論文 / 著書情報
Article / Book Information

論題(和文)	ここまできた音声認識
Title	
著者(和文)	古井 貞熙
Author	Sadaoki Furui
出典(和文)	日本音響学会第15回特別企画：異文化交流チュートリアル, Vol. , No. , pp.
Journal/Book name	, Vol. , No. , pp.
発行日 / Issue date	1999, 3

日本音響学会春季研究発表会特別企画
（1999年3月9日　明治大学）

ここまできた音声認識

古井　貞熙
東京工業大学大学院情報理工学研究科計算工学専攻
〒152 目黒区大岡山2-12-1
Tel/Fax: 03-5734-3480, email: furui@cs.titech.ac.jp

1．はじめに

日本IBMの「Via Voice」、NECの「しゃべっていいメール」、NTTの「腕時計型PHS」、任天堂の「ピカチュウげんきでちゅう」など、いろいろな音声認識製品が注目されるようになってきた。最近では、音声認識を用いてニュース音声を自動的に字幕化する研究も始められている。この背景には、最近の目覚ましい音声認識技術の進歩と、その実現を可能にしたコンピュータやLSIの能力の急速な向上がある。音声認識の形態は、音声のディクテーション（一般には音声ワープロ）と、音声によるコンピュータシステムとの対話に分けられる。それぞれ現在活発に研究が行われているが、技術的に未完成な面も多い。ここでは、音声認識の基本的な技術がどこまで来ているのか、今後どのように展開していくのかについて解説してみたい。なおディクテーションの技術に関しては、日本IBMの西村氏による音響学会誌の解説[1]があるので参照して頂きたい。

2．音声認識の概要

2.1 音声認識の分類

音声認識技術は、次のように分類することができる[2][3]。

(1) (孤立)単語音声認識：区切って発声した単語を認識する。

(2) 単語スポッティング：連続音声中のキーワードだけを認識する。

(3) 連続音声認識：単語を連続して発声した音声を認識する。文音声認識、会話音声認識などとも呼ばれる。

連続音声では単語がなめらかにつながって発声され、物理的な意味での単語境界が存在しないので、仮定した種々の単語列と音声との照合を、単語境界を様々に変えながら試す必要があり、このため照合に必要な計算量が莫大になる。いかにして少ない計算量で正しい単語列を見つけるかが鍵である。

単語毎に区切って発声した文音声を認識するディクテーションでは、音響処理は単語音声認識と同じになり、単語境界の不確定性がないだけ認識が容易になる。ただし、ユーザにとっては単語に区切った発声の負担はかなり大きく、特に日本語では単語の定義が明確でないので、このような発声は難しい。このため、初期の音声ワープロでは単語毎に区切って発声することが要求されたが、最近の機種では連続発声できるものが多い。

2.2 音声認識の原理

音声認識は、観測された音声波形から、発声者が何をしゃべった（しゃべろうとした）のかを推定する問題である。最近では図1に示すように、情報源、通信路、復号部からなるシステムに対する情報処理論・通信理論の問題とし

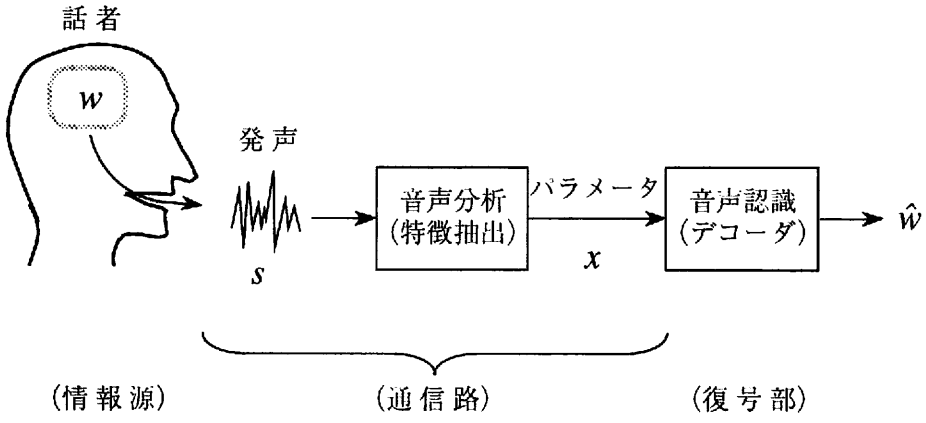
て、確率モデルを用いて定式化することが多い[2][4]。すなわち、発声者（情報源）が頭の中で考えた内容（通常は単語列）をwで表すと、wは発声者の発声器官によって音声波形sに変換される。この音声波形には通常、話者の個人差、付加雑音、伝送歪みなどが重畳している。音声認識システムでは、これをマイクロホンで電気信号に変換した後、波形の数値列としてコンピュータに取り込む。この数値列は、周波数スペクトルの時間変化を表す特徴パラメータ列xに変換される。この特徴パラメータ列から、元の発声者の頭の中にあった単語列wを推定（復号化）する。この際、パターン認識の理論から、事後確率を最大にするwを選べば誤認識が最小になることが分かっているので、図の式に従ってwを決定する。通常、事後確率を直接計算することは困難なので、wは、ベイズ則によって展開した右辺の式を最大化するように推定される。ここで、条件つき確率$P(x|w)$は音素などの音響モデルから特徴パラメータ列が出現する確率として計算され、単語列wが発声される事前確率$P(w)$は言語モデル（文法）によって計算される。

人が音声を聞いて理解しようとするとき、音としての情報だけでなく、語彙に関する情報（どのような単語があるかという情報）を始め、文法、意味など種々の情報を組み合わせて音声を聞き取っている。多くの場合、状況や話題の展開から無意識に単語や句を予測したり、文脈から推し量って認識している。その一つの理由は、音素情報のあいまい性である。すなわち、通常の音声の中では各音素は必ずしもはっきりと発声されず、その部分だけを聞いたのでは何の音素か分からないことがめずらしくない。このため、コンピュータによる音声認識においても、語彙、文法などの情報、すなわち単語列の事前確率$P(w)$の果たす役割が極めて大きい。

3．音声認識の基本技術

3.1 音響モデルと音響処理

条件つき確率$P(x|w)$を求めるためには、単語や音素などの音声の基本的な単位の標準パ



$$\hat{w} = \underset{w}{\operatorname{argmax}}\, P(w \mid x) = \underset{w}{\operatorname{argmax}}\, P(x \mid w) P(w)$$

図1　確率モデルに基づく音声認識システムの構成

ターンあるいはモデルを蓄えておき、仮定した単語列wに従って、その単位を接続したものと、認識すべき音声を音響的に照合する [2]。調音結合、すなわち各音素の物理的特性が前後の音素の影響によって変形する現象があるため、単語を単位とした方が、その変形を自然に取り込めるが、認識対象語彙数が大きくなると、記憶容量と認識のための計算量が膨大になってしまう。このため語彙数が大きいときには、音素を基本単位として用いて、前後の音素に依存した複数のパターンあるいはモデルを用意することにより、調音結合に対処する。

音声の特徴パラメータ系列と、各音素や単語の基本単位との類似の度合いを計算する際に重要な課題の一つは、音声の時間的な伸縮にどう対処するかである。例えば、同じ音素でも、それが含まれる言葉、話者、発声速度などによって長さが変化する。しかも、その変化はゴム紐のように線形（一様）ではなく、部分的に伸びたり縮んだりする非線形な伸縮である。このように非線形に伸縮するパターンを、そのずれを調整しながら照合する方法として、動的計画法（DP; dynamic programming）を用いた方法（DPマッチングと呼ばれる）が広く用いられている [2]。

音声には、時間軸だけでなく、スペクトルの形そのものが、個人差を始めとする多様な要因によって変化する性質がある。この問題に対処するため、近年、図2に示すような隠れマルコフモデル（HMM; hidden Markov model）による方法が広く用いられている [4]。HMMは、確率状態遷移機械であり、各状態遷移確率と、状態遷移に伴う特徴パラメータの観測（出現）確率によって特徴付けられる。これにより、音声の確率的変動が、極めて効率よく表現できる。音素などの音声の各基本単位に対応するHMMをあらかじめ作成しておき、入力音声がそのHMMの接続から生成される確率$P(x|w)$を計算する。この計算も、DPを用いることによって能率よく行なうことができる。

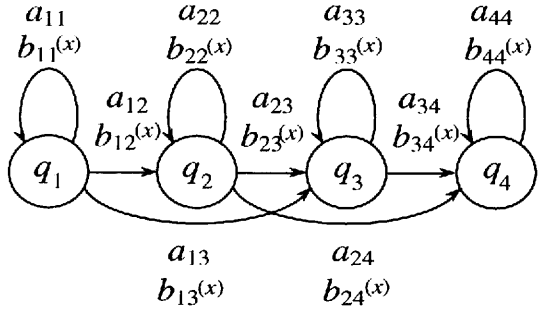


図2　隠れマルコフモデル（HMM）の構成；
q_i：状態、a_{ij}：遷移確率、
$b_{ij}(x)$：スペクトルパターンxの出現確率

連続音声認識の場合には、あらゆる種類の単語の組み合わせ（並び）の候補と入力音声とを照合する必要がある。例えば7桁の数字の組み合わせは　　の膨大な数となり、膨大な計算が必要となる。しかしこの場合も、DPを用いることによって、大幅に計算量を減らすことができる。語彙数がさらに大きい場合には、言語モデル$P(w)$によって候補単語列を効率的に絞ることが必須になる。

3.2 単語スポッティング

音声によるコンピュータシステムとの対話のような場合は、発声者の音声に含まれるキーワードだけを認識し、それ以外の余計な言葉は無視した方が都合がよい。このような目的のために開発されたのが、単語（ワード）スポッティング法である [3]。キーワードを一つだけ含む音声を認識して、そのキーワードを検出す

るためには、図3に示すように、音声を「ガーベジ（ごみ）モデル→キーワード→ガーベジモデル」のつながりによってモデル化する。ガーベジモデルは、あらゆる余計な音声が含まれるように作られる。音声がこのモデル列と照合されると、音声中のキーワードが真ん中のモデルに対応付けられ、それ以外の部分は前後のガーベジモデルに対応付けられる。

この方法は、米国の電話交換手業務の自動化のための音声認識に広く用いられている。具体的には、利用者が発声した自由な話し方の音声から、「コレクトコール」「番号通話」などの5種類のキーワードのいずれかを自動的に検出し、電話サービスの内容を選択するシステムである。

図3　単語（ワード）スポッティングのための音声モデル

3.3 言語モデルと言語処理

近年の音声認識では、言語モデル $P(w)$ として単語の連鎖確率を用いることが多い[5]。統計的言語モデルとも呼ばれる。ただし、単語数が数千あるいは数万種類になると、そのあらゆる種類の単語が多数連なった各系列の出現確率をすべて求めるのは、いくら膨大な文例があっても不可能になる。例えば1万単語の語彙中の任意の4単語からなる単語列の種類は、1万の4乗になってしまう。このため、文中で離れた単語は互いに影響しないと考えて、直前のN-1個の単語のみに依存する確率、N-gramモデルを用いる。N=2の場合がバイグラム（bigram）、N=3の場合がトライグラム（trigram）である。これらを掛け合わせることによって、多数の単語からなる系列の確率 $P(w)$ を計算する。バイグラムやトライグラムの推定には大量な言語データベースと大量な演算が必要であるが、文書の電子化、コンピュータの高性能化などによって、比較的容易に行なえるようになってきた。それでも実際は、言語データベースに見つからないバイグラムやトライグラムが大量に生ずる。これらが認識すべき音声に絶対に生じないとは言えないので、通常、Katzのバックオフ平滑化と呼ばれる方法で確率値を与える[5]。

いくら大規模な文書があっても、それをコンピュータで利用することができなかったら意味がない。日本語に関しては、英語などと違って単語に区切って書く習慣がないため、単語の統計的言語モデルを用いるのが、欧米に比べて大きく遅れた。我々は今から約4年前に、形態素解析プログラムを用いて日本語の文書を形態素に自動的に区切り、形態素を単語の代わりの基本的単位として統計的言語モデルを構築して音声認識に用いる初めての試みを行った[6]。この実験により、日本語においても、統計的言語モデルが英語とほぼ同程度の有効性を持つことが確認され、従来の人手に頼ることを前提としていた文法を用いるよりも認識性能が大きく向上した。

4．大語彙連続音声認識技術の進歩

4.1 ディクテーションと対話音声の認識・理解

最近、単語音声認識や連続数字音声認識は実用化あるいはそれに近いレベルに達しており、大語彙連続音声認識の技術も大きく進歩している[7]。「はじめに」で述べたように、連続音声認識の研究の流れは、ディクテーションと対話

音声の認識・理解を対象とした研究に分けられる。前者はいわゆる音声ワープロに相当し、書き言葉を発声したものを文字に変換するものである。そこで認識の対象とされる言葉は、書き言葉であるから、文法的にはほぼ正確であると想定することができる。そのかわり、認識しなければならない語彙や文体が極めて多様であるという点に難しさがある。後者の対話音声の場合には、対象（タスク）をある分野の情報案内などに限定すれば、ある程度語彙を制限することができるが、話し言葉特有のあいまいな文や、文法的に誤った文も対象としなければならない難しさがある。

4.2 外国の動向

世界で最大の大語彙連続音声認識プロジェクトは、米国のARPA（Advanced Research Projects Agency）によるものである。その内、ディクテーションについては、放送ニュース音声のディクテーションの研究が活発に行なわれており、65000語を対象として、80ないし90%の単語認識率が得られている[8]。対話音声認識については、ATIS（Air Travel Information System）システムの研究が強力に進められた。これは、米国内の航空路線を用いて旅行をすることを想定して、コンピュータによる案内システムに、音声で問い合わせや予約をするシステムである[9]。自由に発声した話し言葉を対象とするため、言い間違い、繰り返しなど難しい問題が多数含まれている。どこまで難しい音声を対象とするかによって異なるが、90～95%程度の単語認識率が得られている。

製品レベルでは、1990年に米国のDragon Systems Onc.が、単語区切りの発声ではあるが、「Dragon Dictate 30K」という初めてのPC用ディクテーションプログラムを発売し、その後、IBMを含めて種々の会社が参入し、1996年頃から連続発声ができる製品が開発されるようになった。これらの製品の多くは、米国英語版だけでなく、ドイツ語、フランス語、イタリア語、スペイン語、イギリス英語版などに発展している[1]。

ヨーロッパでは、列車の時刻案内などを対象とした音声対話システムの開発が、ドイツ、フランスなどを中心に活発に行なわれている。

4.3 日本の動向

日本では、単語区切りの音声を前提としたディクテーションプログラムが、IBMとNECから1997年に発売され、連続発声を受入れる製品へ発展している。研究としては、自動翻訳電話[10]、電話番号案内システム[11]、NHKの放送ニュース音声を認識する研究[12][13]などが進められている。電話番号案内タスクでは、極めて多数の人の名前や住所を、かなり自由な文体で発声した音声の認識ができるように、約80000語を語彙とし、文脈自由文法を用いた認識システムの実験が行なわれている。

ニュース音声の認識では、多数の音声を用いて音素HMMを作成し、約5年分のニュース原稿50万文を用いて、頻度の高い20000語を対象としたトライグラムを作成して用いている。アナウンサーの音声に対しては90%近い単語認識率が得られている。図4に、トライグラムの例を示す。例えば、「日本・の」の後に来る単語としては、「自動車」の確率が最も高く、次が「貿易」であることがわかる。NHKでは、耳の不自由な方などへの放送サービスとして、

ニュース番組に音声認識を用いて字幕をつけることを計画しており、音声認識精度95%を当面の目標にしている[13]。

ビジネスにおいて、音声認識が電話における操作性向上を実現するための技術として注目されるようになってきた。これにより、プッシュボ

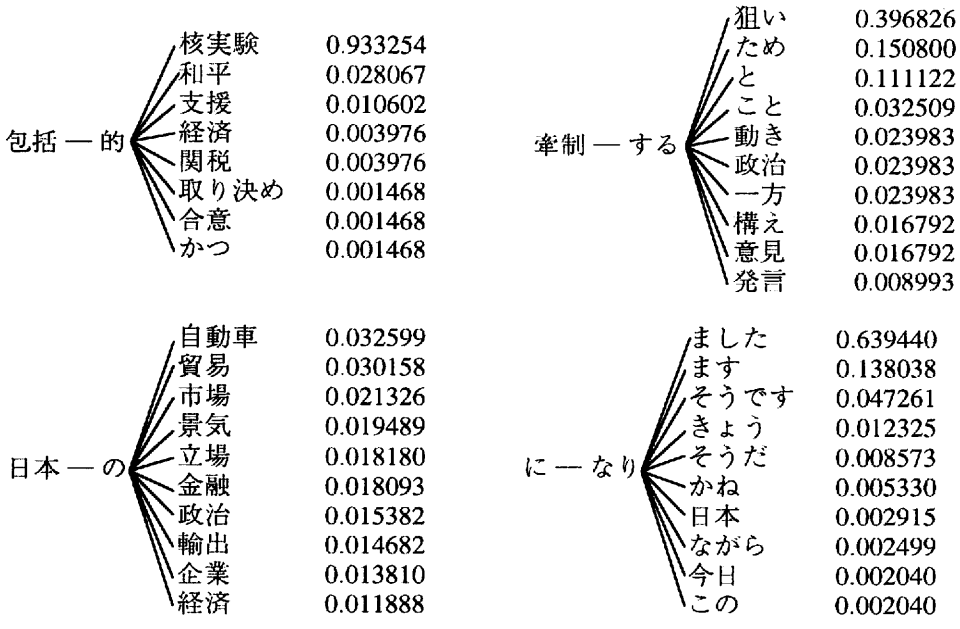


図4　ニュース音声のトライグラムの例

4.4 マルチモーダルシステム

上述のATISシステム や電話番号案内システムでは、基本的にシステムへの入力は音声認識で行い、システムからの情報は画面上の表や文で受け取る形をとっている。マルチモーダル、あるいはマルチメディア環境でのコンピュータとの情報の授受の方法として、これが人にとって最も容易かつ便利な方法と考えられるためである。入力手段として音声と画面のタッチ入力を併用して、使い勝手を向上させる研究も行なわれている[14]。

5．音声認識技術の応用と今後の展開

5.1 今後の応用展開

最近盛り上がりを見せるCTI（computer telephony integration：コンピュータ電話統合）

タン入力よりもはるかに容易に、交通情報、株価情報、各種予約情報などを得ることができる。パソコンのOSやアプリケーションソフトの操作を音声で指示できるようにもなってきた。

大語彙連続音声認識においても、書き言葉に近いきちんと発声された音声であれば、90%程度の単語が正しく文字に変換（ディクテーション）できるようになってきた。この技術を用いて、ニュースなどの音声を文字に変換し、そこから主たる話題（トピックス）を表す言葉や句を自動的に抽出したり、ほしいニュースを検索したりする研究が、最近活発に行なわれている[8][15]。ニュースをデータベース化し検索を容易にする上で、このような技術が役に立つと期待される。音声、音楽、雑音などを自動的に判

別する研究も行なわれており、これを用いれば音声データやビデオデータから、音声区間だけを検出できると期待されている。

CMUの"Informedia Digital Video Library Project"では、自然言語理解、画像処理、音声認識、および情報検索の技術を組み合わせて、膨大な映像と音声データベースを検索するシステムを構築している[8]。音声認識技術は、音声データベースのディクテーションと、音声による検索の両方に用いられる。

現在の音声ワープロの性能は、業務に用いるにはまだ必ずしも十分とは言えず、周囲に他の人がいるときには声が聞こえてしまうので使いにくいという問題もあるが、すでに多数のソフトウェアが販売されており、性能向上とともに普及してくると思われる。

対話音声認識・理解技術の利用としては、WWWブラウザに、マウスやキーボードに加えて音声による選択や呼び出しなどの操作を可能にすることがある。この発展として、コンピュータ中の仮想的な人物がエージェントとなって、インターネット上の情報を自在に検索してくれるシステム（virtual agent system）が考えられている。バーチャルモールの中の店員と対話しながらのショッピングがその一例である[16]。

最近注目されているウェアラブル・コンピュータでは、キーボードを持ち歩くわけにはいかないので、キーボードの代りに音声認識ソフトが活躍することになると予想される。

5.2 今後の課題

以上で述べたように、音声認識の研究は近年大きく進歩しているが、任意の話題についての、あらゆる人による、あらゆる環境での音声が認識できるという目標からは、まだ程遠い状況にある。このためには、まず周囲雑音、電話による歪み、話者の個人差に代表されるような実環境での耐性向上、その具体的方法としての自動適応機能の研究が重要である[7]。

さらに、実際のシステムを実現するだけでなく、研究の効率化のためにも、認識処理速度の向上、いいかえると効率的な認識処理（探索）方法の実現が必須である。また、認識対象（タスク、言語など）が変わったときに始めからシステムを作り直すのでは、無駄が多すぎるので、言語モデルの適応化など、いわゆるportabilityを向上させる研究も重要である。

コンピュータシステムとの対話という、より広くかつより有用な音声認識応用を実現するためには、言い直し、余剰語、未知語（OOV）などを沢山含む対話音声の言語モデルを構築し、自然な音声から、その意味内容を抽出し理解するための研究が極めて重要である。その際、すべての言葉を認識（ディクテーション）しようとするのではなく、意味を理解（意味検出、意味抽出）しようとするアプローチへの移行が必要である。ニュース音声などのディクテーションでも、原稿があらかじめ用意され、それに忠実に発声する場合ばかりではないので、上記の話し言葉特有の問題に対処しなければならない。今後、大規模な話し言葉のデータベースを構築し、これに基づいて、話し言葉の言語モデルを組み立てる必要がある[17]。

最近、毎日新聞を用いた日本音響学会新聞記事読み上げ音声コーパスが構築され、今後日本における大語彙連続音声認識研究の共通データベースとして使われると期待されている。今後

さらにこれを話し言葉へと展開していく必要がある。

6．むすび

音声認識技術が現在どこまで来ていて、どのような問題が残されているかについて概観した。

紙面と時間の都合で触れられなかったが、音声認識に類似した技術として、音声から誰が発声しているかを自動的に判定する話者認識の技術がある[2][3]。この技術も、ネットワークのセキュリティーなどの問題に関連して、関心が高まっている。話者認識の性能も最近向上しているが、音声認識と同様に、雑音が重畳したり、電話を通したり、風邪を引いたりして声が変わっても安定な認識方法の実現が今後の課題である。

今後、これら音声認識技術の一層の進展とともに、画像・映像・テキストなどの異なるメディアや、情報検索・自然言語処理などの異なる分野との技術の融合によって、種々の新しい応用が開けてくるであろう。

参考文献

[1] 西村雅史："音声ワープロ最新事情"、日本音響学会誌、54、3、pp. 229-234 (1998)

[2] 古井貞熙：" 音声情報処理"、森北出版、東京 (1998)

[3] 古井貞熙："音声処理に関する最新技術の原理と応用"、Interface、pp. 82-91 (Aug. 1998)

[4] 中川聖一："確率モデルによる音声認識"、電子情報通信学会、東京（1988）

[5] 北、中村、永田：" 音声言語処理"、森北出版、東京 (1997)

[6] 松岡、大附、森、古井、白井："新聞記事データベースを用いた大語い連続音声認識"、信学論、J79-D-II、12、pp. 2125-2131 (1996)

[7] 古井貞熙：" 大語彙連続音声認識の現状と展望"、音学春季講論、1-6-10 (1998)

[8] Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, Morgan Kaufmann Publishers, Inc., San Francisco (1998)

[9] Proceedings of the DARPA Speech Recognition Workshop, Morgan Kaufmann Publishers, Inc., San Francisco (1997)

[10] ATR国際電気通信基礎技術研究所編：" 自動翻訳電話"、オーム社、東京 (1994)

[11] 南、山田、鹿野、松岡：" 番号案内を対象とした大語い連続音声認識アルゴリズム"、信学論、J77-A、2、pp. 190-197 (1994)

[12] 大附、古井、桜井、岩崎、張：" ニュース音声認識のための言語モデルと音響モデルの検討"、信学技報、SP98-108 (1998)

[13] 安藤彰男："音声認識を用いてニュース音声を字幕化する試み"、NHK技研だより、pp. 2-7 (Dec. 1998)

[14] S. Furui & K. Yamaguchi: "Designing a multimodal dialogue system for information retrieval", Proc. ICSLP, We1R6 (1998)

[15] 高木、桜井、岩崎、古井："ニュース音声を対象とした言語モデルと話題抽出の検討"、信学技報、SP98-33 (1998)

[16] 大室、西、野田、嵯峨山：" インターネットと音声・オーディオ処理技術"、音学誌、52、8、pp. 631-636 (1996)

[17] 古井貞熙："話し言葉の音声理解へ—存在感のある研究を期待して—"、信学技報、SP98-113 (1998)