

研究速報

マルコフ情報源に対し誤り伝搬を抑える VF 符号に関する一考察

木村 勝^{†*} 後藤 正幸[†](正員)
 松嶋 敏泰[†](正員) 平澤 茂一[†](正員)

A Note on Variable to Fixed-Length Codes without Error Propagation for Markov Sources

Masaru KIMURA^{†*}, Nonmember, Masayuki GOTOH[†],
 Toshiyasu MATSUSHIMA[†],
 and Shigeichi HIRASAWA[†], Members

[†] 早稲田大学理工学部経営システム工学科, 東京都

School of Science and Engineering, Waseda University, 3-4-1
 Ohkubo, Shinjyuku-ku, Tokyo, 169-8555 Japan

* 現在, 三菱電機株式会社

あらまし 本論文では, 確率構造が既知であるマルコフ情報源に対する VF 符号について考察する. 通常, 状態遷移確率を符号化に使用すると圧縮レートは高いが誤り伝搬の制御などの VF 符号のいくつかの長所を失ってしまう. 一方, 誤り伝搬を抑えると圧縮レートが劣化する. 本論文では誤り伝搬を抑える方法を提案し, この方法によって生じる冗長度と復号誤り率を示す.

キーワード 情報源符号化, VF 符号, 誤り伝搬

1. まえがき

Tunstall 符号は, パラメータ既知の i.i.d. (independent and identically distributed) 情報源を対象とした VF 符号である. この符号は, properかつ complete^(注1) という性質をもつ辞書を使用して符号化するときに, 一定の辞書サイズで符号長が最小となる辞書の構成法を与えたものである [1]. この意味で, Tunstall 符号は最適な符号化法である^(注2).

本論文を通して確率構造が既知という仮定を置く. VF 符号は, 符号語の先頭位置が容易に分かることにより様々な長所をもつ. すなわち, 復号の際に誤りが混入してもその誤り伝搬を抑えることができること, 及び途中からの復号が可能であること等である. しかしマルコフ情報源を対象とすると, 状態遷移確率を符号化に使用する場合 [2] には, 圧縮レートの面では有利であるが前述の長所を失ってしまう. つまり, 一度誤りが生じると, 次の符号語の状態が正しく特定されないため, それ以降の符号語に誤りが次々と伝搬してしまう. 更に, 同じような理由で途中からの復号も不可能となる. ただし, 定常確率を用いる [3], [8] と, 圧縮レートが劣化する代わりにこれらの長所を保持することができる^(注3). また [4], [6] は, i.i.d. 情報源に対す

る VF 符号を提案しているが, 1 次マルコフ情報源に関しては, 部分系列の先頭の 1 記号の出現確率を定常確率と見立てることによって拡張する方法も提案されている. これらの方法も誤り伝搬を抑えることができるが, 誤り伝搬を抑えることによって生じる冗長度や復号誤り率等については明らかにされていない.

また最近, ユニバーサル符号の立場から誤り伝搬の制御や部分的復号を可能にする符号化法が提案された [7]. この方法は, ある条件のもとで漸近的に最適な圧縮レートを達成する. しかし, 有限時点での符号長や誤り率についての解析はなされていない.

誤り伝搬を抑えることによって生じる冗長度や復号誤り率, あるいは圧縮レートと誤り率の関係等を明らかにすることは, 符号を設計する際に重要である.

そこで, 本論文ではこれらの性質を明らかにすることを目的とする. 既約マルコフ情報源に対する VF 符号において, 符号化に使用する辞書のサイズが有限であるときの誤り伝搬を抑えることによって生じる冗長度を求め, そのために, まず i.i.d. 情報源に対して最適な Tunstall 符号を簡単に修正し, マルコフ情報源に対して誤り伝搬を抑えるアルゴリズムを提案する. そしてその冗長度を定式化し, 数値解析を行った. 更に, 符号系列を 2 元対称通路に通したときの情報源系列の復号誤り率を示し, 数値解析によって符号長と復号誤り率の関係を明らかにした.

2. 従来研究

本論文を通して次の表記法を用いる.

$\mathcal{X} = \{a_0, a_1, \dots, a_{\alpha-1}\}$: 情報源アルファベット.

$\mathcal{Y} = \{0, 1, \dots, \beta-1\}$: 符号語アルファベット.

x : \mathcal{X} の要素を連結して生成される系列

$L(x)$: 系列 x の長さ (記号数).

\mathcal{X}^k : 長さが k である系列全体の集合.

xy : 系列 $x (\in \mathcal{X}^{L(x)})$ と $y (\in \mathcal{X}^{L(y)})$ の連結.

$L_C(x)$: x を符号化した後の対応する符号語の長さ.

また, \log の底は β とする.

2.1 Tunstall 符号

Tunstall 符号は, i.i.d. 情報源を対象とした VF 符号である. この符号化法は, Huffman 符号のように符

(注 1): proper とはある系列が他の系列の suffix になっていないこと (語頭条件), complete とは任意の半無限系列が唯一の suffix をもっていることである.

(注 2): non-proper な辞書を用いたときに Tunstall 符号よりも優れた圧縮レートを達成する可能性があることが [5], [6] で指摘されている.

(注 3): 文献 [3] は FV 符号と VF 符号の符号長の差を解析した論文であり, 特に誤り伝搬についての記述はない.

号木 (辞書) を使用するが, 木の枝に割り振るラベルは符号語アルファベットではなく情報源アルファベットである. そして, 葉の index が符号語に対応する.

ここで, 最終的に構成される M 個の要素数の辞書 $C(M)$ を $C(M) = \{c_0, c_1, \dots, c_{M-1}\}$, $c_i \in \mathcal{X}^{L(c_i)}$ ($i = 0, 1, \dots, M-1$) とする. 以降, 情報源出力系列と区別するために, 辞書の系列は c で表す. また, 辞書構成のステップ数を N とすると, 要素数 M は $M = \alpha + (\alpha - 1)N$ となる.

[Tunstall 符号の辞書の構成法]

ある非負整数 $N (= \frac{M-\alpha}{\alpha-1})$ に対し,

- 1) $C' := \mathcal{X}$, $j := N$ とする.
- 2) $j = 0$ なら $C(M) := C'$ として終了.
- 3) C' の要素の中でその生起確率が最も高いものを c' とするとき,
 $C' := (C' - \{c'\}) \cup \{c | c = c'x, x \in \mathcal{X}\}$
- 4) $j := j - 1$ とし, 2) へ.

普通, $M = \beta^m$ となるように構成する. この $C(M)$ に従って i.i.d. 情報源からの出力系列を部分系列に分解し, 対応する葉の index i を固定長の m けた β 進数に符号化する. また, Tunstall 符号の辞書 C は, proper で complete という性質をもつが, 同じ性質をもつ辞書の中で符号長が最小となる [1].

3. マルコフ情報源に拡張した Tunstall 符号の提案

Tunstall 符号をマルコフ情報源へ拡張する. このとき, 次の二つの方法が考えられる.

- a) あらかじめ各状態ごとに辞書を複数個用意する. 符号化及び復号は, 各時点の状態に対応する辞書を使用して行う.

- b) ただ一つの辞書を用いて符号化, 復号を行う.

ここで, a) に関しては, 通信路において誤りが混入すると, 次のような理由から誤りが伝搬してしまう. すなわち, ある符号語が誤って復号されると次の符号語の状態が誤ってしまうので正しい辞書で復号することができず, それ以降の符号語に次々と誤りが伝搬する. この問題を回避するためには, 符号語の開始状態をその符号語自身に記述すればよい. しかし, 状態数が多くなると状態を記述することによって生じる冗長度が無視できなくなる. 一方, b) は常に一つの辞書を用いて符号化及び復号を行うので誤り伝搬を抑えることができる. 本論文では, b) のみを考え, 辞書構成に定常確率を使用する方法を提案する^(注4).

また, 対象とするマルコフ情報源は状態集合 S を

既約とし, 情報源アルファベットの生起確率行列を $Q(\cdot, \cdot)$, 状態遷移関数を $T(\cdot, \cdot)$ としたとき,

$$\Pr(X_n = x | S_n = s, X_{n-1}, S_{n-1}, \dots) \stackrel{\text{def}}{=} Q(x|s) \quad (1)$$

$$s_n \stackrel{\text{def}}{=} T(x_{n-1}, s_{n-1}) \text{ for } Q(x_{n-1}|s_{n-1}) > 0 \quad (2)$$

を満たす. ただし $n = 1, 2, \dots, x \in \mathcal{X}, s \in S$ とする. また, $x = x_1 x_2 \dots x_L(x), T(x_0, s_0) = s$ としたとき,

$$Q^*(x|s) \stackrel{\text{def}}{=} \prod_{t=1}^{L(x)} Q(x_t | T(x_{t-1}, s_{t-1})). \quad (3)$$

と定義する.

提案する辞書 C_p は, Tunstall 符号の辞書の構成法の 3) における系列の生起確率の代わりに定常確率 $P^*(x)$ を用いて構成する. ここで,

$$P^*(x) \stackrel{\text{def}}{=} \sum_{s \in S} Q^*(x|s) w(s) \quad (4)$$

と定義する. ただし, $w(s)$ は状態 s の定常分布を表す.

このアルゴリズムは, 結果として Ziv アルゴリズム [3] と同様の辞書が構成される. しかし, Ziv アルゴリズムは辞書の要素数をアルゴリズム上で任意に設定できない. 一方, 提案アルゴリズムは, N を動かすことにより要素数 M を $\alpha + (\alpha - 1)N$ となるように自由に設定できる. この意味で提案符号はより自由度が高く, β のべき乗に設定できるため, 圧縮レートにおいても有利である^(注5).

4. 評 価

4.1 圧縮レート

圧縮レート R は以下の式で与えられる.

$$R = \frac{\log M}{EL_M} \quad (5)$$

ここで, EL_M とは, 辞書サイズが M であるときの辞書内の系列 c の長さ $L(c)$ の期待値を表す.

次に, EL_M の計算方法を示すために系列情報源 (segment source) [2] を考える. 系列情報源とは, 次のようなものである. すなわち, 辞書 $C(M)$ が与えられたときに, 符号化は情報源出力系列を辞書 $C(M)$ の

(注4): 定常確率が最も高い状態のもとでの状態遷移確率を使用する方法も考えられるが, 本論文では平均的に圧縮レートがよいと思われる定常確率を使用する.

(注5): 実際に符号化を行うときには, 符号長が $\lceil \log M \rceil$ となるため.

要素で部分系列に区切ることによって行われる．そのため，辞書内の系列 $c_i (i = 0, 1, \dots, M-1)$ を一つの“文字”と見立てることもできる．このように系列をアルファベットと見るのが系列情報源である．したがって，系列情報源の要素の集合 \mathcal{Z}_M^* は， $\mathcal{Z}_M^* = C_p(M)$ である．以後，系列情報源の要素を z で表す．

系列情報源は，以下の式を満足する．

$$n = 1, 2, \dots, s \in \mathcal{S}, z \in \mathcal{Z}_M^* \text{ に対して,}$$

$$\Pr(\mathbf{Z}_n = z | S_n = s, \mathbf{Z}_{n-1}, S_{n-1}, \dots) \stackrel{\text{def}}{=} Q^*(z|s) \quad (6)$$

$$s_{n+1} \stackrel{\text{def}}{=} T(z_n, s_n) \text{ for } Q^*(z_n | s_n) > 0. \quad (7)$$

開始状態の分布は，もとの情報源の定常分布と同じである．しかし，状態集合 \mathcal{S} は既約集合になるとは限らない [2]．既約な部分集合を $G_r (r = 1, 2, \dots)$ ， G_r の集合を G とおくと，提案符号の EL_M は，辞書内の系列と系列情報源の要素が等しいので，

$$EL_M = \sum_{G_r \in G} \Pr(\text{seg} \rightarrow G_r) \cdot \sum_{s \in G_r} q^{\text{seg}}(s|G_r) \sum_{c \in C_p(M)} Q^*(c|s)L(c) \quad (8)$$

によって与えられる．ただし， $\Pr(\text{seg} \rightarrow G_r)$ とは系列が部分集合 G_r に入る確率であり， $q^{\text{seg}}(s|G_r)$ は部分集合 G_r における定常確率である．

4.2 冗長度

定常確率を用いて Tunstall の辞書を構成し，その辞書でロスなく符号化できると仮定する．つまり，Tunstall 符号の本質的な冗長度をひとまず無視する．

ある状態 s のもとでの一系列 (\mathcal{Z}_M^* の要素) 当りの符号長の期待値 $E(L_C(\mathbf{Z}|s))$ は，

$$\begin{aligned} E(L_C(\mathbf{Z}|s)) &= - \sum_{z \in \mathcal{Z}_M^*} Q^*(z|s) \log P^*(z) \\ &= - \sum_{z \in \mathcal{Z}_M^*} Q^*(z|s) \left(\log Q^*(z|s) + \log \frac{P^*(z)}{Q^*(z|s)} \right) \\ &= H(\mathbf{Z}|s) + D(Q^*(\cdot|s) || P^*) \end{aligned} \quad (9)$$

ここで， $H(\cdot)$ はエントロピー， $D(\cdot || \cdot)$ は KL 情報量を表す．各状態で式 (9) の平均をとると，1 系列当りの符号長の期待値 $E(L_C(\mathbf{Z}))$ が得られる．すなわち，

$$\begin{aligned} E(L_C(\mathbf{Z})) &= - \sum_{G_r \in G} \Pr(\text{seg} \rightarrow G_r) \\ &\quad \cdot \sum_{s \in G_r} q^{\text{seg}}(s|G_r) L_C(\mathbf{Z}|s) \\ &= H_\infty(\mathbf{Z}) + D_\infty(Q^* || P^*). \end{aligned} \quad (10)$$

ただし，

$$\begin{aligned} H_\infty(\mathbf{Z}) &= - \sum_{G_r \in G} \Pr(\text{seg} \rightarrow G_r) \sum_{s \in G_r} q^{\text{seg}}(s|G_r) \\ &\quad \cdot \sum_{z \in \mathcal{Z}_M^*} Q^*(z|s) \log Q^*(z|s) \end{aligned} \quad (11)$$

$$\begin{aligned} D_\infty(Q^* || P^*) &= \sum_{G_r \in G} \Pr(\text{seg} \rightarrow G_r) \\ &\quad \cdot \sum_{s \in G_r} q^{\text{seg}}(s|G_r) D(Q^*(\cdot|s) || P^*) \end{aligned} \quad (12)$$

である．式 (10) より次の定理を得る．

[定理 1] 定常確率を用いて Tunstall の辞書を構成し，ロスなく符号化したときの平均冗長度 ρ は，

$$\rho = \frac{D_\infty(Q^* || P^*)}{EL_M} \quad (13)$$

である． \square

これは，提案アルゴリズムの近似的な冗長度である．実際にはすべての $z \in \mathcal{Z}_M^*$ に対して $-\log P^*(z)$ という可変の符号長を割り当てるわけではなく， $\log M (= -\log \frac{1}{M})$ という符号長になる．つまり， $P^*(z) = \frac{1}{M} (\forall z)$ という VF 符号にとって理想的な場合にのみ成り立つ． P^* を，すべての要素 $z \in \mathcal{Z}_M^*$ に対して等確率 ($\frac{1}{M}$) を割り振る分布に置き換えると，提案アルゴリズムの厳密な冗長度が得られる．

提案アルゴリズムは $EL_M \rightarrow \infty (M \rightarrow \infty)$ である． $Q(x|s) > 0 (\forall s, x \in \mathcal{X})$ と仮定すれば， $D_\infty(Q^* || P^*) < \infty$ であるので平均冗長度は $\rho \rightarrow 0 (M \rightarrow \infty)$ となる．これは， P^* をすべての要素に対して等確率 ($\frac{1}{M}$) を割り振る分布とおいても成り立つ．つまり，提案アルゴリズムの平均符号長は漸近的 ($N \rightarrow \infty$) にエントロピーレートに収束する．

4.3 提案符号の復号誤り率

情報源出力系列を符号化し，その符号系列を誤り率 p の 2 元対称通信路 (BSC) に通して復号するときの情報源系列の復号誤り率を評価する．

$\mathcal{Y} = \{0, 1\}$ とし，一つの符号語に 1 bit でも誤りが生じたときに，対応する系列全体が誤って復号される

場合をその上界とする．下界は，対応する系列がたまたま正しく復号される場合があるので，1記号を除いてその他がすべて誤る場合である．

また，ここで定義する復号誤り率とは，情報源系列に含まれる復号誤り記号数を系列全体の記号数で割ったシンボル誤り率とする．

[定理2] 提案符号の復号誤り率 P_{err} は，

$$\frac{1 - (1-p)^{\log M}}{EL_M} \leq P_{err} \leq 1 - (1-p)^{\log M} \quad (14)$$

で与えられる． □

(証明)ここでは上界についてのみ証明を与える(下界についても直ちに導くことができる)．ある符号語(長さ $\log M$)内に誤りが生じる確率は $1 - (1-p)^{\log M}$ である．また長さ n のデータを符号化するとき，誤って復号される部分系列の個数の期待値 $E_{err}(n)$ は，

$$E_{err}(n) = \frac{n}{EL_M} (1 - (1-p)^{\log M}), \quad (15)$$

誤って復号される記号数の期待値は $EL_M \cdot E_{err}(n)$ である．これをデータの長さ n で割ることにより，提案符号の復号誤り率 P_{err} が得られる(この値は n によらないので任意のデータ長に対して成り立つ)． □

P_{err} は辞書のサイズの関数になっている．したがって，BSCの誤り率 p が得られているときには容易に復号誤り率を制御することができる．また， M が大きくなるとともに P_{err} が大きくなる．一方，定理1より M が大きくなると EL_M が大きくなり，圧縮レートが高くなる．つまり，圧縮レートと復号誤り率との間にはトレードオフの関係がある．

5. 数値解析

5.1 圧縮レートと EL_M について

圧縮レートと EL_M の関係を調べるため，1次，2次マルコフ情報源について数値解析を行った(図1~4)．

文献[4]において，1次マルコフ情報源に対しては拡張法が示されているので，1次マルコフ情報源ではこの符号化法(BAC heuristic)とも比較を行った(量子化は“Good probability quantizer”である^{注6)}．

1次マルコフ情報源のパラメータ(図1, 2)

$$(a) \quad Q(0|0) = 0.9 \quad Q(0|1) = 0.1$$

$$(b) \quad Q(0|0) = 0.75 \quad Q(0|1) = 0.4$$

2次マルコフ情報源のパラメータ(図3, 4)

$$(a) \quad Q(0|00) = 0.95 \quad Q(0|01) = 0.08$$

$$Q(0|10) = 0.1 \quad Q(0|11) = 0.9$$

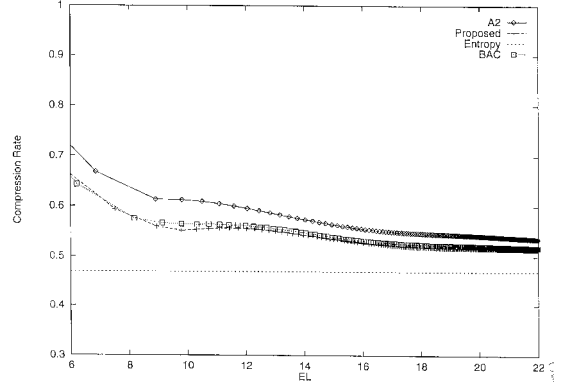


図1 1次マルコフ情報源(a)に対する圧縮レートと EL_M
Fig. 1 Compression rate and EL_M for 1st Markov(a).

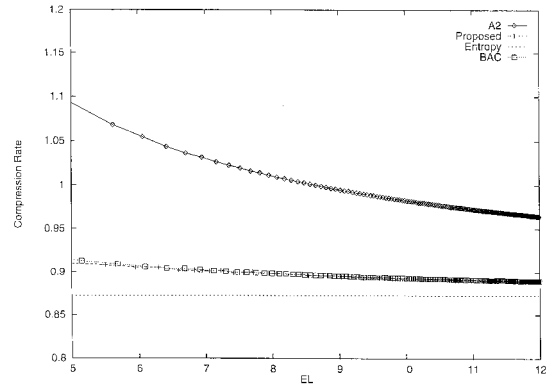


図2 1次マルコフ情報源(b)に対する圧縮レートと EL_M
Fig. 2 Compression rate and EL_M for 1st Markov(b).

$$(b) \quad \begin{aligned} Q(0|00) &= 0.8 & Q(0|01) &= 0.2 \\ Q(0|10) &= 0.7 & Q(0|11) &= 0.6 \end{aligned}$$

図中の A2 とは，3. で示した a) の方法である^(注7)．

図の (a) は，エントロピーが小さく式(13)のダイバージェンスが大きいとき，(b) はエントロピーが大きく，ダイバージェンスが小さいときである．いずれの場合でも， EL_M を大きくすると圧縮レートが向上することがわかる．

また，(a) では提案アルゴリズムと A2 との差が小さいが，(b) では差が大きい．この差は，マルコフ情

(注6): Ziv アルゴリズムは提案アルゴリズムと重なるが，3. で書いたとおり，実際の符号化を考えた提案アルゴリズムの方が圧縮レートにおいて有利である．
(注7): 誤り伝搬対策として符号語の開始状態を符号語自身に記述する方法．

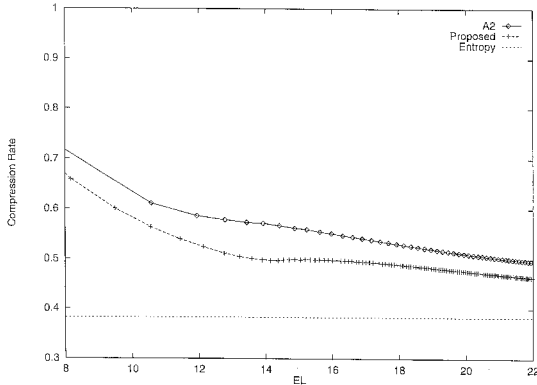


図3 2次マルコフ情報源 (a) に対する圧縮レートと EL_M
Fig.3 Compression rate and EL_M for 2nd Markov(a).

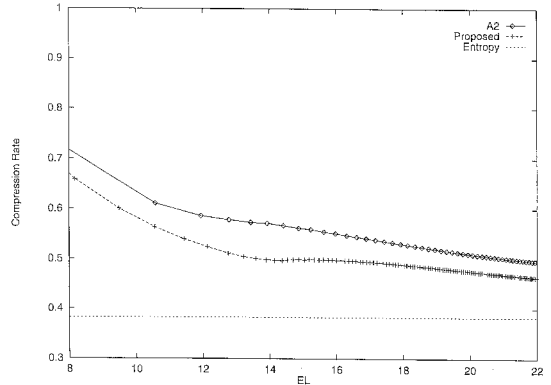


図5 3次マルコフ情報源に対する圧縮レートと復号誤り率
Fig.5 Compression rate and P_{ERR} for 3rd Markov.

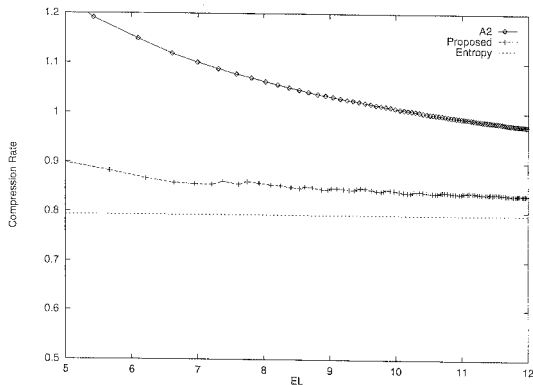


図4 2次マルコフ情報源 (b) に対する圧縮レートと EL_M
Fig.4 Compression rate and EL_M for 2nd Markov(b).

3次マルコフ情報源のパラメータ

$$Q(0|000) = 0.95 \quad Q(0|001) = 0.9 \quad Q(0|010) = 0.1$$

$$Q(0|011) = 0.9 \quad Q(0|100) = 0.8 \quad Q(0|101) = 0.25$$

$$Q(0|110) = 0.9 \quad Q(0|111) = 0.05 \quad p = 10^{-3}$$

辞書の要素数 M を大きくすると EL_M が大きくなり圧縮レートが上がる(図1~4参照)が、その反面、復号誤り率も高くなるというトレードオフの関係があることが読み取れる。これは定理2の結果と一致する。

6. むすび

定常確率を使うことにより、Tunstall 符号を誤り伝搬を抑える性質を有したままマルコフ情報源に拡張する方法を提案した。また、提案法の冗長度を定式化し、数値解析によって圧縮レートの評価を行った。その結果、解析で用いた情報源に対しては従来法 (BAC) や辞書を複数用意する方法よりも提案アルゴリズムが優れた圧縮レートを達成することが示された。

更に、提案符号の符号語を BSC に通したときの復号誤り率を示し、数値解析により圧縮レートと誤り率の関係を明らかにした。

確率構造が既知である場合に VF 符号を設計する際、設計者は本論文で明らかにした式をもとにして容易に誤り率と圧縮レートを制御することができるであろう。

謝辞 本研究を行うにあたり、御検討、御助言をいただきました平澤研究室の皆様へ深く感謝致します。

文 献

- [1] F. Jelinek and K.S. Schneider, "On variable-length-to-block coding," IEEE Trans. on Information Theory, vol.IT-18, no.6, pp.765-774, 1972.
- [2] T.J. Tjalkens and F.M.J. Willems, "Variable to fixed-length codes for Markov source," IEEE Trans. on Inf.

報源の次数が増えるほど増加していくと思われる^(注8)。

5.2 圧縮レートと復号誤り率について

許容される誤り率以内でどこまで圧縮レートを上げることができるか、あるいは辞書のサイズに制限があるときなどにその辞書を用いたときの圧縮レートに対してどの程度の誤りが発生してしまうか等の関係は現在まであまり解析されていない。この関係は、符号を設計するときの重要な指標であると考えられる。

そこで、一例として3次マルコフ情報源について提案符号の圧縮レートと復号誤り率の関係を数値解析によって調べた(図5)。

(注8): 近似的には、定理1の $D_\infty(Q^*||P^*)$ と $\log|S|$ の差で表せる。

- Theory, vol.IT-33, no.2, pp.246-257, 1987.
- [3] J. Ziv, "Variable-to-fixed length codes are better than fixed-to-variable length codes for Markov sources," IEEE Trans. on Information Theory, vol.36, no.4, pp.861-863, 1990.
- [4] C.G. Boncelet, "Block arithmetic coding for source compression," IEEE Trans. on Information Theory, vol.39, no.5, pp.1546-1554, 1993.
- [5] 横尾英俊, "語頭条件を制約としない可変長-固定長符号について," SITA90, pp.209-214, 1990.
- [6] S.H. Cho, J.H. Park, and R. Kohno, "Non-proper variable-to-fixed length arithmetic coding for gray-level image," SITA96, pp.361-364, 1996.
- [7] 岩田賢一, 植松友彦, 岡本栄司, "部分的な復号を可能にする Ziv-Lempel 符号の提案," 信学論 (A), vol.J79-A, no.11, pp.1899-1906, Nov. 1996.
- [8] 木村 勝, 後藤正幸, 松嶋敏泰, 平澤茂一, "マルコフ情報源に対する VF 符号に関する一考察," SITA96, pp.361-364, 1996.
- (平成 10 年 3 月 31 日受付, 9 月 21 日再受付)
-