様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月１９日現在

機関番号：３２７０６
研究種目：若手研究（Ｂ）
研究期間：2008～2011
課題番号：２０７６０２５０
研究課題名（和文）　十分統計量に着目した情報源および通信路に対するユニバーサル符号の理論解析と設計
研究課題名（英文）　A Theoretical Analysis and Designing of Universal Code for Sources and Channels with Sufficient Statistic
研究代表者
有村　光晴（ARIMURA　MITSUHARU）
湘南工科大学・工学部・講師
研究者番号：80313427

研究成果の概要（和文）：
今まで理論解析のあまり成されてこなかった，可変長メッセージ-固定長符号語符号(VF 符号)の圧縮性能の解析を進めたほか，その特別な場合として見ることのできる固定長メッセージ-固定長符号語符号(FF 符号)の圧縮性能の解析を行なった．また，VF 符号の具体例としてタンストール符号の圧縮性能の解析を行ない，これまでとは異なる形で期待値を取った場合でも，平均符号語長が情報源のエントロピーレートに収束することが証明された．

研究成果の概要（英文）：
In this research, some theoretical performance analysis of variable-to-fixed length source codes have been done, and insights about fixed-to-fixed length source coding are obtained. One of the main results is that a new criterion to analyze the compression performance of the Tunstall code is presented, and using this criterion, the Tunstall code is proved to be asymptotically optimal for stationary and memoryless sources.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	900,000	270,000	1,170,000
２００９年度	800,000	240,000	1,040,000
２０１０年度	800,000	240,000	1,040,000
２０１１年度	500,000	150,000	650,000
年度			
総　計	3,000,000	900,000	3,900,000

研究分野：工学
科研費の分科・細目：電気電子工学・通信・ネットワーク工学
キーワード：情報理論，データ圧縮，情報源符号化，ユニバーサル符号，情報スペクトル理論，十分統計量，VF 符号，FF 符号

１．研究開始当初の背景

一般に，データ圧縮に対する研究は，情報理論の研究者によるアルゴリズムの提案および理論的な最適性の証明と，画像や映像，音声などの特定のデータを扱っている研究者によるアルゴリズムの提案およびその実データによる実験的な評価が並行して行なわれてきている．このため，高性能なデータ圧縮アルゴリズムが存在したとしても，その理論的な解析はなかなか行なわれていない．また反対に，理論家によって漸近的に最適であることが証明されたアルゴリズムが実装されるまでには時間がかかっている．

この問題点の存在する理由は，対象とするデータのクラスを特定する方法が上記の両者の研究者の間で異なる，もしくは，対象と

するデータのクラスそのものが異なる，という点に集約することができる．理論的には，より広い情報源クラスに対応したユニバーサル符号を設計する方が良い成果となる．しかし，Rissanen によって提唱された MDL 基準の考え方を用いると，より広いクラスに対応したユニバーサル符号は，その分，情報源クラス内で情報源を指定するための符号語長が増えるため，有限長のデータに対する圧縮性能は悪くなってしまう．よって，単純に理論的に素性の良い情報源クラスに対して，データサイズを増やして行った時の漸近的な最適性を示すことが，実際のデータに対して必ずしも高性能な符号であると示すことでは無い，ということが可能である．

この問題点が顕著に表れたのが，ブロックソート・データ圧縮アルゴリズムである．この符号は実データに対して Lempel-Ziv 法とほぼ同等の実験的な圧縮性能を示したため注目されたが，理論的にその圧縮性能を解析すると，かなり狭い情報源クラスに対してしか漸近的な最適性を示さないことが明らかになった．この現象は，実験的な性能評価と理論的な性能評価が食い違った代表的な例として見ることができる．

よって，単純な定常無記憶情報源やマルコフ情報源，定常エルゴード情報源だけでなく，具体的な画像データや音声データ，その他のコンピュータ上のデータの特性を反映した，データ圧縮アルゴリズムの理論的な性能評価および設計の方法が求められる．

２．研究の目的

上記の背景に対し，この研究は，情報理論側から対象とするデータのクラスを特定し，それに適したデータ圧縮アルゴリズムを設計するための方法論を提案することを目指すものである．

まず，与えられた情報源に対してユニバーサル符号を理論的に設計する際に，これまでに行なわれてきたように，あるアルゴリズムを構成して，対象とする情報源クラスに対するユニバーサル性をその都度証明するのではなく，あらかじめ対象とするクラスに対する漸近十分統計量を構成し，それを用いてユニバーサル符号を構築することを最終的に目指している．

また，これまでに提案されてきたユニバーサル符号の理論的な解析を，改めて漸近十分統計量を用いて行なうことにより，様々なユニバーサル符号のユニバーサル性を，符号それぞれに対して理解するのではなく，汎用的なユニバーサル性の原理を確立することを目指す．

これにより，マルコフ情報源や定常無記憶情報源，定常エルゴード情報源，定常情報源といった，理論的に素性の良い情報源クラスのみならず，特定のデータに適した符号を設計するための理論的な枠組みを構築することで，現在までよりも高性能なデータ圧縮アルゴリズムを提案することを目指す．

さらに，この方法論を通信路符号化にも適用することで，情報源と同様に，対象とする特定の通信路に適したユニバーサル通信路符号を設計する方法を構築することを目指している．

３．研究の方法

情報源符号化アルゴリズムは，符号化する元データの集合が固定長か可変長か，また符号化された符号語集合が固定長か可変長かによって，FF 符号，FV 符号，VF 符号，VV 符号に分けることができる．本研究では，このうち FF 符号および VF 符号の圧縮性能について集中的に理論的な解析を行なった．この理由として，これまでに提案されている多くのアルゴリズムが FV 符号であり，VF 符号についてはほとんど研究が進んでいないこと，また，FV 符号として解析されることの多い Lempel-Ziv 符号が，実は VF 符号として理解する方が素直であると考えられることが挙げられる．

この解析において，本研究では情報スペクトル的方法を用いて，一般情報源に対する性能評価を行なった．一般情報源は，定常性やエルゴード性を仮定しない情報源ということができ，情報源の特性に依存しない，汎用的な理論を構築することが可能である．また，これまでのシャノン流情報理論では，データ圧縮の限界として，情報源の確率構造から計算されるエントロピー，さらにデータサイズを大きくしたときのエントロピーの極限として導出されるエントロピーレートが用いられている．これは自己情報量の期待値として解釈することが可能であるが，情報スペクトル理論では，この自己情報量の期待値ではなくて確率分布を解析の対象としている．そのため，平均的な振る舞いだけでなく，より細かい性能の評価が可能となることが期待される．

４．研究成果

これまでに，以下のような研究成果が得られた．

まず，タンストール符号の平均符号語長に関する評価を行なった(雑誌論文①，学会発表⑩)．具体的には，VF 符号の代表的なアルゴリズムであるタンストール符号について，これまでとは期待値の取り方を変えた漸近的な符号化レートの評価を行ない，今までの期待値の取り方によって得られる極限値と同じ値に収束することを証明した．実は，これまでの VF 符号の性能解析における平均符号語長の定義は，符号語長をブロック長の期

待値で割ったものになっており，１シンボルあたりの符号語長に対してその期待値を取ったものとはなっていない．この期待値の定義方法は，符号語を１回だけ送信するワンショット符号化だけでなく，系列を複数の可変長のブロックに区切って，それらを VF 符号化するという実際のアプリケーションにも対応できるという利点が存在する．一方で，FV 符号で普通に用いられている評価とは期待値の取り方が異なるため，FV 符号と VF 符号の性能評価を対応させることができないという難点も存在する．そこで本研究では，通常の FV 符号と同様の，１シンボルあたりの平均符号語長に対して，その期待値を取ったものを理論的に評価し，これまでの期待値の取り方と同じく，１シンボルあたりの符号語長が定常無記憶情報源のエントロピーに収束することを理論的に証明した．この証明の中では，特に，期待値の収束と確率収束のつながり，符号化レートと冗長度のつながりなどが直接表れるため，証明における理論的な見通しが，これまでに比べて大幅に増したと考えられる．

次に，固定長メッセージ-固定長符号語符号(FF 符号)の冗長度なるものを定義し，その漸近的な評価を行なった(学会発表①④⑥⑦)．1993 年に Han-Verdu によって情報スペクトル理論が構築されて以降，さまざまな情報スペクトル的な評価が行なわれて来ているが，その中で，古賀によって 2000 年に「情報源の漸近的な幅」なるものが定義されている．このとき古賀は Homophonic Coding と呼ばれる符号化の問題を考えるためにこの量を定義している．本研究では，この量が FF 情報源符号化の冗長度に対応することを理論的に証明した．この FF 情報源符号化の評価は，情報スペクトルの研究の中でも最も基本的な問題の一つであるが，これまではその符号化レートの評価が行なわれているのみで，冗長度の評価は行なわれてこなかった．本研究では，FF 符号の冗長度の評価という問題を定義することによって，古賀によって定義された量が，最も基本的な情報スペクトルの問題に対応していることを示した．

また，上記２つのテーマを組み合わせた形として，可算無限アルファベットに対する VF 符号の評価を行なった(学会発表⑤⑧⑨)．特に，１シンボルあたりの符号語長がスペクトル上エントロピーレート以下にはならないという，いわゆる逆定理が，FF 符号に対する証明をそのまま拡張したものとして表わせることを示した．

最後に，最近提案された Compression via Substring Enumeration (CSE)法について，その圧縮性能を一部明らかにした(学会発表②③)．このアルゴリズムは，与えられたデータの中に存在する全ての部分列を数え上げて符号化するというアルゴリズムで，提案されたもののその理論的な解析はまだ十分には行なわれてはいない．本研究ではこの符号化アルゴリズムに着目し，そのユニバーサル性の証明を試みている．その結果の一つとして，提案されたオリジナルのアルゴリズムのままでは，情報源のエントロピーまでデータを圧縮できないような情報源が存在することが明らかになり，その改良アルゴリズムを提案することができた．

以上のように，まだ研究当初の目的を達成するには，まだ時間がかかりそうであるが，VF 符号を中心としてその圧縮性能やユニバーサル性に関してある程度の結果を得ることができた．

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Mitsuharu Arimura, “On the Average Coding Rate of the Tunstall Code for Stationary and Memoryless Sources,” IEICE Trans. Fundamentals, Vol. E93-A, No. 11, pp. 1904-1911, Nov., 2010.（査読あり）

〔学会発表〕（計１０件）

① Hiroki Koga, Mitsuharu Arimura, and Ken-ichi Iwata, “Coding Theorems on the Worst-Case Redundancy of Fixed-Length Coding for a General Source,” Proc. 2011 IEEE International Symposium on Information Theory (ISIT2011), pp. 1434-1438, Saint Petersberg, Russia, July 31-Aug. 5, 2011.

② 嶋優希，岩田賢一，有村光晴，Lossless Data Compression via Substring Enumeration におけるある改良，電子情報通信学会技術研究報告，No. IT-2011-1, pp. 1-6, 大阪市立大学, May 20, 2011.

③ Ken-ichi Iwata and Mitsuharu Arimura, “An Improvement in Lossless Data Compression via Substring Enumeration,” Proc. 10th IEEE/ACS International Conference on Computer and Information Science (ICIS2011), pp. 219-223, Sanya, Hainan Island, China, May 16-18, 2011.

④ 古賀弘樹，有村光晴，岩田賢一，情報スペクトルの幅と固定長符号化の最悪冗長度，電子情報通信学会技術研究報告，No. IT-2010-83, pp. 93-98, 大阪大学, March 3-4, 2011.

⑤ Mitsuharu Arimura and Ken-ichi Iwata, “A Converse Coding Theorem for Variable-to-Fixed Length Source Coding of General Sources,” Proc. of the 33rd Symposium on Information Theory and its Applications (SITA2010), pp. 349-352, Matsushiro, Nagano, Japan, Nov. 30-Dec.3, 2010.

⑥ Mitsuharu Arimura and Ken-ichi Iwata, “The Minimum Achievable Redundancy Rate of Fixed-to-Fixed Length Source Codes for General Sources,” Proc. 2010 International Symposium on Information Theory and its Applications (ISITA2010), pp. 595-600, Taichung, Taiwan, Oct. 17-20, 2010.

⑦ 有村光晴, 岩田賢一, 一般情報源に対する無歪み FF 符号の最小達成可能冗長度レート, 電子情報通信学会技術研究報告, No. IT2010-21, pp. 57-62, 工学院大学, July 22-23, 2010.

⑧ Mitsuharu Arimura and Ken-ichi Iwata, “On the Achievable Redundancy Rate of Fixed Length Source Code for General Sources,” Proc. 2010 IEEE International Symposium on Information Theory (ISIT2010), pp. 126-130, Austin, TX, USA, June 13-18, 2010.

⑨ 有村光晴, 岩田賢一, 一般情報源に対する FF 符号および VF 符号の冗長度レートについて, 電子情報通信学会技術研究報告, No. IT2009-136, pp. 413-418, 信州大学, March 4-5, 2010.

⑩ 有村光晴, 定常無記憶情報源に対する Tunstall 符号の個別冗長度, 第 32 回情報理論とその応用シンポジウム (SITA2009)予稿集, pp. 619-624, 山口県山口市, Dec. 1-4, 2009.

〔図書〕(計１件)

① 白木善尚編, 村松純, 岩田賢一, 有村光晴, 渋谷智治著, 情報理論, オーム社, 2008 年 9 月.

６．研究組織

(1)研究代表者

有村　光晴 (ARIMURA　MITSUHARU)

湘南工科大学・工学部・講師

研究者番号：80313427