

# 高性能汉语数码串语音识别

李虎生, 刘 加, 刘润生

(清华大学电子工程系, 北京 100084)

**摘 要:** 本文给出了一个高性能汉语数码串非特定人连续语音识别系统, 其声学模型基于 Mel 倒谱系数和连续 HMM, 识别时采用多候选帧同步搜索算法, 并采用了 MCE 算法进行训练以提高系统的区分能力, 实验证明该系统的识别率为 94.8% (不定长数字串) 和 96.8% (定长数字串)。为增强系统的实用性, 本文还研究了基于 MAP 算法的说话人自适应算法和基于置信度的拒识算法。在进行自适应后, 误识率可相对下降 40% 以上, 在拒绝掉 5% 的正确语音时, 系统识别率可以上升到 96.9% (不定长数字串) 和 98.7% (定长数字串)。

**关键词:** 汉语数码串; 语音识别; 说话人自适应; 置信度

**中图分类号:** TN912 **文献标识码:** A **文章编号:** 0372-2112 (2001) 05-0595-05

## High Performance Mandarin Digit String Speech Recognition

LI Hu-sheng, LIU Jia, LIU Run-sheng

(Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

**Abstract:** A high performance mandarin digit string speaker-independent speech recognition system is given. The acoustic model is based on the Mel Frequency Cepstrum Coefficient and the continuous hidden Markov model (HMM). The multi-candidate frame synchronous search algorithm is adopted in the recognition stage with the MCE algorithm as the training approach. Experiments demonstrate that the correct recognition rate of the system is 94.8% (unknown length) and 96.8% (known length). In order to enhance the flexibility of the system, this paper also conduct research on the maximum a posteriori (MAP) based speaker adaptation and confidence measure based rejection. More than 40% recognition errors can be removed after adaptation and the recognition rate can be improved to 96.9% (unknown length) and 98.7% (known length) when 5% of the correct results are rejected.

**Key words:** mandarin digit string; speech recognition; speaker adaptation; confidence measure

### 1 引言

非特定人汉语数码语音识别 (Mandarin Digit Speech Recognition, MDSSR) 在电话语音拨号、信息查询、家电语音遥控、工业控制等诸多领域具有广泛的应用价值<sup>[1]</sup>, 因此在小词表语音识别中有着特殊的重要意义。孤立词的发音方式<sup>[2]</sup>不符合人的自然发音习惯, 给识别系统的使用者带来了很大不便, 因此对基于连续发音方式的数码串语音识别系统 (Mandarin Digit String Speech Recognition, MDSSR) 的研究是 MDSR 得到更广泛应用的前提。但目前 MDSSR 的性能距实用要求尚有一定距离, 目前文献报道的 MDSSR 非特定人无拒识不定长数字串的识别率尚无超过 95% 者<sup>[1]</sup>, 而与 MDSSR 相对应的英语数码串语音识别的串识别早已超过 99%<sup>[3]</sup>, 造成这一差距的主要原因在于汉语数码语音间的混淆程度远高于英语数码语音, 其中的“2”、“8”、“9”、“6”、“1”、“7”、“0”、“6”等语音极易产生混淆, 而且在连续发音时, 汉语数码语音的协同发音现象更为严重, 极易产生“11”被识别为“1”、“55”被识别为“5”或“22”被识别为“2”的情况。因此直接利用英语数字语音识别的技术

是无法获得较高的识别性能的。

本文提出了一个高性能的 MDSSR 系统, 采用了一系列方法提高了系统的识别性能, 并研究了说话人自适应和拒识算法。对测试语音, 在无拒识时系统的识别率为 94.8% (不定长数字串) 和 96.8% (定长数字串); 拒绝掉 5% 的正确语音后, 系统识别率可以上升到 96.9% (不定长数字串) 和 98.7% (定长数字串); 进行说话人自适应后, 识别错误可以减少 40% 以上。

### 2 识别系统的声学模型和识别搜索算法

#### 2.1 声学模型

MDSSR 系统的声学模型包括语音特征参数的提取和语音的建模。早期的实验表明在数码语音识别中 Mel 频标倒谱系数 (Mel Frequency Cepstrum Coefficient, MFCC) 参数的性能明显优于线性预测倒谱系数 (Linear Prediction Cepstrum Coefficient, LPCC) 参数<sup>[2-4]</sup>, 因此本文采用 MFCC 参数为语音特征参数。为体现语音的动态特性及能量对语音的区分作用, 还在

收稿日期: 2000-04-04; 修回日期: 2000-09-10

基金项目: 国家自然科学基金 (No. 69975007); 国家 863 项目 (No. 863-306ZD13-04-6)

语音特征中加入了一阶差分 MFCC 及其一阶能量和一阶差分能量,其中能量参数用语音平均能量进行了归一化.整个特征参数是一个 26 维的矢量,包括 12 维 MFCC 参数,12 维差分 MFCC 参数,归一化能量和一阶差分能量.

数码语音模型一共有 13 个,其中包括 11 个数字(“1”有 [yi]和[yao]两种念法),一个静音模型(Silence Model)和一个暂停模型(Pause Model).本文采用基于整词的上下文无关连续隐含马尔科夫模型(Hidden Markov Model, HMM)作为数码语音识别模型<sup>[5]</sup>.HMM 模型的输出概率分布为混合高斯密度函数.数码语音的 HMM 状态数为 7,静音模型和暂停模型的状态数为 1.实验证明最优的高斯混合分量数为 9.本文采用 Baum - Welch 算法训练基本的 HMM 参数.

用 HMM 描述语音的缺陷之一在于:HMM 中段长(包括状态长度和词的长度)的分布为几何分布,这是与实际情况不符合的<sup>[5]</sup>.造成 HMM 对段长的描述的不合理的原因在于传统 HMM 中状态跳转概率与当前状态已持续时间无关.通常可以采用与状态持续时间相关的状态跳转概率.目前主要采用分布、高斯来描述段长概率分布,或直接采用直方图描述段长分布信息.考虑到前两者均有人为的假设,而存储直方图所需的存储量与输出概率分布参数的存储量相比是很小的,因此本文采用直方图描述.即采用实际统计的段长数据来估计状态转移概率  $a_{ii}(d=)$ .  $a_{ii}(d=)$  表示当状态  $i$  已持续  $-1$  帧的条天下,在  $1$  帧仍然保持在状态  $i$  的概率.本文称  $a_{ii}(d=)$  为持续时间为  $d$  的状态跳转概率.  $a_{ii}(d=)$  是一个时变量.如果用  $p_i(d \geq)$  表示在  $i$  状态的持续时间  $d$  大于 的概率.根据  $a_{ii}$  的物理意义可以得出:

$$a_{ii}(d=) = \frac{p_i(d \geq)}{p_i(d \geq - 1)} \quad (1)$$

$p_i(d \geq)$  估计过程比较简单,只需用 Viterbi 算法将训练语音完成对各状态的对准,然后统计各状态持续时间的分布即可.在获得段长概率分布后,即可按式(1)计算出持续时间为  $d$  的状态跳转概率  $a_{ii}(d=)$ .考虑统计数据的不足,因此必须对小概率结果进行修正,这样就得到公式(2):

$$a_{ii}(d=) = \begin{cases} \frac{P_i(d \geq)}{P_i(d \geq - 1)}, & \text{if } > \\ 1 - \sum_{i=1}^{d-1} P_i(d=i), & \text{if } < \end{cases} \quad (2)$$

其中  $p_i(d=)$  表示统计所得的状态  $i$  的持续时间为 的概率, 为一个很小的正数.实验证明考虑段长分布后,识别性能会有明显的提高.

### 2.2 识别搜索算法

由于在识别数字串时,识别系统并不知道每个数字的交界点在什么地方,需要用搜索算法来获得识别的最佳数码串.本文采用基于 Viterbi 译码的多候选帧同步搜索算法.搜索的网络结构如图 1、图 2 所示.搜索网络由声学节点,即 HMM 各状态和句法节点(图中的圆圈所示)以及节点间的连接关系所构成.搜索时,在声学节点计算当前语音帧对于该状态的输出概率,选择前续节点中的前  $N$  个最优路径,并更新其路径信息;在句法节点则记录下路径的内容,即数字的序列,并将其

传递下去.当语音结束时,最后一个句法节点中前  $N$  最优路径信息中存储的数字序列即前  $N$  个识别候选结果.算法的具体流程参见文献[7].该算法可以输出多个候选结果由用户确认,提高系统的可靠性.并且可以在多候选的基础上实现最小分类错误算法和拒识等.

### 3 MCE 算法

在训练基本 HMM 参数时,通常采用最大似然(Maximum Likelihood, ML)准则.实验证明根据 ML 准则所得的系统的识别性能并不非常好.HMM 模型仅仅能够对语音进行近似的描述,除段长分布的不合理外,还表现在:(1)HMM 将语音划分为若干状态,在每一个状态内是稳态的,但语音只是准平稳信号,无法用有限的几个状态来描述.对于一些发音连续变化的语音,如“2”中的儿化音,HMM 则很难对其作准确的描述.(2)连续 HMM 假定状态输出概率为混合高斯分布函数,而真正的语音的分布是非常复杂的,很难用简单的形式表征.因此利用 ML 准则训练 HMM 参数时,只能获得次优的识别性能.尤其对于易混语音,很难达到满意的识别效果.90 年代以来,美国 Bell 实验室的 Biing-Hwang Juang, Chir-Hui Lee 等学者提出了最小分类错误(Minimum Classification Error Rate, MCE)准则<sup>[8]</sup>.该算法能够有效提高 HMM 模型对易混语音的区分能力.本文采用 MCE 算法进一步提高 MDSSR 的识别性能.

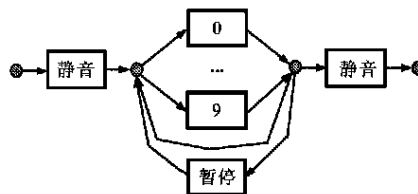


图 1 不定长数字串搜索网络

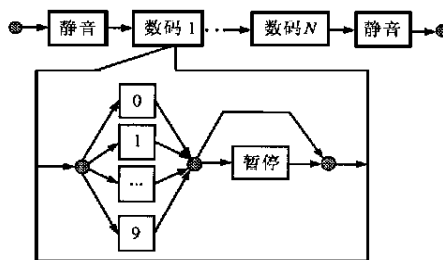


图 2 定长数码串的搜索网络

首先定义语音  $X$  的类误识测度(class misclassification measure)为公式(3):

$$d(X) = -s_r(X; ) + \frac{1}{N} \log \left[ \frac{1}{N} \sum_{j \in \text{错误的集合}} \exp(s_j(X; )) \right] \quad (3)$$

其中  $s_j(X; )$  表示语音  $X$  在参数为 的识别系统中候选的匹配分数(一般采用累积概率的对数);  $r$  表示正确的候选; 表示错误的候选的集合,其大小为  $N$ ; 为正常数.利用 Sigmoid 函数将式(3)值限制在 0 至 1 之间:

$$l(d(X)) = \frac{1}{1 + \exp(-d(X) + )} \quad (4)$$

在此基础上,定义 MCE 算法的目标函数为:

$$E[l(d(X))] = E\left[\frac{1}{1 + \exp(-d_i(X) + \mu)}\right]$$

$$= \frac{1}{M} \left[ \sum_{i=1}^M \frac{1}{1 + \exp(-d_i(X) + \mu)} \right] \quad (5)$$

其中  $E$  代表数学期望;  $i$  表示第  $i$  个识别语音模型;  $M$  表示识别语音个数;  $\mu$  为常数. 一般设为 0. 当  $d(X)$  远小于 0 时, 意味着识别的正确, 而  $d(X)$  大于 0 这意味着识别结果是错误的. 而  $l(d(X))$  是  $d$  的单调递增函数, 因此目标函数是误识率的单调递增函数. 目标函数的降低即直接意味这系统识别性能的改善.

MCE 算法训练的过程即对式 (5) 定义的目标函数的优化过程. 一般采用广义概率下降 (Generalized probabilistic descent, GPD) 法进行目标函数的优化. 考虑到均值矢量是识别系统最关键的参数, 因此本文仅对均值矢量采用 MCE 算法, 而保持协方差矩阵和混合分量系数不变. 采用 GPD 法时, 参数的调整满足:

$$t_{i+1} = t_i - \alpha \nabla l(X|s_i) \quad (6)$$

其中  $t$  表示迭代的节拍,  $\alpha$  为步长调节系数. 为计算方便起见, 令

$$s_j(X) = \sum_{t=1}^T \log(b_s^j(x_t)) \quad (7)$$

其中  $\{s_j^i\}_{i=1,2,\dots,T}$  为第  $j$  个候选串的最优状态序列,  $s_1(X)$  为正确的识别序列, 其它为错误识别序列. 这可以通过 Viterbi 算法对准后获得. 对均值矢量  $\mu_s^m$  优化时有:

$$\nabla_{\mu_s^m} l(d) = \frac{\partial l(d)}{\partial d} \nabla_{\mu_s^m} (d) \quad (8)$$

其中:

$$\frac{\partial l(d)}{\partial d} = l(d) (1 - l(d)) \quad (9)$$

$$\nabla_{\mu_s^m} d = - \sum_{t=1}^T (s_t^1 - s) \frac{b_s^1}{b_s^1} (x_t - \mu_s^m)$$

$$+ \sum_{n=2}^N r_n \sum_{t=1}^T (s_t^n - s) \frac{b_s^n}{b_s^n} (x_t - \mu_s^m) \quad (10)$$

其中  $\delta_{ij}$  为 Kronecker 函数.  $x_t$  为第  $t$  帧语音的特征矢量,  $b_s^m$  与  $b_s$  分别为状态  $s$  的总输出概率和第  $m$  分量的输出概率. 而  $r_n$  满足:

$$r_n = \exp(s_n) / \sum_{j=2}^N \exp(s_j) \quad (11)$$

在优化过程中, 目标函数式 (5) 中常数  $\mu$  的选择决定着目标函数与误识率是否直接相关, 因此对识别性能有较大的影响. 由式 (9) 可见, 当  $\mu$  较大时, 均值矢量的调整量主要由  $d$  的绝对值较小的识别结果 (即处于识别正确错误的临界点上的数字串) 决定; 而当  $\mu$  较小时, 均值矢量的调整量主要由  $d$  的绝对值较大 (即识别结果非常可靠或错误非常严重的数字串) 的识别结果决定. 由于对识别结果非常可靠的数字串进行调整对误识率的改进不大, 因此需要设置较小的  $\mu$ . 但此时又会忽略对识别错误严重的数字串的调整, 所以采用统一的是不合理的. 必须采用动态调整  $\mu$  的方案, 即对不同数值的  $d$  采用不同的  $\mu$ . 本文采用以下简单的方法:

$$d = \begin{cases} C_1, & d > 0 \\ C_2, & d \leq 0 \end{cases} \quad (C_1 < C_2) \quad (12)$$

其中  $C_1 = 0.01, C_2 = 0.03$ .

在连续数码串的识别中, 由于各种错误的情况很复杂, 在第二选以后的结果也包含了丰富的识别错误的信息, 例如数字串“259”识别时的第二、三选为“859”和“256”时, 第二、三选可以分别反应易混语音对“8”和“2”以及“9”和“6”的信息. 所以在 MCE 算法优化过程中, 考虑更多的候选可以更充分的利用信息, 进一步提高识别性能. 实验证明候选数大于 2 的情况优于候选数为 2 的情况.

### 4 说话人自适应

为提高汉语数码串语音识别性能, 本文引入了最大后验概率 (Maximum a posteriori, MAP) 说话人自适应算法, 利用少数的特定用户的语音, 使该用户的数码语音识别率有所提高<sup>[9]</sup>. 用 MAP 算法对数码串作自适应的过程包括: (1) 利用前向-后向算法计算各帧语音属于各状态的概率. (2) 利用式 (13) 更新 HMM 参数. 其中  $\mu$  为某状态某高斯分量自适应后的均值矢量,  $x_t$  为第  $t$  的语音特征矢量,  $m$  为自适应前的 (即非特定人) 均值矢量,  $\alpha$  为恒定常数.

$$\mu = \frac{\sum_{t=1}^T (x_t - m)}{T} + m \quad (13)$$

### 5 拒识算法

识别结果的错误往往会给用户造成不良的后果, 必须对识别结果的置信度进行判断, 提高系统识别结果的可靠性. 本文采用了一种基于多候选结果的置信度计算方法<sup>[10]</sup>. 对每一个识别出的数字和整个识别出的数字串都计算出其置信度, 然后根据其是否大于某阈值决定是否将其拒绝掉. 该置信度的计算使用了对每个候选路径的累积分数  $s^n, n = 1, 2, 3, \dots, N$ , 其中  $N$  为候选数. 设识别结果中需要计算置信度的语音单位的起止帧分别为  $t_s$  和  $t_e$ . 对于识别结果的某一语音单位, 令  $L$  为在范围  $R_s = [\max(1, t_s - L/2), \min(t_e + L/2, T)]$ ,  $l = (t_e - t_s)/2$  内存在相同语音单位的候选路径的集合. 则置信度可以定义为:

$$C = \frac{\sum_{n=1}^N \exp(-s^n)}{\sum_{n=1}^N \exp(-s^n)} \quad (14)$$

其中  $\alpha$  为一个经验常数. 该置信度计算方法的优点在于计算非常简单, 易于实时实现, 而且实验证明有良好的计算性能.

### 6 实验结果

实验中所用的语音数据为包括 40 个说话人的汉语数码语音库. 其中每人念 80 个数字串 (平均串长为 4), 语音采样频率为 16kHz. 测试识别率时, 一共测试 4 遍, 每一遍用不同的 10 个说话人进行测试, 识别系统参数由其他 30 人训练数据所得, 最后的识别性能为 4 遍测试的平均结果.

#### 6.1 段长信息的作用

表 1 给出在模型中增加段长参数后系统的识别性能. 在

无段长信息的情况下,状态转移概率都设为 0.5.由表 1 可见,对于不定长数字串的情况,引入段长信息后识别率提高了约 1 个百分点;而对定长数字串的作用则是非常有限的.

表 1 引入段长信息后系统性能的变化

识别率	有段长信息		无段长信息	
	不定长数字串	定长数字串	不定长数字串	定长数字串
	91.2 %	95.4 %	90.1 %	95.2 %

6.2 MCE 算法性能

首先验证动态调整 对 MCE 算法的影响.图 3 给出了动态调整 (情况 2) 和恒定 (情况 1) 两种情况下优化过程中误识率的变化曲线.图中 TRAIN 表示训练集,TEST 表示测试集.实验表明采用了动态调整后,优化最终结果和收敛速度都有了明显的改善.其次验证了增加候选数对 MCE 算法性能的影响.图 4 给出了不同候选数时的误识率的变化.其中情况 1 中候选数为 2,情况 2 中候选数为 5.从图中可以看出多候选有助于 MCE 算法性能的改善与提高.

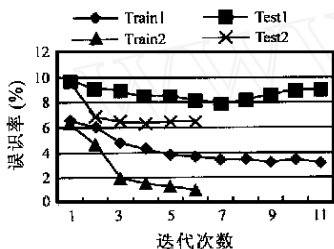


图 3 不同调整方式下 MCE 算法误识率比较

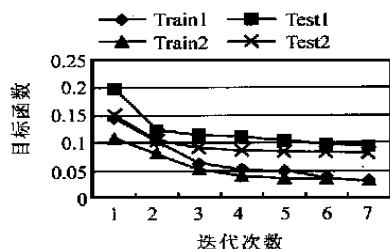


图 4 不同候选数时 MCE 算法误识率比较

识别系统的整体性能见表 2 与表 3.表 2 为首选识别率;表 3 为多候选的识别性能.从表 2 与表 3 中可以看出定长数码串的识别率高于不定长数码串的识别率.对不定长数码串识别,识别的替换错误占很大比例;对定长数码串识别,识别的删除错误占很大比例.

表 2 系统首选识别性能

串识别率	不定长数字串			串识别率	定长数字串		
	各种错误所占百分比				各种错误所占百分比		
	替换	插入	删除		替换	插入	删除
94.8 %	49.7 %	24.2 %	26.1 %	96.8 %	4.46 %	19.6 %	75.9 %

表 3 系统多候选识别性能

不定长数字串				定长数字串			
第一选	第二选	第三选	第四选	第一选	第二选	第三选	第四选
94.8 %	97.3 %	98.2 %	98.4 %	96.8 %	98.8 %	99.1 %	99.2 %

6.4 说话人自适应性能

在实验中还测试了不同自适应语音数量时的不定长数码串自适应性能.测试结果为语音库中的 10 个说话人的分别作自适应后的性能平均所得,其结果见图 5,图中 SI 表示非特定人误识率,SA 表示自适应后的误识率.可以看到,当自适应语音较多时(40 个数字串),系统对于特定说话人的误识率下降了 44 %,而且自适应语音越多,自适应的效果越好.

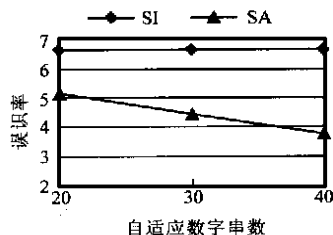


图 5 自适应算法性能

6.5 拒识性能

本文还测试了利用置信度对识别结果作拒识后的识别性能.其中正确接收率表示接收(未作拒识)正确结果的百分比,而错误接收率表示接收错误结果的百分比.串和数字的识别率是对于接收语音的识别正确结果的百分比.由表 4 和表 5 可见,采用拒识算法后可以拒绝掉大量的错误识别结果,有效提高系统的识别率和可靠性.

表 4 整句拒识性能

	不定长数字串			定长数字串		
	正确接收率	错误接收率	串识别率	正确接收率	错误接收率	串识别率
正确接收率	100 %	95 %	90 %	100 %	95 %	90 %
错误接收率	100 %	60 %	42 %	100 %	42 %	35 %
串识别率	94.8 %	96.9 %	97.8 %	96.8 %	98.7 %	98.9 %

表 5 单个数字拒识性能

	不定长数字串			定长数字串		
	正确接收率	错误接收率	串识别率	正确接收率	错误接收率	串识别率
正确接收率	100 %	95 %	90 %	100 %	95 %	90 %
错误接收率	100 %	64 %	53 %	100 %	36 %	27 %
串识别率	98.7 %	99.2 %	99.3 %	99.1 %	99.7 %	99.8 %

7 结论

本文提出了一个高性能的 MDSSR 系统.达到相对可观的识别性能,满足初步实用的要求.系统首选识别率为 94.8 % (不定长数字串)和 96.8 % (定长数字串),而多选识别率可达 97 % (不定长数字串)和 98 % (定长数字串)以上;采用 MAP 算法进行自适应(自适应语音为 40 个数字串)后,对于 10 个说话人其平均误识率下降 44 %;采用置信度进行拒识后,在拒绝掉 5 % 的正确数字串时,系统识别率为 96.9 % (不定长数字串)和 98.7 % (定长数字串),在拒绝掉 5 % 的正确数字时,识别错误的数字可以拒绝掉 36 % (不定长数字串)和 64 % (定长数字串).

虽然本文提出的 MDSSR 系统已经达到相当好的识别性能,但与英语数码串非特定人语音识别系统相比还有距离.进一步的研究工作将包括改进识别模型结构,增加音调模型等细化工作.

中科院自动化所为本文提供了实验所用的数码语音库,

在此表示衷心感谢。

#### 参考文献:

- [ 1 ] 顾良. 汉语数码语音识别方法研究及 DSP 系统设计 [D]. 清华大学工学博士论文, 1997, 1.
- [ 2 ] 李虎生, 刘加, 刘润生. 高性能汉语数码语音识别算法 [J]. 清华大学学报(自然科学版), 2000, 40(1): 32 - 34.
- [ 3 ] Y. Normandin et al. High-performance connected digit recognition using maximum mutual information estimation [J]. IEEE Trans. Speech and Audio Processing, 1994, 2(2): 299 - 311.
- [ 4 ] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Trans. on ASSP, 1980, 28(4): 357 - 366.
- [ 5 ] 赵庆卫, 肖熙, 王作英, 陆大. 段长信息在连续语音识别中的应用研究 [J]. 声学学报, 2000, 25(2): 175 - 181.
- [ 6 ] C. H. Lee, L. R. Rabiner. A Frame-synchronous network search algorithm for connected word recognition [J]. IEEE Trans. on Acoustics, Speech and Signal Processing, 1989, 37(11): 1649 - 1658.
- [ 7 ] 李虎生. 汉语数码串语音识别及说话人自适应 [D]. 清华大学工学硕士论文, 2000, 6.
- [ 8 ] B. H. Juang, W. Chou, C. H. Lee. Minimum classification error rate methods for speech recognition [J]. IEEE Trans. on Speech and Audio Processing, 1997, 5(3): 257 - 265.
- [ 9 ] 李虎生, 杨明杰, 刘润生. 汉语数码语音自适应算法 [J]. 电路与系统学报, 1999, 4(2): 1 - 6.
- [ 10 ] Bernhard Rueber. Obtaining confidence measures from sentence probabilities [A]. ICASSP '97 [C]: 1200 - 1204.

#### 作者简介:

**李虎生** 1998年7月获得清华大学电子工程系无线电技术专业学士学位, 2000年7月获得清华大学电子工程系通信与电子系统硕士学位. 目前在美国 Princeton 大学攻读博士学位.

**刘加** 1983年7月获得清华大学电子工程系无线电技术专业学士学位, 1986年7月获得清华大学电子工程系通信与电子系统硕士学位, 1990年4月获得清华大学电子工程系通信与电子系统博士学位. 1990年4月~1992年8月中科院遥感卫星地面站从事美国6号陆地卫星图象处理系统开发工作. 1992年9月~1994年2月在英国剑桥大学作博士后, 从事语音识别系统研究工作. 现任清华大学电子系副教授. 中国电子学会高级会员, 美国 IEEE 会员. 目前研究方向包括: 语音识别、语音合成、语音编码、语音识别专用芯片设计, 多传感器融合技术, 以及多媒体数字通信系统.