言模型进行判决。直观地想，m 值越大，即保留的候选越多，越不容易遗漏正确的识别结果，识别率很可能越高。基于此，就参数 m 对识别率的影响问题，我们进行了实验，结果如图 4 所示。

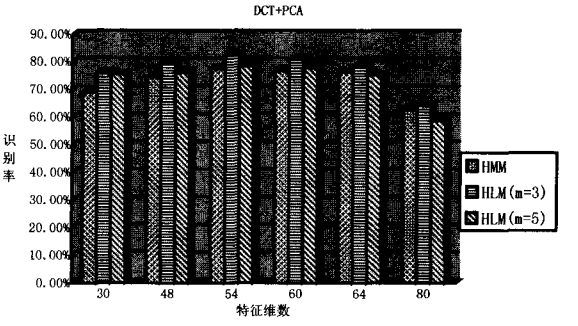

图 4 DCT+PCA 方法下的性能比较

与预想的不同，$m=3$ 时的识别率反而高于 $m=5$ 时的识别率。分析原因，我们认为，在没有建立能够提供完整先验信息的语料库之前，应该以 HMM 识别结果为主，而语言的先验知识作为辅助手段。由于语言模型在语料有限时往往训练不充分，如果候选的 HMM 过多，概率排名靠后的 HMM 结合了语言模型后，联合概率可能突然变大，从而造成误识。当然，这样的结论还有待我们在更完整的数据库上求证。值得注意的是，即使参数不同，使用 HLM 方法的总体性能还是优于 HMM。

综上所述，两种不同的特征提取方法下，应用 HLM 的识别率均比 HMM 高。原因在于 HLM 充分利用语言的先验知识，对 HMM 下误识的音进行一定程度的校正，使识别更加精确。我们的数据库中有 935 个汉字、96 个音，尽管由训练语料库得到的语言模型并没有给出完美的先验信息，但足以证明语言模型所带来的积极作用。表 1、表 2 是反映语言模型作用的例子。

表 1 “启动电子计时器”中“qi”的识别过程

	1	2	3	4	5
HMM 概率最大的前 5 个音	zi	qi	ying	ji	ni
HMM 下的条件概率	0.393	0.353	0.225	0.026	0.01
拼音出现在句首的概率	0.01	0.025	0.015	0.035	0.045
HLM 概率	0.0039	0.0088	0.0033	0.0009	0.0004

表 2 “dong”的识别过程

	1	2	3	4	5
HMM 概率最大的前 5 个音	duo	dong	zuo	chong	huo
HMM 下的条件概率	0.540	0.272	0.116	0.070	0.01
拼音在“qi”之后的概率	0.01	0.047	0.01	0.01	0.01
HLM 概率	0.0054	0.012	0.0011	0.0007	0.0001

显然，利用语言模型的先验知识后，一些 HMM 下概率最大，但是错误的候选音得到了一定的校正。

4.2 语言模型背景下的文字流解析

图 5 给出了 HLM 方法的音识别率和在此基础上利用语言模型进行文字流解析的实验结果。数据库中有 935 个汉字、96 个音，如不借助语言的先验知识，平均每个音对应 10 个汉字，盲目解析的正确率只有 10%。而应用语言模型的正确率最高可达 70.5%，对单一视觉流的文字解析而言是可观的。这充分说明了语言模型在文字流解析环节的强大作用。

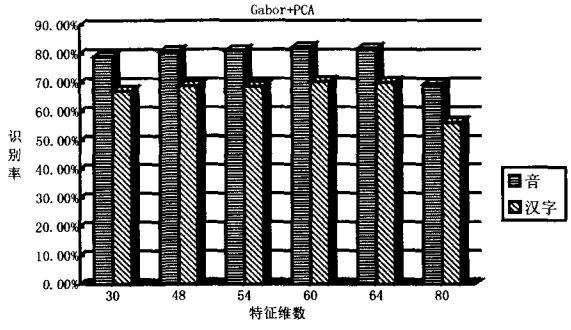

图 5 音识别与文字流解析性能比较

综上，语言模型对唇读的拼音识别和文字流解析环节都有积极的作用。但是，简单的语言模型存在一些问题，即当某个音识别错误时，很可能紧邻其后的音也被识别错误，类似于“多米诺骨牌效应”，尤其是句首音误识时，表 3 是一个例子。因此，语言模型有待我们改进成可回溯的模型，比如，加入反馈校正环节。

表 3 “类多米诺骨牌效应”

正确汉字	正确拼音	HMM	HLM
吃	chi	chi	shi
点	dian	dian	ye
东	dong	nong	nong
西	xi	xin	xin

结束语 本文针对唇读中口型序列和语言序列的一对多映射问题，主要研究语言模型对唇读的作用，突破单纯采用声学后验概率进行识别的传统框架，建立融合 HMM 和语言背景知识的新模型 HLM，并应用语言模型进行文字流解析。实验表明，建立语言模型能校正部分 HMM 识别错误的音，对单一视觉流的说话内容识别和文字流解析起着重要积极作用。它是由唇动序列得到语言序列唯一解的有效途径，必然成为未来实用唇读系统中不可或缺的一部分。语言模型在唇读中的应用虽有不少问题尚待解决，但并非不可解决。比如，解决“类多米诺骨牌效应”问题，可通过加入回馈环节等。

参 考 文 献

[1] Potamianos G, et al. Audio-Visual Automatic Speech Recognition: An Overview [M]. MIT Press, 2004

[2] Potamianos G, Graf H P, Cosatto E. An Image Transform Approach for HMM Based Automatic Lipreading[C]// Proc. Int. Conf. Image Processing. 1998, 1: 173-177

[3] Potamianos G, Neti C. Improved ROI and Within Frame Discriminant Features for Lipreading[C]// Proc. Int. Conf. Image Processing. Thessaloniki, Greece, 2001, 3: 250-253

[4] 姚鸿勋，高文，王瑞，等. 视觉语言——唇读综述[J]. 电子学报，2001, 29(2): 239-246

[5] Potamianos G, et al. Recent Advances in the Automatic Recognition of Audio-visual Speech[C]. Proc. of the IEEE, 2003, 91(9): 1306-1326

[6] Rosenfeld R. A Maximum Entropy to Adaptive Statistical Language Learning[C]. Computer Speech and Language, 1996, 10(3): 187-228

[7] Chomsky N. Aspects of the Theory of Syntax [M]. Cambridge: MIT Press, 1965

[8] Chomsky N. Syntactic structures [M]. Mouton, 1964

[9] 黄昌宁，张小凤. 自然语言处理技术的三个里程碑[J]. 外语教学与研究，2002, 34(3): 180-187

[10] 王晓龙，关毅. 计算机自然语言处理[M]. 北京：清华大学出版社，2005: 47-68

[11] Hong Xiaopeng, Yao Hongxun, Wan Yuqi, et al. A PCA Based Visual DCT Feature Extraction Method for Lip-reading[C]// Int. Conf. on Intelligent Information Hiding and Multimedia Signal Processing. 2006

[12] Katz S M. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer[C]. IEEE Transaction on Acoustic, Speech and Signal Processing, 1987, 35(3): 400-401

[13] Kneser R, Ney H. Improved Backing-off for M-gram Language Modeling// Proceedings of the IEEE[C]. Int. Conf. on Acoustics, Speech and Signal Processing. Detroit, MI, USA, 1995: 181-184

[14] Rosenfeld R. Two Decades of Statistical Language Modeling: Where Do We Go from Here? [C]// Proceedings of the IEEE. 2000, 88(8): 1270-1278

(上接第 136 页)

[33] Dong J X, Krzyzak A, Suen C Y. Statistical result of human performance on USPS database. Technical Report, CENPARMI. Concordia University, 2001

[34] 芮挺，沈春林，丁健. 基于最佳鉴别变换的 HMM 手写数字字符识别. 中国图像图形学报，2004, 9(8): 1008-1013

[35] 芮挺，沈春林，丁健，等. 基于主分量分析的手写数字字符识别. 小型微型计算机系统，2005, 26(2): 289-292

[36] Garrett S M. How Do We Evaluate Artificial Immune Systems? Evolutionary Computation, 2005, 13(2): 145-178