



情報理論とその応用学会ニュースレター

「情報源符号化ワークショップ」技術特集

名誉会員紹介

- 会員へのメッセージ 嵩 忠雄 (名誉会員)
 会員へのメッセージ 丸林 元 (名誉会員)
 文脈木重みづけ法とその周辺 川端 勉 (電気通信大学情報通信工学科)
 文法変換に基づく圧縮 喜田拓也 (九州大学附属図書館研究開発室)
 ソートを用いた圧縮法 有村 光晴 (電気通信大学大学院システム学研究科)
 音声と楽音の情報圧縮
 守谷 健弘 (NTT サイバースペース研究所・東京大学大学院情報理工学系研究科客員教授)
 量子情報源符号化 松本 啓史 (科学技術振興事業団 今井量子計算機構プロジェクト)
 国際会議のお知らせ

名誉会員紹介

会員へのメッセージ

名誉会員 嵩 忠雄

1. 最終経歴: 広島市立大学教授 2003 年 3 月退職
2. 大阪大学, 奈良先端科学技術大学院大学名誉教授
3. 会へのメッセージ: 学会も、大学同様、国際競争に直面していると言われています。幸い、SITA は ISITA の実績も上がりつつあります。また、SITA 固有の軟構造の特色を生かして、余りコストをかけない独自の会員サービス*でしぶとく競争を生き抜いていかれると期待しております。

*一例をあげますと、原稿締切期日を発表期日にできるだけ近づける。また新規参加者も気軽に参加できる情報交換の場を広げる。

4. 現在のメールアドレス: kasami@cs.hiroshima-cu.ac.jp



会員へのメッセージ

名誉会員 丸林 元

私は NTT 時代以来一貫して通信方式の研究を行ってきました。情報理論を勉強したのは PCM 方式が台頭しつつあった 1955 年頃で、私の中では情報理論はあくまでアナログとデジタルの橋渡しをした理論という位置付けで、情報理論そのものの研究は私の研究対象外でした。本学会には正統な情報理論の研究者としてではなく、“とその応用”に関する研究者として参加し、いつもお歴々の先生方の研究に対する真摯な取り組みに刺激を受けてきました。私のようないささか羽根の色の違う人間と暖かくお付き合い頂き、そのうえ名誉会員にまでご推挙頂きまして大変有難く感謝しております。

もう定年まで残り 1 年有余という今日になって、やっと誰からも強制されることなく自由に研究できる楽しみを味

わっている今日この頃です。

末筆ながら本会のますますの発展を祈っております。

1994 年 3 月 長岡技術科学大学定年退官
 現在 創価大学教授、長岡技術科学大学名誉教授



文脈木重みづけ法とその周辺

川端 勉（電気通信大学情報通信工学科）

無歪みユニバーサル圧縮に関する情報理論研究者の長年の夢は、定常情報源に対する実用的符号化法、特に木情報源を対象とし究極的な冗長度を達成するエレガントな方法、の発見であろう。マイルストーンの一つは文脈木重み付け (CTW) 法である。本講演ではまず、基礎となる算術符号とバイズ符号からはじめ、CTW 法を分かりやすく解説する。次に、この方法の問題点を述べ関連した諸研究について触れる。

発表の内容

- 情報源符号化
- Shannon-Fano-Elias-Pasco-Rissanen のデータ圧縮理論
- 符号化の性能
- バイズ符号
- Clarke and Barron(1992) の結果
- 木情報源（例による定義）
- 木情報源に対するバイズ符号
- CTW のアルゴリズム
- CTW 法のためのデータ構造とアルゴリズム（二進）
- CTW 法の性能（個別冗長度）
- 増分文脈集合による CTW 法（初期系列が未知であることへの対処, $|A| = m$ ）
- 増分文脈アルゴリズムの冗長度性能
- PPM(blending) と CTW の関係
- 非定常情報源への対処（有限窓 CTW法）
- ミニマックス情報量基準

情報源符号化

- A : 情報源アルファベット
 - 例: $\{0, 1\}$ 二進アルファベット, $\{a, b, \dots, z\}$ 英子文字アルファベット
 - A の要素の列を $x_1, x_2, \dots, x_t, \dots, x_n = x_1^n$ データ列（ファイル）と考え、 x_1^{t-1} を時刻 t における過去系列という。
 - x_1^n を x^n とかく。
 - $|A|$: 集合 A の要素数
- 情報源符号化（続き）

- 符号

$$c: A^n \rightarrow \{0, 1\}^* = \{\lambda, 0, 1, 00, 01, 10, 11, 000, \dots\}$$

- 符号語と符号長符号語の例: $c(x^n) = 0110$ 符号語 $c(x^n)$ の符号長を $l(x^n)$ と表す。例えば $l(x^n) = 4$ である。
- 符号語の語頭と語尾
- 符号 c が語頭符号

$$\iff [\forall x^n \neq \forall y^n \Rightarrow c(x^n) \text{ は } c(y^n) \text{ の語頭でない.}]$$

- **定理:**

$$\text{符号 } c \text{ が語頭符号} \Rightarrow \sum_{x^n} 2^{-l(x^n)} \leq 1 \text{ (クラフトの不等式)}$$

- **逆定理:** クラフトの不等式を満足するような符号長関数に対して対応する語頭符号が存在する。
- データ列 X^n を劣確率関数 q で符号化する。
- **系** 劣確率関数 $q: A^n \rightarrow [0.0, 1.0)$ (即ち不等式 $\sum_{x^n} q(x^n) \leq 1$ を満足する関数) に対して, $l(x^n) = \lceil -\log q(x^n) \rceil$ なる符号長関数を持つ語頭符号が存在する。(逆定理からの帰結)

Shannon-Fano-Elias-Pasco-Rissanen のデータ圧縮理論 [2]

- 入力文字列の辞書順と出力文字列の辞書順が同じである符号をアルファベティック符号という。
- $l(x^n) = \lceil -\log q(x^n) \rceil + 1$ なる符号長関数を持つ, 語頭かつアルファベティック符号が存在する。これを本節で示す。
- A^n での辞書順において, $\text{succ}(x^n)$ で x^n の次のデータ列を表す。
- 方法: 分布関数

$$F(x^n) := \sum_{y^n \prec x^n} q(y^n)$$

について区間 $[F(x^n), F(\text{succ}(x^n))]$ の中点

$$F(x^n) + \frac{1}{2}q(x^n)$$

を区間幅の半分 $(1/2)q(x^n)$ の精度の二進固定小数点で表現する。小数点以下

$$l(x^n) = \lceil -\log q(x^n) \rceil + 1$$

ビットで表現すれば, 互いに区別可能で, かつ符号は語頭かつアルファベティックである。

- 分布関数は

$$F(x^n) = \sum_{l=1}^n q(x^{l-1}) \sum_{y_l \prec x_l} q(y_l | x^{l-1}), \text{ ただし } q(y_l | x^{l-1}) = \frac{q(x^{l-1}y_l)}{q(x^{l-1})}$$

とかけるので, 逐次的に計算ができる。

- 列確率関数 q が

$$q(a|x^l) = 2^{-D} \text{ の自然数倍} \\ \sum_{a \in A} q(a|x^l) \leq 1$$

を満足するならば, Shannon-Fano-Elias 符号において

$$q(x^l) = \prod_{i=1}^l q(x_i | x^{i-1})$$

を有効数字 K 桁の浮動小数で下から近似することにより, $c(x^n)$ の計算時間が, すなわちの x^n 符号化復号時間が $O(n)$, 符号長の増分がせいぜい $-n \log(1 - 2^{-(1-K)})$ となるように修正ができる。(Pasco-Rissanen)

符号化の性能

- 個別評価

$$-\log q(x^n) = -\log p(x^n) + \log \frac{p(x^n)}{q(x^n)}$$

- 期待値評価

$$E_p[-\log q(x^n)] = H^{(n)}(p(\cdot)) + D^{(n)}(p(\cdot) \| q(\cdot))$$

バイズ符号 [1][8]

- $\mathcal{M}$: 離散モデル族

- $M \in \mathcal{M}$: モデル
- Θ_M : M に付随する連続なパラメータ空間
- $p(x^n|\theta_M, M)$: M と $\theta_M \in \Theta_M$ によって決まる確率密度

とする.

- 従来の考え方: $x^n (\stackrel{\text{def}}{=} x_1 \dots x_n \stackrel{\text{def}}{=} x_1^n)$ から $(\hat{M}, \hat{\theta}_{\hat{M}})$ の対を推定し, $p(x^n|\hat{M}, \hat{\theta}_{\hat{M}})$ で符号化.
- Bayes 符号: $\cup_{M \in \mathcal{M}} \Theta_M$ 上の事前分布 $w(\theta_M, M)$ に関する混合 $m(x^n) = E_w[p(x^n|\theta_M, M)]$ により符号化する.
- 逐次的ベイズ符号: w_n は x^n を観測した後の θ_M, M に関する Bayes 事後確率密度とすると,

$$\begin{aligned} m(x_{n+1}|x^n) &= m(x^{n+1})/m(x^n) \\ &= E_{w_n}[p(x_{n+1}|x^n, \theta_M, M)] \end{aligned}$$

を用いて逐次的に符号化ができる.

Clarke and Barron(1992) の結果 [5][6]

- 考えているモデルの中に情報源が含まれるものとする.
-

$$\begin{aligned} &\text{min-max 冗長度} \\ \stackrel{\text{def}}{=} &\min_m \max_{\theta_M, M} E_{X^n}[-\log \frac{p(X^n|\theta_M, M)}{m(X^n)}] \end{aligned}$$

- 事前分布 $w(\theta_M, M)$ に関するベイズ符号器とは平均冗長度の事前分布に関する期待値を最小にする符号器 $m(X^n)$ である.
- この最小平均冗長度は入力分布 $w(\theta_M, M)$ と通信路 $p(x^n|\theta_M, M)$ により決まる相互情報量となる.
- 上記の MinMax を与える事前分布はこの相互情報量を最大にする入力分布となる (Gallager).
- モデルが一つの場合: 連続パラメータ θ をもつ確率モデルについて

$$\begin{aligned} \sup_w I(\theta; X^n) &= \min_{q^n} \max_{\theta} D(p(X^n|\theta) \| q^n) + o(1) \\ &= \frac{d}{2} \log \frac{n}{2\pi e} + L + o(1) \end{aligned}$$

が成り立つ. ただし, d は θ の次元で, L は $p(x|\theta)$ の Fisher 情報量行列 $J(\theta)$ を用いて

$$L = \log \int \sqrt{\det J(\theta)} d\theta$$

で定義される量である. 言い換えると, MinMax を与える事前分布はパラメータ空間上情報計量に関する一様分布である.

- 離散有限な添字 $M \in \mathcal{M}$ のみで指定される確率モデル $p(x^n|M)$ について,

$$\begin{aligned} I(M; X^n) &= \sum_{M, X^n} w(M) p(X^n|M) \log \frac{p(X^n|M)}{\sum_{\tilde{M}} w(\tilde{M}) p(X^n|\tilde{M})} \\ &= \sum_M w(M) \log \frac{1}{w(M)} + o(1) \end{aligned}$$

が成り立つ. ただし, $w(M)$ は $\mathcal{M}$ 上の一様分布である.

木情報源 (例による定義) [7]

- 二進アルファベットのベルヌーイ情報源

$$p(x^n) = \theta_\lambda^{n^0} (1 - \theta_\lambda)^{n^1}$$

を $\partial T = \{\lambda\}$ を葉集合とする木 $T = \phi$ で表す.

- 一次マルコフ情報源

$$p(x^n|x_0) = \theta_0^{n_0^0} (1 - \theta_0)^{n_0^1} \theta_1^{n_1^0} (1 - \theta_1)^{n_1^1}$$

を $\partial T = \{0, 1\}$ を葉集合とする木 $T = \{\lambda\}$ で表す.

- 二次マルコフ情報源

$$p(x^n|x_0, x_1) = \prod_{s \in \{00, 10, 01, 11\}} \theta_s^{n_s^0} (1 - \theta_s)^{n_s^1}$$

を $\partial T = \{00, 10, 01, 11\}$ を葉集合とする木 $T = \{\lambda, 0, 1\}$ で表す.

- 一部縮退 ($\theta_0 0 = \theta_1 0 := \theta_0$) した二次マルコフ情報源

$$p(x^n|x_0, x_1) = \prod_{s \in \{0, 01, 11\}} \theta_s^{n_s^0} (1 - \theta_s)^{n_s^1}$$

を $\partial T = \{0, 10, 11\}$ を葉集合とする木 $T = \{\lambda, 1\}$ で表す.

- 上記の例のように、一般に多重マルコフ情報源で、木 T および、その外部葉集合 ∂T の上での確率パラメータの組 $\{\theta_s^0 := p(0|s)\}_{s \in \partial T}$ が与えられることにより定まるマルコフ情報源を木情報源という.

木情報源に対するベイズ符号 [10]

- $n_s^a := \{l : 1 \leq l \leq n, \text{ such that } (s_{l-1}, x_l) = (s, a)\}$
- $n_s := \sum_{a \in A} n_s^a$
- $n := \sum_{s \in \partial T} n_s$

と定義する.

- $0 < \theta_s^a < 1$ を仮定する. (これにより、すべての情報源は既約なマルコフ連鎖となる.)
- Markov 情報源に対して Clarke-Barron 理論を拡張する研究が重要である.
- $\det J(\theta_T)$ の計算までは順調.

$$\det J(\theta_T) = \prod_{s \in \partial T} p(s)^{m-1} \prod_{a \in A} (\theta_s^a)^{-1}$$

, ただし $p(s)$ は状態 s の定常分布である.

- **一般に正しくない仮定:** 「 θ_T にはによらず $p(s)$ は一定とする。」の下で Jeffreys 事前確率は,

$$w_2(\theta_T|T) = \prod_{s \in \partial T} \left\{ \frac{\Gamma(\frac{m}{2})}{\Gamma(\frac{1}{2})^m} \prod_{a \in A} (\theta_s^a)^{-1/2} \right\}$$

となる. (Jeffreys 事前確率に基づく予測推定量は Takeuchi-Kawabata[13])

- この事前分布によって系列 $x_1, \dots, x_n$ の確率密度 $\prod_{s \in \partial T} \prod_{a \in A} (\theta_s^a)^{n_s^a}$ を平均すると,

$$m(x^n|s_0) = \prod_{s \in \partial T} \frac{\prod_{a \in A} (n_s^a + \frac{1}{2})(n_s^a - \frac{1}{2}) \cdots (\frac{1}{2})}{(n_s + \frac{m}{2} - 1)(n_s + \frac{m}{2} - 2) \cdots \frac{m}{2}}$$

となり, 従って x_{n+1} の Bayes 事後予測確率は

$$m(x_{n+1}|x^n) = \frac{n_s^{x_{n+1}} + \frac{1}{2}}{n_s + \frac{m}{2}}$$

となる. (これは Krichevsky and Trofimov[4] が (i.i.d. の場合に) 与えた予測確率であり, これが CTW 法の重要な要素になる)

- 一方、二進木 T を語頭符号化する符号長関数 l について、クラフト不等式により、符号化確率 $w(T) = 2^{-l(T)}$ が定義されるが、これが、木に対する一つの事前分布である。

- とくに、Willems らの事前分布：

$$w_1(T) = 2^{-l(T)} = 2^{-|T \cup \partial T| + |\partial T \cap A^d|}$$

は、木の前順踏査に付随して得られる、深さが $(d-1)$ 以下の木に対する符号に付随したものである。

- 上記の Krichevsky and Trofimov 推定量をこの事前分布によって重み付けしたものが CTW 法であり、両者の持つ再帰性をうまく利用している。

CTW のアルゴリズム [7]

- 再帰関係

$$p_w(s) \stackrel{def}{=} \begin{cases} \frac{p_e(s) + \prod_{a \in A} p_w(as)}{2} & \text{depth}(s) < d \\ p_e(s) & \text{depth}(s) = d \end{cases}$$

によって、 $p_w(\lambda)$ が $\sum_{T \in \mathcal{M}} w_1(T) m(x^n | s_0, T)$ を計算する。

- 例 ($d = 2$)

- $\mathcal{T} := \{\phi, \{\lambda\}, \{\lambda, 0\}, \{\lambda, 1\}, \{\lambda, 0, 1\}\}$
- $\partial \mathcal{T} := \{\{\lambda\}, \{0, 1\}, \{00, 10, 1\}, \{0, 01, 11\}, \{00, 10, 01, 11\}\}$
- $c(\mathcal{T}) := \{0, 100, 110, 101, 111\}$
-

$$\begin{aligned} p_w(\lambda) &= \sum_{T \in \mathcal{T}} 2^{-l(T)} \prod_{s \in \partial T} p_e(s) \\ &= 2^{-1} p_e(\lambda) + 2^{-3} p_e(0) p_e(1) + 2^{-3} p_e(00) p_e(10) p_e(1) \\ &\quad + 2^{-3} p_e(0) p_e(01) p_e(11) + 2^{-3} p_e(00) p_e(10) p_e(01) p_e(11) \\ &= 2^{-1} p_e(\lambda) + 2^{-1} \{2^{-1} p_e(0) + 2^{-1} p_e(00) p_e(10)\} \times \\ &\quad \{2^{-1} p_e(1) + 2^{-1} p_e(01) p_e(11)\} \\ &= 2^{-1} p_e(\lambda) + 2^{-1} p_w(0) p_w(1) \end{aligned}$$

- この再帰関係により $p_w(\lambda)$ を得るには 1 シンボル当り木の深さだけの計算量しか必要としない

CTW 法の一つの一般化 [7][10]

- $\cup_{l=0}^d A^d$ 上の関数 γ を,

$$\begin{cases} s \in \cup_{l=0}^{d-1} A^d \text{ のとき} & 0 < \gamma(s) < 1, \\ s \in A^d \text{ のとき} & \gamma(s) = 1 \end{cases}$$

を満たす関数と定義する。

- また、木 $T \in \cup_{l=0}^{d-1} A^d$ に関し,

$$w_1^{(\gamma)}(T) = \prod_{s \in T} (1 - \gamma(s)) \prod_{s \in \partial T} \gamma(s)$$

と定義する。

- このとき

$$\sum_{T \in \mathcal{M}} w_1^{(\gamma)}(T) m(x^n | s_0, T)$$

は再帰関係

$$p_w(s) \stackrel{def}{=} \begin{cases} \gamma(s) p_e(s) + (1 - \gamma(s)) \prod_{a \in A} p_w(as) & \text{depth}(s) < d \\ p_e(s) & \text{depth}(s) = d \end{cases}$$

により、 $p_w(\lambda)$ として、計算される。

CTW 法のためのデータ構造とアルゴリズム (二進) [7]

- $s \in \cup_{l=0}^d A^l$ 上

$$\begin{cases} n_s^0 & (= 0) \\ n_s^1 & (= 0) \\ p_e(s) & (= 1.0) \\ p_w(s) & (= 1.0) \end{cases}$$

(括弧内は初期値) を用意する.

- $n = 1, 2, 3, \dots$, について,
 - x_{n+1} を読む.
 - 次に過去系列 $x_{-\infty}^n$ の語尾である文脈 s をその長さを $0, 1, \dots, d$ と変えながら,
 - * $p_e(s)$ に $\frac{n_{s^{n+1}} + \frac{1}{2}}{n_s + 1}$ を乗ずる.
 - * $n_{s^{n+1}}, n_s$ をカウントアップする.
 - 次に, 過去系列 $x_{-\infty}^n$ の語尾である文脈 s をその長さを $d, d-1, d-2, \dots, 0$ と変えながら,

$$p_w(s) \stackrel{\text{def}}{=} \begin{cases} \frac{p_e(s) + \prod_{a \in A} p_w(as)}{2} & \text{depth}(s) < d \\ p_e(s) & \text{depth}(s) = d \end{cases}$$

を行う.

CTW 法のためのデータ構造とアルゴリズム (二進)

- 上の再帰式を計算するのに, 素朴な方法では p_w 同士の掛算が $(m-1)$ 回, γ との掛算が 2 回, 足算 1 回で十分である.
- しかし m が大きい時にはその対数程度の回数で計算できる方法がある. 従って入力をビット長で計れば二進文脈木の深さ程度で計算できる.
- 一方, 算術符号を行なうので, 壘積確率も必要になる. 二進の場合, これは易しいが多進の場合工夫がいる. このことも考慮して, 多シンボル用に状態毎に壘積確率計算木を用意すれば計算量が減少する. また再帰式式の中に現れる $p_w(as)$ の確率積の計算も状態毎に確率積計算木を用意することで解決し, 結局全てがうまくゆく (具体的には [12], [9], 特に [11] では算術符号として実現する上で不可欠な確率の有限桁近似の後の符号化アルゴリズムの正当性を論じている).

CTW 法の性能 (個別冗長度)[7]

$$\begin{aligned} & x^n \text{ の符号器 } m \text{ による算術符号長} + \log p(x^n | \theta_T, T) \\ &= (\lceil -\log m(x^n) + 1 \rceil + \log m(x^n)) \\ &\quad - \log \frac{m(x^n)}{m(x^n|T)} - \log \frac{m(x^n|T)}{p(x^n|\theta_T, T)} \\ &\leq 2 + \{2|\partial T| - 1\} \\ &\quad + \{|\partial T| + \frac{|\partial T|}{2} \log \frac{n}{|\partial T|}\} \end{aligned}$$

不等式の意味

- 第 1 項は算術符号長に固有の冗長度
- 第 2 項は $-\log w_1(T)$ により上界される
- 第 3 項は不等式 $-\log p_e(s) \leq 1 + \frac{1}{2} \log n_s$ と, $\log$ の凹性から得られる.

増分文脈集合による CTW 法 [17] (初期系列が未知であることへの対処, $|A| = m$)

- $T_0 := \phi$

- For $t := 1, 2, 3, \dots$,
 - Input x_t
 - Create node $v_t :=$ the unique prefix of $x_1^{t-1} \in \partial T_{t-1}$
 - Add v_t into T_t
 - Initialize $p_e(v_t) := p_w(v_t) (= 1/m)$ and $\forall a \in A$ let

$$n_{v_t}^a := \begin{cases} 1 & \text{if } x_t = a \\ 0 & \text{otherwise} \end{cases}$$

- For every proper postfix s of v_t , in a bottom up way, do
 - * multiply $\frac{n_s^{x_t} + 1/2}{n_s + m/2}$ to $p_e(s)$
 - * $p_w(s) := \frac{p_e(s) + m^{-1} \prod_{b \in A} p_w(bs)}{2}$,
where $p_s(s') = 1$ if $s' \notin T_t$
 - * count up $n_s^{x_t}$
- use $p_w(\lambda)$ as $p(x^t)$.

増分文脈アルゴリズムの冗長度性能 [17]

- 個別冗長度に対する評価が可能
- 定常エルゴート情報源に対するユニバーサル性が証明可能

PPM(blending) と CTW の関係

•

$$m(l|x^n) := \frac{\gamma(s)m(x^n|s)}{\sum_{s: x^n \text{ の語尾}} \gamma(s)m(x^n|s)},$$

ただし $|s| = l$

•

$$\frac{m(x^{n+1})}{m(x^n)} = \sum_{l=1,2,\dots} m(l|x^n) \frac{n_{x_n-l+1}^{x_{n+1}} + \frac{1}{2}}{n_{x_n-l+1}^{x_n} + \frac{m}{2}}$$

非定常情報源への対処（有限窓CTW法） [14] [15][16]

有限の過去のデータのみに基づいて次のシンボルを予測する。次のような特徴をもつ。

- Krichevski-Trofimov 推定を基礎にする。
- CTW 法として組み込むことができる
- データ圧縮法としてのアルゴリズムの正当性がいえる
- CTW 法推定の差の評価に基づき性能の解析が可能である。
- ゼロ確率など構造を有する場合には Laplace 推定量を用いることにより冗長度性能を向上することができる。

ミニマックス情報量基準 [10]

問題：一般的な場合、すなわち離散有限モデル族 $\mathcal{M}$ および $M \in \mathcal{M}$ によって決まるパラメータ空間 Θ_M 上の同時事前分布 $w(\theta_M, M)$ については？

- $|\mathcal{M}| < \infty$
- $\mathcal{M} = \cup_d \mathcal{M}_d$, ただし $\mathcal{M}_d$ は Θ_M の次元が d であるような $\mathcal{M}$ の集合。
- $\Theta_d := \cup_{M \in \mathcal{M}_d} \Theta_M$

- $\theta_M \in \Theta_M$ は i.i.d. 情報源 $p(x^n|\theta_M, M) = \prod_{l=1}^n p(x_l|\theta_M, M)$ を定義するものとし, Θ_M は Clarke and Barron ,92 の正規性条件を満足するものとする. また全ての確率分布は (Mixture 結合に関する) 狭義凸集合の境界集合に含まれるものとする.
- $d_{max} := \max_{M \in \mathcal{M}} \dim \Theta_M$
- $\forall M, \exists \tilde{M} \in \mathcal{M}_{d_{max}}, \Theta_M \subset \Theta_{\tilde{M}}$
- $L_M := \int \sqrt{\det J(\theta_M)} d\theta_M$

定理 1 (川端)

$$w_1^*(M) := \begin{cases} \frac{L_M}{\sum_{M \in \mathcal{M}_{d_{max}}} L_M} & M \in \mathcal{M}_{d_{max}} \\ 0 & M \notin \mathcal{M}_{d_{max}} \end{cases}$$

$$w_2^*(\theta_M|M) := \frac{\sqrt{\det J(\theta_M)}}{L_M}$$

とするとき, $w_1^*(M)w_2^*(\theta_M|M)$ は次の意味で漸近的にミニマックスな冗長度を与える $\cup_M \Theta_M$ の上の *least favorable* な事前分布である. すなわち,

$$\begin{aligned} & \text{ミニマックス冗長度} \\ &:= \min_{q^n} \max_{M, \theta_M} D(p(X^n|M, \theta_M) \parallel q^n) \\ &= \log Z + o(1) \end{aligned}$$

ただし Z は規格のための定数であり

$$Z \stackrel{\text{def}}{=} \left(\frac{n}{2\pi e}\right)^{d_{max}/2} \sum_{M \in \mathcal{M}_{d_{max}}} L_M$$

である.

$\log Z$ は M によらず (もちろんパラメータ θ_M にもよらない) モデル族のみの関数である. $\log Z$ をミニマックス情報量規準と呼ぼう (モデル選択のための規準ではなくモデル族選択のための規準の一つ).

まとめ

定常エルゴート過程に対するデータ圧縮を考えると, 実際の近似マルコフモデルとして木モデルが標準的な役割を担っている. このモデルに対してもっとも有効な方法が文脈木重み付け法と考えられる. 本発表では, この原理とアルゴリズムを中心に述べた. また, 関連した研究として, 筆者の関係したものの選んで述べた. これら以外にも, 思いつくものが多くあるが今回はとり上げなかった. 別の機会に委ねたい.

参考文献

- [1] L.D.Davisson, "Universal noiseless source coding," *IEEE Trans. on Information Theory*, vol.IT-19, no.6, pp.783-795, Nov., 1973.
- [2] Pasco, "Source coding algorithms for fast data compression," Ph.D. Thesis, Department of Electrical Eng. Stanford University, 1976.
- [3] J.Rissanen, "Complexity of string in the class of markov sources," *IEEE Trans. on Information Theory*, vol.IT-32, no.4, pp.526-532, July, 1986.
- [4] R.E.Krichevsky and V.K.Trofimov, "The performance of universal encoding," *IEEE Trans. on Information Theory*, vol.IT-27, no.2, pp.199-207, March, 1981.
- [5] B. S. Clarke and A. R. Barron, "Information-Theoretic Asymptotics of Bayes Methods," *IEEE Trans. on Information Theory*, vol.IT-36, no.3, pp.453-471, May, 1990.

- [6] B. S. Clarke and A. R. Barron, "Jeffreys prior is asymptotically least favorable under entropy risk," *Journal of Statistical Planning and Inference*, vol.41, pp.37-60, 1994.
- [7] F.M.J.Willems, Y.M.Shtarkov and T. J. Tjalkens, "The context tree weighting method: basic properties," *IEEE Trans. on Information Theory*, vol. 41, pp.653-664, May 1995.
- [8] T.Matsushima,H.Inazumi and S. Hirasawa,"A class of distortionless codes designed by Bayes decision theory," *IEEE Trans. on Information Theory*, vol.IT-37, no.5, pp.1288-1293, September, 1991.
- [9] 川端勉,"CTW (2進情報源, 事前分布=(a universal prior for tree, パラメータ:Jeffreys, 深さ制約=d) 符号器," 及び"多シンボル ($|A| = 2^3$) のCTWのための文脈 s のデータ構造," 電気通信大学情報理論研究会, 93/5/6 発表メモ.
- [10] 川端 勉, "ベイズ符号と文脈木重みづけ法," **電子情報通信学会研究技術報告**, IT93-121, Mar. 1994.
- [11] 横尾直広, 川端勉,"文脈木重みづけ法の簡素な実現ならびにその正当性の証明," 電子情報通信学会技術研究報告,IT, Mar. 1994.
- [12] 横尾直弘, 川端 勉, "多値情報源に対する CTW 法の実現," **電子情報通信学会秋季大会講演論文集**, Vol.1, SA-7-4, p271-272, Sep. 1994.
- [13] J.-I. Takeuchi and T. Kawabata, "Approximation of Bayes code for Markov sources," *Proceedings 1995 IEEE International Symposium on Information Theory*, p.391, Whistler, Sep. 1995.
- [14] 坂口浩章, 川端 勉, "有限窓を用いた文脈木重みづけ法," **電子情報通信学会論文誌**, Vol.J80-A, No.12, pp.2155-2163, Dec. 1997
- [15] 川端 勉, 田崎尚久, "マルコフ情報源に対する有限窓 Krichevsky-Trofimov 推定量の冗長度," **電子情報通信学会論文誌**, Vol. J82-A, No.1, pp.112-118, Jan. 1999
- [16] 川端 勉, 渡辺太一, Zulhani Bin Rasin, "マルコフ情報源に対する有限窓 Laplace 推定量の冗長度解析," **電子情報通信学会論文誌**,Vol.J84-A, No.6, pp.705-713, Jun. 2001.
- [17] T. Kawabata and F. M. J. Willems, "A context tree weighting algorithm with an incremental context set," *IEICE Transactions on Fundamentals*, Vol.E83-A, No.10, pp.1898-1903, Oct. 2000.

文法変換に基づく圧縮

喜田拓也（九州大学附属図書館研究開発室）

文法変換に基づく圧縮

九州大学附属図書館
研究開発室
喜田拓也



講演内容

前半:

- 文法変換に基づく圧縮とは？
- SEQUITURアルゴリズム
- Kieffer と Yang の研究

後半:

- 圧縮テキストに対する文字列照合

情報源符号化ワークショップ／文法変換に基づく圧縮

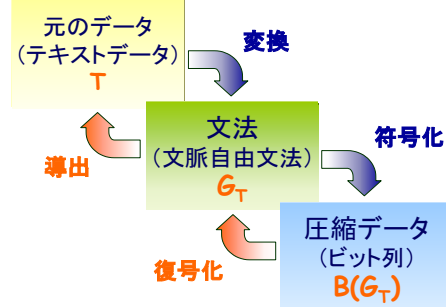


文法変換に基づく圧縮とは？

- 文法変換に基づく圧縮とは？
 - 文法変換に基づく圧縮
 - 文脈自由文法
 - これまでの研究
- SEQUITURアルゴリズム
- Kieffer と Yang の研究
- 圧縮テキストに対する文字列照合

文法変換に基づく圧縮

John C. Kieffer and En-hui Yang, 2000.



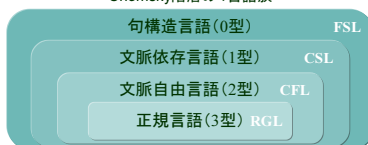
情報源符号化ワークショップ／文法変換に基づく圧縮



文脈自由文法

- 4つ組 $G=(N, \Sigma, P, S)$ によって定義される。
 - N : 非終端アルファベットと呼ばれる空でない有限集合。
 - Σ : 終端アルファベットと呼ばれる空でない有限集合。
 - $P: N \times (N \cup \Sigma)^*$ の有限部分集合。
 P の要素 (A, α) は生成規則と呼ばれ、 $A \rightarrow \alpha$ と書かれる。
 - $S: S \in N$ で、開始記号という。
- BNF(Backus Naur Form)は文脈自由文法の表記法の一つ。

Chomsky階層の4言語族



情報源符号化ワークショップ／文法変換に基づく圧縮



文脈自由文法の例(1)

文脈自由文法 $G=(N, \Sigma, P, S)$

$N=\{S\}, \Sigma=\{a, b\}, P=\{S \rightarrow ab, S \rightarrow aSb\}$

$S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aaabbbb$

$L(G)=\{a^n b^n \mid n \text{ は } 1 \text{ 以上の整数}\}$

G が表す文字列の集合



構文木:

情報源符号化ワークショップ／文法変換に基づく圧縮



これまでの研究 (C. Kieffer ら調べによる)

[G固定: 構文木を圧縮]

- R. Cameron [1988]
- E. Kawaguchi and T. Endo [1980]
- E. Kourapova and B. Ryabko [1995]

[$L(G)=\{x\}$ となる G を圧縮]

- C. Cook, A. Rosenfeld, and A. Aronson [1976]
- J. Storer and T. Szymanski [1995]
- C. Nevill-Manning and I. Witten [1997]

情報源符号化ワークショップ/文法変換に基づく圧縮



文脈自由文法の例 (2)

文脈自由文法 $G=(N, \Sigma, P, S)$

$N=\{S, A, B\}, \Sigma=\{a, b\},$

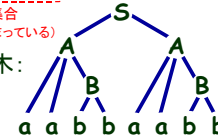
$P=\{S \rightarrow AA, A \rightarrow aaB, B \rightarrow bb\}$

$S \Rightarrow AA \Rightarrow aaBaaB \Rightarrow aabbaabb$

$L(G)=\{aabbaabb\}$

G が表す文字列の集合
(この場合は一意に決まっている)

構文木:

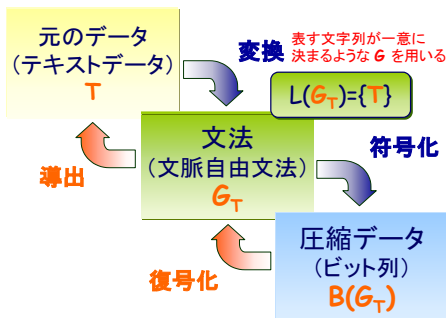


情報源符号化ワークショップ/文法変換に基づく圧縮



文法変換に基づく圧縮

John C. Kieffer and En-hui Yang, 2000.



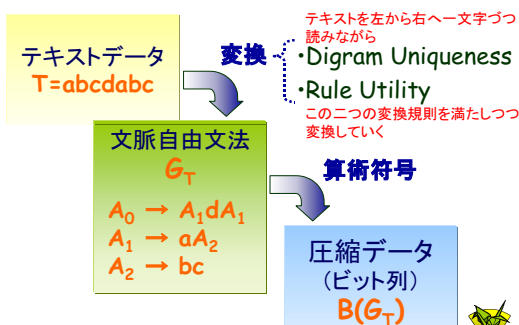
情報源符号化ワークショップ/文法変換に基づく圧縮



SEQUITUR アルゴリズム

- 文法変換に基づく圧縮とは?
- SEQUITUR アルゴリズム
 - アルゴリズムの概要
 - SEQUITUR の特徴
 - 性能について
- Kieffer と Yang の研究
- 圧縮テキストに対する文字列照合

アルゴリズムの概要



情報源符号化ワークショップ/文法変換に基づく圧縮



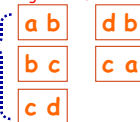
— Digram Uniqueness —

- Digram とは?

$T = a b c d b c a b c d b c$

digram
(隣り合う2文字の組)

この T には以下の5つの digram が含まれている



情報源符号化ワークショップ/文法変換に基づく圧縮



— Digram Uniqueness —

- 同じ digram が、生成規則の右辺に2回以上現われてはいけない

$T = a b c d b c a b c d b c$

$S \rightarrow a b c d b c$
 $A \rightarrow b c$

$b c$ が2回でたので、 $b c$ を生成する規則 A を加える

- 1) 新たに出てきた digram が、以前に出てきた digram と同じものだった場合

情報源符号化ワークショップ/文法変換に基づく圧縮



— Digram Uniqueness —

- 同じ digram が、生成規則の右辺に2回以上現われてはいけない

$T = a b c d b c a b c d b c$

$S \rightarrow a A d A$
 $A \rightarrow b c$

そして $b c$ を A に置換

- 1) 新たに出てきた digram が、以前に出てきた digram と同じものだった場合

情報源符号化ワークショップ/文法変換に基づく圧縮



— Digram Uniqueness —

- 同じdigramが、生成規則の右辺に2回以上現われてはいけない

T = a b c d b c a b c d b c

S → a A d A a b c
A → b c また b c が出現

- 2) 新たに出てきたdigramに対する生成規則がすでにある場合

情報源符号化ワークショップ／文法変換に基づく圧縮

15/58

— Digram Uniqueness —

- 同じdigramが、生成規則の右辺に2回以上現われてはいけない

T = a b c d b c a b c d b c

S → a A d A a A
A → b c b c を A に置換

- 2) 新たに出てきたdigramに対する生成規則がすでにある場合

情報源符号化ワークショップ／文法変換に基づく圧縮

16/58

— Digram Uniqueness —

- 同じdigramが、生成規則の右辺に2回以上現われてはいけない

T = a b c d b c a b c d b c

S → a A d A a A
A → b c 置換したことで、新たな digram が出現

- 1) 新たに出てきたdigramが、以前に出てきた digramと同じものだった場合
- 2) 新たに出てきたdigramに対する生成規則がすでにある場合

情報源符号化ワークショップ／文法変換に基づく圧縮

17/58

— Digram Uniqueness —

- 同じdigramが、生成規則の右辺に2回以上現われてはいけない

T = a b c d b c a b c d b c

S → B d A B
A → b c 規則 B を作って置換
B → a A

- 1) 新たに出てきたdigramが、以前に出てきた digramと同じものだった場合
- 2) 新たに出てきたdigramに対する生成規則がすでにある場合

情報源符号化ワークショップ／文法変換に基づく圧縮

18/58

— Rule Utility —

- S を除くすべての生成規則は2回以上使われなければならない

T = a b c d b c a b c d b c

S → B d A B d
A → b c 規則 C を作って置換
B → a A
C → B d

情報源符号化ワークショップ／文法変換に基づく圧縮

19/58

— Rule Utility —

- S を除くすべての生成規則は2回以上使われなければならない

T = a b c d b c a b c d b c

S → C A C
A → b c
B → a A すると...
C → B d 規則 B は右辺に一回しか使われていないことになる!

情報源符号化ワークショップ／文法変換に基づく圧縮

20/58

— Rule Utility —

- S を除くすべての生成規則は2回以上使われなければならない

T = a b c d b c a b c d b c

S → C A C
A → b c 規則 B を消して、代わりに C の右辺の B に入れる
C → a A d

- 1) 1回しか使われない生成規則は省略しよう!
(大局的にはそれが最適とは限らないが...)

情報源符号化ワークショップ／文法変換に基づく圧縮

21/58

—変換アルゴリズムの結果—

- テキストの最後まで変換すると...

T = a b c d b c a b c d b c

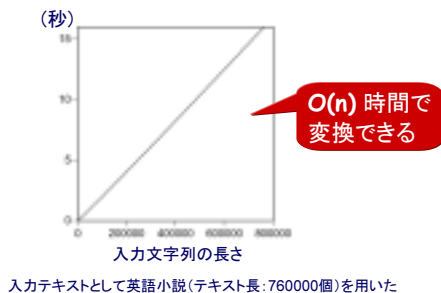
S → D D
A → b c
D → a A d A

たったこれだけで T を表すことができる

情報源符号化ワークショップ／文法変換に基づく圧縮

22/58

アルゴリズムの実行時間



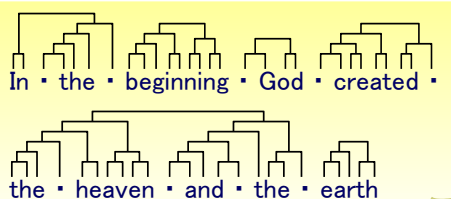
情報源符号化ワークショップ/文法変換に基づく圧縮

23/58

SEQUITURの特徴(1)

- 生成された文法 G_T は、元のテキストが持っている特徴(文法構造やよく出るフレーズ)を階層的に表現する

自然言語の文章の場合



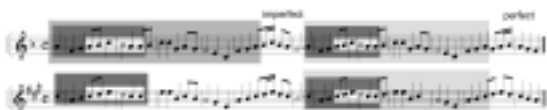
情報源符号化ワークショップ/文法変換に基づく圧縮

24/58

SEQUITURの特徴(2)

- 生成された文法 G_T は、元のテキストが持っている特徴(文法構造やよく出るフレーズ)を階層的に表現する

音楽(楽譜)の場合



情報源符号化ワークショップ/文法変換に基づく圧縮

25/58

生成文法の例(1)

- 構文木の階層が最も深い文法を構成するとき

$T = ababcababcdabcdeabcdef$

$S \rightarrow ABCDDf$

$A \rightarrow ab$

$B \rightarrow Ac$

$C \rightarrow Bd$

$D \rightarrow Ce$

一つ前の非終端記号を含み、右辺の長さが2

$O(\sqrt{n})$

- 最も大きい文法を構成するとき

$T = aabacadae...bbcbdbe...$

$S \rightarrow aabacadae...$

同じdigramが現れないので、生成規則ができない

n

情報源符号化ワークショップ/文法変換に基づく圧縮

26/58

生成文法の例(2)

- 最も小さな文法を生成するとき

$T = aaaaaaaaaaaaaaaaaa...$

$S \rightarrow DD$

$A \rightarrow aa$

$B \rightarrow AA$

$C \rightarrow BB$

$D \rightarrow CC$

$O(\log n)$

- 生成規則の数が最も多くなるとき

$T = aaaaaababacacdad$

$S \rightarrow AABBCDD$

$A \rightarrow ab$

$B \rightarrow ab$

$C \rightarrow ac$

$D \rightarrow ad$

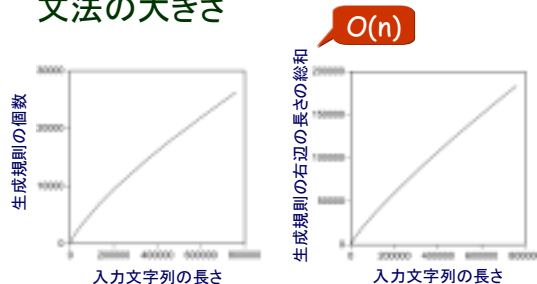
右辺に非終端記号が現れない

$n/4$

情報源符号化ワークショップ/文法変換に基づく圧縮

27/58

文法の大きさ



入力テキストとして英語小説(テキスト長:760000個)を用いた

情報源符号化ワークショップ/文法変換に基づく圧縮

28/58

性能について

- 圧縮率(圧縮データ長/元のテキスト長)
- 圧縮・展開の実行時間

		SEQUITUR	Compress	Gzip	Gzip -9
Brown Corpus (6.83Mbyte)	圧縮率(%)	34	44	39	39
	圧縮時間	35.93	1.02	3.54	5.07
	解凍時間	8.93	0.61	0.33	0.33
Genbank (17.11Mbyte)	圧縮率(%)	21	27	23	22
	圧縮時間	197.14	1.91	13.79	62.03
	解凍時間	17.47	1.30	0.57	0.55

時間はかかるが、圧縮率は単純な Gzip よりも良くなる！ 時間の単位は秒

情報源符号化ワークショップ/文法変換に基づく圧縮

29/58

Kieffer と Yang の研究

- 文法変換に基づく圧縮とは？
- SEQUITURアルゴリズム
- Kieffer と Yang の研究
 - 研究の成果
 - Admissible Grammars
 - Grammar Transform
- 圧縮テキストに対する文字列照合

研究の成果

J. C. Kieffer and E. Yang, Grammar-Based Codes: A New Class of Universal Lossless Source Codes, IEEE Transactions on Information Theory, Vol. 46, No. 3, May 2000.

- ゆるい制約のもとで、文法変換に基づく圧縮はユニバーサル符号であることを示した。
- 効率のよい変換の設計手法を示した。
 - Admissible grammar と二種類の文法変換の方法
 - Admissible grammar の判定方法
 - L-system (A. Lindenmayer [1968], G. Rozenberg and A. Salama [1980]) との対応づけ
 - Grammar の符号化方法
 - 生成規則の効率の良い削り方

情報源符号化ワークショップ／文法変換に基づく圧縮

31/58

Admissible Grammar

- Admissible grammar は文脈自由文法のサブクラスの一つ
- Admissible grammar G は $L(G)=\{x\}$ を保証する

Admissible grammar G の定義

1. G は deterministic である。
任意の変数 A について、 $A \rightarrow \alpha$ となる生成規則がただ一つだけある。
2. どの生成規則の右辺にも空語が現れない。
 $A \rightarrow \epsilon$ となる生成規則が存在しない。
3. $L(G) \neq \emptyset$ である。
4. G は使用しない記号を持たない。
任意の変数・文字 Z ($Z \neq S$) について、 $S \Rightarrow^* x$ の導出の間に少なくとも一回は出現する。

$A \rightarrow aA$
このような規則を許さない

情報源符号化ワークショップ／文法変換に基づく圧縮

32/58

文法変換の方法

Asymptotically compact grammar transforms

- x を表現する Gx は Admissible である
- $\lim_{n \rightarrow \infty} \max_{x \in \Sigma^n} (|Gx|/|x|) = 0$

→ Lempel-Ziv grammar transform
→ Bisection grammar transform

Irreducible grammar transforms

- x を表現する Gx は Admissible である
- 異なる変数は異なる文字列を生成する
- どのような記号のペア (digram) も生成規則の右辺に2回以上現れない
- すべての非終端記号は少なくとも2回は Gx の生成規則の右辺に現れる

→ Sequitur transform

情報源符号化ワークショップ／文法変換に基づく圧縮

33/58

圧縮テキストに対する文字列照合

- 文法変換に基づく圧縮とは？
- SEQUITUR アルゴリズム
- Kieffer と Yang の研究
- 圧縮テキストに対する文字列照合
 - 文字列照合問題とは？
 - 研究背景と目的
 - 目標1の達成
 - コーラージュ・システム
 - 目標2の達成

文字列照合問題とは？

- テキスト T 中に含まれるパターン P の出現を求める問題
 - KMP法 (Knuthら[1974])
 - BM法 (BoyerとMoore[1977])
 - ビットパラレル手法 (Abrahamson[1987], Baeza-YatesとGonnet[1992])

パターン **compress**

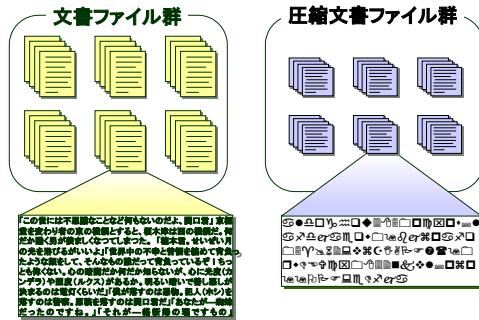
テキスト

We introduce a general framework which is suitable to capture an essence of **compressed** pattern matching according to various dictionary based **compressions**. The goal is to find all occurrences of a pattern in a text without **decompression**, which is one of the most active topics in string matching. Our framework includes such **compression** methods as Lempel-Ziv family, (LZ77, LZSS, LZ78, LZWF), byte-pair encoding, and the static dictionary based method. Technically, our pattern matching algorithm extremely extends that for LZW **compressed** text presented by Amir, Benson and Farach [Amir94].

情報源符号化ワークショップ／文法変換に基づく圧縮

35/58

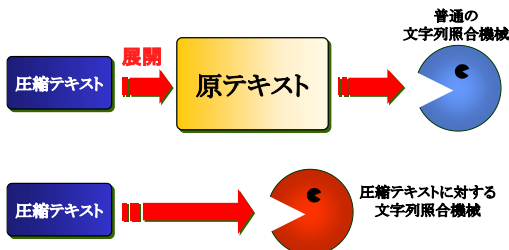
研究目的



情報源符号化ワークショップ／文法変換に基づく圧縮

36/58

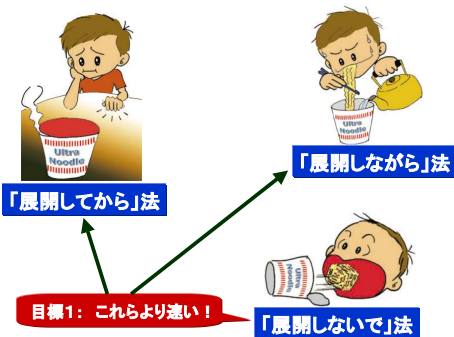
圧縮されたデータに対する文字列照合



情報源符号化ワークショップ／文法変換に基づく圧縮

37/58

この問題に対する3つの手法



情報源符号化ワークショップ／文法変換に基づく圧縮

38/58

圧縮方法	照合アルゴリズム
Run-length	Eilam-Tzoreff & Vishkin (1988)
Run-length (two dim)	Amir <i>et al.</i> (1992, 1997); Amir & Benson (1992)
LZ77 family	Farach & Thorup (1995); Gasieniec, <i>et al.</i> (1996); Klein & Shapira (2000)
LZ78 family	Amir <i>et al.</i> (1996); Kida <i>et al.</i> (1998, 1999); Navarro & Tarhio (2000); Kärkkäinen <i>et al.</i> (2000); Navarro <i>et al.</i> (1999)
LZ family	Karpinski <i>et al.</i> (1997); Miyazaki <i>et al.</i> (1997); Hirao <i>et al.</i> (2000)
Straight-line programs	Fukamachi <i>et al.</i> (1998); Klein & Shapira (2001); Miyazaki <i>et al.</i> (1998)
Huffman	Takeda (1997)
Finite state encoding	Moura <i>et al.</i> (1998)
Word based encoding	Manber (1994); Shibata <i>et al.</i> (1998)
Pattern substitution	Shibata <i>et al.</i> (1999)
Antidictionary based	

情報源: 符号化ワーキングショップ / 文法変換に基づく圧縮

39/38

圧縮方法	照合アルゴリズム
Run-length	Eilam-Tzoreff & Vishkin (1988)
Run-length (two dim)	Amir <i>et al.</i> (1992, 1997); Amir & Benson (1992)
LZ77 family	Farach & Thorup (1995); Gasieniec, <i>et al.</i> (1996); Klein & Shapira (2000)
LZ78 family	Amir <i>et al.</i> (1996); Kida <i>et al.</i> (1998, 1999);
LZ family	Navarro & Tarhio (2000); Kärkkäinen <i>et al.</i> (2000); Navarro <i>et al.</i> (1999)
Straight-line programs	Karpinski <i>et al.</i> (1997); Miyazaki <i>et al.</i> (1997); Hirao <i>et al.</i> (2000)
Huffman	Fukamachi <i>et al.</i> (1998); Klein & Shapira (2001); Miyazaki <i>et al.</i> (1998)
Finite state encoding	Takeda (1997)
Word based encoding	Moura <i>et al.</i> (1998)
Pattern substitution	Manber (1994); Shibata <i>et al.</i> (1998)
Antidictionary based	Shibata <i>et al.</i> (1999)

情報源符号化ワークショップ／文法変換に基づく圧縮

39/58

40/58

目標1の達成

2-bit 長さ (パターン)	AlphaStation XP1000 (秒)	compress(LZW)+KMP (秒)	gunzip(LZ77)+KMP (秒)
5	0.85	0.65	0.55
10	0.85	0.65	0.55
15	0.85	0.65	0.55
20	0.85	0.65	0.55
25	0.85	0.65	0.55
30	0.85	0.65	0.55

AlphaStation XP1000
(Alpha21264: 667MHz)
Tru64 UNIX V4.0F
Genbank (DNA塩基配列)
17.1Mbyte

「展開しながら」法
compress(LZW)+KMP
gunzip(LZ77)+KMP

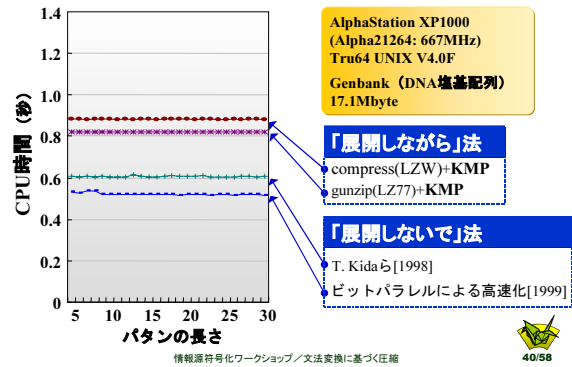
「展開しないで」法
T. Kidaら[1998]
ビットパラレルによる高速化[1999]

CPU時間 (秒)

2-bit 長さ (パターン)

情報源符号化ワークショップ/文法変換に基づく圧縮

40/58



圧縮テキストに対する 文字列照合

- 文法変換に基づく圧縮とは？
- SEQUITURアルゴリズム
- Kieffer と Yang の研究
- 圧縮テキストに対する文字列照合
 - 文字列照合問題とは？
 - 研究背景と目的
 - 目標1の達成
 - コラーージュ・システム
 - 目標2の達成

- 文法変換に基づく圧縮とは？
- SEQUITURアルゴリズム
- Kieffer と Yang の研究
- 圧縮テキストに対する文字列照合
 - 文字列照合問題とは？
 - 研究背景と目的
 - 目標1の達成
 - コラージュ・システム
 - 目標2の達成

論文の衝突

- 第一次ショック (at CPM'99)
 - T. Kida, et al., Shift-And Approach to Pattern Matching in LZW Compressed Text
 - G. Navarro and M. Raffinot, A General Practical Approach to Pattern Matching over Ziv-Lempel Compressed Text
- 第二次ショック (at CPM2000)
 - Y. Shibata, et al., A Boyer-Moore type algorithm for compressed pattern matching
 - G. Navarro and J. Tarhio, Boyer-Moore string matching over Ziv-Lempel compressed text

G. Navarro とその家族



情報源符号化ワークショップ／文法変換に基づく圧縮



42/58

- **第一次ショック (at CPM'99)**
 - T. Kida, et al., Shift-And Approach to Pattern Matching in LZW Compressed Text
 - G. Navarro and M. Raffinot, A General Practical Approach to Pattern Matching over Ziv-Lempel Compressed Text
- **第二次ショック (at CPM2000)**
 - Y. Shibata, et al., A Boyer-Moore type algorithm for compressed pattern matching
 - G. Navarro and J. Tarhio, Boyer-Moore string matching over Ziv-Lempel compressed text

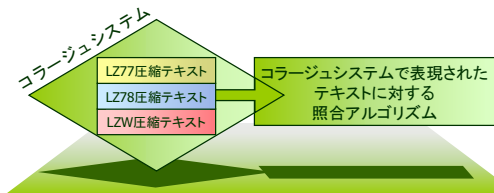
G. Navarro とその家族

情報源符号化ワークショップ／文法変換に基づく圧縮

42/58

- コラージュ・システム(Collage system)
T. Kida, et al. A Unifying Framework for Compressed Pattern Matching,
In Proc. 6th International Symp. On String Processing and Information Retrieval,
pp. 89-96. IEEE Computer Society, 1999.
- 辞書式圧縮法によって圧縮されたテキストを表現
するための形式的体系

辞書式圧縮法によって圧縮されたテキストを表現するための形式的体系



情報源符号化ワークショップ／文法変換に基づく圧縮

43/58

コラージュ・システムの例

$X_1 = a$;
 $X_2 = b$;
 $X_3 = X_1 * X_2$; (連接) ab
 $D: X_4 = X_2 * X_1$; (連接) ba
 $X_5 = (X_3)^3$; (繰り返し) $ababab$
 $X_6 = {}^3X_5$; (切り落とし) bab

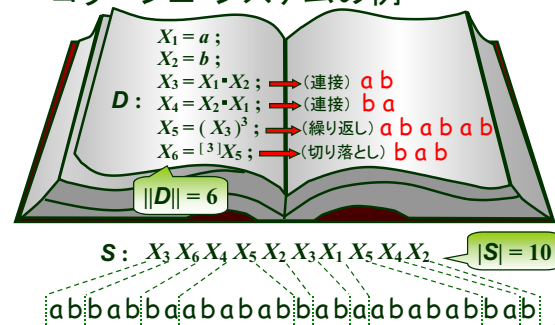
$\|D\| = 6$

$S: X_3 X_6 X_4 X_5 X_2 X_3 X_1 X_5 X_4 X_2$ $|S| = 10$

$abababbaabababbaabababbaabababba$

情報符号化ワークショップ／文法変換に基づく圧縮

44/58



情報源符号化ワークショップ／文法変換に基づく圧縮

44/58

コージュ・システムの定義

- コージュシステムとは、組 (D, S)
- D : トークン割当の集合 (辞書に相当)
 - $X_1 = \text{expr}_1; X_2 = \text{expr}_2; \dots; X_n = \text{expr}_n$
各 expr_k は以下のいずれか
 - $a \quad a \in \Sigma \cup \{ \epsilon \}$ 一文字割当
 - $X_i \cdot X_j \quad i, j$ は整数で、 $i, j < k$ 連結
 - $(X_i)^j \quad i, j$ は整数で、 $i < k$ 繰り返し
 - $[i]X_j \quad i, j$ は整数で、 $i < k$ 前切り落とし
 - $X_j^{[j]} \quad i, j$ は整数で、 $i < k$ 後切り落とし
 - $\|D\| = n : D$ 中のトークンの個数
 - $X.u$: トークン X が表す文字列
- $S : D$ で割当てられたトークンの列 (符号列に相当)
 - $X_{i_1} X_{i_2} \dots X_{i_l} \quad (X_i \text{ は } D \text{ 中のトークン})$
 - $|S| = l$: トークンの列の長さ



45/58

- コラージュシステムとは、組 $\langle D, S \rangle$
- D : トークン割当の集合 (辞書に相当)
 - $X_1 = \text{expr}_1; X_2 = \text{expr}_2; \dots; X_n = \text{expr}_n$
各 expr_k は以下のいずれか
 - a 一文字割当
 - $X_i \cdot X_j$ 連結
 - $(X_i)^j$ 繰り返し
前切り落とし
後切り落とし
 - $X_i^{[j]}$ 後切り落とし
 - $\|D\| = n$: D 中のトークンの個数
 - $X.u$: トークン X が表す文字列
- S : D で割当てられたトークンの列 (符号列に相当)
 - $X_1 X_2 \dots X_{l_i}$ (X_i は D 中のトークン)
 - $|S| = l$: トークンの列の長さ

情報源符号化ワークショップ／文法変換に基づく圧縮

45/58

照合アルゴリズムの計算量

定理

コーラージュシステム $\langle D, S \rangle$ で表現されたテキストに対する文字列照合問題は、
 $O(\|D\| + |P|^2)$ 領域を用いて
 $O((\|D\| + |S|) \cdot \text{height}(D) + |P|^2 + r)$ 時間で解決できる。

$|P|$ はパタンの長さ

r はパタンの出現回数

圧縮テキスト長

D が切り落とし操作を含まない場合は、
 $O(\|D\| + |S| + |P|^2 + r)$ 時間で解決できる。

情報源符号化ワークショップ/文法変換に基づく圧縮

46/58

コーラージュシステム $\langle D, S \rangle$ で表現されたテキストに対する文字列照合問題は、 $O(\|D\| + |P|^2)$ 領域を用いて $O((\|D\| + |S|) \cdot \text{height}(D) + |P|^2 + r)$ 時間で解決できる。

圧縮テキスト長

$|P|$ はハタンの長さ

f はハタンの出現回数

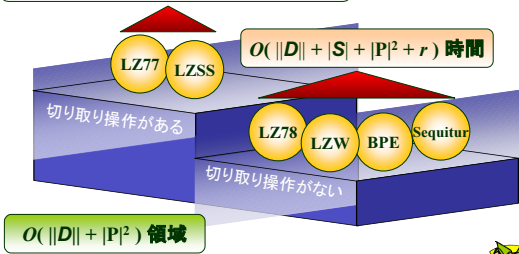
D が切り落とし操作を含まない場合は,
 $O(\|D\| + |S| + |P|^2 + r)$ 時間で解決できる.

情報源符号化ワークショップ／文法変換に基づく圧縮

46/58

得られた知見

$$O((\|D\| + |S|) \cdot \text{height}(D) + |P|^2 + r) \text{ 時間}$$



情報源符号化ワークショップ／文法変換に基づく圧縮

47/58



情報源符号化ワークショップ／文法変換に基づく圧縮

48/58

圧縮文字列照合する理由は？



情報源符号化ワークショップ／文法変換に基づく圧縮

49/58

圧縮文字列照合する理由は？

目標2 新たな目標

原テキスト上の照合時間

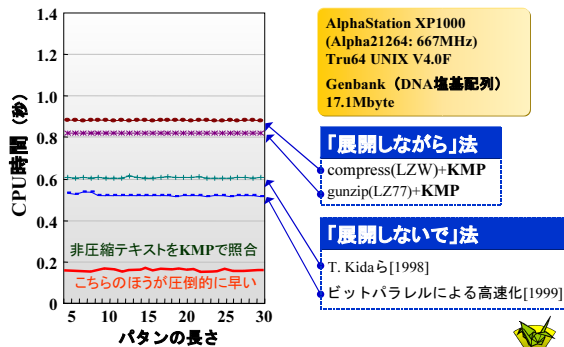
圧縮テキスト上の照合時間



情報源符号化ワークショップ／文法変換に基づく圧縮

50/58

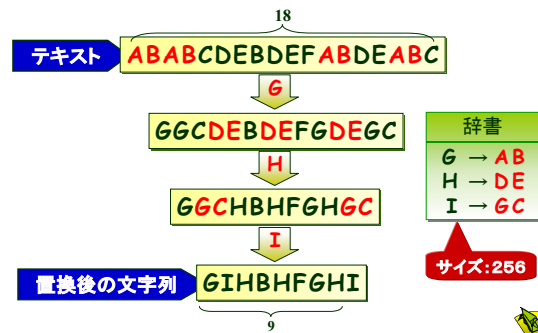
新たな目標(目標2)



情報源符号化ワークショップ／文法変換に基づく圧縮

51/58

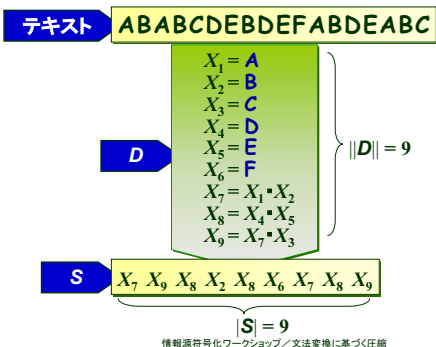
Bite Pair Encoding(BPE)圧縮法



情報源符号化ワークショップ／文法変換に基づく圧縮

52/58

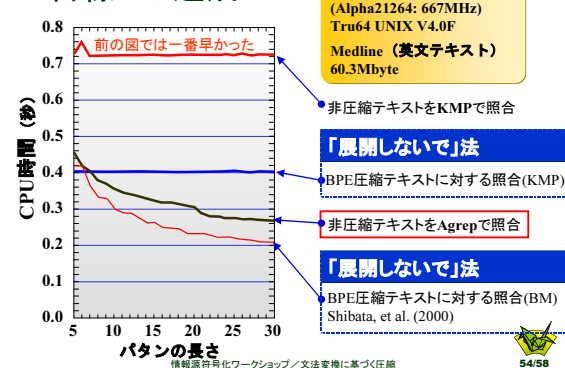
コラージュ・システムによるBPEの表現



情報源符号化ワークショップ／文法変換に基づく圧縮

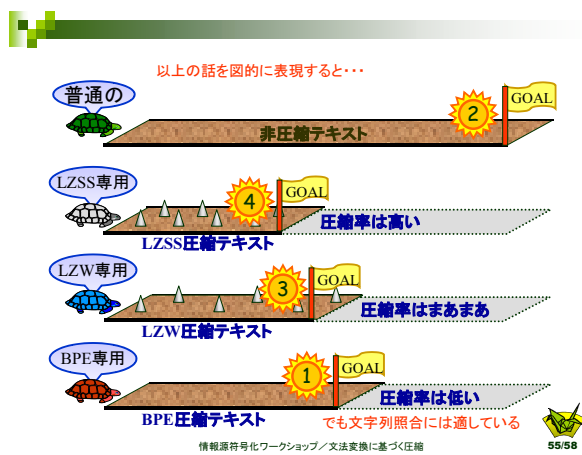
53/58

目標2の達成



情報源符号化ワークショップ／文法変換に基づく圧縮

54/58



その他の結果

- SEQUITUR圧縮テキスト上の文字列照合
 - S. Mitarai, et al., **Compressed Pattern Matching for SEQUITUR**, In *Proc. Data Compression Conference 2001*, pp. 469-478, IEEE Computer Society, 2001.
- 圧縮テキストに対する近似文字列照合
 - T. Matsumoto, T. Kida, et al., **Bit-parallel approach to approximate string matching in compressed texts**, In *Proc. 7th International Symp. On String Processing and Information Retrieval*, pp. 221-228, IEEE Computer Society, 2000.
 - G. Navarro, T. Kida, et al., **Faster Approximate String Matching over Compressed Text**, In *Proc. Data Compression Conference 2001*, pp. 459-468, IEEE Computer Society, 2001.

情報源符号化ワークショップ／文法変換に基づく圧縮



56/58

まとめ

- SEQUITURアルゴリズム
 - テキスト → 文法 G_T → (算術符号) → 圧縮データ
 - 圧縮率はGzipよりもよいが、圧縮・展開に非常に時間がかかる
 - 生成された文法 G_T は、元のテキストが持っている特徴を階層的に表現する
- Kieffer と Yang の研究
 - ゆるい制約のもとで、文法変換に基づく圧縮はユニバーサル符号であることを示した
 - 効率のよい変換の設計手法を示した
- 圧縮テキストに対する文字列照合
 - Collage systemを提案し、その上での文字列照合アルゴリズムを開発した
 - BPE圧縮されたテキストに対する文字列照合を開発した
 - 元のテキストを照合するより高速
 - 文字列処理を高速化するためのデータ圧縮

情報源符号化ワークショップ／文法変換に基づく圧縮



57/58

ソートを用いた圧縮法

有村 光晴（電気通信大学大学院システム学研究科）

ソートを用いた圧縮法

電気通信大学
大学院情報システム学研究科
有村光晴
E-mail: arimura@ieee.org

0. 無歪データ圧縮に使用されるツール

- 文字列の一致
 - Lempel-Ziv法
- 確率の計算
 - Huffman符号、算術符号とその拡張
- 木構造、グラフ
 - Huffman符号、Lempel-Ziv符号、文法圧縮、DMC、CTW
- 文脈自由文法
 - 文法圧縮
- ソート
- 置換 (permutation)

2

2002年11月11日



情報源符号化ワークショップ

1. ソートを用いたデータ圧縮

1. ブロックソート法 (bzip2プログラム) (Burrows-Wheeler)
2. 文脈ソート法 (横尾、高橋)
3. 文脈の条件つきで辞書を分割したLZ法 (Protnik-Weinberger-Ziv, 大峰-山本)
4. ACB法 (acbプログラム) (Buyanovski)
5. LZ78法に対する文脈ソート法 (有村-山本)

3

2002年11月11日



情報源符号化ワークショップ

2. ブロックソート法 (bzip2)

- a. Burrows-Wheeler変換
- b. Move-To-Front+Run-Length変換
- c. Huffman符号化

4

2002年11月11日



情報源符号化ワークショップ

2a-1. Burrows-Wheeler変換の概要

1. 長さnのブロックXが与えられたとする
2. 巡回シフト
3. 左から順に各行を見ながら辞書順にソート
4. 各行の最後(一番右)のシンボルを上から順に取り出した物をYとする

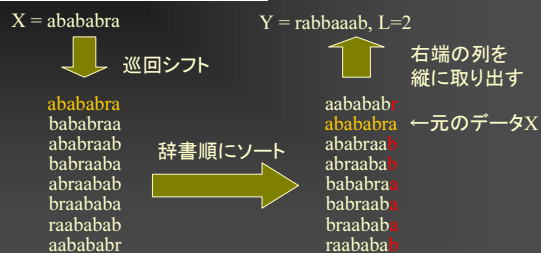
5

2002年11月11日



情報源符号化ワークショップ

2a-2. Burrows-Wheeler変換の例



6

2002年11月11日



情報源符号化ワークショップ

2a-3. Burrows-Wheeler変換の性質

X = abababra

巡回シフト

abababra
bababraa
ababraab
babraaba
abraabab
braababa
raababab
aabababr

- BW変換の出力Yは、入力系列Xの置換(permutation)
- 各シンボルは、その直後のシンボルによってソート

aababab
abababra
ababraab
abraabab
bababraa
babraaba
braababa
raababab
raababab

←元のデータX

7

2002年11月11日

情報源符号化ワークショップ

2a-4. bzip2におけるBurrows-Wheeler変換

1. ファイルからブロックを切り出す
command line option: -1... -9
-1: 100Kbytes単位
-9: 900Kbytes単位
 - デフォルトは900Kbytes単位
 - このサイズ未満のファイルの場合には、ファイル全体をひとつのブロックとして扱う
2. それぞれのブロックに対してBW変換を行なう

8

2002年11月11日

情報源符号化ワークショップ

2b-1. Move-To-Front (MTF) 法

(Book-Stack法、Move-To-Front法、Recency-Rank法、超整理法)

- 出現し得るシンボルの集合をリストにしたものが与えられているとする ({a, b, c, d})
 - 1. シンボルを順に見ながら、リストの何番目にあるかを出力する (b ⇒ 2番目なので0-originだと1)
 - 2. 見つかったシンボルをリストの先頭に移動する (b を先頭に移動 ⇒ {b, a, c, d})
 - 3. 次のシンボルを見る
- 以上の操作を繰り返す

9

2002年11月11日

情報源符号化ワークショップ

2b-2. bzip2におけるMTF+RL変換アルゴリズム

- MTF変換の例(アルファベット{a, b, c})
 - S[0] = {a, b, c}
 - X[1] = b ⇒ Y[1] = 1, S[1]={b, a, c}
 - X[2] = b ⇒ Y[2] = 0, S[2]={b, a, c}
 - X[3] = b ⇒ Y[3] = 0, S[3]={b, a, c}
- Y[i] = 0 の場合は連続して出現した個数を2進表示したものを出力 ⇒ 0,1の列
- Y[i] ≠ 0 は1足したものを出力 ⇒ 2以上の値の列

10

2002年11月11日

情報源符号化ワークショップ

2b-3. bzip2におけるMTF+RL変換の実際

```
void generateMTFValues()
{
    配列 mtfv[] : 実際に MTF+RL変換で出力される値の列
    配列 mtfFreq[] : mtfv[]の中の各シンボルの出現回数を記憶するテーブル
    配列 yy[] が MTF list となっている

    for ( i = 0; i <= last; i++) {
        データを見て MTF list yy[] を更新する
        rank が 0 であったら、0のRun Lengthを増やして次のシンボルへ
        rank が 0 以外であったら、
        ・それまで0のRunが続いていたなら、Run Lengthを2進表示した際の
        0, 1を少し加工して、RUNA, RUNB(実際の値は0と1)というシンボル
        としてmtfv[]に登録する
        ・"今回のシンボルの文字コード+1"を mtfv[] に登録する
        ・mtfFreq[]を更新
    }
}
```

11

2002年11月11日

情報源符号化ワークショップ

2c-1. Huffman符号化

- アルファベット{a,b,c,d,e}の場合の例

- 各シンボルの相対頻度が以下の通り
a : 0.4, b : 0.2, c : 0.1, d : 0.1, e : 0.2
- このとき、最も小さい値を2個取り出してまとめる
a : 0.4, b : 0.2, (c,d) : 0.2, e : 0.2
- これを繰り返す
a : 0.4, b : 0.2, ((c,d), e) 0.4
(a,b) : 0.6, ((c,d), e) : 0.4
((a,b), ((c,d), e)) : 1.0



12

2002年11月11日

情報源符号化ワークショップ

2c-2. bzip2におけるHuffman符号

- 0~257のシンボルを、大体同じ頻度になるようにグループ分けする

データサイズが ~199	2 groups
201~600	3 groups
601~1200	4 groups
1201~2400	5 groups
2400~	6 groups

- 各グループに重みを付けたHuffman木をそれぞれ作成する
- 50シンボル単位のブロックで、どのHuffman木を用いるとコストが小さくなるかを計算する
- 各ブロックに対してどのHuffman木を用いるかを示す列 selector[]と、実際にHuffman木で作成した符号を送る

13

2002年11月11日

情報源符号化ワークショップ

2d-1. bzip2によってなぜ良く圧縮できる?

- 実は完璧に理論化されている訳ではない
- BW変換については、文脈ソートとの関連で解析(横尾)
- シンボル拡大によってマルコフ情報源のエントロピーレートを達成(有村、山本)
その後、Effrosや村松によってエントロピーレートを達成できる情報源クラスが広がられている
- MTF変換+算術符号化の性能
 - 有村、山本によって解析
 - マルコフ情報源のエントロピーレートを達成できるような情報源の条件が分かった

14

2002年11月11日

情報源符号化ワークショップ

2d-2. MTF法によって良く圧縮できる情報源

- 有村、山本による結果
- マルコフ情報源において、遷移確率が

$$\begin{pmatrix} 1-(A-1)p & p & \cdots & \cdots & p \\ p & 1-(A-1)p & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & p \\ p & \cdots & \cdots & p & 1-(A-1)p \end{pmatrix}$$

のような場合に、情報源のエントロピーレートを達成(=理論的な限界まで圧縮可能)

- ここでpの値が小さい時には、同じシンボルが連続して出現し、またエントロピーは小さい値になる
⇒同じシンボルが連続して出現する場合、MTF法で良く圧縮できる

15

2002年11月11日



情報源符号化ワークショップ

2d-3. BW変換+MTF変換で良く圧縮できるデータ

データ: abababra

abababra
bababraa
ababraab
babraaba
abraabab
braababa
raababab
aabababr

ソート

aabababr
abababra
ababraab
abraabab
babraaba
braababa
raababab
aabababr

シンボル'a'の直前に
出たシンボル
シンボル'b'の直前に
出たシンボル

BW変換の出力をMTF法で良く圧縮できる

⇒ ソートによって、類似したシンボルが連続して出やすくなっている筈である

16

2002年11月11日



情報源符号化ワークショップ

2d-4. ブロックソート法でなぜ良く圧縮できる?

■ BW変換:

- データの各シンボルを並べ替え
- 同じシンボルが連続して出やすいデータに変換
- ソートによって、うまくデータの文脈依存性を取り出している

■ MTF法:

- 同じシンボルが連続して出現しやすいデータを効率良く符号化

17

2002年11月11日



情報源符号化ワークショップ

3. 文脈ソート法

直後のシンボル列が同じようなシンボルがまとめられて、連続してMTF法に入力

適応型アルゴリズムに書き換え、文脈をソートしてシンボルを整数に変換

文脈ソート法(高橋-横尾)

18

2002年11月11日



情報源符号化ワークショップ

3a. 文脈ソート法のアルゴリズム

...abbabacdaceabc g
...abbabacdaceabc g a
...abbabacdaceabcga s
...abbabacdaceabcgas g
...abbabacdaceabcgasg a

ソート

5 ...abbabacdaceabcga s
3 ...abbabacdaceabc g
1 ...abbabacdaceabc a
...abbabacdaceabcgas a
2 ...abbabacdaceabcgas g

過去に出現した
シンボルとその文脈

←これから処理する
シンボルと文脈

文脈でソートして番号付け
⇒一致するシンボルは'1'
⇒小さい値が出やすいので
良く圧縮できる

19

2002年11月11日



情報源符号化ワークショップ

4. 文脈によって辞書を分割するLempel-Ziv法

- Protnik-Weinberger-Zivによって理論的な評価
- 大峰-山本によって実験的な評価
- 山崎による実装あり(C Magazine 2002年11月号)

◆オリジナルのLempel-Ziv法:

辞書に登録する単語は、過去のデータから作成されたもの

◆文脈によって辞書を分割するLempel-Ziv法:

辞書を複数作成し、登録する単語はそれぞれ直前のシンボル列(文脈)によって別々の辞書に登録

20

2002年11月11日



情報源符号化ワークショップ

4a. 文脈によって辞書を分割するとどうなる?

■ LZ78法の場合を考える

データからブロックを切り出して、次の3つの値をブロックに符号として割り当てるアルゴリズム

1. 最大一致ブロック
2. 一致したブロックの次の一致しないシンボル

■ 辞書を分割

⇒辞書の指定は直前の文脈から分かる
⇒細かく分割された辞書の中で最大一致ブロックを指定すれば良い
⇒この部分に関して符号化を効率化

21

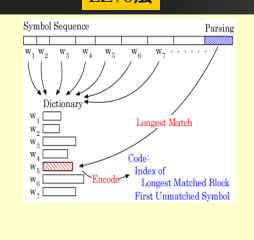
2002年11月11日



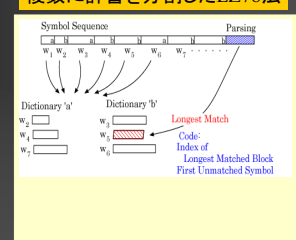
情報源符号化ワークショップ

4b. 直前の1シンボルで辞書を分割した時の符号化

LZ78法



複数の辞書を分割したLZ78法



22

2002年11月11日



情報源符号化ワークショップ

5. ACB法

■ LZ77法の拡張

LZ77法は、データからブロックを切り出して、次の3つの値をブロックに符号として割り当てるアルゴリズム

1. 最大一致ブロックの位置
2. 最大一致長
3. 一致したブロックの次の一致しないシンボル

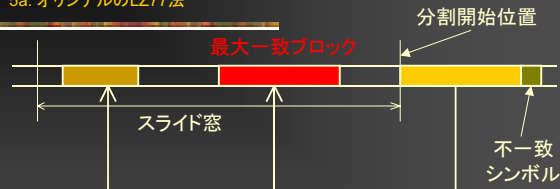
ACB法では、上の1.~3.のそれぞれに対して工夫をしている。

23

2002年11月11日

情報源符号化ワークショップ

5a. オリジナルのLZ77法



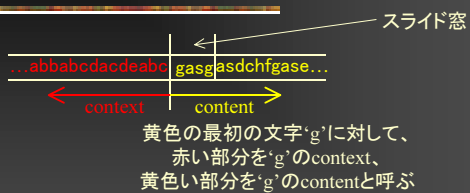
- 1) 分割開始位置から先のブロックと最も長く一致するブロックをスライド窓の中から選ぶ。
- 2) 一致長だけブロックを切り出し、(開始位置, 長さ, 一致していない次のシンボル)をそのブロックに対する符号とする。

24

2002年11月11日

情報源符号化ワークショップ

5b-1. ACB法での最大一致ブロックの開始位置の指定方法



スライド窓の中の各シンボルに対して、そのcontextとcontentのペアを取り出し、contextでソートする

25

2002年11月11日

情報源符号化ワークショップ

5b-2. ACB法での最大一致ブロックの開始位置の指定方法



- ソート
- 5 ...abbabdacdeabcbgsgasdchfgase... contextでソートして番号付け
3 ...abbabdacdeabcbgsgasdchfgase...
1 ...abbabdacdeabcbgsgasdchfgase... →最大一致は'-1'
2 ...abbabdacdeabcbgsgasdchfgase...

26

2002年11月11日

情報源符号化ワークショップ

5c. ACB法での一致しないシンボルの指定方法

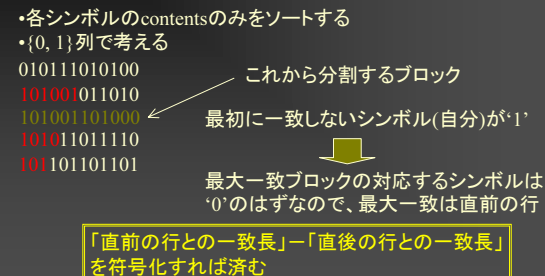
- ACB法では、2値系列に対して処理
- '0'もしくは'1'の1bitで指定可能

27

2002年11月11日

情報源符号化ワークショップ

5d. 最大一致長の指定方法



28

2002年11月11日

情報源符号化ワークショップ

5e. ACB法によってなぜ良く圧縮できる?

■ LZ77法と比較して

- 最大一致ブロックの開始位置
- 最大一致長
- 一致していないシンボル

それぞれの符号を効率化

- ソートによって、文脈およびブロックの類似性をうまく取り出して符号化に反映させている

29

2002年11月11日

情報源符号化ワークショップ

6. 文脈ソートをLZ78法に適用した場合

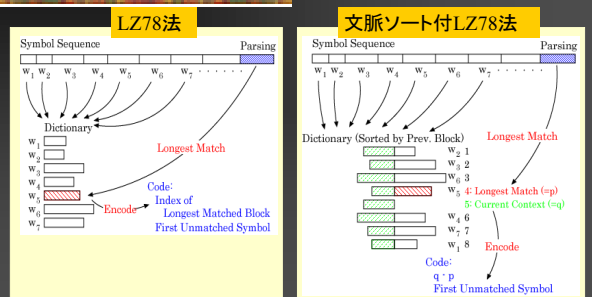
- 有村-山本
- ACB法では、ソートするための文脈の長さは $\log n$ に固定している
- ソートするのだから、同じ文脈が複数無い方が楽
- LZ78では、分割されたブロックは全て異なる
⇒一つ前に分割されたブロックでソート
(木を用いたデータ構造を使用可能)

30

2002年11月11日

情報源符号化ワークショップ

6a. 直前の分割ブロックでソートするLZ78法



31

2002年11月11日

情報源符号化ワークショップ

6b. 文脈ソート+LZ法のバリエーション

- 基本的にはLempel-Ziv法の全てのバリエーションに文脈ソート法を組み合わせ可能
- LZ法のバリエーションで長さの符号化に関して最も頑張っているのはFiala-Greene (ACB法よりも良い事は簡単に示せる)
- では、文脈ソート+Fiala-Greeneがベスト？ (誰か実験、実装してください)

32

2002年11月11日

情報源符号化ワークショップ

7a. 実際の圧縮性能

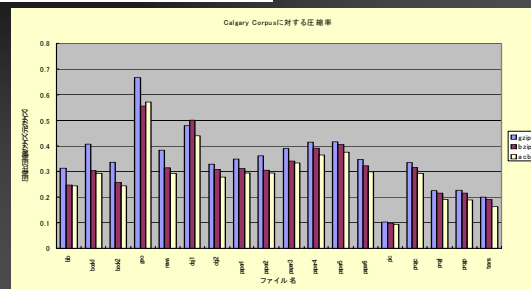
- 以下のツールで圧縮率を比較
 - gzip (Lempel-Ziv 77法)
 - bzip2 (ブロックソート法)
 - ACB (Lempel-Ziv 77法の改良)
- 圧縮率は各ファイルに対して (圧縮後のサイズ)/(元のサイズ) を表示 (全く圧縮できない状態が1.0)

33

2002年11月11日

情報源符号化ワークショップ

7b. Calgary Corpus に対する圧縮率

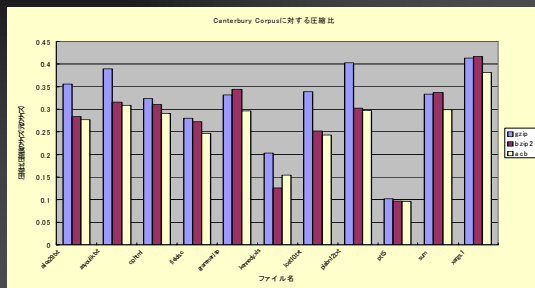


34

2002年11月11日

情報源符号化ワークショップ

7c. Canterbury Corpus に対する圧縮率

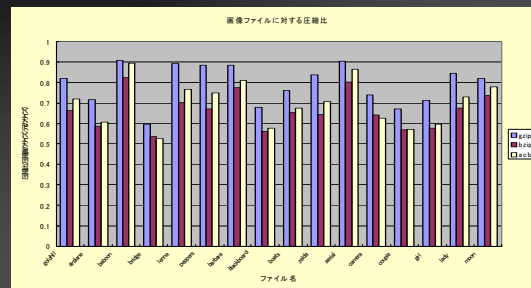


35

2002年11月11日

情報源符号化ワークショップ

7d. 画像データに対する圧縮率



36

2002年11月11日

情報源符号化ワークショップ

7e. 実験に対する考察

- ACB法はLZ77法の改良版なので、gzipに比べて圧縮率は良くなる
- ACBとgzipは、もしデータが定常であれば、Window Sizeを大きくした時に情報源のエントロピーレートまで圧縮できるはずである
- ところが、bzip2は理論的には、かなり限られた情報源でしか情報源のエントロピーレートを達成できない筈
- ということは、
 - データのサイズが、LZ法で十分にエントロピーに近く圧縮できるほど大きくない？
 - 実は、サンプルファイルが、ブロックソート法でエントロピーレートまで圧縮できるような特別なファイル？
 - そもそも定常でない場合には、圧縮限界はどうなる？

37

2002年11月11日

情報源符号化ワークショップ

8. まとめ(ソートを用いた圧縮の今後)

- 基本的に、現在のソートの使用法は
 - A) 類似した文脈を集める
⇒ 類似性で偏りを作成して符号化を効率化
 - B) 類似したブロックを集める
⇒ 類似性によってブロック長の符号化を効率化
- 他のソートの用い方は無いのか？
- 他の離散数学的ツールを使えないか？
置換、グラフ、木構造、素数、etc...

38

2002年11月11日

情報源符号化ワークショップ

音声と楽音の情報圧縮

守谷 健弘 (NTT サイバースペース研究所・東京大学大学院情報理工学系研究科客員教授)

講演内容

音声と楽音の情報圧縮

NTTサイバースペース研究所
東京大学大学院情報理工学系研究科客員教授

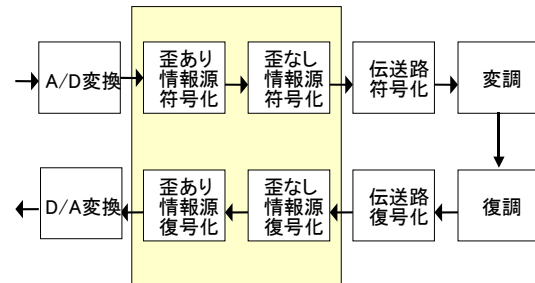
守谷 健弘

1. 符号化の基本技術と代表的方式
 - 予測、量子化、変換、聴覚特性
2. 音声符号化の標準化動向
 - ITU-T, IMT-2000, PDC
3. 楽音符号化の標準化動向
 - MPEG-1, 2, 4, 7, MP3, AAC, スケーラブル
4. 開発事例
 - TwinVQ と ロスレス符号化
5. 応用とデモ

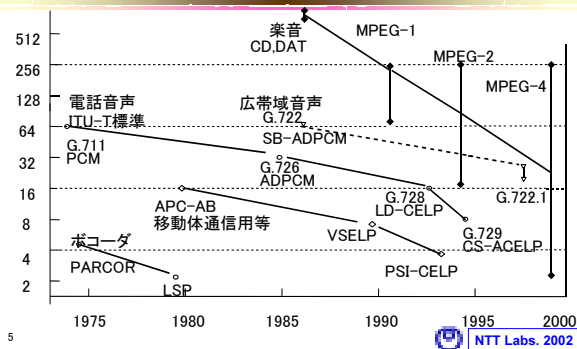
1. 符号化の基本技術と代表的方式

- 量子化歪削減
 - 線形予測とパラメータ
 - 予測と変換
 - スカラ量子化とベクトル量子化
- モデルの利用
- 聴覚特性の利用

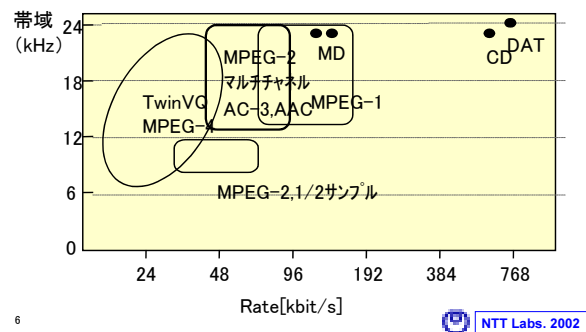
1. 1 情報圧縮の枠組み



1. 2 情報圧縮の歴史



1. 3 音楽の符号化のビットレートと帯域



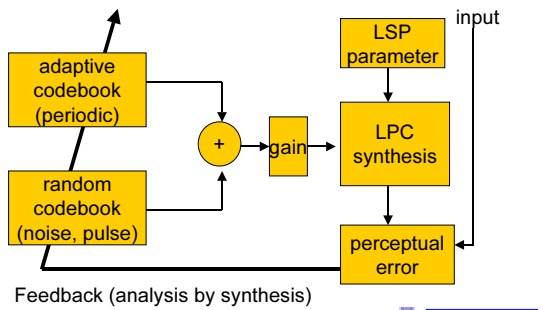
1. 4 歪の削減

- 予測符号化と変換符号化
 - 予測パラメータ
 - DCTとMDCT
 - 適応ビット配分
 - 変換利得と予測利得
- ベクトル量子化
 - 共役ベクトル量子化
 - 励振ベクトルとCELP
- 適応化

7

NTT Labs. 2002

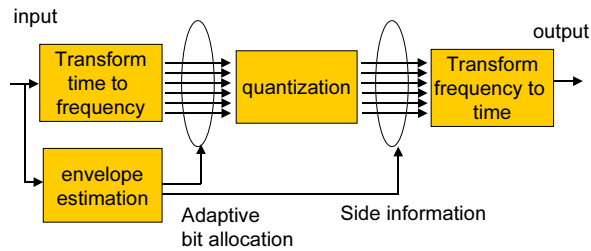
1. 5 音声符号化 CELP



8

NTT Labs. 2002

1. 6 変換符号化



9

NTT Labs. 2002

1. 7 予測符号化と変換符号化

	時間領域(予測)	周波数領域(変換)
相関小さい	予測不可	スペクトル平坦
効果	予測利得	変換利得
利得	波形エネルギー 予測誤差エネルギー	スペクトル相乗平均 スペクトル相乗平均
手段	閉ループ量子化	適応割り当て 適応重み
相関大きい	予測可能	スペクトル偏り

10

NTT Labs. 2002

1. 8 音楽符号化の要素技術

	変換	量子化
MPEG-1L1,L2	帯域分割	適応配分
MPEG-1 L3 (MP3)	帯域分割+MDCT	適応可変長
ATRAC	帯域分割+MDCT	適応配分
AC-3	MDCT	適応可変長
AAC	MDCT	適応可変長
TwinVQ	MDCT	重み付ベクトル

11

NTT Labs. 2002

1. 9 時間領域の適応化

- 無音と有音
- 無声音と有声音
- 音声と非音声 →MPEG-4
- 分析長または変換窓長 →MPEG-1,2,4

12

NTT Labs. 2002

1. 10 周波数領域の適応化

- 無声音と有声音 →IMBE, MELP
- MSステレオ →MPEG-1,2,4
- PNS (Perceptual Noise Substitution) →MPEG-4
- LTP (Long Term Prediction) →MPEG-4
- FSS (Frequency Selective Switch) →MPEG-4

13

NTT Labs. 2002

2. 音声符号化の標準化動向

- ITU-T G.シリーズ
- IMT-2000
- 日本国内携帯電話
 - PDC full/half rate
 - cdmaOne
 - PDC enhanced full rate
- 世界の携帯電話
 - GSM
 - TIA

14

NTT Labs. 2002

2. 1 ITU-T標準化の動向 (G.729)

- CS-ACELP
(Conjugate Structure Algebraic CELP)
- G.729 8 kbit/s 8 kHz 10 msフレーム
- G.729A 低演算量 (10 Mips)
- G.729B Voice Activity Detection
- G.729C floating point
- G.729D 6.4 kbit/s
- G.729E 11.8 kbit/s

15



2. 2 ITU-T標準化の動向

- 広帯域音声 16kHzサンプル
 - 24 kbit/s, 32 kbit/s
Picture Tel方式 → G.722.1
 - 16 kbit/s AMR標準と共通化
- 4 kbit/s 募集
 - 15機関の応募
 - 3チームによる競争⇒2チームの競争
 - 最終案決裂
 - Connexant, NTT, 松下, AT&T, FT, DT案

16



2. 3 IMT-2000標準化の動向

- 3GPP (3rd Generation Partnership Project)
(ARIB, TTC, T1, ETSI, TTA)
- W-CDMA
- 双方向電話音声CODEC
AMR (Advanced Multi Rate)
- video phone (H.263)
- オーディオ、低ビット音声

17



2. 4 AMRの概要

- 8kHz サンプル 20msフレーム ACELP
- 可変レート符号化
12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15, 4.75 kbit/s, +
無音情報
- 伝送路に適応させた連続制御が可能
- GSM (EFR:enhanced full rate)互換
- GSMとしてサービス開始1999年
- Wideband AMR 2001年
- NTT DoCoMo Forma 2001年

18



2. 5 PDCの標準

- Full rate TDMA 1990–
VSELP (6.7 + 4.5[訂正符号] kbit/s)
- Half rate TDMA 1993–
PSI-CELP (3.45 + 2.15 kbit/s)
- cdmaOne CDMA 1997–
EVRC (8/4/0.8 kbit/s IS-127互換)
- Enhanced full rate TDMA 1998–
G.729 (8.0 kbit/s) + 3.2 kbit/s
ACELP (6.7 kbit/s AMR互換) + 4.5 kbit/s

19



3. 楽音符号化の標準化動向

- MPEG-1 1992
- MPEG-2 multi-channel/half sampling 1994
AAC 1997
- MPEG-4 Version 1 1999
Version 2 2000
- MPEG-7 2001
- <http://www.cselt.it/mpeg/documents>
- <http://www.tnt.uni-hannover.de/mpeg/audio>
- AC-3 (Dolby), ATRAC (Sony)

20



3. 1 MPEG-1/audio

- Sampling rate: 32, 44.1, 48 kHz
- Channel: mono/stereo
- Bit rate: 32 – 448 kbit/s (Layer-I)
32 – 384 kbit/s (Layer-II)
32 – 320 kbit/s (Layer-III)
- Algorithm:
Layer-I 32帯域分割適応割当
Layer-II +量子化改善
Layer-III +MDCT+可変長符号+ビット貯蔵, ++

21



3. 2 MPEG-2/multi channel

- Sampling rate: 32, 44.1, 48 kHz
- Channel: mono/stereo/5+1
(正面、左右(2)、斜め後ろ(2)、低周波)
- Bit rate: MPEG-1+拡張
- Algorithm:
Layer-I, -II, -III と後方互換
古いMPEG-1デコーダでもステレオとして
再生可能

22



3. 3 MPEG-2/half sampling rate

- Sampling rate: 16, 22.05, 24 kHz
- Channel: mono/stereo
- Bit rate: 32 – 256 kbit/s (Layer I)
8 – 160 kbit/s (Layer II,III)
- Algorithm:
Layer-I, -II, -III と後方互換
古いMPEG-1デコーダでも再生可能

23



3. 4 MPEG-2/Layer III(MP3)

- MDCT
 - 高周波数分解能
 - エリアジング補正
 - バーク尺度に比例する帯域分割
- 可変窓長 (192 または 576点)
- 非線型ステップ+可変長符号
- ビット貯蔵
- ハイブリッドステレオ (MS+インテンシティ)

24



3. 5 MPEG-2/AAC

- プロファイル
 - メイン
 - LC (Low Complexity) ★
 - SSR (Scalable Sampling Rate)
- sampling rate: 32, 44.1, 48 kHz, +2倍、1/2, 1/4
- channel数: 1-48
- ビットレート: 8-576 kbit/s/ch
- これまでのMPEG-1,2とは互換性なし
- ARIB規格 BSデジタル放送用

25



3. 6 MPEG-2/AAC (メイン, LC)

- MDCT可変窓長 (128 または 1024点)
- バーク尺度に比例する帯域分割
- バーク尺度毎のMS+インテンシティ ステレオ
- 非線型ステップ+可変長符号
(2または4次元ハフマン符号)
- ビット貯蔵
- TNS(Time domain Noise Shaping)
- 後方フレーム間予測 (メインプロファイルのみ)

26



3. 7 MPEG-2/AAC (SSR)

- 帯域分割+MDCT可変窓長 (32 または 256点)
- 量子化基本構成はLCプロファイル と共通
- +利得正規化
- +低域のみのデコード可能
 - 帯域スケラブル
 - デコーダ処理量スケラブル

27



3. 8 MPEG-4オーディオの概要

- MPEG-4オーディオ (Version 1) は1999年にISO/IECの国際標準として成立
- 音声、音楽の圧縮符号化だけでなく、合成音楽や合成音声の規格を含む
- MPEG-2 AACを含む
- 広範囲のサンプリング周期、ビットレートに対応
- ビットレートスケラビリティが実現されている

28



3. 9 MPEG-4 (圧縮符号化)

- 低ビット音声 HVXC
(Harmonic Vector eXcitation Coder)
- 音声(電話、広帯域) CELP
(Code Excited Linear prediction)
- 低ビット楽音 TwinVQ
(Transform domain Weighted Interleave VQ)
- 楽音 AAC (MPEG-2の拡張)
(Advanced Audio Coder)

29



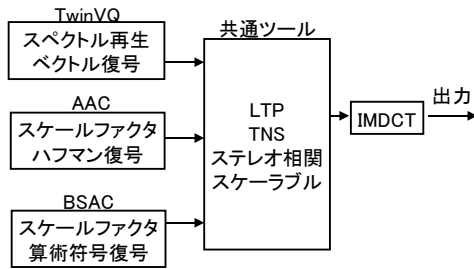
3. 10 MPEG-4 AAC

- LTP (Long Term Predictor)の追加
- PNS (Perceptual Noise Substitution)の追加
- FSS (Frequency Selective Switch)によるスケラブル符号化
 - AAC+AAC...
 - .CELP+AAC...
 - TVQ+AAC....

30



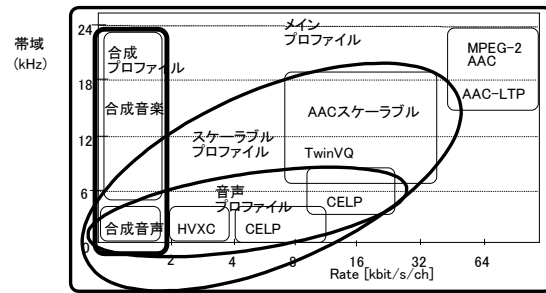
3.11 楽音符号化



31

NTT Labs. 2002

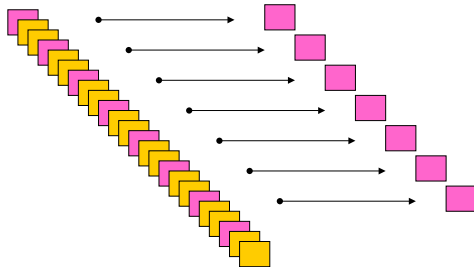
3.12 MPEG-4オーディオのプロファイル



32

NTT Labs. 2002

3.13 MPEG-4のスケラビリティ



33

NTT Labs. 2002

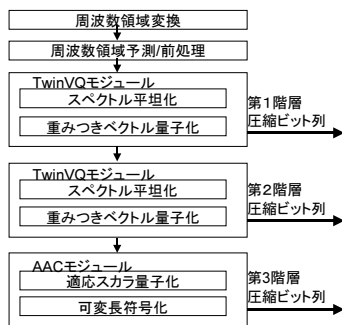
3.14 MPEG-4スケラビリティの利用

	低層	高層
情報量	低ビット	高ビット
操作性	容易	不要
伝送形態	1 対 N	1 対 1
誤り対策	不均一訂正	再送
課金	無料	有料

34

NTT Labs. 2002

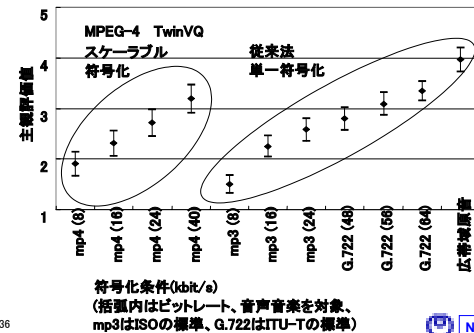
3.15 スケラブル楽音符号化の構成例



35

NTT Labs. 2002

3.16 スケラブル楽音符号化の品質



36

NTT Labs. 2002

3.17 MPEG-4 Version 2

- 新機能の追加
 - BSAC(Bit Slice Arithmetic Coding)
 - HILN (Harmonic Individual Line + Noise)
 - AAC Low-delay
 - 無音圧縮 (CELP, HVXC)
 - 誤り耐性
 - 3D音場 (両耳効果、環境特性)
- トランスポートストリームの定義
- 新たなプロファイル

37

NTT Labs. 2002

3.18 MPEG-7

- データベース検索支援formatのみの標準
 - 内容から特徴パラメータの抽出法
 - 特徴パラメータを利用する高速検索は関連する技術であるが標準化対象外
- DLL (Description Definition Language)
 - 記述言語 (XML, LISPなど)
- DS (Description Scheme)
 - 記述のモデル、枠組み
- D (Descriptor)
 - 特徴パラメータ、作曲者、曲名など

38

NTT Labs. 2002

3. 19 MPEG-21の課題

- コンテンツを集散的に扱う統一的で柔軟な記述
- コンテンツの識別と記述のためのフレームワーク
- コンテンツの著作権を管理保護する方法
- バリューチェーンにわたりコンテンツを統一的に扱うインターフェースとプロトコル
- ネットワークや端末を通じ、相互接続を可能とするコンテンツへのアクセス方法
- あらゆるメディアリソースの統一的で柔軟な表現
- 報告可能イベントの性能をユーザが知るための標準尺度とインターフェース

39



4 開発事例

- TwinVQ
 - 低ビット楽音符号化
 - モバイルマルチメディア
- ロスレス符号化 MPEGで標準化技術公募
 - ブロードバンド化
 - 高品質化

40



4. 1 TwinVQの特徴 —その1

- 楽音符号化と音声符号化の利点を融合
- ミニディスクやMP3(MPEG-1 layer III)より圧縮率が高く、低ビットでも品質の劣化が少ない
- チャンネルあたり8 kbit/s 以上で動作し、原音に対する圧縮率は1/10から1/100
- 復号器の処理は軽い
- 符号器の処理は調整可能で、並列パイプライン化が可能

41



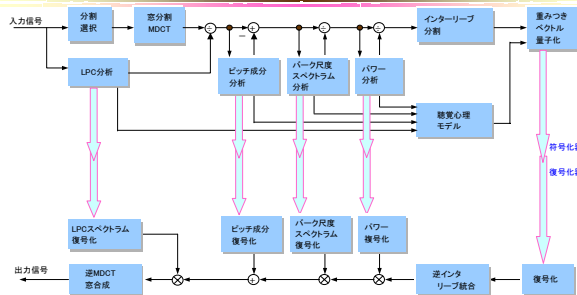
4. 2 TwinVQの特徴 —その2

- ビット列編集、ランダムアクセス、飛び越し、逆転再生が容易
 - ビット配分固定、固定フレーム
- 符号誤りやパケット消失に強い
 - 可変長符号を使わない
 - 誤り耐性のある共役ベクトル量子化
 - フレーム間予測は1フレーム移動平均のみ
- 不均一誤り訂正による効果的保護が可能
 - 補助情報の比率が小さい

42



4. 3 TwinVQの基本処理



43



4. 4 ビット列の構成の比較



44



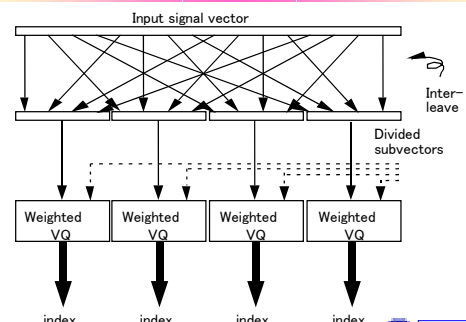
4. 5 TwinVQの主要要素技術

- MDCT係数の適応的ベクトル量子化
 - 重み付き尺度によるパターン照合
 - 時空間インターリーブ
 - 共役構造による処理量メモリ量削減
- 線形予測分析とLSP (Line Spectrum Pair) パラメータの予測ベクトル量子化
- バーク尺度スペクトル推定
- 周期的ピーク成分量子化

45



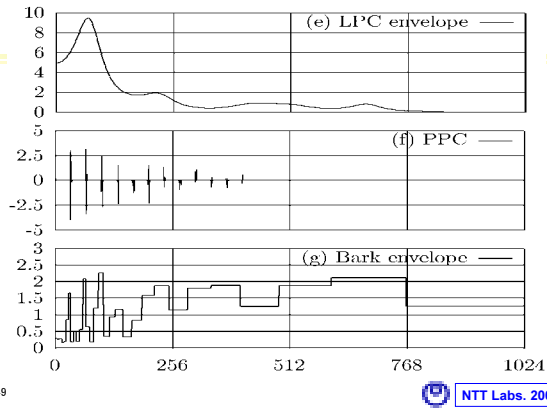
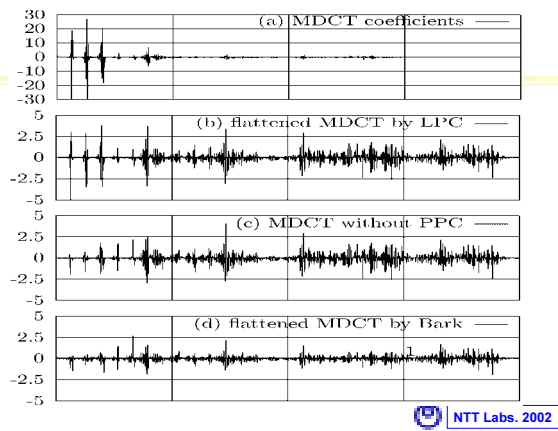
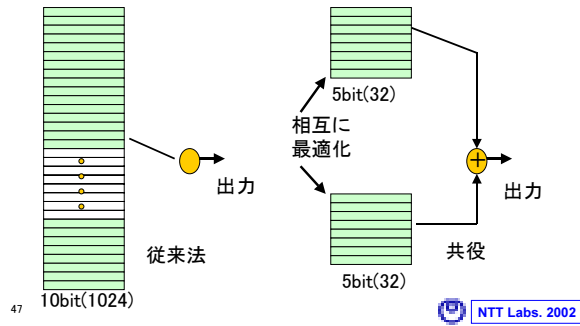
4. 6 TwinVQの基本量子化



46



4.7 共役ベクトル量子化



4.8 ロスレス符号化の要求条件

- ロスレス符号化 (圧縮効率)
- スケーラビリティ
 - ビットレート
 - サンプリングレート
 - Fine granularity (near lossless)
- 後方互換性
- ランダムアクセス
- 演算量
- MPEG-4との共通性

50

NTT Labs. 2002

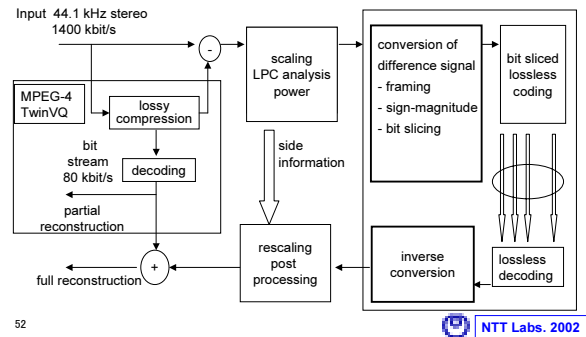
4.9 研究的興味

- ロッシー・ロスレスによるハンディキャップ
- スケーラブル化によるハンディキャップ
- 補助情報による改善
- ニアロスレスの品質
- IPの優先制御との組み合わせ
- 暗号やすかしとの組み合わせ

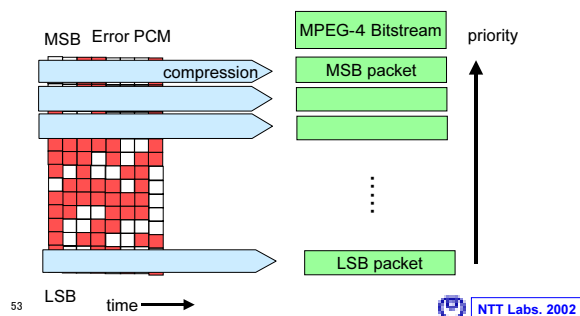
51

NTT Labs. 2002

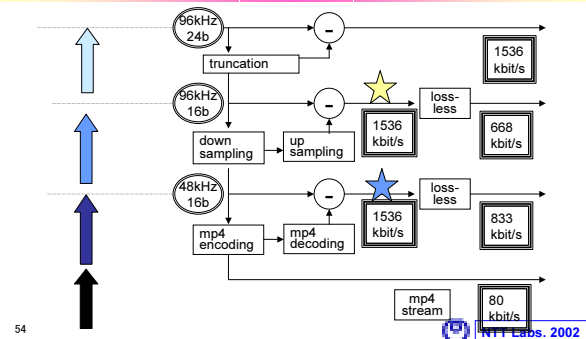
4.10 ロッシー・ロスレス符号化



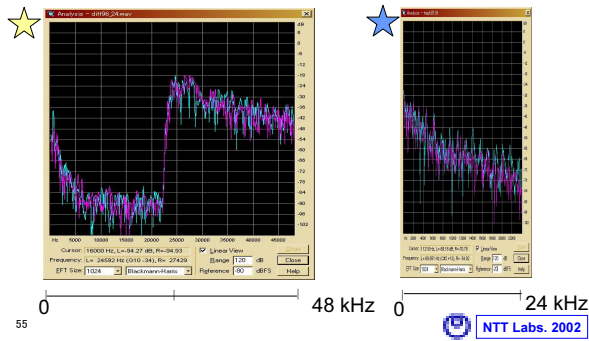
4.11 水平ビットスライス



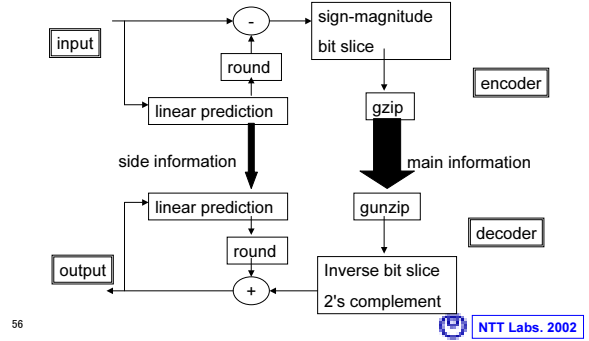
4.12 サンプルレートスケーラブル



4. 13 誤差信号のスペクトル



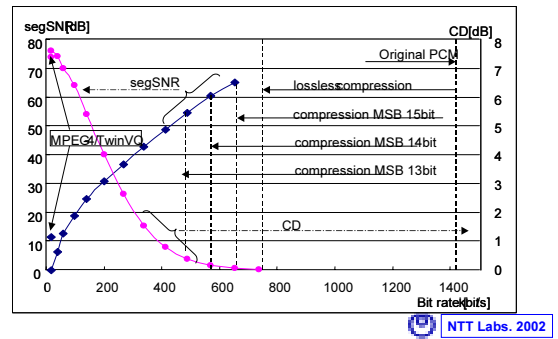
4. 14 基本ロスレス符号化の例



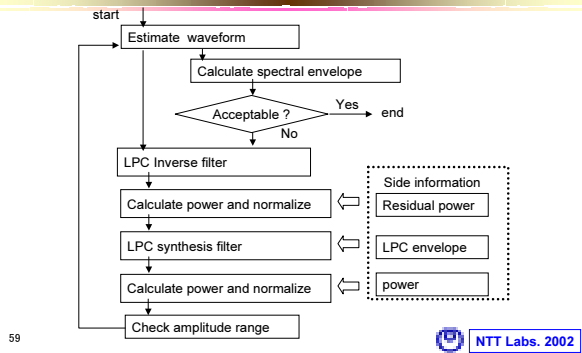
4. 15 圧縮率比較

	scalable data rate (kbit/s)	accumulated data rate (kbit/s)	stand-alone compression (kbit/s)	original data rate (kbit/s)
96 kHz 24 bit	1536	3117	2868	4560
96 kHz 16 bit	668	1581	1420	3072
48 kHz 16 bit	833	913	881	1536
lossy coding	80	80	80	80
total	3117	3117	5177	9248

4. 16 ニアロスレスの性能



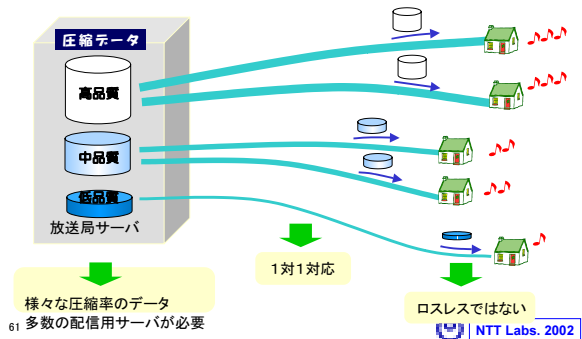
4. 17 ニアロスレスの後処理例



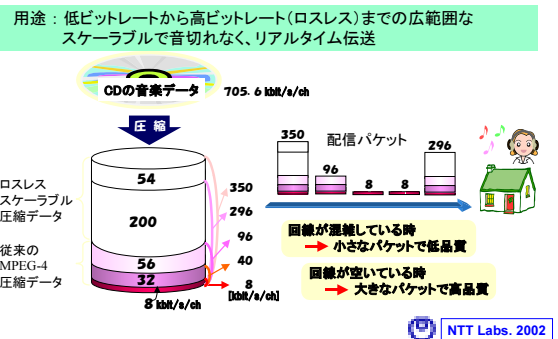
5 応用と課題

- ストリーミング
- 著作権保護と普及のバランス

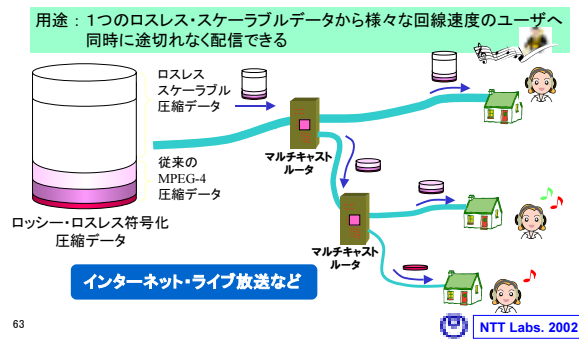
5. 1 現状のストリーミング



5. 2 想定用途例(動的レート制御)



5.3 想定用途例(マルチキャスト)



63

5.4 電子すかし

- 画像、映像は実用段階
- 音楽の市場は大きくコピーの問題も大きい
- 音楽のすかしの問題点
 - 音楽は劣化に敏感
 - 音楽データにフレームの概念がない
 - 圧縮符号化、帯域制限、位相処理
- 複数回のコピーを禁止するフラグの埋め込み

64

NTT Labs. 2002

5.5 圧縮符号と電子すかし

- 圧縮符号ビット列に埋め込む
→ 再生信号には残らない
- 圧縮符号の符号化過程で埋め込む
再生信号にもすかしが残るように
→ 圧縮符号化と一体化
- 圧縮符号で使わない聴覚特性利用

65

NTT Labs. 2002

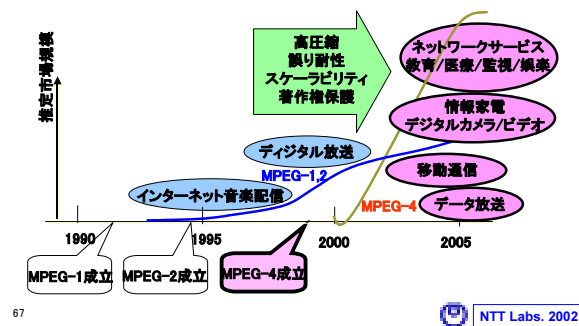
5.6 楽音符号化の今後の課題

- MPEG-4 の枠組みでの開発
 - スケーラブル符号化
 - マルチメディア用途の開拓
- 著作権保護システム
 - 電子すかし
 - シリアル番号管理
- データベース探索 MPEG-7
- 放送、携帯電話、LAN、家電製品の共通規格

66

NTT Labs. 2002

5.7 応用分野と市場規模



67

5.8 音楽デモ

- 48 kHz ステレオ 原音
- MPEG-4スケーラブル 8+16+16 kbit/s/ch
- MPEG-4 スケーラブル 8+16 kbit/s/ch
- MPEG-4 スケーラブル 8 kbit/s/ch

68

NTT Labs. 2002

5.9 音声デモ

- ITU-T G.711 64 kbit/s
- ITU-T G.726 32 kbit/s
- ITU-T G.728 16 kbit/s
- ITU-T G.729 8 kbit/s
- PDC Full 6.7 kbit/s
- PDC Half 3.45 kbit/s
- MPEG4 HVXC 2 kbit/s
- MPEG4 TwinVQ 8kbit/s

69

NTT Labs. 2002

5.10 誤り耐性デモ

- 原音 32 kHz サンプル, モノラル
- スケーラブル 16 kbit/s
誤りなし
- スケーラブル 16 kbit/s
0.1% バースト誤り
- スケーラブル 16+16 kbit/s
誤りなし
- スケーラブル 16+16 kbit/s
0.1%バースト / 0.5%バースト

70

NTT Labs. 2002

- 実時間encoder+transport+decoder
- 原音: 44.1 kHz ステレオ 16bit
- コア: MPEG-4 TwinVQ
- 16 kbit/s - 1 Mbit/s (ロスレス)
- Encoder側からの制御
- Decoder側、ネットワークからの制御

71



量子情報源符号化

松本 啓史 (科学技術振興事業団 今井量子計算機構プロジェクト)

Abstract

量子情報源圧縮の全般的な概説ののち、特に量子 universal source code の研究を紹介する。量子力学の観測は対象の破壊を伴い、全般的な傾向としては、明瞭な情報を引き出そうとすればするほど、破壊の程度も激しくなる。一方、最適な universal code は情報源のエントロピーを明らかにしてしまうので、このような操作は著しく情報源を破壊してしまい、復号可能な圧縮にはならなさそうに一見、思われる。この見かけの背景を如何に回避するかが本講演の中心的な話題である。

1 はじめに

量子力学が、日常的直観に反する帰結を生み出すことは、通俗科学書などを通じて広く知られていると思う。通常、情報科学は古典力学的な世界観に立脚しているが、量子力学の上に立脚した情報科学を量子情報科学とよぶ。特に、観測や系の状態といった概念は、量子力学では根本的な改変をうける。

量子情報科学は 1960 年代の創始以来、細々と研究が続けられてきたが、最近、急激に幅広い関心を集めるようになった、その第一の理由は、古典的 (以後、非量子的の意味で、「古典的」という言葉を用いる) な情報処理では不可能であったことが、量子情報処理で可能なことがわかってきたからである。

1990 年代中ごろ、P. Shor は離散対数と素因数分解を確率的に多項式時間 (後の改良版では、線形時間) で解く量子アルゴリズムを発表し¹、アカデミックな世界のみならず、各国の国防関係者にも衝撃を与えた。もしも量子計算機が実現すれば、現在使われている公開鍵暗号システムが崩壊することを意味するからである。

それより早く、1980 年代初頭、Wiesner により、また後に独立に Bennett, Brassard により、量子鍵配布プロトコルが提案された。これは、盗聴を確実に検知することで、100%安全な秘密鍵の共有を行なうものである。現在、古典のプロトコルの安全性は計算量的仮定 (それは量子計算が出現すれば破れる) や、何らかの人間的な事情 (メモリーの有限性など) に依拠している。一方、量子鍵配布の安全性は、自然法則である量子力学そのものに依拠しており、盗聴者がどのような全能の操作が可能でも、決して安全性を崩すことが出来ないのである²。近年は、実験的研究も進み、都市内であればコストを別にして、実用的に使うことができるレベルに達している。

このような応用を陰で支えているのは、量子通信理論その他の、量子情報の基礎理論である。例えば、量子鍵配布の安全性の証明では、盗聴者が得る情報量を評価する必要がある、それにはシャノン理論の量子版が活躍する。また、量子計算機の実現の為には、古典計算機の場合と同じく、誤り訂正符号が設計されなければならない。

本講演では、量子通信理論の一部である、量子情報情報源符号化について解説をする。一番単純な場合からはじめて、最新の話である、universal 符号を目指して解説をする [6, 7, 11]。

量子情報理論は、勿論その固有の応用と意義があるのだが、一方、従来思いつかなかった角度で古典情報理論を見直すことにも繋がる。たとえば、で述べる混合状態情報源の圧縮は、確率分布情報源圧縮、という今まで考えられていなかった古典情報理論の問題を提起する。また、ここでは述べなかったが、visible compression [2, 17] なる概念をもちいて、相互情報量を情報源符号化の立場から特徴付けることもできる他、Reverse Shannon Theorem など、興味深いアプローチが多くなされて

¹ この業績にはニルバンナ賞がおくられた

² 安全性の証明は、ノイズのあるチャンネルを仮定して、極めて定量的、厳密、かつ厳しい仮定の下に行なわれている。

いる。

この解説を出発点にして、一人でも多くの方が量子情報理論に参入されることを望むものである。

2 量子力学の基礎

2.1 状態, 測定

量子力学では、実験の結果を確実に予言することはできず、その確率分布のみが理論的に予言できる。また、測定により系の状態は一般には変化してしまう。この確率分布や測定後の状態は、以下の二つの要因でま

1. 行なった測定
2. 測定直前の系の状態

上で系の**状態, 測定**という用語を用いてきたが、量子力学では、これらはある複素ベクトル空間上の行列（線形作用素）で表現されている。このベクトル空間を**表現空間**という。量子力学の慣習では、縦ベクトルを $|u\rangle = (u_1 \ u_2 \ \dots)^T$ 、その複素共役転置を $\langle u| = (u_1^* \ u_2^* \ \dots)$ で表わす。ある特定の物理系の理論を展開するとき、表現空間の次元をどうとるか、物理的な考察によって個別に決められる。例えば、電子などの粒子は、軌道運動の他に内部的な「回転」に対応する自由度（スピン）を持つと考えられている。この自由度に対応する表現空間は $\mathbf{C}^2$ である [16]。

系の状態は**密度行列 (密度作用素)** という

$$\rho = \rho^*, \quad \text{tr } \rho = 1, \quad \rho \geq 0, \quad (1)$$

を充たす行列で表され、測定は、**instrument** という行列の組 $\{A_\omega\}_{\omega \in \Omega}$ で表現される。ここで、 ω は測定値、 Ω は可能な測定値全体の集合であり、 A_ω は

$$\sum_{\omega \in \Omega} A_\omega^* A_\omega = \text{Id} \quad (2)$$

を充たすとする。(Id は単位行列。)

先に述べたように、測定結果は確率的にのみ定まり、この確率分布は系の状態と測定の関数である。 ρ で表現される状態の系に測定 $\{A_\omega\}_{\omega \in \Omega}$ を施すと、確率

$$\text{tr } A_\omega^* A_\omega \rho \quad (3)$$

で観測値 ω が得られ、状態は

$$\frac{A_\omega \rho A_\omega^*}{\text{tr } A_\omega^* A_\omega \rho} \quad (4)$$

へと変化する。(* は複素共役転置 (adjoint))。

先ほどの電子のスピンの場合、物理で 'z 軸方向のスピン測定' といわれる測定があるが、それは測定値として $\uparrow$ (upspin), $\downarrow$ (downspin) を返し、対応する instrument は、

$$A_\uparrow = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A_\downarrow = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

と書ける。この測定は、不均一磁場を通過する電子の軌道を観測することで実現できる [16]。状態が $\rho = [\rho_{i,j}]$ であった場合、確率 ρ_{11} で $\uparrow$ が、確率 ρ_{22} で $\downarrow$ が観測され、事後の状態は各々の場合、以下のようになる。

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

状態は、その密度行列 ρ のランクが 1 である時、**純粋状態** であるという。純粋状態の密度行列は、あるベクトル $|\Psi\rangle$ を用いて $\rho = |\Psi\rangle\langle\Psi|$ と書ける。このとき、簡潔の為に状態 $|\Psi\rangle$ などと言うことがあり、このベクトル $|\Psi\rangle$ を**状態ベクトル**という。

純粋状態でない状態を**混合状態**という。混合状態は、純粋状態の確率的重ね合わせとすることができる。なぜならば、任意の密度行列は、

$$\rho = \sum_i p_i |\Psi_i\rangle\langle\Psi_i|$$

のように純粋状態の凸結合でかけるからである。

2.2 系の時間的变化, 量子情報処理で可能な操作

本小節では系の時間的变化についての理論を述べるが、これは同時に量子情報処理で可能な状態操作の特徴付けにもなっている。

初期状態を ρ とすると、(2) を充たすある行列の組 $\{A_\omega\}_{\omega \in \Omega}$ を用いて、変化後の状態 ρ' は、

$$\rho' = \sum_{\omega \in \Omega} A_\omega \rho A_\omega^* \quad (5)$$

と書ける。ただし、この場合、 ω には測定値という意味合いはなく、単に和をとる為のインデックスである。

以上のべたことから、状態に対して行える操作を特徴付けることができる。すなわち、

符号化その他、状態操作はすべて (5) を充たさなければならない。さらに、状態から情報を引き出す操作、

すなわち測定は節 2.1 で述べられたものに限るから、併せて量子情報処理で何が可能であるかが公理的に規定されていることになる。こういった理論は 1980 年代の前半までにほぼ完成され、それにより量子情報科学は堅固な基礎を持つに至った [15]。

2.3 合成系

表現空間が $\mathcal{H}_A, \mathcal{H}_B$ となる 2 つの系をまとめて一つの系とみると、これを**合成系**といい、その表現空間は、以下に定義する $\mathcal{H}_A, \mathcal{H}_B$ のテンソル積空間 $\mathcal{H}_A \otimes \mathcal{H}_B$ である。

ベクトル空間 $\mathcal{H}_A, \mathcal{H}_B$ の正規直交基底系を $\{|e_i^A\rangle\}, \{|e_j^B\rangle\}$ とする。そして、形式的な積 $|e_i^A\rangle \otimes |e_j^B\rangle$ を考え、基底 $\{|e_i^A\rangle \otimes |e_j^B\rangle\}$ で張られる空間が $\mathcal{H}_A \otimes \mathcal{H}_B$ である。 $\mathcal{H}_A \otimes \mathcal{H}_B$ の次元は $\dim \mathcal{H}_A \times \dim \mathcal{H}_B$ に等しい。

2.4 古典系の記述

建前としては、古典力学は量子力学から導かれることになっている。実際にその導出が成功しているか否かは別にして、量子的な記述の枠組で古典的な情報処理が記述できることは確かである。つまり、量子情報科学は（少なくとも形式上は）古典情報科学の拡張になっている。

まず、ある基底 $\{|e_i\rangle\}$ を固定する。そして、行なう観測および操作は $A_j = \sum_i a_{i,j} |e_i\rangle \langle e_j|$ と書けるものに制限する。そして、密度行列は基底 $\{|e_i\rangle\}$ で対角化されているもののみを考える。すると、密度行列 $\sum_i p_i |e_i\rangle \langle e_i|$ は確率分布 $\mathbf{p} = (p_1, \dots, p_{\dim \mathcal{H}})$ に、TPCP 写像 $\{A_j\}$ は確率遷移行列 $[|a_{i,j}|^2]$ を持つマルコフチェーンにそれぞれ対応する。

感覚的にいうと、量子は行列の世界であるので、非可換である。それを可換な部分に制限したものが、古典の世界なのである。

3 量子情報源符号化

3.1 問題設定

今、確率分布 p_x に従って、状態 ρ_x を生成する量子情報源から、 n 個毎回独立に状態が出てくるとする。これを、状態を壊さずになるべく少ないメモリーを用いて記録しておきたい。例えば、電子のスピンを用いて記録すると、なるべく電子の数を少なくすませた

い。さきほど述べたように、電子のスピンは 2 次元複素ベクトル空間だったから、 m 個のスピンでは 2^m 次元のベクトル空間が準備できる。

問題を数学的にきちんと書くと、以下ようになる。 $\mathbf{x}_n = (x_1, \dots, x_n)$ として、確率

$$p_{\mathbf{x}_n} := \prod_{i=1}^n p_{x_i}$$

で

$$\rho_{\mathbf{x}_n} := \bigotimes_{i=1}^n \rho_{x_i}$$

なる状態が生成されてくる。 ρ_x が住んでいる空間の次元を d とすると、とすると、 $\rho_{\mathbf{x}_n}$ は d^n 次元の空間に住んでいる。これを、TPCP 写像 $\mathbf{E}_n$ で $d'_n (\leq d^n)$ 次元の空間に住んでいる密度行列に移してやる (符号化)。そして、必要なときには、また別の TPCP 写像 ($\mathbf{D}_n$) を用いて、 d^n 次元の空間に戻してやる (復号)。ここで、復号した結果が、もとの状態とほとんど差がないことを要請する。すなわち、 $\text{dist}(\cdot, \cdot)$ を密度行列の空間に定義された適当な距離として、

$$\sum_{\mathbf{x}_n} \text{dist}(\rho_{\mathbf{x}_n}, \mathbf{D}_n \circ \mathbf{E}_n(\rho_{\mathbf{x}_n})) \rightarrow 0 \quad (n \rightarrow \infty) \quad (6)$$

を要請する。そして、この拘束条件のもとで、 $\mathbf{E}_n$ や $\mathbf{D}_n$ を最適化して

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log d'_n$$

を最小化するのである。

距離 $\text{dist}(\cdot, \cdot)$ は、本稿では Bure の距離

$$\text{dist}(\rho, \sigma) = B(\rho, \sigma) := 1 - \sqrt{\text{tr} \rho^{\frac{1}{2}} \sigma \rho^{\frac{1}{2}}} \quad (7)$$

を用いる (これは Hellinger の距離の量子版である。) が、他のものを用いてもよい。

3.2 純粋状態情報源

さて、以上の問題設定の下、次の定理が成り立つ [1, 18]。

Theorem 1 もしも ρ_x がすべて純粋状態ならば、

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log d'_n \geq S(\bar{\rho}) \quad (8)$$

であり、等号が達成可能である。

ここで、 $S(\rho)$ はフォン・ノイマンエントロピーで、

$$S(\rho) = -\text{tr } \rho \log \rho$$

であり、 $\bar{\rho}$ は、情報源から出てくる密度行列の平均で、次式で定義される。

$$\bar{\rho} := \sum_x p_x \rho_x$$

不等式の証明 [1] はテクニカルなので省く。等号は、以下に説明する Schumacher 符号で達成できる [18] が、これは古典の圧縮とよいアナロジーがある。古典の場合は、典型系列に着目するが、量子では後で定義する典型部分空間に着目するのである。

$\bar{\rho}$ の固有値をまとめて $\mathbf{q} := (q_1, \dots, q_d)$ と書くと、 $\sum_j q_j = 1$ であり、各 q_j は非負なので、この並びは確率分布として解釈できる。 $\bar{\rho}^{\otimes n}$ の固有値は、 $\prod_{j=1}^d q_j^{n_j}$ の形をしており、各々の固有値の重複度は $\frac{n!}{n_1! \dots n_d!}$ である。つまり、独立に確率分布 $\mathbf{q} := (q_1, \dots, q_d)$ から n 回サンプリングしたときの確率分布が、そのまま $\bar{\rho}^{\otimes n}$ の固有値の系列になっている。そして、この i.i.d の系列のタイプを指定すると、 $\bar{\rho}^{\otimes n}$ の固有値、固有空間が指定される。そこで、 $\mathbf{n} := (n_1, \dots, n_d)$ として、 $\mathcal{H}_{\mathbf{n}}$ で固有値 $\prod_{j=1}^d q_j^{n_j}$ に対応する固有空間を表すことにする。そして、典型系列に対応する概念として、**典型固有空間**を次のように定義する：

$$\bigoplus_{\mathbf{n}: \|\frac{\mathbf{n}}{n} - \mathbf{q}\| < \epsilon} \mathcal{H}_{\mathbf{n}}.$$

また、 P_n^ϵ を典型固有空間への射影であるとする。

今、もしも ρ_x が $\bar{\rho}$ の固有空間への射影に等しかったとしよう。つまり、 $\rho_x \rho_y = \delta_{x,y}$ が成立するとする。この場合は、適当に添字を付け直すと、 $p_x = q_x$ となっていて、さらに、節 2.4 にもある通り、 x が確率分布 $p_x = q_x$ で出てくる古典情報源に等しくなる。であるから、 $x_1 \dots x_n$ が典型系列になっている場合のみを保存すればよい。これは、符号化の TPCP 写像 $\mathbf{E}_n$ を、

$$\mathbf{E}_n(\rho) = P_{n,\epsilon} \rho P_{n,\epsilon} \quad (9)$$

と定義することで実現される。復号化は、なにも操作をくわえず、そのまま符号化された状態を差し出せばいい。

実は、 ρ_x が $\bar{\rho}$ の固有空間への射影になっていなくても、上記の方法がうまくいくことが簡単に証明できる。

ただし、一般には、 $x_1 \dots x_n$ が典型系列に入っている場合を抜き出す、という操作はこの符号化に一致しないし、最適にもならないのである。

まず、 $H(\mathbf{q})$ をシャノンエントロピーとすると、任意の δ に対して、 ϵ を十分小さくとると、

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log d'_n = H(\mathbf{q}) + \delta = S(\bar{\rho}) + \delta \quad (10)$$

であることは、古典の情報源符号化と同様に証明される。問題は、拘束条件 (3.1) が満たされているかであるが、それは以下のようにチェックされる。

$$\begin{aligned} & \sum_{x_n} \text{dist}(\rho_{x_n}, \mathbf{D}_n \circ \mathbf{E}_n(\rho_{x_n})) \\ &= 1 - \sum_{x_n} p_{x_n} \text{tr} \sqrt{\rho_{x_n} P_n^\epsilon \rho_{x_n} P_n^\epsilon \rho_{x_n}} \\ &= 1 - \sum_{x_n} p_{x_n} |\text{tr} P_n^\epsilon \rho_{x_n}| \\ &\leq 1 - |\text{tr} P_n^\epsilon \sum_{x_n} p_{x_n} \rho_{x_n}| \\ &= 1 - |\text{tr} P_n^\epsilon \bar{\rho}^{\otimes n}|. \end{aligned}$$

ここで、 $\text{tr} P_n^\epsilon \bar{\rho}^{\otimes n}$ は典型系列が出る確率に等しく、1 に漸近する。よって、拘束条件 (3.1) が確かめられた。

3.3 混合状態情報源

混合状態情報源の場合は、Koashi-Imoto により解決された。

この問題を古典にやきなおすと、確率分布の系列の圧縮を扱っていることになる。つまり、確率 p_x で確率分布 $r_x(j)$ を選び、 $r_x(j)$ に従って j を振り出す。これを繰り返して系列 $x_1 \dots x_n$ と $j_1 \dots j_n$ が生成される。今、 $j_1 \dots j_n$ のみが与えられ、これを圧縮する。ただし、復号にあたっては $j_1 \dots j_n$ を復元する必要はなく、確率分布 $\prod_{x=1}^n r_x(j)$ に従う系列を生成できればよい、とするのである。

勿論、 $\sum_x p_x r_x(j)$ で j を生成する情報源を考え、通常の情報源符号化を行えば、この目的は達成されるが、一般にはこの方法は無駄が多い。例えば、確率 1 で確率分布 $r(j)$ が振り出されるとすると、何も記録する必要がない。復号は、確率分布 $r(j)$ に従って j を生成すればいいのである。

以上の考察から、平均の状態 $\bar{\rho}$ が同じでも、混合状態の場合、純粋状態よりも圧縮できる場合のあること

がわかる．詳細は原論文 [12] を参照されたい．ただ，この問題は，古典の場合も以前は解かれていなかったことを強調したい．

3.4 Universal 符号

3.5 固定長 universal 符号

今後，再び ρ_x が純粋状態である場合を考える．今までの議論では，情報源が既知の場合であった．情報源が未知の場合，どうすればいいだろうか．ところで，情報源の詳細はわかっていなくとも， $\bar{\rho}$ さえわかっているならば符号が構成できることをに注意されたい．その意味では，Schumacher の符号は元々ある程度の Universality があるのである．

本節では，固定長 universal 符号を考える．つまり，エントロピー $S(\bar{\rho}) = H(q)$ がある値 R 以下であることがわかっている場合を考える．

まず，さらに $\bar{\rho}$ の固有ベクトルがわかっている，固有値が未知の場合を考える．このとき，Schumacher の符号の P_n^ϵ を

$$\bigoplus_{n: H(\frac{\bar{\rho}}{n}) < R + \epsilon} \mathcal{H}_n. \quad (11)$$

への射影でおきかえればよい．すると，レート R が達成され，状態は壊れないことが，先と同様に証明できる．

次に， $\bar{\rho}$ の固有ベクトルもわかっていない場合には，先の符号で，(11) への射影を下記の空間への射影 $Q_{n,R}^\epsilon$ で置き換えたもの考える：

$$\text{span}_{\mathbb{C}}\{U^{\otimes n}|\phi\rangle : U : \text{a unitary}, |\phi\rangle \in \text{subspace (11)}\}. \quad (12)$$

つまり，固有ベクトルがどんな方向であっても，確実にカバーされるように大きめに空間をとっておくのである．この符号は，Jozsa と Horodecki 一家³により提案された [10]．まず，射影の台を大きめにとってあるので，この符号で状態が破壊されないことはわかる．問題はレートを達成するのだが，上記の空間の次元を評価せねばならず，群の表現論の知識が必要になるので，詳細には立ち入らない．

³R. Horodecki が父親で残りの二人は息子．

3.6 可変長 universal 符号-構成の方針

Jozsa と Horodecki の符号は，エントロピーの上界がわかっていないと役に立たない．エントロピーが全く未知の場合は，なんとか与えられた状態系列からそれを推定してやらないといけない．しかし，量子力学では，測定で一般に状態は破壊されてしまう．一般的な傾向として，精度の高い測定ほど，状態を激しく破壊する．

そこで，我々は状態の破壊が微小ですむ程度にエントロピーを推定し，それを用いて Jozsa と Horodecki の符号を適応することにする⁴．

3.7 状態を破壊しないで推定する

しばらく符号化の話から離れて，測定結果から状態を推定する話をする．測定 $\{A_\omega\}_{\omega \in \Omega}$ を行なったときに ω が観測される確率は (3) で与えられる．もしも，同じ状態の系のコピーが $n (> 1)$ 個与えられていると，それらすべてに測定 $\{A_\omega\}_{\omega \in \Omega}$ を施すと，確率分布 (3) に従うデータが n 個得られることになる．であるから，古典の統計学に従って，その結果から $\text{tr } A_\omega^* A_\omega \rho$ を， $\frac{\# \text{ of } \omega}{n}$ で推定でき，誤り確率は n が大きくなると指数的に 0 に漸近し，その指数は古典の大偏差原理を用いて評価できる．従って， $\hat{\rho}$ を ρ の推定量とすると，連立一次方程式

$$\text{tr } A_\omega^* A_\omega \hat{\rho} = \frac{\# \text{ of } \omega}{n} \quad (13)$$

を解いて $\hat{\rho}$ が求まる．ただし，この連立一次方程式が一意の解を持つように測定をうまく選ぶ．すなわち， Ω の要素の数が十分大きく，又， $\{A_\omega^* A_\omega\}$ が線形独立になるように選ぶ．⁵

さて，問題は，このような測定を行なうと，一般には状態が著しく変化してしまうことである．では，どのようにすれば状態の変化をさけることが出来るだろうか．答えは，「粗い測定」をすればいいのである．

以下に，「粗い測定」の構成を考える．今までは，推定のプロセスを，測定と測定結果からの推定値の計算に分けて考えてきたが，古典的な情報処理もまとめて，

⁴文献 [6, 7, 11] に現れる手法も，見かけはやや違って見えるかもしれないが，本質的にはここで述べる符号と同じことをしている．

⁵最適な推定を行なおうとすれば，測定の最適化を考えなければならない．通常は，さらに適応的に測定を変えたり，標本の間の相互作用を考えたりする [8, 9]．しかし，本稿では推定の最適化は考えない．

測定の中にいれこんでしまうことも考えられる。すなわち、 $\mathbf{C}^{d \otimes n}$ 上の、測定値が $\hat{\rho}$ を与える測定 $\{B_{\hat{\rho}}\}$ を考える ($B_{\hat{\rho}}$ は n にも依存するが、その依存性は繁雑をさけるために省く)。こうすると、まず、同じ推定値を与える中間的な測定値を区別しないことになるので、測定が目が粗くなっている (勿論、推定値自身はかわらない)。 $B_{\hat{\rho}}$ は正定値対称であるとする、

$$\begin{aligned} & \sum_{\hat{\rho}} (\text{tr } B_{\hat{\rho}}^2 \rho) \text{dist} \left(\frac{B_{\hat{\rho}} \rho B_{\hat{\rho}}}{\text{tr } B_{\hat{\rho}}^2 \rho}, \rho \right) \\ &= 1 - \sum_{\hat{\rho}} \sqrt{\text{tr } B_{\hat{\rho}}^2 \rho} \text{tr} \sqrt{\rho^{\frac{1}{2}} B_{\hat{\rho}} \rho B_{\hat{\rho}} \rho^{\frac{1}{2}}} \\ &= 1 - \sum_{\hat{\rho}} \sqrt{\text{tr } B_{\hat{\rho}}^2 \rho} \text{tr} \rho^{\frac{1}{2}} B_{\hat{\rho}} \rho^{\frac{1}{2}} \\ &\leq 1 - \sum_{\hat{\rho}} \sqrt{\text{tr } B_{\hat{\rho}}^2 \rho} \text{tr} \rho^{\frac{1}{2}} B_{\hat{\rho}}^2 \rho^{\frac{1}{2}} \\ &= 1 - \sum_{\hat{\rho}} (\text{tr } \rho B_{\hat{\rho}}^2)^{\frac{3}{2}} \end{aligned}$$

を得る。ここで、4 行目の不等式には、次の事実を用いた：

$$0 \leq B \leq I \Rightarrow 0 \leq B^2 \leq B.$$

従って、ある点列 $\{\rho_n\}$ があって、 $\lim_{n \rightarrow \infty} \text{tr} \rho B_{\hat{\rho}}^2 = 1$ であったとすると、状態は破壊されずにすむ。今、上記にのべたように適切に測定を構成すれば、 $n \rightarrow \infty$ では、 $\hat{\rho}$ は ρ の回りに集中して分布する。であるから、いくつかの $\hat{\rho}$ を区別しない測定を用いれば、状態を破壊しないですむであろう。

より具体的には、推定値の空間を小さなセル $U_{\delta, \bar{\rho}}$ に分割して、そのセルに $\hat{\rho}$ が入った場合は、推定値 $\bar{\rho}$ を返すことにする。 $U_{\delta, \bar{\rho}}$ は $\bar{\rho}$ を中心とする半径 δ の球を含むように設計される。こうして新しい測定 $\{C_{\bar{\rho}}\}$,

$$C_{\bar{\rho}}^2 := \alpha \sum_{\hat{\rho} \in U_{\delta, \bar{\rho}}} B_{\hat{\rho}}^2 \quad (14)$$

が定義される。ここで、 α は規格化条件 (2) を満たすための定数である⁶。なお、 $C_{\bar{\rho}}$ 、 α は n, δ に依存するのであるが、繁雑を避ける為にその依存性を省く。

⁶詳細は省くが、 $\bar{\rho}$ の動く範囲や、セルの形が、定数 α だけの調節で (2) が満たされるよう、巧妙に設計しておく。

この測定での状態の破壊の程度は、以下のように評価できる。

$$\begin{aligned} & \sum_{\bar{\rho}} (\text{tr } C_{\bar{\rho}}^2 \rho) \text{dist} \left(\frac{C_{\bar{\rho}} \rho C_{\bar{\rho}}}{\text{tr } C_{\bar{\rho}}^2 \rho}, \rho \right) \\ &= 1 - \sum_{\bar{\rho}} \sqrt{\text{tr } C_{\bar{\rho}}^2 \rho} \text{tr} \rho^{\frac{1}{2}} C_{\bar{\rho}} \rho^{\frac{1}{2}} \\ &= 1 - \alpha \sum_{\bar{\rho}} \sqrt{\text{tr} \sum_{\hat{\rho} \in U_{\delta, \bar{\rho}}} B_{\hat{\rho}}^2 \rho} \text{tr} \rho^{\frac{1}{2}} \sum_{\hat{\rho} \in U_{\delta, \bar{\rho}}} B_{\hat{\rho}}^2 \rho^{\frac{1}{2}} \\ &\leq 1 - \alpha \sum_{\bar{\rho}} (\text{tr} \sum_{\hat{\rho} \in U_{\delta, \bar{\rho}}} B_{\hat{\rho}}^2 \rho)^{\frac{3}{2}} \end{aligned}$$

この測定の構成で、 δ を大きくとれば、状態はよりよく保存され、一方で、推定の精度が落ちる。

いささか細かい議論ののち、次のことがわかる：

$$\lim_{n \rightarrow \infty} \text{tr} \sum_{\bar{\rho} \in U_{\delta, \bar{\rho}}} B_{\bar{\rho}}^2 \rho = 1$$

を充たす $\bar{\rho}$ の「数」 (一般には測度) を β とし、 δ を n の増大とともに適当な早さで 0 に漸近させると、

$$\beta \alpha \rightarrow 1 (n \rightarrow \infty)$$

となる。従って、この測定では状態は破壊されないのである。また、推定の誤差は、 δ の増加関数であり、これは $n \rightarrow \infty$ で 0 に収束する。

3.8 可変長 universal 符号

前節で述べたように、非破壊な測定を用いて、いくらでも小さい誤差で状態を推定できるのであった。であるから、まずこの測定を用いて生成された状態を計り、 $\bar{\rho}$ を推定し、しかるのちに、Jozsa-Holodezki の符号を用いればよい。また、 R はフォン・ノイマンエントロピーよりもわずかに大きくとっておく。これにより、最適なレート達成する符号が構成できるのである。また、復号の為には、このレート R も記憶しておかなければいけないが、これは適当な精度まで、やや大きめに評価して記録しておけばよい。

なお、情報源符号化の場合、平均すると $\bar{\rho}$ になる状態を生成しているので、前節の状態の推定と少し話しが異なるが、以下のように、推定の場合とほぼ同じよ

うに状態の破壊の程度を見積もることが出来る:

$$\begin{aligned}
& \sum_{\mathbf{x}_n} p_{\mathbf{x}_n} \sum_{\tilde{\rho}} (\text{tr } C_{\tilde{\rho}}^2 \rho_{\mathbf{x}_n}) \text{dist} \left(\frac{C_{\tilde{\rho}} \rho_{\mathbf{x}_n} C_{\tilde{\rho}}}{\text{tr } C_{\tilde{\rho}}^2 \rho_{\mathbf{x}_n}}, \rho_{\mathbf{x}_n} \right) \\
& \leq 1 - \alpha \sum_{\mathbf{x}_n} p_{\mathbf{x}_n} \sum_{\tilde{\rho}} (\text{tr } \sum_{\hat{\rho} \in U_{\delta, \tilde{\rho}}} B_{\hat{\rho}}^2 \rho_{\mathbf{x}_n})^{\frac{3}{2}} \\
& \leq 1 - \alpha \sum_{\tilde{\rho}} (\text{tr } \sum_{\hat{\rho} \in U_{\delta, \tilde{\rho}}} B_{\hat{\rho}}^2 \sum_{\mathbf{x}_n} p_{\mathbf{x}_n} \rho_{\mathbf{x}_n})^{\frac{3}{2}} \\
& = 1 - \alpha \sum_{\tilde{\rho}} (\text{tr } \sum_{\hat{\rho} \in U_{\delta, \tilde{\rho}}} B_{\hat{\rho}}^2 \tilde{\rho})^{\frac{3}{2}}
\end{aligned}$$

3.9 符号の性質

ここで構成した符号は、確かに最適なレートに漸近的に達成する。しかし、古典的な可変長 universal 符号と異なり、歪みを0にすることはできない。実は、universal でなくとも、量子的場合には、歪みが0の圧縮は出来ないことが証明されているのである [13]。

また、可変長符号であるので、漸近的に0にいく確率ではあるが、エントロピーレートを上回ってしまう。この確率が減少する指数を最適化するには、文献 [6, 7] にあるように、群の表現論を用いる必要がある。

References

- [1] H. Barnum, C. A. Fuchs, R. Jozsa and B. Schumacher, Phys. Rev. A, **54**, 4707-4711 (1996).
- [2] W. Dür, G. Vidal, and J.I. Cirac, "Visible compression of commuting mixed states," Phys. Rev. A **64**, 022308 (2001).
- [3] C. W. Helstrom, Minimum Mean-Square Error Estimation in Quantum Statistics, Physics Letters Vol.**25A**, pp.101-102 (1967).
- [4] C. W. Helstrom, Quantum Detection and Estimation Theory, Academic Press, New York (1976).
- [5] A. S.Holevo, Probabilistic and Statistical Aspects of Quantum Theory, North-Holland, Amsterdam (1982).
- [6] M. Hayashi and K. Matsumoto, "Quantum universal variable-length source coding," Physical Review A **66**, 022311 (2002).
- [7] M. Hayashi and K. Matsumoto, "Simple construction of quantum universal variable-length source coding," LANL eprint <http://xxx.lanl.gov/abs/quant-ph/0209124> (2002).
- [8] 林正人, 松本啓史, "量子系における統計的推測の最近の発展," 応用数理, Vol. 11, No. 3 pp. 27-48 (2001).
- [9] 林正人, "量子系の統計的推測 (理論と応用)," 電子情報通信学会誌 vol. 85, No. 8 pp. 600-605 (2002)
- [10] R. Jozsa, M. Horodecki, P. Horodecki and R. Horodecki, "Universal Quantum Information Compression," Phys. Rev. Lett. **81**, 1714 (1998);
- [11] Stuart Presnell and Richard Jozsa, "Universal quantum compression with minimal prior knowledge," EQIS 2002 (2002).
- [12] Masato Koashi and Nobuyuki Imoto, "Compressibility of Mixed-State Signals," LANL eprint <http://xxx.lanl.gov/abs/quant-ph/0103128>(2001).
- [13] Masato Koashi and Nobuyuki Imoto, "Quantum information is incompressible without errors," LANL eprint <http://xxx.lanl.gov/abs/quant-ph/0203045> (2002).
- [14] M. A.Nielsen and I. L.Chuang, Quantum Computation and Information, Cambridge University Press, (2000).
- [15] 小澤正直, 観測理論の数理, 数理物理への誘い3 (江沢洋編), 遊星社 (2000), pp. 163-189.
- [16] J. J.Sakurai, 現代の量子力学 (上), 吉岡書店 (1985), pp. 2-30.
- [17] E. Soljanin, "Compressing Mixed-State Sources by Sending Classical Information," LANL eprint <http://xxx.lanl.gov/abs/quant-ph/0110130> (2001).
- [18] B. Schumacher, "Quantum coding," Phys. Rev. A **51**, 2738-2747 (1995)

国際会議のお知らせ

以下のご案内する内容につきましては、変更になっている場合もありますので、ご自身でのご確認をお願い致します。
最新情報は以下のサイトをご覧ください。

- IEEE Information Theory Conference and Workshop : <http://www.itsoc.org/confs/frcon.html>
- IEEE Communication Society Conference : <http://www.comsoc.org/confs/>

MILCOM 2004

日時 2004 年 10 月 31 日-11 月 03 日
場所 Monterey, CA, USA
URL
原稿 締切 (TBD)

Globecom 2003

日時 2003 年 12 月 01 日-05 日
場所 San Francisco, USA
URL <http://www.globecom2003.com/>
原稿 締切 2003 年 03 月 01 日

ISSSTA 2004

日時 2004 年 08 月 30 日-09 月 02 日
場所 Leura, Australia
URL <http://www.isssta2004.org/>
原稿 締切 2004 年 02 月 01 日

WPMC 2003

日時 2003 年 10 月 19 日-22 日
場所 Yokosuka, Japan
URL <http://www1.ilcc.com/WPMC/index.html>
原稿 締切 2003 年 05 月 16 日

ISIT 2004

日時 2004 年 06 月 27 日-07 月 02 日
場所 Chicago, USA
URL <http://www.isit2004.org/>
原稿 締切 (TBD)

TENCON 2003

日時 2003 年 10 月 14 日-17 日
場所 Bangalore, India
URL <http://www.ewh.ieee.org/r10/bangalore/tencon2003/index.php>
原稿 締切 2003 年 03 月 15 日

ICC 2004

日時 2004 年 06 月 20 日-24 日
場所 Paris, France
URL <http://www.icc2004.org/>
原稿 締切 2003 年 09 月 01 日

MILCOM 2003

日時 2003 年 10 月 13 日-10 月 16 日
場所 Boston, MA
URL <http://www.milcom.org/2003/cfp.htm>
原稿 締切終了 (2003 年 02 月 14 日)

VTC-2004 Spring

日時 2004 年 05 月
場所 Genoa, Italy
URL
原稿 締切 (TBD)

VTC-2003 Fall

日時 2003 年 10 月 04 日-09 日
場所 Orlando, Florida
URL <http://www.vtc2003.org>
原稿 締切 2003 年 03 月 10 日

WCNC 2004

日時 2004 年 03 月 22 日-03 月 24 日
場所 Dallas, USA
URL
原稿 締切 (TBD)

PIMRC 2003

日時 2003 年 9 月 7 日-10 日
場所 Beijing, China
URL <http://www.pimrc2003.org>
原稿 締切 2003 年 03 月 31 日

INFOCOM 2004

日時 2004 年 03 月 07 日-11 日
場所 Hong Kong
URL <http://www.ieee-infocom.org/2004/>
原稿 締切 2003 年 07 月 01 日

Information Theory Workshop

日時 2003 年 07 月 06 日-07 月 10 日
場所 Hong Kong
URL <http://personal.ie.cuhk.edu.hk/info-itw/>
原稿 締切 2003 年 03 月 15 日

ISIT 2003

日時 2003 年 06 月 29 日-07 月 04 日
場所 Yokohama, Japan
URL <http://www.isit2003.org/>
原稿 締切終了 (2002 年 11 月 01 日)

ICC 2003

日時 2003 年 05 月 11 日-05 月 15 日
場所 Anchorage, USA
URL <http://www.icc2003.com/>
原稿 締切終了 (2002 年 08 月 15 日)

VTC-2003 Spring

日時 2003 年 4 月 22 日-25 日
場所 Seoul, Korea
URL <http://www.vtc2003spring.org/>
原稿 締切終了 (2002 年 09 月 15 日)

ICASSP 2003

日時 2003 年 04 月 06 日-10 日
場所 Hong Kong
URL <http://www.eie.polyu.edu.hk/~icassp03/>
原稿 締切終了 (2002 年 10 月 25 日)

Information Theory Workshop

日時 2003 年 03 月 30 日-04 月 04 日
場所 Paris, France
URL <http://itw2003.enst.fr/>
原稿 締切終了 (2002 年 11 月 30 日)

INFOCOM 2003

日時 2003 年 03 月 30 日-04 月 03 日
場所 San Francisco, USA
URL <http://www.ieee-infocom.org/2003/>
原稿 締切終了 (2002 年 07 月 11 日)

WCNC 2003

日時 2003 年 03 月 16 日-20 日
場所 New Orleans, USA
URL <http://www.comsoc.org/confs/wcnc/2003/index.html>
原稿 締切終了 (2002 年 09 月 01 日)

編集後記

SITA Newsletter 第 47 号をお送りします。

今回は昨年 11 月 11 日に行なわれました情報源符号化ワークショップ特集号です。先生方からかなりの力作の原稿を頂戴したため、今回はボリュームの大きい号となりました。お忙しい中執筆にとりくんでくださった講演者の皆

様に、この場をお借りして御礼申し上げます。

本年はインフルエンザが大流行し、私の周辺でも強健な方でも高熱で倒れられる例がでています。年度末を控え、論文、報告書等で忙殺される時期です。皆様どうぞご自愛ください。
(大橋正良)

編集担当者

杉村立夫（編集理事）

〒 380-8553 長野市若里 4-17-1

信州大学工学部電気電子工学科

Tel. 026-269-5237

Fax. 026-269-5220

Email tsugimu@gipwc.shinshu-u.ac.jp

野上保之（編集幹事）

〒 700-8530 岡山市津島中 3-1-1

岡山大学工学部通信ネットワーク工学科

Tel. 086-251-8128

Fax. 086-251-8127

E-mail nogami@cne.okayama-u.ac.jp

大橋正良（編集理事）

〒 356-8502 上福岡市大原 2-1-15

(株)KDDI 研究所無線エンジニアリンググループ

Tel. 0492-78-7862

Fax. 0492-78-7521

E-mail ohashi@kddilabs.jp

村田英一（編集幹事）

〒 152-8552 東京都目黒区大岡山 2-12-1

東京工業大学大学院 理工学研究科

電気電子工学専攻

Tel./Fax. 03-5734-3573

E-mail murata@mobile.ss.titech.ac.jp

情報理論とその応用学会事務局

〒 194-8610 東京都町田市玉川学園 6-1-1

玉川大学工学部情報通信工学科

通信システム工学研究室内, 山崎浩一 気付

Tel: 042-739-8442 (直通)

Fax: 042-739-8858

E-mail: sita-o.ce@hn.is.uec.ac.jp