音声自動認識に関する情報工学的諸考察†

中川聖一†† 坂井利之†††

本論文は，音声自動認識についてのさまざまな問題点を列挙し，それらに対する諸考察を実験結果を基に述べている．それらは，一貫して音声自動認識をシステム的にとらえたもので，より本質的な能率のよい効果的な認識方法を見い出すことに主眼をおいている．

主な研究結果として，音韻識別へのグルーピング手法の導入，音韻識別部の能力が単語識別率に与える影響評価，単語音声の記述法とそれに基づいた単語音声認識手法の分類と評価，単語音声の認識時間を短縮するために単語音声の大局・局所的特徴を用いた予備選択（前照合）の概念の導入とその可能性の検討，認識対象語彙の認識困難度の評価，韻律情報の音声認識への利用法の検討，準最適な単語列を得るための新しい木探索法の提案，音声理解に適した文解析法の提案，言語情報の有効性の評価などがあげられる．

1. はじめに

音声波から言語情報を抽出するという狭義の意味での音声自動認識では，1950 年代より大きく分けて次の3つのレベルで研究がなされてきた[1]．

ⅰ）音韻・音節の認識

ⅱ）単語音声の認識

ⅲ）連続（単語）音声の認識・理解

また，韻律情報の研究は，これら各レベルと密接な関係にある．さらに，不特定話者を対象とする音声認識においては，話者認識の研究と切り離して考えることはできない．**表1**に，各レベルにおける研究上の諸問題を挙げた．

最近，ようやく限定語彙の単語音声の認識が実用化された．しかし，従来から音声認識を困難なものにすると考えられている音声パターンの変動要因である個人差や調音結合の問題が解決されたわけではなく，使用者ごとにシステムを適応させ，また，単語単位のパターンマッチング法により調音結合の問題を回避して実現されたに過ぎない．

音声認識の最終目標は音声タイプライタの実現であるが，この実現はきわめて難しく，新しく音声理解という概念が生まれた[2]．

タイプ入力を対象とした自然言語の理解研究の現状から察すれば，この音声理解の研究は時期早尚の観がある．しかし，認識対象を小世界に限れば（語彙数でいうなら数百単語程度）十分取り扱える範囲となり，また，実用面からいっても価値が多いと思われる[3]．

筆者らは，会話音声の認識モデルとして，階層モデルの立場を採る．階層モデルは，各レベルの処理がモジュール化され，柔軟な見通しのよいシステムになるという単なる便宜上のほかに，音韻認識，単語認識という閉じた研究を含み，学問的・実用的面からも興味あるモデルである．人間の音声認識過程は，まだ不明な点が多いが恐らく階層モデルに基づいていると思われる．一方，各レベルを一つのネットワークモデルに埋め込みシステム全体を一つの簡単なモデルとしてとらえる方法がある．これを仮りにネットワークモデルと呼ぶ．単語音声の認識レベルでいえば，音韻認識を介する方法が階層モデル，単語単位のパターンマッチングに基づく方法がネットワークモデルといえる．ネットワークモデルは，計算量が比較的大きくなるが，階層レベルで見られるような各レベルでの情報の損失とその伝搬の問題が回避できる利点がある．現在の技術レベルでは，認識率からいえば，ネットワークモデルの方がすぐれている．階層モデルがネットワークモ

表1 音声認識の主な問題点
Table 1 Significant problems on automatic speech recognition.

レベル	主な問題点	選択性
音韻・音節	特徴抽出，調音結合，個人差，セグメンテーション	個人差の正規化／学習，認識単位
単語	時間長の非線形，単語辞書の作成	情報圧縮法，語彙（数と内容）
連続単語	word juncture，単語境界の検出	木探索法
文	構文・意味・プラグマティクス，談話	タスク，解析方向，知識の表現法

† Some Considerations on Automatic Speech Recognition from View Points of Information Science by SEIICHI NAKAGAWA (School of Information Engineering, Toyohashi University of Technology), and TOSHIYUKI SAKAI (Department of Information Science, Faculty of Engineering, Kyoto University).

†† 豊橋技術科学大学情報工学系

††† 京都大学工学部情報工学科

情報量/解中間表現	10^5 bit	10^4 パラメータ時系列	10^2 ラティス表現	20 ラティス表現 (タスクによる)	10 (タスクによる)
階層レベル	音声波	特徴抽出	音韻・音節レベル	単語レベル	連続単語レベル・会話文レベル
冗長な情報	雑音情報 位相情報	音源情報 話者情報	継続時間情報 振幅情報	単語辞書に矛盾する音韻の組合せ	構文・意味・プラグマティクス 情報に矛盾する組合せ
使える情報		音声発生のモデル	音韻体系 音形規則 各音韻に対する音響的性質 韻律情報	単語辞書 韻律情報	構文・意味・プラグマティクス ユーザモデル 談話モデル 韻律情報
本論文のテーマ		話者のクラス分け	音韻のクラス分け 出力形式 音韻識別能力の評価	情報圧縮法 単語のクラス分け 単語認識の困難度	単語境界の検出 木探索法 文解析手順 言語情報の評価

図1 階層的モデルに基づく音声認識システムの情報工学的解釈
Fig. 1 Interpretation of a speech recognition system based on a hierarchical model from a view point of information science.

デルに認識率の上で優るためには，表1に挙げた諸問題を解決しなければならない．

莫大な音声データから言語情報のみを抽出する問題は，事前のさまざまな知識と処理過程で得られる知識を用いて，各階層レベルでの音声の持つ冗長な情報の圧縮とそれに伴う言語情報の損失の低減・回復を図る多段階認識過程である．**図1**は，この過程を示したものである．

筆者らは，この8年間，音声認識の全分野にわたって研究を進めてきた．その研究精神は，一口にいえば，情報工学的な（音声情報の本質に基づいたより能率的なより効果的，よりシステム的な）アプローチである．これらの研究成果の一部はすでに学会誌に発表済みである[5)~8)]．本論文では，新たに得られた最近の研究成果に基づいて，音声認識の全レベルについて問題点を指摘し，それを解決するための系統的な見解を示す．

2. 音韻認識

2.1 話者と音韻のクラス分け

（a） 話者のクラス分け

特徴抽出の段階で，音韻認識に必要な特徴のみを抽出することは難しく，通常話者の個人性を表わす特徴も同時に抽出される．音声が言語内容の伝達手段であることを考えれば，個人性を表わす特徴パラメータの存在が音韻認識を困難にしている大きな要因であることは事実である．話者による音声パターンの変動に対処するため，学習法と正規化法が研究されてきた．学習法は，あらかじめ発声された音声から，その発声者の声質を学習しようとするもので，能率のよい学習法が研究課題である．筆者らは，その究極的な方法として，単独5母音による学習法と教師なし学習法の可能性を検討してきた[5),7)]．一方，正規化法は，話者間に共通な，音韻間になんらかの不偏的構造の存在を仮定するものである．ほとんどの場合が声道形の相似形を仮定している．しかし厳密には，老若男女によってその構造は異なっており[9)]，この仮定には無理がある．また，学習法でも標準話者とかけ離れた話者の学習には，学習速度や学習誤りの問題が生じる．これらは，そもそも構造の異なる発生源から生じたパターンを一つの共通のモデルで処理することから生じた問題である．そこで筆者らは，あらかじめ各話者を構造の異なるものに分類してから，個々の同一の構造内で学習ないしは正規化をする方法を提案する．

広くパターン認識の分野ではクラスタリングの問題として取り扱われているが，ここでは見通しの良さや将来の実用性（性別や年令，身長等をシステムに教えてやる）を考え，性別と年令の層別によって生じる個人差の除去のために話者のクラス分けを行う．層別として，子供（11才前後），青年（20才前後），中年(40才以上）のそれぞれの男女，計6クラスを考えれば十分と思われる．各クラスについて，それぞれ20名の単独5母音を分析対象としてクラス分けの効果を調べた．種々のクラス分類によるクラス識別率と母音認識率を**表2**に示す．2クラスとは男・女，3クラスとは子供・成人の男・成人の女である．これらの結果から，人間の発声器官の構造の違いとして3クラスを設ければ十分と思われる．

（b） 音韻のクラス分け

話者の個人差と調音結合の存在が音声の自動認識を困難にしている双壁である．調音器官の連続的な運

表 2 話者クラス分けによる日本語5母音の認識
Table 2 Recognition of Japanese five vowels by clustering of speakers.

分割クラス数		分割なし	2クラス	3クラス	6クラス
話者クラス認識率			72.4%	80.7%	63.8%
母音認識率	話者クラス未知		92.5%	94.4%	94.1%
	話者クラス既知	92.6%	94.7%	96.4%	98.1%

特徴パラメータは，ピッチと1～10次のパコール係数（対数断面積比に変換）認識はマハラノビスの汎距離に基づく．

動によって生じる調音結合の現象は，聞き手にとってはむしろ好都合な現象である．この調音結合が音声のセグメンテーションを困難にし，しいては音声認識の最適な認識単位が問題となってくる．調音結合の問題を解決するべきであるという立場に立たないならば，認識単位は調音結合の影響をあまり受けない程度の時間長を選ぶのが自然で，音節[10]やVCV単位[11]（V＝母音，C＝子音）が考えられてきた．音節の種類が少ない日本語にとっては，この種のアプローチは大変有望である．これらは広義の音韻のクラス分けといえよう．ここでは，母音を例にとって狭義の音韻のクラス分けについて述べる．

各母音はコンテキストによってその音響的特徴は異なるが大きく分けて鼻音化母音と非鼻音化母音に分けられ，各母音を2つのクラスに分類するのが音声生成モデルとの対応の点からも妥当と考えられる（ここでは，無声化母音は考えない）．我々は，この仮説を確めるために次の実験を行った．音声資料は，5名に10数字を5回発声してもらったものである．ほかの10名の話者から作成した5母音の標準パターン（スペクトル）を初期値として，

① 数字中の母音スペクトルを使って，その話者の標準パターンの学習を行う．

② 各母音スペクトルパターンを2分類し，安全のために一方のクラスは初期値のスペクトルパターンを残し，他方のクラスのみ学習を行う．

③ 各母音スペクトルパターンを2分類しながら学習を行う．

10数字の認識実験から教師なし学習・教師あり学習共に，③，②，①の順にすぐれていることが明らかとなった[7]．このことから，連続音声中の母音を静的な特徴で認識するためには，少なくとも各話者の各母音に対して，2つの標準パターンを設けることが必要であるといえる．

2.2 音韻認識率と単語認識率との関係

現在の技術レベルでは，連続音声を音韻の系列に正しく変換することは不可能である．それゆえ，音声認識システムにおける音韻認識部の役割は，単語認識の前処理と考えてよい．それでは，どの程度の音韻認識率があれば，どの程度の単語認識率が得られるだろうか．これに答える研究例はほとんどなく，ただセグメンテーションが完全な場合に対する文字認識における解析的な研究があるに過ぎない[12]．音声分野では，音韻識別誤りがあるとほかの単語とどの程度混同するかという調査があるだけで[13),14]，実際の限定単語の認識率に与える影響についてはあまり考察されていない．いずれにしても，セグメンテーションが不完全な場合には，解析的な考察は不可能に近く，我々はシミュレーションによってこの問題を検討した．

まず，2つの音韻系列の一致度を評価するため距離を定義する必要がある．音韻認識には，通常脱落，挿入，置換の3種の誤りがある．最も簡単な，しかもしばしば使われる尺度として，これらの誤りをすべて等しく取り扱う方法がある[14),15]．

ここでは，実際のシステムになるべく忠実なシミュレーションを行うために次の実験を行った．以下順次述べる．

i）単語識別の方法

筆者らが開発した単語音声の識別法[5]をそのまま用いるが，システムに依存しない不偏的な結論を得るようにできるだけ考慮した．音韻識別部で得られる識別音韻列と単語辞書で与えられる音韻列とのマッチングを行い，最もよく整合のとれる単語辞書に対応する単語を識別結果とする．マッチングの際に，辞書中の音韻と識別音韻列中の音韻の対応度は，あらかじめ与えられている音韻間類似度によって評価する．ただし，この対応づけには，次の制限が設けてある．

① 辞書中のすべての音韻は，識別音韻列の連続する2音韻以内と対応づける．

② 識別音韻列中のすべての音韻は，辞書中の連続する2音韻以内と対応づける．

これらの制限を満たす範囲で，あらゆる可能な対応づけについて整合値を比較し，最もよく整合のとれる対応づけに対する整合値をその単語の尤度とする．最も尤度の高い単語を識別結果とする．これは，ダイナミックプログラミングの手法を用いて能率よく計算できる．

ii）音韻識別部のシミュレーション

音韻識別部で取り扱う音韻のカテゴリは，5母音(a, i, u, e, o)，撥音 (N)，半母音 (y, w)，有声子音

(m, n, ŋ, b, d, g, r, z), 無声子音 (s, c, h, p, t, k), 促音 (Q) の 23 種類である．ここでは，撥音と半母音も有声子音として取り扱う．また，/p, t, k/ 相互間の識別はきわめて難しいので，これらを同一の音韻カテゴリとして取り扱う．音韻識別部のシミュレーションで用いたパラメータは，音韻識別率，音韻脱落確率，音韻挿入確率，音韻置換確率マトリックスである．音韻置換確率は，誤って識別された場合，どの音韻に置換されるかを表わす確率で，実際のシステムと同じふるまいをするように音韻間類似度（コンフュージョンマトリックスに対応）から作成した．促音に関しては，実際のシステムでも確実に認識できるので，識別率 100%，脱落確率・挿入確率はともに 0% であるとした．上記の確率を用いて，擬似識別音韻列を生成する．

ⅲ）実験結果

音韻脱落確率と挿入確率は 0%，10%，20%の 3 通りについて，音韻の識別率は，母音と無声子音については，80% と 90% の 2 通り，有声子音については，40% と 60% の 2 通りについて，合計 45 通りのパラメータの組合せについて調べた．各パラメータの組に対して，各単語の擬似識別音韻列の発生回数は 10 回とした．実験結果を**表 3** に示す．＊印の欄は，セグメンテーションが完全であることをシステムに知らせた場合である．表には示していないが，脱落確率が 10%，20% の場合で，挿入確率が 0% のとき，システムにそのことを知らせると，単語識別率は約 10% 向上した．実験結果から次のことが明らかとなった．

① 脱落確率が単語識別率に与える影響はきわめて大きい．

② 挿入確率は，単語識別率にさほど影響を与えず，むしろ脱落確率と同程度以上に挿入確率がある方が好ましい．

③ 母音と子音の識別率が単語識別率に与える影響に大差はない．

④ 100 単語程度の語彙数に対して，95% 以上の識別率を得るためには，脱落確率を少なくとも数%程度以下に押える必要がある．

⑤ 音韻識別よりもセグメンテーションの方が重要な研究課題である．

表 3 シミュレーションによる 100 都市名の認識実験結果
Table 3 Recognition results of 100 city names by simulations.

音韻識別率		母音	80	90	80	80	90
セグメンテーション		有声子音	40	40	60	40	60
脱落	挿入	無声子音	80	80	80	90	90
20	20		58.6	61.0	61.7	60.7	62.7
10	20		73.5	75.5	74.8	73.2	78.4
0	20		94.8	96.8	94.9	95.5	98.2
20	10		52.4	54.5	54.7	53.3	59.3
10	10		73.8	75.6	77.4	74.9	80.2
0	10		93.1	97.1	96.3	94.7	98.1
20	0		53.2	54.1	53.7	52.2	56.4
10	0		72.2	76.1	74.9	71.7	77.3
0	0		94.1	95.7	94.6	94.3	98.4
0	0		96.9	98.3	98.7	98.2	99.8

（音韻間類似度等の言語情報を利用した場合）

2.3 音韻認識結果の出力形式

通常，音声認識システムにおける音韻識別部の出力は，複数個の候補からなる音韻ラティスの形式が多い[17),18)]．鹿野は，この音韻ラティスに，相互情報量なる概念を用いて，音韻ラティスと同じ情報量をもつ通常の音韻系列の音韻識別率を求める方法を提案し，音韻ラティス表現の有効性を検討している[8)]．しかし，実際の場合，識別された各音韻は，信頼度（識別スコア）を持っているのが普通である．この場合，上述のように実質的な識別率を導出するのは難しい．我々は，さまざまな出力形式と単語識別率との関係を調べた．

実験には，我々が開発した音声理解システム LITHAN[6)]を用いた．LITHAN の音韻識別部の出力は，各セグメントについて，第 1 候補音韻，第 2 候補音韻，第 1 候補と第 2 候補の信頼度の比，継続時間長である．男性話者 5 名が普通の速度で発声した算術文 80 文の認識結果を**表 4** に示す．実験結果は，システムの音韻識別部の能力にもよるが，第 1 候補だけを用いる方法や信頼度，継続時間長を用いない方法はよくないといえる．また，入力音声間の音響的類似性と相関の小さい音韻間類似度（表 4 には弁別特徴から導出したものも示してある）を用いた場合は極端に単語や文の認識率は悪くなる．

表 4 音韻識別部の出力形式と単語・文認識率との関係
Table 4 Relationship between output forms of phoneme recognitions and word or sentence recognition rate in arithmetic expressions.

第一候補音韻	第二候補音韻	信頼度	継続時間長	音韻間類似度	単語認識率	文認識率
○	○	○	○	スペクトル距離	93.5%	60.0%
○	○	×	○	〃	89.4	40.0
○	×	×	○	〃	91.3	53.0
○	○	○	×	〃	91.8	58.0
○	○	○	○	弁別特徴距離	84.4	45.0

（○印：使用する，×印：使用しない）

3. 単語認識

3.1 単語音声認識法の分類

単語音声の認識は，単語発声区間全体にわたる入力パターンと標準パターンとのマッチングによるのが一般的である．このとき，入力パターンと標準パターンに何を選ぶか（特徴パラメータ列か音韻列か等）によって，各種の手法が提案されている．マッチングは，ダイナミックプログラミングの手法が使えるアルゴリズムが良くほぼ定着化しつつある．**図2**は，入力パターンあるいは標準パターンとして考えられる時系列パターンの種類を示している．**表5**は，それぞれの組合せに対する長所と短所をまとめたものである．

不特定話者を対象とする単語音声の認識システムでは，話者によらない標準パターンの作成が可能なものはf型のみである．また，音声パターンを音韻系列に変換する（F型）には複雑な処理を経なければならないから，情報の損失や認識時間にやや不利な点があるが，一旦，音韻系列に変換されると単語への変換過程はあまり時間を要しない利点がある．それゆえ，語彙数が増大した場合（数百単語）のメモリ量や処理量を考慮すれば，不特定話者を対象とする単語認識システムは F-f 型が（その変形も含めて）最も有望と考えられる．

3.2 単語音声の大分類と認識時間の短縮

認識対象語彙数が増大すれば，並列処理でない限り当然認識時間も増大する．そこで，認識時間を短縮するために，認識対象語彙の中から，入力単語と類似している少数の候補単語だけを予備選択することが考えられる．この予備選択は，認識時間を短縮することが目的であるから瞬時に実行できるものでなければならない．文字認識の分野では，この種の数多くの研究があるが音声認識の分野ではほとんどない．我々は，この目的に適した単語音声の局所的特徴と大局的特徴を導入し，その有効性を検討した．それぞれの特徴パラメータは，予備選択アルゴリズムの高速化のためすべて2値パターンで表現されている．

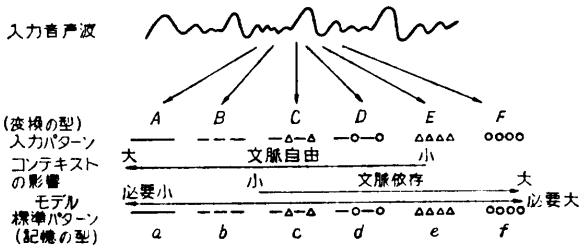

図2 パターン・マッチング法における入力パターンと標準パターンの種類

Fig. 2 Kinds of input pattern and reference pattern on pattern matching.

— フレーム単位の特徴パラメータで記述された時系列
△ 特徴パラメータで記述されたセグメント
○ 各セグメントを音韻に対応させたもの
おおまかにいえば，$A(a)$ は入力音声全体を，$B(b)$ は音節単位の系列を，$C(c)$，$D(d)$ は過渡音と定常音の系列を，$E(e)$，$F(f)$ は音韻の系列を表わす．

表5 単語音声認識手法の分類と特徴

Table 5 Classification and characteristics of spoken word recognition methods.

方法	長所	短所
A-a	限定話者に対しては識別能力大　アルゴリズムが簡単　ハード化が容易	必要な記憶容量大　処理量大　個人差の学習が困難
A-b B-b	限定話者に対しては識別能力かなり大	必要な記憶容量やや大　処理量やや大　個人差の学習が困難　(セグメンテーション必要)
A-c C-c D-d	A-a と F-f の両者の長所	A-a と F-f の両者の短所
A-e E-e	必要な記憶容量小　個人差の学習がやや容易	調音結合の影響がやや大　(セグメンテーションが必要)
E-f	最も一般的（学問的価値大）　必要な記憶容量小　個人差の学習が容易　言語情報の導入が容易	調音結合の影響が大　セグメンテーションが必要

（a） 単語音声の局所的特徴

単語の頭部および尾部のスペクトル（20 チャネル・1/4 オクターブのフィルタ群による分析によって得る）を単語音声の局所的特徴を表わす特徴として用いる．頭部では，エネルギーレベルがある閾値を越えたところから8フレーム分 (80 ms) のスペクトルを取り正規化した後，子音部と母音部を分離するためにスペクトルの過渡部で二分し，それぞれ平均した後，ある閾値との大小比較により各チャネルごとに二値化し，40 ビットの 0, 1 列に変化する．この場合，母音や無声摩擦音で始まる単語は，8フレームとも同一の音韻であるが，同じように二分している．

尾部では，エネルギーレベルが上述の閾値より小さくなる直前の6フレーム分を取り正規化した後，チャネルごとに平均して 20 ビットの 0, 1 列に変換する．

頭尾部は，単語の発声速度に関係なく正確にとらえることができ，しかもコンテキストの影響が少ない．これらのパターンは，認識対象語彙に関係なく日本語の全単音節からあらかじめ作成できる．

（b） 単語音声の大局的特徴

表 6 単語音声の大分類
Table 6 Clustering of spoken words.

(a) 音韻数差の分布 (%)

音韻数差	0	1	2	3	4	5	6	7	8以上
比率	17.7	25.9	21.7	16.9	10.6	4.0	2.1	0.7	0.4

(b) 音韻列パターンの距離差分布 (%)

距離	0	1	2	3	4	5	6	7	8	9	10	11以上
音韻の有無	4.1	13.4	24.1	25.9	19.3	9.7	2.9	0.6	0.1			
音韻数考慮	1.3	1.0	3.9	8.6	14.2	16.5	18.4	14.1	11.5	5.8	3.3	1.5

(c) 頭尾部スペクトル二値パターンおよび音声パワーレベル変化パターンの距離差分布 (%)

距離	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24以上
頭部スペクトル	4.6	0.3	0.2	0.1	0.7	1.4	1.2	1.7	1.6	2.7	2.5	2.6	2.3	2.3	6.0	4.5	1.4	6.4	1.9	6.0	7.3	6.6	9.1	5.6	21.0
尾部スペクトル	9.0	6.0	2.6	5.8	2.7	2.4	5.1	6.6	7.0	8.1	8.0	3.5	4.9	10.5	11.5	5.4	0.9	0.0	0.0	0.0	0.0				
音声パワー	1.0	0.1	0.1	0.6	1.1	2.4	3.2	4.9	7.7	9.4	10.7	11.7	11.4	10.8	8.7	6.4	4.3	2.8	1.6	0.7	0.3	0.1	0.0	0.0	0.0

(i) 単語を構成する音韻数

発声速度のいかんにかかわらず，発声された音韻数は常に一定であるから，実際に識別された音韻数によって発声された単語を推定する．このとき，システムのセグメンテーション能力が不完全であることに注意を要する．

(ii) 音韻列の全体的特徴

システムによって識別された音韻列中に，各母音，無声子音，有声子音が含まれているかどうかで，音韻列の全体的特徴を記述し，8ビット（5母音に5ビット，無声子音に2ビット，有声子音に1ビット）の0,1列に変換する．あるいは，もう少し厳密に各音韻が存在する個数まで考慮する（**表 6**(b)の下段）．このとき，音韻の出現順序は無視する．

(iii) エネルギーレベル変化の全体的特徴

発話区間を 14 区間に分割し，各区間に2ビットを割り当て，各区間でのエネルギーレベルが上昇か下降か，またエネルギーレベルの最大値の上半分にあるか下半分にあるかをそれぞれ 0,1 で表わし，全体で 28 ビットの0,1列で表わす．

(c) 100 都市名により実験結果

単語音声の頭尾部のスペクトル二値パターンとエネルギーレベル変化パターンは，男性8名にそれぞれ日本語単音節と100都市名を発声してもらい作成した．表 6(a) はすべての二単語対の組合せでの音韻数差の分布，表 6(b)は，音韻列パターンによるハミング距離差の分布(上段)と各音韻の個数まで考慮した（例えば，/a/ が2個ある→0011，無声子音が3個ある→0111；撥音・半母音は有声子音とみなした）距離差の分布，表 6(c)は，頭尾部のスペクトル二値パターンの距離差分布とエネルギーレベル変化パターンの距離差分布を示す．これらのパターンは，話者による個人差があるため，認識システムに組込む際には，余裕をもったしきい値で利用する必要がある．実際の実験では（用いたしきい値は表6の縦棒で示す），候補単語を 1/10 に絞ることができ，この予備選択に要した時間は数ミリ秒程度であった[16]．

表 7 語彙数・単語長等と単語認識率との関係
Table 7 Relationship between size of vocabulary or word length and word recognition rate.

語彙	語彙数	平均単語長	単語長の分散	(脱落確率，挿入確率) % (10,10)	(0,10)	(10,0)	(0,0)	認識困難度
都市名	10	6.8	2.9	92.5	99.7	91.5	98.5	0.080
	20	6.8	2.9	88.7	99.7	87.1	98.6	0.104
	30	6.8	2.9	85.2	98.6	83.2	97.4	0.118
	50	6.8	2.9	81.5	97.6	80.2	96.0	0.136
	70	6.8	2.9	78.0	97.1	77.0	95.7	0.148
	100	6.8	2.9	77.4	96.3	74.9	94.6	0.160
	30	4.4	0.6	76.0	95.0	75.5	94.0	0.161
	30	6.5	0.3	84.0	97.5	83.0	97.8	0.133
	30	8.6	1.0	87.3	98.0	86.7	98.0	0.109
数字	10	3.1	0.7					0.128
50音	50	1.9	0.1					0.274

音韻認識率：母音 80%，有声子音 60%，無声子音 80%

3.3 認識対象語彙と単語認識率との関係

単語の認識率は，一般に認識対象語彙数の増加とともに減少するが，語彙数だけでなく，単語間の音響的類似性にも大きく依存する．認識対象語彙の認識困難度を定量的に表現する方法が二，三試みられてはいるが[19),20)]，いずれも音韻（系列）間の類似性の与え方に依存し必ずしも客観的とはいえない．

我々は，少し観点を変えて，認識困難度の要因として，語彙数，平均単語長（音韻数）およびその分散を考え，2.2 節で述べたシミュレーション実験によって

検討した．ここでは音韻認識率を固定し（母音＝80% 有声子音＝60%，無声子音＝80%），脱落確率，挿入確率のみ 0% と 10% の2通りで実験した．種々の語彙数と100都市名を単語長によって3グループに分けて行った実験結果を表7に示す．予想されるように，語彙数の増加，平均単語長やその分散の減少が単語認識率を減少させる要因であることがわかる．表7から，対象語彙の認識困難度は，語彙数を n，平均単語長を m，その標準偏差を σ とすれば，次式で近似できることがわかる（ただし，この場合は，単語間の音響的な類似性を考慮しておらず，語彙がランダムに選ばれていると仮定している）．

$$\text{単語認識困難度}=\log n/(a+b\cdot m+c\cdot\delta)$$

a, b, c は係数

これは，単語を連結し，2連続単語の認識を考えると理解できる．つまり，2連続単語を新たな語彙と考えると，その平均単語長は $2m$，標準偏差は $\sqrt{2}\sigma$，語彙数は n^2 となり，もとの認識困難度より，困難になる．（注：連続単語はランダムに選ばれた孤立単語とはみなし難い）実験結果から，$a=4$，$b=c=1$ で大体の目安を与えると考えられる．これから，10 数字 $n=10$，$m=3.1$，$\sigma=0.7$）の認識は，大体 40 語彙の都市名と同程度の困難度といえる．

4. 連続単語音声の認識

4.1 連続単語音声の認識実験

本節では，連続単語音声全体を一つのパターンとみなして，孤立単語の認識法をそのまま適用する方法には限界があることを指摘し，単語境界の検出結果をなんらかの形で利用することが必要であることを述べる．

2.2 節で述べた手法を用い，2連続都市名単語の認識実験を行った．100 都市名から 10 都市名を選び，これからすべての可能な2連続単語 100 個を作り，これらを新たに孤立単語とみなし次の3種類の認識実験を行った．

i）入力単語が2連続単語とわかっている場合（語彙数 100）

ii）入力単語が孤立単語か2連続単語のどちらかだとわかっている場合（語彙数 110）

iii）単語境界が正確にわかっている場合（語彙数 10）

音韻の認識率は固定し，脱落と挿入確率のみ変化させて実験した（3.3 節と同じ値）．結果を表8に示す．

表 8 シミュレーションによる連続都市名の認識（平均単語認識率）

Table 8 Recognition of two connected city names by simulations (recognition rate/word).

認識条件	（脱落確率，挿入確率）%			
	(10,10)	(0,10)	(10,0)	(0,0)
2連続単語だとわかっている場合	91.5	99.6	90.3	99.3
2連続単語だとわかっていない場合	91.0	99.6	89.6	99.3
単語境界がわかっている場合	95.0	100.0	93.0	100.0

音韻認識率：母音 80%，有声子音 60%，無声子音 80%

表から明らかなように，連続単語を孤立単語とみなして認識する手法には限界があることがわかる．これは，単語境界が明確でないためにあいまい性が増大するためである．

4.2 単語境界の検出

単語境界の検出が必要であることは次のことを考えれば十分であろう．“yamatokōriyama” と発声した場合を我々が聞いた場合，それは「大和」と「郡山」の2連続単語か「大和郡山」の1単語かの区別がつく．これは，我々が韻律情報と呼ばれるピッチパターンや振幅，継続時間長などの情報を手がかりとして単語のまとまりを知覚しているからである．そこで我々は，ピッチパターンと振幅（音声パワー，エネルギーレベル）パターンのみで，どの程度単語のまとまりが知覚できるかを調べた．

話者2名がランダムに，1単語，2連続単語，3連続単語を発声し，その音声からピッチパターンと音声パワーを抽出した．正弦減衰波をそのピッチパターンに対応するように伸縮しながら繰り返し，また，その振幅を音声パワーに対応させて，もとの音声の韻律情報だけを保存した擬似音声を作った．すなわち，音韻性は除去され「ウー」に近い連続音に変換した（無声音の区間は，白色雑音で近似した）．これを7名の被験者に聞かせ，その擬似音声が何単語からなっているかを判定してもらった．単語として，10 数字と 100 都市名を選んで実験したところ，いずれも 95% 程度の正確さで単語数を知覚することができた．このことは，韻律情報が単語のまとまりを与える重要な手がかりとなっていることを示している．

この事実に基づいて，我々はピッチパターンと音声パワーだけを用いた単語境界の検出アルゴリズムの開発をすすめている．現在のところ，10 数字の連続音声で，単語境界検出率は約 90% である[21]．韻律情報は，文音声の認識のときには，文節境界の検出[22]や文

型の推定に用いることができるといわれてきた．これらのことは，現在知覚実験と自動認識の両方向から検討をすすめている．

4.3 木探索法

単語境界を完全に正しく検出することは，実際的には期待できず，単語境界の候補点を検出しておいて，単語の認識への評価に利用するのが最も妥当であると思われる．そこで，一般に連続単語の認識は，可能な単語列の集合となり，それは木（ツリー）で表現できる．これらの集合のうち，最も評価基準の高い単語列を認識結果とするのが普通である．しかし，その単語列の数は莫大になるのが普通で，すべてを評価するのは得策でない．それゆえ，能率よく可能性のある単語列のみ探索する必要がある．これは木探索の問題に帰着できる．人工知能研究では，種々の木探索法が提案されているが，音声認識に利用できる，すなわち，最も評価基準の高い単語系列を能率よく見い出す探索法はない．我々は，あらゆる可能な系列をすべて展開することなく，準最適な系列を能率よく導出する並列展開型木探索法を開発した[23]．

（a） 並列展開型木探索法（$\alpha-\beta-\gamma$ 法）

木探索の問題を次のように設定する．入力が l 個のシンボル系列よりなっているとする．そのとき，1番目から l 番目までの各シンボルに対して，その場所に出現しうるすべてのシンボルに対し，そのシンボルが存在する可能性が得点で与えられているとする．このとき，1番目のシンボルから l 番目のシンボルに至る道のうちで，最も得点合計（または平均値）の高いものを入力系列とする．

並列展開型木探索法で使用するパラメータ（α, β, γ）とアルゴリズムの手順を以下に示す．

α：ある段階で同時に展開する節点の個数

β：1つの節点から展開されたものの中で残す節点の個数

γ：木の展開途中で残しておく節点の総数

ステップ 1：記憶しておく節点の集合に根節点を入れる．

ステップ 2：記憶されている節点の中から，その節点に至る道の平均得点の高いものから α 個展開する．展開しうる節点が α 個未満なら全部展開する．展開すべき節点が終節点ならば，ステップ5へゆく．

ステップ 3：ステップ2で展開した各節点から生じた新しい節点のうち，もとの節点ごとに β 個得点の高いものを記憶しておく節点集合に入れる．β 個未満ならば，すべて節点集合に入れる．

ステップ 4：記憶している節点の個数が γ 個を越えていれば，そのうち各節点に至る道の平均得点の高いものから γ 個選び，ほかの節点は消去してステップ2へゆく．

ステップ 5：根節点からその終節点に至る道が，この場合の入力系列とみなされる．終了．

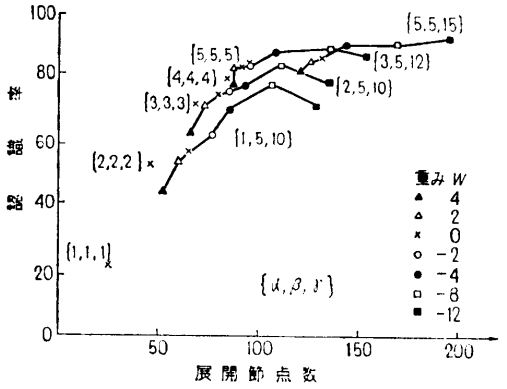

図 3 文認識率と展開節点数の関係
Fig. 3 Relationship between recognition rate and number of expanded nodes.
木の深さ＝5，分岐数＝5

（b） シミュレーションによる実験結果

木の深さ＝5，各節点での分岐数＝5 の木における各節点の得点（0〜200）を一様乱数によって生成し，100 個の木を作った．探索中の部分系列の評価基準として，根節点から各節点に至る道の平均値として重みを考慮したもの（展開コスト）も実験した（W＝重み係数，n＝シンボル数）．

修正平均得点＝平均得点＋$W\times n$

種々の $\{\alpha, \beta, \gamma\}$ に対する認識率（最高得点の得る道を見い出した割合）と節点展開個数の関係を**図3**に示す．図からもわかるように，従来の1つずつ展開していく（$\alpha=1$）方法よりも並列して複数個同時に展開していく方がすぐれていることがわかる．また当然の結果として，W が負なら，認識率と節点展開個数は増加し，正なら減少する．これらの結論は，実際に算術文の音声認識に適用した実験でも確かめられている[6]．

5. 文音声の認識

5.1 文解析手順

システムの評価として，単語認識システムでは，認識対象語彙の認識複雑度の量的表現が重要であったのと同様に，音声理解システムでは，タスクの複雑度の

量的表現が必要である．Goodmanは，平均分岐数 (average branching factor) と探索空間の大きさ (search space size) なる概念を導入し，タスクに用いられる文を生成する文法の複雑度を定義した[20]．しかし，この定義によるタスクの複雑度は，文の解析方向によって異なってくる．我我は，この事実を逆に利用し，与えられたタスクに対する最適な解析方向があることを示す．まず初めにGoodmanが導入した諸量の定義を述べる[20]．

平均文長 ASL (average sentence length)

$$ASL=\sum_{t}\sum_{s\in NFS} p(s/t).$$

動的平均分岐数 $DABF$ (dynamic average branching factor)

$$DABF=2^{\{\sum_{t}\sum_{s\in NFS} p(s/t)\cdot\log BR(s)\}/ASL}.$$

探索空間の大きさの対数 $LSSS$ (logarithm of search space size)

$$LSSS=ASL\times\log_2[DABF].$$

ここで，$NFS=\{s/NEXT(s)\ni\varepsilon\}$ で，$NEXT(s)$ は状態 s に後続する状態の集合を示す．$BR(s)$ は，状態 s での分岐数を示し，$p(s/t)$ は t 個の単語列を生成して状態 s に達する確率を示す．各分岐点での分岐確率は，タスクによって定義する必要があるが，等確率に選ぶのが最も簡単であり，実際の認識システムでも可能なものはすべて等確率として扱うのが普通である．

図4に簡単な日本文の例を示す．この例について，文解析方向別の上の諸量を求めたものが**表9**である．表より，この例では右から左方向へ解析，すなわち文末から文頭に向って解析していく方が，左から右への解析方向に比べて能率的であるといえる．テキスト入力の場合は，解析方向は処理能率の差として現われるが，音声理解の場合は，探索空間の減少は，処理能率の向上のほかに認識率の向上にもつながる．**表10**は，4.3節で述べたシミュレーションの手法を用いて種々の木の大きさと認識率との関係を示したものである．ただ，文末から文頭への解析は，単語の予測アルゴリズムやほかの種々の言語知識の導入による文解析などを考慮するとやや複雑になりあまり得策でない．その

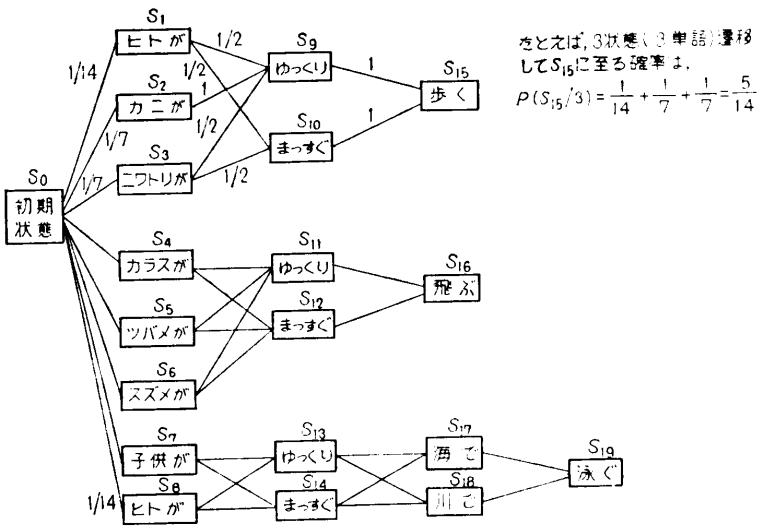

図4　簡単な日本文を生成するネットワーク表現
Fig. 4　Network representation of generation of simple Japanese setences.

表9　解析方向別の文法（タスク）の複雑度
Table 9　Degree of complexity of the grammars (tasks) corresponding to various parsing directions.

タスク	解析方向	ASL	$DABF$	$LSSS$
図 4	左→右	3.24	2.45	4.17
	右→左	3.33	2.26	3.94
	述語同定後，左→右	3.33	2.34	4.09
計算機網	左→右	9.31	2.89	14.27
	述語同定後，左→右	9.36	2.76	13.73

ASL：平均文長，$DABF$：動的平均分岐数
$LSSS$：探索空間の大きさの対数

表10　探索空間の大きさと認識率との関係
Table 10　Relationship between search space size and recognition rate.

木の深さ ASL	分岐数 $DABF$	$LSSS$	並列型木探索法 (α, β, γ)			
			(1, 5, 10)	(2, 5, 10)	(3, 5, 12)	(5, 5, 15)
4	5	9.3	77%	81%	88%	88%
5	5	11.6	58	74	80	83
6	5	13.9	64	70	75	81

(4.3 節のシミュレーションによる実験)

点，述語を同定してから，その文型に対応する文法を用いて左から右へ解析を進めていく方法は，日本語の場合，述語が語尾にあり同定しやすいことを考えれば好ましい方法といえる．もちろん，タスクによって多少変わるが，通常の場合は述語を先に同定する方がよい．実際，我々が実験に用いた「計算機網」と称するタスクにおいても上述のことが確められている（表9

表 11 音声理解システム LITHAN のシステム評価
Table 11 System evaluation of the LITHAN speech understanding system.

タスク			数字	算術文	カレンダー	計算機網
語彙数			10	24	30	101
平均分岐数（静的）			10	10	8	5
話者数			20	5	12	10
実験に用いた延べ文数				80	120	200
実験に用いた延べ単語数			1500	560	647	1983
用いた言語情報別認識率（%）	音響	文		20	8	—
		単語	97	78	54	—
	音響・構文	文		71	40	58
		単語		94	83	91
	音響・構文 意味・プラグマティクス	文			70	66
		単語			90	93
音韻認識率（%）	母音・撥音		76(85)	86(92)	70(80)	74(88)
	半母音		12	30	15	10
	有声子音		10(14)	34(42)	22(37)	28(42)
	無声子音		65(87)	81(90)	64(82)	44(63)
有声／無声分離率（%）			95	97	97	95
有声子音検出率（%）			29	53	50	55
発声スピード（音韻／秒）			8.9	9.4	10.7	10.4
(α, β, γ)				(5, 5, 15)	(5, 5, 5)	(2, 5, 10)

参照）．英語の場合は，述語が文の前部分にあるので，左から右への解析の方が，右から左への解析よりも能率的である[24]．

5.2 言語情報の有効性

音声理解システムの研究成果の一つとして，言語情報の利用が音声認識システムにとって重要であることを示したことがあげられる．しかし，構文情報や意味プラグマテックス情報が，実際どの程度認識率に寄与したかの実験例は少ない．

表 11 は，我々が開発した音声理解システム LITHAN を種々のタスクに適用した結果をまとめたものである．“数字”は孤立数字音声の認識，“算術文”は，たとえば「11+2×3（ジューイチタスニカケルサン）」のような算術文，“カレンダー”は「1月1日月曜日（イチガツツイタチゲツヨービ）」のような月日と曜日，“計算機網”は「計算機言語の磁気ドラム装置から磁気ドラムをはずせ」のような計算機網への指令および状態問合せである．音声認識率の括弧内の数字は，第二候補まで含めた場合を示す．発声スピードは，1秒間に発声された平均音韻数を示している．表より明らかなように，言語情報，特に構文情報の認識率への貢献度が大きいことがわかる．表 11 は，今後の音声理解システムを開発する際の一つの目安を与えるであろう．

6. むすび

本論文は，音声自動認識に関する諸問題に対して情報工学的な考察を加えたものである．ただ紙面の都合で，音声スペクトルの特徴分析[8]や有声子音の知覚実験[25]に基づく検討，個人差の正規化や学習[7]，調音結合[7]等の問題については省いた．本研究で得られた成果についてまとめると，

① 不特定話者を対象とする音声認識では，まず発声者の音声生成構造の違いに基づいて，話者を3クラス程度（子供，成人男性，成人女性）に分類すべきである．以後クラス内で，話者の正規化[26]あるいは学習によって個人差を除去する方法が最適だと思われる．

② 連続音声中の母音を静的な特徴で認識するためには，各話者，各母者に対して少なくとも2つの標準パターンを設ける必要がある．

③ 音韻識別では，脱落誤りが単語識別率に与える影響が最も大きく，挿入誤りの影響は小さい．

④ 音韻識別部の出力形式として音韻ラティスが有効であり，しかも各音韻の信頼度や継続時間長を利用する方がよい．

⑤ 不特定話者の数百単語の単語認識システムでは，音韻認識を介する手法が有望である．

⑥ 認識時間の短縮のために，認識対象語彙から入力単語と類似している単語候補を予備選択することが有効である．単語音声の大局・局所的特徴を使用することによって，1/10 以上の時間短縮が期待できる．

⑦ 限定語彙の単語認識困難度は，認識対象語彙の単語間の音響的類似度に依存するが，単語がランダムに選ばれているとすれば，語彙数の対数に比例し，平均単語長とその標準偏差の和に反比例する．

⑧ 連続単語音声の認識では，単語境界を検出し，これをなんらかの形で利用することが必要である．

⑨ 韻律情報は，連続単語音声では各単語のまとまりを知覚させる役割を果しており，これらを用いることにより単語境界の検出が可能である．

⑩ 単語列集合から，最も評価基準の高い単語列を見い出す方法として並列展開型木探索法（$\alpha-\beta-\gamma$ 法）がすぐれている．

⑪ 文解析の方向によって探索空間の大きさが変わる．日本語文を扱うタスク内では，述語を同定してか

ら，文頭より文末に向って解析を進めていくのがよいと考えられる．

⑫ 実用に耐えうる音声理解システムを開発するには，音韻識別部の能力を向上させる必要があり，有声／無声の分離を完全（95% 以上）にし，有声子音の検出率60%以上，音韻識別率80%程度を得る必要がある．

⑬ いずれのタスクにおいても，言語情報の利用は大変有効である．特に構文情報の貢献が大きい．

謝辞 本論文中の実験的研究は，京都大学情報工学教室坂井研究室の音声グループの皆様，特に，白方博教氏，山尾雅利氏，内海雅行氏，水上治雄氏，寺西功氏，中瀬弘巳氏，森元学氏のご協力を得ました．ここに感謝します．

参考文献

1) 中川聖一：機械による会話音声の認識・理解研究の動向，情報処理，Vol. 19, No. 7, pp. 675-685 (1978).
2) Newell, A. et al.: Speech Understanding Systems: Final Report of a Study Group, North Holland (1973).
3) Lea. W. A. and Shoup, J. E.: Gaps in the Technology of Speech Understanding, Conference Record of International Conference on ASSP, pp. 405-408 (1978).
4) Pierce, J. R.: Whiter Speech Recognizer? JASA, Vol. 46, No. 4, pp. 1049-1051 (1969).
5) 坂井，中川：不特定話者・連続音声向き単語音声の識別，情報処理，Vol. 17, No. 7, pp. 650-658 (1976).
6) Sakai, T. and Nakagawa, S.: A Speech Understanding System of Simple Japanese Sentences in a Task Domain, 電子通信学会論文誌，Vol. E 60, No. 1, pp. 13-20 (1977).
7) 中川，坂井：個人差の種々の学習機能をもつ実時間単語音声識別システム，電子通信学会論文誌，Vol. 61-D, No. 6, pp. 395-402 (1978).
8) 中川，坂井：日本語音声スペクトルの特徴分析および音声認識・話者認識への考察，日本音響学会誌，Vol. 35, No. 3, pp. 111-117 (1979).
9) Fant, G.: Speech Sounds and Features, MIT Press (1973).
10) Fujimura, O.: Syllable as a Unit of Speech Recognition, IEEE Trans. Vol. ASSP-23, No. 1, pp. 82-87 (1975).
11) 中津，好田：VCV 音節を単位とした単語音声の認識，電子通信学会論文誌，Vol. 61-A, No. 5, pp. 464-471 (1978).
12) 阿部，秦野，福村：辞書を利用する文字認識系の能力の評価，電子通信学会論文誌，Vol. 52-C, No. 6, pp. 305-312 (1969).
13) 板橋，鈴木，城戸：単語中の幾つかの子音の辞書による識別，電子通信学会論文誌，Vol. 54-C, No. 1, pp. 10-17 (1971).
14) 牧野，城戸：近距離単語間の識別に必要な音素対の性質，電子通信学会論文誌，Vol. 62-D, No. 8, pp. 507-514 (1979).
15) 関口，重永：グラフを利用して誤りを含む記号列を分類する方法とその音声認識への応用，情報処理，Vol. 19, No. 9, pp. 831-838 (1978).
16) 中川，内海，坂井：単語音声の大局・局所両特徴による前照合法と個人差の学習法，日本音響学会音声研資 S 78-23 (1978).
17) Woods, W. A.: Motivation and Overview of SPEECHLIS—An Experimental Prototype for Speech Understanding Research, IEEE Trans. Vol. ASSP-23, No. 1, pp. 2-10 (1975).
18) 鹿野：音韻ラティスの評価システム，電子通信学会論文誌，Vol. 63-A, No. 3, pp. 181-188 (1980).
19) Moore, R. K.: Evaluating Speech Recognizers IEEE Trans. Vol. ASSP-25, No. 2, pp. 178-182 (1977).
20) Goodman, R. G.: Analysis of Languages for Man-Machine Voice Communication, Ph. D. thesis, Stanford University (1976).
21) 中川，中瀬，坂井：連続数字音声中の単語境界の検出，日本音響学会春季大会，3-2-10 (1979).
22) 三嶋，浮田，中川，坂井：日本語文音声の韻律情報による文節境界の自動検出法，日本音響学会春季大会，3-2-11 (1979).
23) Nakagawa, S. and Sakai, T.: A Parallel Tree Search Method, Proceedings of the 6th International Joint Conference on Artificial Intelligence (1979).
24) Nakagawa, S. and Sakai, T.: On Parsing Direction and Tree Search in the LITHAN Speech Understanding System, Joint Meeting of ASA and ASJ in Honolulu (1979).
25) Nakagawa, S. and Sakai, T.: Some Properties of Japanese Sounds through Perceptual Experiments and Spectral Analysis, Studia Phonologica, Vol. XI, pp. 48-64 (1977).
26) 中川，神谷，坂井：単語音声スペクトルの個人差に基づく単語音声の正規化，電子通信学会パターン認識と学習技報，PRL 79-62 (1979).

(昭和54年8月31日受付)
(昭和55年6月19日採録)