

基于停止集的喷泉编码有限长性能估计

林广荣 依那 董明科 梁庆林
(北京大学电子学系 北京 100871)

摘要: 喷泉编码是一类基于删除信道、面向数据分组的前向纠错编码技术。该文分析了停止集的尺度分布对固定码率喷泉编码解码性能的影响,提出了一种估算低误码条件下喷泉编码有限长性能的方法以及一种低复杂度的停止集尺度分布搜索算法。比较结果表明,该文给出的喷泉码解码性能上下界与实际仿真结果非常接近。

关键词: 喷泉编码; 停止集; 前向纠错编码; 尺度分布

中图分类号: TN911.22

文献标识码: A

文章编号: 1009-5896(2008)11-2634-04

Finite Length Analysis of Fountain Codes Based on Stopping Set

Lin Guang-rong Yi Na Dong Ming-ke Liang Qing-lin
(Dept. of Electronics, Peking University, Beijing 100871, China)

Abstract: Digital Fountain is a packet oriented FEC technology for erasure channels. And the performance of fountain codes is limited by the distribution of stopping sets. In the paper, a pair of bounds was proposed to estimate the decoding failure probability of finite length fountain codes. And a search algorithm of low complexity was presented to get the stopping set distribution. The comparisons show that the upper and lower bounds are very close to the simulation results.

Key words: Fountain code; Stopping set; Forward error correction code; Size distribution

1 引言

喷泉编码是近年来发展起来的一类前向纠错编码技术,它将带有检错机制的分组交换信道等效为删除信道,将传统纠错编码的处理对象拓展到了数据分组,形成了一种高效可靠的前向纠错机制^[1-3]。要得到喷泉编码的解码性能,最常用的方法是采用传统的蒙特卡罗仿真,但蒙特卡罗方法的计算量较大,当信道误码较低时需要耗费大量仿真时间。

密度推演是目前较为完善的一类分析喷泉编码性能的理论方法,如文献[4]提出的基于异或树的分析方法可以根据固定码率喷泉编码变量节点及校验节点的维度分布计算出其在码长无限长、各节点满足独立性假设的理想条件下的解码性能。但喷泉编码的性能受码长变化的影响相当大,而上述方法对码长有限条件下的性能估计则无能为力。

文献[5]给出了一种分析有限长 LDPC 码在 BEC 下性能的方法,该方法可用于估算给定二分图维度分布和码长的 LDPC 码集的平均性能。平均性能,指的是符合上述参数的所有码集解码性能的期望。考虑到即便参数相同,不同构造方法得到的喷泉码性能也会存在较大差异。因此,对于估计已给定喷泉码的性能,上述方法同样存在较大的估计偏差。

本文通过分析喷泉编码的停止集尺度分布对其解码性能的影响,提出了一种估算喷泉编码有限长性能的方法。本

文在第2节首先讨论了上述性能估计方法所需参数的取值范围,并据此给出了一组喷泉码解码性能的上下界,可用于计算有限长喷泉码解码的失败概率。第2节还给出了一种基于维度分布的搜索算法,用于搜索给定喷泉码停止集的尺度分布。本文的第3节给出了上述解码性能界计算的理论与喷泉码实际仿真结果的比较。

2 基于停止集尺度分布的喷泉码性能估计

喷泉编码一般采用置信度传播算法进行解码^[1,6,7]。将固定码率喷泉码的 H 矩阵进行映射,可以得到形如图1的生成图。其中,左侧圆形节点称为变量节点,对应 H 矩阵的一列;右侧矩形节点称为校验节点,对应 H 矩阵的一行。在删除信道模型中,生成图的每个变量节点(分别对应一个数据包)或者被删除,或者确知其正确值,因此对应的置信度也只有“0”或“ $\pm\infty$ ”三种取值。引入停止集的概念可以描述上述解码特性,停止集^[5] S 是满足如下条件的变量节点的集合:任意与 S 所包含变量节点相连的校验节点,都连接到 S 至少2次。停止集所包含的变量节点的数量称为停止集的尺度。如图1中集合 $\{v_2, v_6, v_9\}$ 是尺度为3的停止集。当一个停止集所包含变量节点全部被删除时,其所对应的所有校验关系均至少有2个变量节点被删除,因此该停止集完全无法被恢复。

定义 $N(i)$ 表示任意 i 个变量节点组成的集合当中,至少包含1个非空停止集的节点集合的数量,则喷泉码解码的失败概率可以表示为

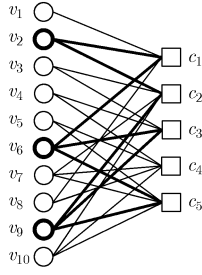


图 1 $\{v_2, v_6, v_9\}$ 为停止集

$$p_{\text{解码失败}} = \sum_i P\{i \text{ 个变量节点被删除}\} N(i) / C_n^i \quad (1)$$

其中 n 为喷泉码的码长， $N(i) / C_n^i$ 表示在 i 个变量节点被删除的条件下，解码失败的概率。容易看出， $N(i)$ 的取值范围在 $0 \sim C_n^i$ 之间，因此式(1)的一般形式又可写成

$$\begin{aligned} p_{\text{解码失败}} &= 1 - \sum_i C_n^i (1-\delta)^{n-i} \delta^i \left(1 - \frac{N(i)}{C_n^i}\right) \\ &= 1 - \sum_i C_n^i (1-\delta)^{n-i} \delta^i - \sum_i (C_n^i - N(i)) (1-\delta)^{n-i} \delta^i \end{aligned} \quad (2)$$

其中 δ 为解码前误包率，即变量节点的删除概率。

2.1 对 $N(i)$ 分布的估计

要根据式(2)计算喷泉码的解码性能，需要先获知 $N(i)$ 的分布(即 i 取不同值时 $N(i)$ 的取值)。但精确计算 $N(i)$ 的分布需要遍历搜索全部 C_n^i 个变量节点的集合，其复杂度会随 i 的增大指数上升。

当喷泉码停止集的尺度分布 $S(i)$ 已知时，可以用 $S(i)$ ， $i \leq j$ 来估计 $N(j)$ 的取值。其中 $S(i)$ 表示尺度为 i 的停止集的数量。定义 $M(i, j)$ 为满足下述条件的集合： $M(i, j)$ 中的每个元素都是由 j 个变量节点组成的集合，且每个元素至少包含 1 个尺度为 i 的停止集。又定义 $\text{size}(A)$ 表示集合 A 中元素的个数，回顾 $N(i)$ 的定义，容易看出

$$N(j) = \text{size}\left(\bigcup_{i=1}^j M(i, j)\right) \quad (3)$$

对给定的 $i, j (i < j)$ ， $\text{size}(M(i, j))$ 表示包含任一尺度 i 停止集的 j 个节点的取法的总数。对任一确定的尺度 i 停止集，包含其 j 个节点有 C_{n-i}^{j-i} 种取法，考虑其中被重复统计的取法数量，可以得到

$$S(i)C_{n-i}^{j-i} - C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)} \leq \text{size}(M(i, j)) \leq S(i)C_{n-i}^{j-i} \quad (4)$$

式(4)的两边省略了 $\text{size}(M(i, j))$ 展开式中对被重复统计的取法数量的部分修正。其中 $C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)}$ 项为同时包含 2 个尺度 i 停止集导致被重复统计的取法数量的修正值，该修正值在所包含的 2 个停止集仅有 1 个变量节点不同时取最大值，此时包含上述 2 停止集的取法共有 $C_{n-(i+1)}^{j-(i+1)}$ 种。更高阶的修正项即对同时包含 3 个或更多停止集而被重复统计的取法数量的修正在式(4)两边均被忽略。若 $i = j$ ，则有

$$\text{size}(M(j, j)) = S(j) \quad (5)$$

由式(3)-式(5)可得

$$\begin{aligned} \max_{1 \leq i < j} (S(i)C_{n-i}^{j-i} - C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)}, S(j)) \\ \leq N(j) \leq \min\left(\sum_{i=1}^j S(i)C_{n-i}^{j-i}, C_n^j\right) \end{aligned} \quad (6)$$

上式中 $\max_{1 \leq i < j} (A(i), B)$ 表示 $A(i)|_{i=1,2,\dots,j-1}$ ， B 中的最大值， $\min(A, B)$ 则表示 A, B 中的最小值。

2.2 基于停止集尺度分布的解码性能上下界

将式(6)代入式(2)，可以得到喷泉码解码失败概率的上下界如下，其中 k 为原始信息节点的数量。

$$p_{\text{upper}} = 1 - (1-\delta)^n - \sum_{j=1}^{n-k} [C_n^j - N_U(j)] (1-\delta)^{n-j} \delta^j \quad (7)$$

$$N_U(j) = \min\left(\sum_{i=1}^j S(i)C_{n-i}^{j-i}, C_n^j\right) \quad (8)$$

$$p_{\text{lower}} = 1 - (1-\delta)^n - \sum_{j=1}^{n-k} [C_n^j - N_L(j)] (1-\delta)^{n-j} \delta^j \quad (9)$$

$$N_L(j) = \max_{1 \leq i < j} (S(i)C_{n-i}^{j-i} - C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)}, S(j)) \quad (10)$$

下面将对 $S(i)$ 的取值范围进行讨论。如图 2 所示，在停止集的尺度分布 $S(i)$ 中， i 实际的取值范围应为 $[0, n]$ ，对于任意给定(\mathbf{H} 矩阵已知)的喷泉码，可以得到 j_0 和 j_2 满足下述条件：

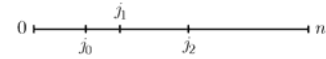


图 2 i 的取值范围

当 $i < j_0$ 时， $S(i) = 0$ ，则根据式(6)可知， $N(i) = 0$ ；(例： $S(0) = 0$ ， $S(1) = 0$)。

当 $i > j_2$ 时， $N(i) = C_n^i$ ，因此无需再关注 $S(i)$ 的取值；(例：当 $j_2 = n - k$ 时，只要被删除变量节点数大于 j_2 ，则解码必然失败，因此 C_n^i 种取值皆包含停止集)。

当 $j_0 \leq i \leq j_2$ 时， $0 \leq S(i) < C_n^i$ ，且 $0 < N(i) < C_n^i$ 。此时需要根据式(6)估计 $N(i)$ 的取值范围。

再引入变量 j_1 满足 $0 \leq j_1 \leq j_2 + 1$ 。假定当 $j_1 \leq i \leq j_2$ 时， $S(i)$ 未知，此时令 $S(i)$ 分别等于 C_n^i 或 0，代入式(8)，式(10)，则喷泉码解码失败概率的上下界可写成

$$\begin{aligned} p_{\text{upper}} &= 1 - \sum_{j=0}^{j_0-1} C_n^j (1-\delta)^{n-j} \delta^j \\ &\quad - \sum_{j=j_0}^{j_1-1} [C_n^j - N_U(j)] (1-\delta)^{n-j} \delta^j \end{aligned} \quad (11)$$

$$N_U(j) = \min\left(\sum_{i=j_0}^j S(i)C_{n-i}^{j-i}, C_n^j\right) \quad (12)$$

$$\begin{aligned} p_{\text{lower}} &= 1 - \sum_{j=0}^{j_0-1} C_n^j (1-\delta)^{n-j} \delta^j - \sum_{j=j_0}^{j_2} [C_n^j - N_L(j)] (1-\delta)^{n-j} \delta^j \end{aligned} \quad (13)$$

$$N_L(j) = \max_{j_0 \leq i < j} (S(i)C_{n-i}^{j-i} - C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)}, S(j)) \quad (14)$$

下面给出有关上述解码失败概率上下界的一些性质。

性质 1 Iff $j_1 > j_0$ 时, $\lim_{\delta \rightarrow 0} [(p_{\text{upper}} - p_{\text{lower}})/p_{\text{lower}}] = 0$;

证明

$$\frac{(p_{\text{upper}} - p_{\text{lower}})/p_{\text{lower}} = \sum_{j=j_0}^{j_1-1} [N_U(j) - N_L(j)](1-\delta)^{n-j} \delta^j + \sum_{j=j_1}^{j_2} [C_n^j - N_L(j)](1-\delta)^{n-j} \delta^j}{\sum_{j=j_0}^{j_2} N_L(j)(1-\delta)^{n-j} \delta^j + \sum_{j=j_2+1}^n C_n^j (1-\delta)^{n-j} \delta^j} \quad (15)$$

当 $j_1 > j_0$ 时, $N_U(j_0) = N_L(j_0) = S(j_0) > 0$, 因此, 令式(15)分子分母同时除以 δ^{j_0} , 则有

$$\lim_{\delta \rightarrow 0} [(p_{\text{upper}} - p_{\text{lower}})/p_{\text{lower}}] = \lim_{\delta \rightarrow 0} \left\{ \frac{\sum_{j=j_0+1}^{j_1-1} [N_U(j) - N_L(j)] (1-\delta)^{n-j} \delta^{j-j_0} + \sum_{j=j_1}^{j_2} [C_n^j - N_L(j)] (1-\delta)^{n-j} \delta^{j-j_0}}{S(j_0)(1-\delta)^{n-j_0}} \right\} = 0$$

当 $j_1 \leq j_0$ 时, $\forall j_0 \leq j \leq j_2$, $N_U(j) = C_n^j$, $N_L(j) = 0$, 代入式(15), 得

$$\lim_{\delta \rightarrow 0} [(p_{\text{upper}} - p_{\text{lower}})/p_{\text{lower}}] = \lim_{\delta \rightarrow 0} \left[\frac{\sum_{j=j_0}^{j_2} C_n^j (1-\delta)^{n-j} \delta^{j-j_0}}{\sum_{j=j_2+1}^n C_n^j (1-\delta)^{n-j} \delta^{j-j_0}} \right] = +\infty$$

\therefore Iff $j_1 > j_0$ 时, $\lim_{\delta \rightarrow 0} [(p_{\text{upper}} - p_{\text{lower}})/p_{\text{lower}}] = 0$. 证毕

性质 2 对给定构造的喷泉码, 即 j_0 和 j_2 确定, $(p_{\text{upper}} - p_{\text{lower}})$ 随着 j_1 的增大而递减。

证明 令 $\Delta p(k) = (p_{\text{upper}} - p_{\text{lower}})|_{j_1=k}$.

当 $k_2 \geq k_1$ 时,

$$\begin{aligned} \Delta p(k_1) - \Delta p(k_2) &= (p_{\text{upper}}|_{j_1=k_1} - p_{\text{upper}}|_{j_1=k_2}) + (p_{\text{lower}}|_{j_1=k_2} - p_{\text{lower}}|_{j_1=k_1}) \\ &= \sum_{j=k_1}^{k_2-1} [C_n^j - N_U(j)](1-\delta)^{n-j} \delta^j \\ &\quad + \sum_{j=k_1}^{j_2} [N_L(j)|_{j_1=k_2} - N_L(j)|_{j_1=k_1}](1-\delta)^{n-j} \delta^j \quad (16) \end{aligned}$$

$\therefore N_L(j)|_{k_1 \leq j \leq j_2, j_1=k_2} = \max_{j_0 \leq i \leq \min(j, k_2)} (S(i)C_{n-i}^{j-i} - C_{S(i)}^2)$
 $\cdot C_{n-(i+1)}^{j-(i+1)}$.

$N_L(j)|_{k_1 \leq j \leq j_2, j_1=k_1} = \max_{j_0 \leq i \leq k_1} (S(i)C_{n-i}^{j-i} - C_{S(i)}^2 C_{n-(i+1)}^{j-(i+1)})$

\therefore 当 $k_1 \leq j \leq j_2$, $N_L(j)|_{j_1=k_2} - N_L(j)|_{j_1=k_1} \geq 0$;

又 \therefore 当 $k_1 \leq j \leq k_2 - 1$, $C_n^j - N_U(j) \geq 0$, 代入式(16),

可得

$$\Delta p(k_1) - \Delta p(k_2) \geq 0 \quad \text{证毕}$$

由性质 1 可知, 只要至少 1 个大于零的 $S(i)$ 已知(本文假定尺度小的 $S(i)$ 总是被优先知道, 即若 $S(i_0)$ 已知, 则 $S(i) | i < i_0$ 也已知。考虑到实际应用中 $S(i)$ 计算复杂度与 i 的关系(参见 2.3 节, 可以认为上述假定是合理的), 将已知的 $S(i)$ 代入式(12), 式(14), 即可得到喷泉码解码失败概率

的上下界, 且随着信道误包率的降低, 该上下界将快速收敛至解码失败概率的真实值。

由性质 2 可知, 上述失败概率上下界的估计误差会随着 j_1 的增大(即已知的 $S(i)$ 越多)逐渐减小。

2.3 停止集尺度分布的获取

要获取停止集的尺度分布 $S(i)$, 也可以采用遍历搜索的方式, 但这种方法的计算量过于巨大。文献[8]给出了一种在给定维度分布及码长的图码码集中估算停止集分布期望的方法。但即便维度分布、码长等参数均相同, 不同的构造方法产生的停止集尺度分布也存在较大差异, 从而导致了解码性能的差异。本文希望能够对给定构造喷泉码的性能作出估计, 因此需要统计精确的停止集尺度分布。

根据停止集的定义, 容易证明, 当 $j_0 \leq i < 2j_0$ 时, 同一停止集中任意两个变量节点必然可以找到一条路径互相连接, 且该路径所经过的所有变量节点均包含于该停止集。定义满足上述条件的停止集为连通停止集。而当 $i \geq 2j_0$ 时, 一个停止集或者本身即是连通停止集, 或者是若干个交集为 ϕ 的连通停止集的并集。

对连通停止集, 可以采用下述算法简化搜索的计算量(假定喷泉码生成图中变量节点数为 n , 校验节点数为 k , 当前搜索尺度为 j):

(1)对生成图进行预搜索, 得到每个变量节点 v_i 的相邻变量节点集(以下简称邻点集)。相邻变量节点, 是指该节点与 v_i 连接到同一校验节点, 参见图 3, $\{v_1, v_2, v_4, v_5, v_6\}$ 是 v_3 的邻点集。

(2)对尺度为 j 的变量节点集合: $\{s_0, s_1, \dots, s_{j-1}\}$, 按下述步骤遍历:

(a)令 s_0 遍历 v_i , $i = 0, 1, \dots, n-1$;

(b)对任意 s_0 , 令 s_1 遍历 s_0 邻点集中下标大于 s_0 的所有变量节点: 如图 3, 当 $s_0 = v_3$ 时, $s_1 = v_i, v_i \in \{v_4, v_5, v_6\}$;

(c)类似地, 对任意 $\{s_0, s_1\}$, 令 s_2 遍历 s_0, s_1 邻点集的交集中下标大于 s_0 的所有变量节点: 如图 4, 当 $s_0 = v_3, s_1 = v_5$ 时, $s_2 = v_i, v_i \in \{v_4, v_6, v_7, v_8\}$;

(d)重复上述步骤直至完成对 s_{j-1} 的遍历;

(e)对任意 $\{s_0, s_1, \dots, s_{j-1}\}$, 检验其是否为停止集:

令 $c_i = 0$, $i = 0, 1, \dots, n-k-1$;

对所有 $v_i \in \{s_0, s_1, \dots, s_{j-1}\}$ 循环;

对上述 v_i 所有邻接校验节点 c_m 循环, 令相应 $c_m + +$;

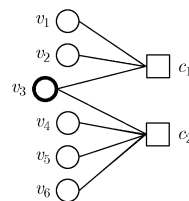


图 3 v_3 的邻点集

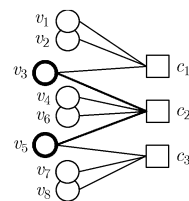


图 4 v_3, v_5 邻点集的交集

if $c_m \neq 1$, 对 $m = 0, 1, \dots, n - k - 1$ 都成立, 则 $\{s_0, s_1, \dots, s_{j-1}\}$ 为停止集;

(f)完成检验。

(3)完成搜索。

上述搜索算法可以将尺度为 j 的停止集的计算量由 $O(n^j)$ 降低至 $O(n(l \cdot r)^{j-1})$, 其中 l 和 r 分别为喷泉码生成图中变量、校验节点的平均维度。由于喷泉码生成图通常为稀疏图, 即 l 和 r 远小于 n , 因此该算法可以有效降低搜索的计算量。

3 基于停止集的喷泉码性能界与仿真结果比较

图 5, 图 6 给出了对 2.2 节性能界的计算结果与喷泉码性能实际仿真结果的比较。其中, 喷泉码参数为 $(n, k) = (256, 224)$, 码率 $7/8$, 变量节点维度分布取典型值 $0.125x + 0.75x^2 + 0.125x^3$, 采用 PEG 算法^[9]构造。

图 5 为当 $j_1 = 2$ 时, 由停止集尺度分布上下界式(11), 式(13)求得的解码失败概率。由于 $j_1 = j_0 = 2$, 上下界在低误码时并不收敛, 符合对性质 1 的分析; 图 5 中同时给出了用密度推演求得的给定码率和维度下无限码长喷泉码的理想性能, 可以看出与有限长性能存在很大偏差。

图 6 为当 j_1 依次取 3, 4, 5, 6 时, 由上下界公式求得的解码失败概率。由图中可以看出, 当满足 $j_1 > j_0$ 时, 该上下界随误包率的降低逐步收敛到真实的解码失败概率, 并且 j_1 越大(即已知的 $S(i)$ 越多), 收敛速度越快。当 j_1 增大时, 下界曲线的变化幅度很小, 在图 6 中几乎完全重合。由比较结果可以看出, 本文给出的停止集尺度分布上下界, 与实际的仿真结果非常接近。

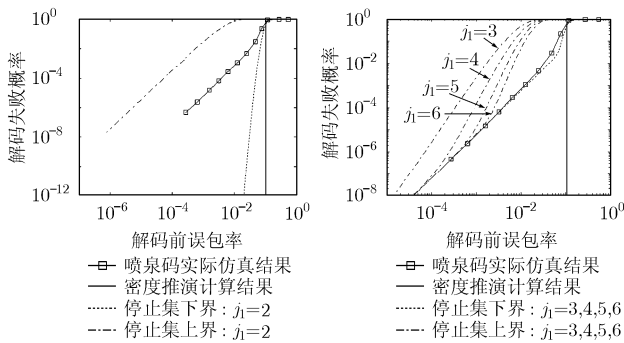


图 5 $j_1=2$, 上下界不收敛

图 6 $j_1=3, 4, 5, 6$, 上下界逐步收敛

4 结束语

本文给出的喷泉码解码性能上下界, 可以通过小尺度停止集分布, 精确估计出固定码率喷泉码在低误码率时的解码性能, 解决了传统的蒙特卡罗仿真方法在低误码率时仿真计算量巨大的问题。文中提出的基于维度分布的停止集搜索算法, 有效降低了获取小尺度停止集分布的计算量, 从而降低了上述性能估计方法的复杂度。

参考文献

- [1] Byers J W, Luby M, and Mitzenmacher M, *et al.* A digital fountain approach to reliable distribution of bulk data. Proc. ACM SIG-COMM, Vancouver, BC, Canada, 1998: 56-67.
- [2] Luby M, Mitzenmacher M, and Shokrollahi A, *et al.* Practical loss-resilient codes. Proc. 29th ACM Symposium on Theory of Computing, Texas, 1997: 150-159.
- [3] Luby M, Mitzenmacher M, and Shokrollahi A, *et al.* Efficient erasure correcting codes. *IEEE Trans. on Inform. Theory*, 2001, 47(2): 569-584.
- [4] Luby M, Mitzenmacher M, and Shokrollahi A. Analysis of random processes via and-or tree evaluation. Proc. 9th Annu. ACM-SIAM Symp. Discrete Algorithms, San Francisco, CA, 1998: 364-373.
- [5] Di C, Proietti D, and Telatar I E, *et al.* Finite-length analysis of low-density parity-check codes on the binary erasure channel. *IEEE Trans. on Inform. Theory*, 2002, 48(6): 1570-1579.
- [6] Shokrollahi A. Raptor codes. *IEEE Trans. on Inform. Theory*, 2006, 52(6): 2551-2567.
- [7] MacKay D J C. Fountain codes. *IEEE Proc-Commun*, 2005, 152(6): 1062-1068.
- [8] Orlitsky A, Viswanathan K, and Zhang J. Stopping set distribution of LDPC code ensembles. *IEEE Trans. on Inform. Theory*, 2005, 51(3): 929-953.
- [9] Hu X Y, Eleftheriou E, and Arnold D M. Regular and irregular progressive edge-growth tanner graphs. *IEEE Trans. on Inform. Theory*, 2005, 51(1): 386-398.

林广荣: 男, 1980年生, 博士生, 研究方向为卫星与无线通信。
 依那: 男, 1974年生, 讲师, 研究方向为卫星与无线通信网络。
 董明科: 男, 1973年生, 讲师, 研究方向为数字通信和信号处理。
 梁庆林: 男, 1941年生, 教授、博士生导师, 研究方向为卫星与无线通信。