基于段长分布的 HMM 语音识别模型

王作英, 肖 熙
(清华大学电子工程系, 北京 100084)

摘 要: 本文针对齐次 HMM 语音识别模型在使用段长信息时存在的缺陷, 形式化地定义了一种适合语音信号描述的自左向右非齐次隐含马尔科夫模型, 证明了这种模型的状态转移概率表示与状态段长表示的等效性, 并在此基础上提出了基于段长分布的 HMM 模型(DDBHMM). 非特定人连续语音实验结果表明, 仅仅利用状态段长信息的 DDBHMM 语音识别模型比经典 HMM 模型的性能有了明显的提高(误识率降低了 17.8%), 展示了 DDBHMM 的良好的性能, 为语音信号的时长、语速、时间断续性以及语音特征的相关性等重要特征的描述和利用开辟了空间.

关键词: 段长; 语音识别; DDBHMM
中图分类号: TN912.34 文献标识码: A 文章编号: 0372-2112 (2004) 01-0046-04

Duration Distribution Based HMM Speech Recognition Models

WANG Zuo-ying, XIAO Xi
(*Dept. of Electronic Engineering, Tsinghua University, Beijing 100084, China*)

Abstract: In order to overcome the defects of the duration modeling of homogeneous HMM in speech recognition, a Duration Distribution Based HMM (DDBHMM) is proposed in this paper based on a formalized definition of a left-to-right inhomogeneous Markov model, which has been demonstrated that it can be identically defined by either the state duration or the state transition probabilities. The speaker independent continuous speech recognition experiments have shown that, by only modeling the state duration in DDBHMM, a significant improvement (17.8% error rate reduction) has been achieved comparing with the classical HMM. The ideal properties of DDBHMM will give promise to many aspects of speech modeling, such as the modeling of the state duration, speed variation, speech discontinuity and the inter frame correlation.

Key words: duration; speech recognition; DDBHMM

1 引言

在经典的HMM[1]语音识别模型中, 为了方便研究, 假设模型中的状态转移具有齐次 Markov 性, 并且模型的观测特征是帧间独立的, 从而可以使用 Baum-Welch 算法[2]进行 HMM 模型的训练和识别. 在这种齐次 Markov 性假设的前提下, 经典HMM 模型状态 i 的驻留概率 $a_{i,i}$ 为常数, 系统进入状态 i 后在该状态连续驻留的时间 τ(即段长) 是服从几何分布

$$P(\tau)= a_{i,i}^{\tau-1}(1- a_{i,i}),\ \tau \geq 1 \qquad (1)$$

许多实验统计都表明, 经典 HMM 模型这种段长的几何分布形式不能很好地描述语音的段长特征[3-5]. 为此, Ferguson 和 Levinson 等人对模型进行改进, 将状态段长概率直接引入到 HMM 模型中, 相继提出了 VDHMM 模型[6]和 CVDHMM 模型[7]. VDHMM 和 CVDHMM 模型的状态驻留概率由决定, 但是状态间的转移仍然定义为常数, 保留了齐次 Markov 性的特点, 这种模型又常被称作 HSMMs 模型(Hidden Semi-Markov Models)[8,9]. 常用的段长分布形式有 Gamma 分布, Gaussian 分布, Poisson 分布和均匀分布等. 也可以应用指数簇类中的若干个函数的组合来拟合实际的段长分布[10]. 此外, 也有学者研究用 ESHMM(Expanded State HMM) 模型[11,12]来增强 HMM 模型对段长信息的表达能力.

研究表明在 VDHMM 和 CVDHMM 模型中, 由于 a_{ij} 是常数, 因此不能保证在任意时刻 τ, Markov 链中的概率公式 $a_{ii}(\tau)+\sum_{j=i+1}^{N} a_{ij}=1$ 总是成立. 所以这种模型存在明显的理论上缺陷. 而 ESHMM 模型是属于齐次的 HMM 模型, 所能描述的段长分布形式有限, 当子状态的结构复杂时, 不利于构造快速算法. 显然比较彻底的解决方案是采用非齐次的 HMM 模型[13]. 本文作者[14]和 Vaseghi[15,16]等都分别先后提出在自左向右的 HMM 模型中利用段长驻留概率表示状态的转移概率的方法, 我们认为这是一种较好的表示非齐次 HMM 模型的方法. 但是 Vaseghi 提出的方案仅局限于无状态跨越假设的特殊情况, 存在应用上的局限性. 就此问题, 在本文的第二部分对采用段长分布表示非齐次 HMM 模型的方法进行了理论研

收稿日期: 2002-04-15; 修回日期: 2003-06-08

究, 提出了基于段长分布的 HMM 模型(简称 DDBHMM 模型). 在本文的第三部分给出了采用 DDBHMM 模型的实验结果, 在第四部分总结了 DDBHMM 模型的特点.

2　基于段长分布的 HMM 模型

对于自左向右的Markov 模型, 定义状态的驻留概率为

$$\begin{aligned} a_{i,i}(k) &= \text{Prob}(k \text{ 时间系统仍处于状态} i \mid \text{系统在} k-1 \text{时刻处于状态 } i, h_i) \\ &= \text{Prob}(\tau_i \geq k \mid \tau_i \geq k-1, h_i), \\ & \qquad i = 1, 2, \Lambda, N; k \geq 1 \end{aligned} \tag{2}$$

在式(2)中, $a_{i,i}(k)$ 为系统在离散时刻 k 停留在状态 i 的驻留概率, τ_i 为系统在状态 i 连续驻留的时间长度, 简称为段长. k 从系统进入状态 i 后开始计时, h_i 表示系统进入状态 i 之前所经历的历史事件. 定义式(2)把 Markov 链的状态驻留概率与段长分布概率直接联系起来了, 状态驻留概率可以用段长分布概率来确定, 即

$$a_{ii}(k) = \begin{cases} \dfrac{P_i(\tau_i \geq k \mid h_i)}{P_i(\tau_i \geq k-1 \mid h_i)}, & k > 0 \\ 1, & k = 0 \end{cases} \tag{3}$$

在下面的推导中, 为了书写简明, 在不影响理解和公式证明的情况下, 我们将 $P_i(\tau_i \geq k \mid h_i)$ 写成 $P_i(\tau_i \geq k)$, 将 $P_i(\tau_i = k \mid h_i)$ 写成 $P_i(\tau_i = k)$等.

为了定义系统的转移概率, 我们把从状态 i 到状态 $i+m$ $(m>1)$的跨越转移等效于系统离开状态 i 后, 滑过状态 $i+1$、状态 $i+2$、……一直到状态 $i+m-1$(其驻留时间长度为零), 而在状态 $i+m$ 发生了停留(即 $\tau_{i+m} \geq 1$)的这一系列事件. 因此我们定义状态的转移概率为:

$$\begin{aligned} a_{i,i+m}(k) &= Prob(\tau_i = k-1 \mid \tau_i \geq k-1) \cdot \prod_{k=i+1}^{i+m-1} P_k(\tau_k = 0) \\ &\quad \cdot P_{i+m}(\tau_{i+m} \geq 1) \\ &= \frac{P_i(\tau_i = k-1)}{P_i(\tau_i \geq k-1)} \prod_{k=i+1}^{i+m-1} P_k(\tau_k = 0) \\ &\quad \cdot (1 - P_{i+m}(\tau_{i+m} = 0)) \end{aligned} \tag{4}$$

在式(4)中 $Prob(\tau_i = k-1 \mid \tau_i \geq k-1)$ 表示在 $k-1$ 时刻系统处于状态 i 的条件下, 在 k 时刻系统离开状态 i 这个事件发生的概率. 我们需要特别强调的是, 这里把"从状态 i 到状态 $i+m$ $(m>1)$的跨越转移"与"滑过状态 $i+1$、状态 $i+2$、……一直到状态 $i+m-1$(其驻留时间长度均为零)"这样两个事件紧密地联系起来, 造成了这里讨论的模型与主流 HMM 模型中的自左向右模型的本质差别. 后者把这两件事情看成是完全独立的. 但是我们知道, 在实际的语音产生过程中这两件事情却是同时发生的. 由于主流 HMM 模型中没有包含这一重要信息, 那么从信息论的观点看就肯定会给识别系统的性能带来损失. 为了区别于目前流行的自左向右模型的概念, 我们把此处定义的模型称为严格自左向右的 Markov 模型.

上述定义的严格自左向右 Markov 模型是非齐次的 Markov 模型, 它具有下述性质:

(1) 状态驻留概率(或转移概率)与状态段长概率是一一对应的. 式(3)、(4)是表明状态驻留概率和转移概率可以用段长概率表示. 反之状态段长概率也可以用状态驻留概率表示, 即

$$\begin{aligned} P_i(\tau_i = k) &= P_i(\tau_i \geq k) - P_i(\tau_i \geq k+1) \\ &= a_{ii}(1)\, a_{ii}(2) \Lambda a_{ii}(k)(1 - a_{ii}(k+1)) \end{aligned} \tag{5}$$

这种一一对应说明: 一个严格自左向右的 Markov 链既可以用状态驻留概率和转移概率来表示, 也可以用状态驻留的段长概率来表示.

(2) 状态转移概率不是独立变量而是由状态驻留概率唯一确定. 由式(4)、(5)可得

$$a_{i,i+m}(k) = (1 - a_{i,i}(k)) \prod_{j=i+1}^{i+m-1} (1 - a_{j,j}(1)) \cdot a_{i+m,i+m}(1) \tag{6}$$

这个性质说明: 若在 HMM 中采用本文定义的严格自左向右的 Markov 结构, 不能也无需对状态的转移概率和驻留概率分别进行训练, 只需从训练数据中获得段长概率的分布就可以唯一地确定这个非齐次Markov 模型.

(3) 容易证明

$$\sum_{j=i}^{N} a_{ij}(k) = 1 - \frac{P_i(\tau_i = k-1)}{P_i(\tau_i \geq k-1)} \prod_{j=i+1}^{N} P_j(\tau = 0) \tag{7}$$

这表明, 如果不是所有的状态都不可跨越的($P_i(\tau = 0) = 0$, $i = 1, 2, \Lambda, N$), 则系统在某个状态驻留或转移的概率和小于 1, 特别是当 $i = N$ 时上式退化成为 $a_{NN} = 1$, 系统永远停留在状态 N. 此时如果用状态 N 来表示语音单元, 则意味着该状态的段长是无限大的, 这当然是违背事实的. 因此, 必须在系统中补充一个吸收态($N+1$). 其中

$$a_{i,N+1}(k) = \begin{cases} \dfrac{P_i(\tau_i = k-1)}{P_i(\tau_i \geq k-1)} \prod_{j=i+1}^{N} P_j(\tau_j = 0), & i = 1, \Lambda, N-1 \\ 1 - a_{NN}(k) = \dfrac{P_N(\tau_N = k-1)}{P_N(\tau_N \geq k-1)}, & i = N \\ 1, & i = N+1 \end{cases} \tag{8}$$

这样可以使式(9)成立

$$\sum_{j=i}^{N+1} a_{ij}(k) = 1, \quad i = 1, \Lambda, N+1 \tag{9}$$

现在我们用段长概率来表示本文所定义的这个严格自左向右的非齐次 Markov 模型, 将它应用于语音识别, 把模型中的状态与语音中的音素或音节等语音单元对应起来, 而把这些语音单元读音的语音信号特征作为对应语音单元的观测量, 我们就得到了一个基于段长分布的 HMM 模型(Duration Distribution Based Hidden Markov Models), 简称为 DDBHMM 模型(图 1). 显然这是一个非齐次隐含 Markov 模型. DDBHMM 的模型参数为 $\lambda = (B, D)$, 其中矩阵 $B = [b_1(o), b_2(o), \Lambda b_N(o)]$ 是状态的特征观测概率矩阵, 矩阵 $D = [d_1(\tau), d_2(\tau), \Lambda, d_N$

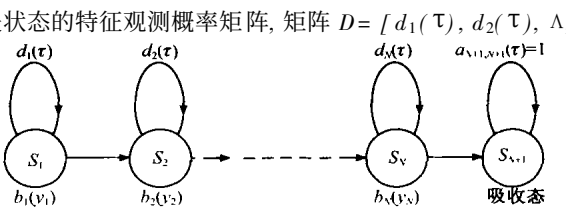

图 1　DDBHMM 模型结构

$(\tau)]$是状态的段长概率矩阵, N 是模型状态的个数.

为了计算 DDBHMM 模型产生观测矢量 $O_1^T=[o_1, \Lambda, o_T]$ 的概率, 假设 $S_1^T=[s_1, \Lambda, s_T]$ 为观测矢量 O_1^T 所对应的状态序列, τ_i 为系统在状态 i 的段长, 用 $t_i=\sum_{k=1}^{i}\tau_k(1\le i\le N)$ 表示状态序列 S_1^T 的分割点$(t_0=0, t_N=T)$. 于是对于给定的 DDBHMM 模型 λ, 可以推导得出式(10)

$$\begin{aligned}P(O_1^T, S_1^T|\lambda) &= a_{1,1}(1)\prod_{i=1}^{N}\prod_{k=2}^{\tau_i}a_{i,i}(k)\cdot a_{i,i+1}(\tau_i+1)\\ &\cdot\prod_{i=1}^{N}b(o_{t_{i-1}+1}^{t_i}|o_1^{t_{i-1}})\\ &=\prod_{k=1}^{\tau_i}d_k(\tau_k|\tau_1^{k-1})\,b(o_{t_{k-1}+1}^{t_k}|o_1^{t_{k-1}})\end{aligned}\qquad(10)$$

这是一个相当一般化的描述: 它既包容了状态段长之间的相关, 也包含了帧间特征之间的相关, 而且段长分布函数 $d_k(\tau_k|\tau_1^{k-1})$ 可以取任何形式$(k=1, 2, \Lambda, N)$. 因此 DDBHMM 模型具备了很强的对语音的描述能力. 在实际的语音识别应用中可以进行各种简化. 例如, 假设观测矢量是段(状态)间不相关时, 式(10)可以写成

$$P(O_1^T, S_1^T|\lambda)=\prod_{k=1}^{N}d_k(\tau_k)\,b_k(o_{t_{k-1}+1}^{t_k})\qquad(11)$$

我们看到式(11)结果与 SM[17]语音分段模型相似. 事实上, SM 模型只是 DDBHMM 模型不考虑状态跨越的一个特例. 当选择 DDBHMM 模型中的段长分布为几何分布时, 经典的 HMM 模型是 DDBHMM 模型的特殊情形, 即

$$d_i(\tau)=(1-a_{i,i})a_{i,i}^{\tau},\ i=1, \Lambda, N;\ \tau\ge 0\qquad(12)$$

这时系统的状态驻留概率为常数 $a_{i,i}$, $i=1, 2, \Lambda, N$. 状态间的转移概率

$$\begin{aligned}a_{ij}=&(1-a_{i,i})(1-a_{i+1,i+1})\Lambda(1-a_{j-1,j-1})a_{j,j},\\ &j=i+1, i+2, \Lambda, N+1\end{aligned}\qquad(13)$$

值得注意的是, 这里的转移概率 $a_{i,j}, j=i, i+1, \Lambda, N+1$ 是驻留概率的函数, 而不是像在目前广为流行的 HMM 模型中那样把转移概率看成是独立变量.

DDBHMM 模型中的非齐次 Markov 过程是有后效性的过程, 不能使用诸如 Viterbi、Baum-Welch 等基于 Bellman 动态规划理论的解码算法和参数重估算法, 需要有新的、高效的模型训练和解码算法. 在文[18]中给出了获得式(10)中的最佳分割点 t_1, t_2, Λ, t_N 快速算法. 此算法采用帧同步算法搜索最优路径. 在对每一帧进行剪枝时, 保证最优路径不会被错误地删除. 在连续语音识别中采用此方法进行最优路径搜索的计算量可以比全搜索的方法下降 3 个数量级以上. 识别时采用 MLSS 算法[19], 通过搜索和比较各个模型下的最优状态路径, 采用式(14)就可以完成语音识别任务.

$$\lambda=\arg\max_{(\lambda,S)}P(O_1^T, S_1^T|\lambda)$$
$$=\arg\max_{(\lambda;t_1,t_2,\Lambda,t_N)}\prod_{k=1}^{N}d_k(\tau_k|\tau_1^{k-1})\,b_k(o_{t_{k-1}+1}^{t_k}|o_1^{t_{k-1}})\qquad(14)$$

至此, 本文给出了完整的 DDBHMM 语音识别模型. 它能适用于采用任何段长概率分布的自左向右的 HMM 模型, 在理论上它能包容目前常用的经典 HMM 模型. 经典 HMM 所固有的段长几何分布形式的局限性已经被 DDBHMM 模型很好地解决.

3 DDBHMM 语音识别试验

为了验证 DDBHMM 语音识别模型的性能, 我们进行了非特定人孤立字和连续语音识别试验. 用一个 6 状态的 DDBHMM 模型来表示一个汉字的发音. 语音信号的采样率为 16KHz, 语音的帧长为 20mS, 帧移为 10mS. 每帧语音的归一化能量及 14 维 MFCC 系数连同它们的一阶差分、二阶差分系数共同构成了 45 维的特征矢量. 特征的观测概率采用 45 维全协方差阵的高斯分布函数, 段长分布采用 1 维的高斯分布.

用于非特定人孤立字语音识别试验的数据是汉语 1254 孤立字全音节的录音数据, 共有男、女声数据各 50 人. 采用轮流训练和识别的方法, 即对男、女声数据都分别用其中 49 人的录音数据用于模型训练, 剩余 1 人的录音数据用于识别测试, 这样就分别得到男、女声各 50 个非特定人的识别结果. 表 1 给出了非特定人孤立字识别结果的平均值. 可以看出采用 DDBHMM 算法的实验结果较为理想, 平均的汉字音节的识别率达到 90.68%, 与经典的齐次 HMM 模型相比, 音节识别的正确率提高了 0.85%, 识别错误率下降了 8.36%.

表 1 1254 全音节孤立字 DDBHMM 模型识别率

实验模型	男声识别率(%)	女声识别率(%)	平均识别率(%)
齐次 HMM	89.80	89.85	89.83
DDBHMM	90.41	90.95	90.68

在连续语音识别试验中, 采用的是"863"计划提供的男、女声各 83 人的连续语音录音数据. 其中, 男声数据共有 48348 个句子, 包含 591925 汉字音节, 女声数据有个 48372 句子, 包含 588082 个汉字音节. 在实验中对男、女声数据分别采用其中的 70 人数据作为训练集数据, 剩余的 13 人作为测试集数据进行声学层的语音识别实验(没有使用语言模型).

表 2 为采用 DDBHMM 模型进行连续语音识别的实验结果. 正如我们所预期的, 在采用了语音段长信息后, 删除错误

表 2 非特定人 DDBHMM 模型连续语音识别率(%)

模 型	齐次 HMM			DDBHMM			改善值		
	男 声	女 声	平均值	男 声	女 声	平均值	男 声	女 声	平均值
正确率(%)	73.06	69.68	71.37	75.65	74.02	74.84	2.59	4.34	3.47
替换错误(%)	25.79	29.66	27.73	22.25	24.39	23.32	3.54	5.27	4.41
插入错误(%)	2.23	4.45	3.34	0.90	1.10	1.00	1.33	3.35	2.34
删除错误(%)	0.96	0.59	0.78	2.09	1.60	1.85	-1.13	-1.01	-1.07
总错误率(%)	28.98	34.70	31.84	25.24	27.09	26.17	3.74	7.61	5.68

将有所增加,而插入错误会减少. 与经典的齐次 HMM 识别模型相比,本文提出的 DDBHMM 模型对男、女声非特定人连续语音识别的正确率平均提高了 3.47%,识别总错误率的绝对值下降了 5.68%,相对值下降了(5.68/31.84) * 100% = 17.8%,由此可见 DDBHMM 语音识别模型具有良好的识别性能.

4 总结

本文提出了新的基于段长分布的 DDBHMM 模型,这是一个具有严格自左向右拓扑结构的一般非齐次隐含马尔科夫模型. 由于 DDBHMM 模型对语音信号产生的顺序性和实际语音的段长分布都做了合理的描述,它不仅在理论上解决了经典 HMM 模型在描述段长信息方面存在的缺陷,而且给语音信号的时长、语速、时间断续性以及语音特征信号的相关性等重要特征的描述和利用开辟了空间. 非特定人连续语音实验结果表明,仅仅利用状态段长信息的 DDBHMM 语音识别模型就比经典的 HMM 模型的误识率降低了 17.8%,展示了 DDBHMM 的良好的性能. 这说明在语音识别中采用 DDBHMM 模型在理论与实践上是和谐的.

参考文献:

[1] Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition [J]. Proceedings of the IEEE, 1989, 77(2): 257–286.

[2] S E Levision, L R Rabiner, M M Sondhi. An introduction to the application of the theory of probabilistic fucntions of a markov process to automatic speech recognition [J]. The Bell System Technical Journal, April 1983, 62(4): 1035–1074.

[3] 齐士玲, 张家录. 汉语普通话辅音音长分析 [J]. 声学学报, 1982, 7(1): 8–13.

[4] Burshtein D. Robust parametric modeling of durations in hidden Markov models [A]. Proceedings of IEEE ICASSP [C]. Detroit, MI, USA, ICASSP-95, 1995. 1: 548–551.

[5] Wei-wen HUNG, Hsiao-chuan WANG. Improvement of noisy speech recognition using a proportional alignment decoding algorithm in the training phase [J]. Computer Speech and Language, 1998, 12(3): 165–192.

[6] J D Ferguson. Variable duration models for speech [A]. Proc. of Symposium on the Application of Hidden Markov Models to Text and Speech[C]. Princeton, New Jersey: Institute for Defense Analyses Communications Research Division, 1980. 143–179.

[7] S E Levinson. Continuously variable duration hidden markov models for automatic speech recognition [J]. Computer Speech and Language, 1986, 1: 29–45.

[8] M J Russell, R K Moore. Explicit modeling of state occupancy in Hidden Markov Models for automatic speech recognition [A]. Proceedings of IEEE ICASSP[C]. ICASSP-85, Tampa, Florida, USA, 1985. 5–8.

[9] Ratnayake N, Savic M, Sorensen J. Use of semi-Markov models for speaker-independent phoneme recognition [A]. Proceedings of IEEE ICASSP [C]. San Francisco, CA, USA, ICASSP-92, 1992, 1. 565–568.

[10] Mitchell C D, Jamieson L H. Modeling duration in a hidden Markov model with the exponential family [A]. Proceedings of IEEE ICASSP [C]. Minneapolis, MN, USA, ICASSP-93, 1993, 2. 331–334.

[11] Martin J Russel, Anneliese E Cook. Experimental Evaluation of Duration Modeling Techniques for Automatic Speech Recognition [A]. Proceedings of IEEE ICASSP[C]. ICASSP-87, Dallas, USA, 1987. 2376–2379.

[12] Bonafonte A, Vidal J, Nogueiras A. Duration modeling with expanded HMM applied to speech recognition [A]. Proceedings of the Fourth International Conference on Spoken Language, Philadelphia [C]. PA, USA, ICSLP 96, 1996, 2. 1097–1100.

[13] Ramesh P, Wilpon J G. Modeling state durations in hidden Markov models for automatic speech recognition [A]. Proceedings of IEEE ICASSP [C]. San Francisco, CA, USA, ICASSP-92, 1992, 1. 381–384.

[14] 王作英. 基于段长分布的HMM 语音识别模型 [A]. 第二届全国汉字、汉语识别会议 [C]. 庐山, 1989, 9.

[15] S V Vaseghi. Hidden markov models with duration-dependent state transition probabilities [J]. Electronics Letters, 1991, 27(8): 625–626.

[16] Vaseghi. State duaration modeling in hidden moarkov models [J]. Signal Processing, 1995, 41: 31–41.

[17] Ostendorf M, Digalakis V V, Kimball O A. From HMM's to segment models: a unified view of stochastic modeling for speech recognition [J]. IEEE Transactions on Speech and Audio Processing, Sept. 1996, 4(5): 360–378.

[18] Zuo-ying WANG, Hong-ge GAO. An inhomogeneous HMM speech recognition algorithm [J]. Chinese Journal of Electronics, January 1998, 7(1): 73–77.

[19] Neri Merhav, Yariv Ephraim. Hidden Markov Modeling Using the Most Likely State Sequence[A]. Proceedings of IEEE ICASSP [C]. Toronto, Ont., Canada, ICASSP-91, 1991, 1. 469–472.

作者简介:

王作英　男, 1935 年 8 月出生于江西省赣县, 1959 年于清华大学无线电电子学系毕业, 1963 年毕业于苏联莫斯科鲍曼高等工业学校制造系, 获博士学位, 自 1963 年至今在清华大学电子工程系任教, 现为该系教授, 博士生导师, 中国通信学会通信理论委员会副主任, 获国务院特殊津贴专家, 研究领域为信号和信息处理, 主要从事语音信号处理研究.

肖　熙　男, 1967 年 10 月出生于福建省福州市, 1990 年清华大学电子工程系毕业获学士学位, 1992 年清华大学电子工程系毕业获硕士学位, 1992 年至今在清华大学电子工程系任教, 现为该系副研究员, 主要从事语音信号处理研究.