学 位 論 文

音声を媒体とした情報の受容に関する基礎研究

平成6年12月

峯 松 信 明

学位論文

音声を媒体とした情報の受容に関する基礎研究

1994 年 12 月 20 日 提出

指導教官 広瀬 啓吉 教授

工学系研究科 電子工学専攻 27123

峯松 信明

目次

図目次

表目次

第 1 章

序 論

人間同士が行なう情報交換の媒体として音声の占める役割は非常に大きい。もちろん音声以外にも，視覚・触覚・嗅覚などの感覚器官を使用した情報交換も行なわれているが，送信及び受信という情報交換の両側面における効率，容易さ，及び伝達可能な情報の種類の豊富さ (音響的情報，言語的情報，感情/意図に関する情報，性別/話者性に関する情報など) を考えると，音声が人間にとって最も基本的/根源的な情報交換手段であると言って良いであろう[1]。この音声の音響的および言語的構造は，人間の知的活動と深い相関を持っており，文化や社会の発展に密接に結び付いている。事実，世界的に文化の高度に発達した地域は電話網の発達した地域とよく一致しているとの報告もある[1]。

この音声を媒体とした情報伝達を，人間を中心に，非人間間との情報伝達をも含めて考えると，概略的には図 1.1 のように考えることができる。この図に示される "人間" が関与する情報交換において，人間による情報の送信は "音声の生成"，人間による情報の受信は "音声の知覚" と呼ばれることが多い。この両者における処理過程の分析/解明は医学・生理学・心理学・哲学・認知科学・理学・工学など種々の分野の研究対象となっている。更に近年では，大学院大学などにより，各分野間の境界領域を埋めるべく研究環境も積極的に整えられつつある。

一方，近年の高度な計算機技術の発展にも支えられて，従来ディスプレイやキーボードなどの装置を通して行なわれてきた "人間対機械" 間の情報交換を，音声を媒体として行なうことは出来ないかと，長年に渡り研究が行なわれてきた。キーボードのような装置と異なり，(母国語の) 音声を媒体とすることが出来れば，機械への情報入力に関しては何らの練習も不要となる。更に当然のことながら，機械への情報の送信及び機械からの情報の受信に利用される感覚器官は主に「口，耳」だけとなり，手・足などの異なる器官を用いた作業を並行して行なうことも可能となる。このように，音声を媒体とした情報交換手段は，人間にとって非常にタスクの小さな，いわゆる，"楽な" 手段であることは明らかである。この音声を媒体とした情報交換の実現を目指し，機械に音声を送信させる "音声合成" 技術 (図 1.1 参照)，機械に音声を受信させる "音声認識" 技術 (図 1.1 参照)，及びこれら両者を支援すべく，"音声分析" 技術を大きな 3 本の柱として，種々の技術が開発され，現在に至っている。

本研究は，以上のような性質を有する音声を媒体として情報が送信され[1]，人間或は機械がその情報を受信する場合，各々,「どのような処理が行なわれているのか」そして「どのような処理を行なわせれば良いのか」と言う観点に立ち，"人間" 及び "機械" における

[1] 送信側としては，人間を前提としている。

音声を媒体とした情報の受容に関する一連の基礎研究を行なったものである。その結果，本研究の内容は大きく，

- 人間による音声の知覚過程の分析とそのモデル化 (研究 **A**)
- 計算機による音声の認識手法の高精度化 (研究 **B**)

の 2 つに分かれる。両者の接点を十分に解明できれば，学術的には非常に興味深い結果が得られると考えられる。本研究でも試験的にではあるが，両者を結び付ける試みを行なっている。なお，上記しているように，本論文では混乱を避けるため，原則的に，人間による音声を媒体とした情報の受容を "知覚" と呼び，機械による音声を媒体とした情報の受容を "認識" と呼ぶことにする。但し，混乱を来たさないと十分に判断される箇所は，その範囲内では無いことを断っておく。

音声認識技術に目を向けてみると，近年の計算機処理能力の驚異的な発展，及びデジタル信号処理技術の向上に支えられて，その処理技術は飛躍的に改善された。日本・アメリカ・ドイツ間での翻訳電話の実験試行も記憶に新しいところである[2]。しかし，これらの技術を実用面から考えてみると，残念ながら現状では十分に社会的要求を満足させるだけのレベルに達しているとは言い難い。現在までに公開された認識システムを概観すると，話者の限定，性別による認識能力の差，語彙・統語構造の制限，発話環境の整備，認識 (及び応答) 速度等，ユーザーに課す負担はまだまだ重い[2][3]。

音声認識の初期の研究においては，同一ラベル (音素) の音声には必ずある一定不変の音響的特徴が存在するとの仮定の下，各ラベル毎の音響的不変量が追及された。しかし，人間が発声する音声パターンは，同一ラベルのものでも種々の要因により，絶えず変動しているのが事実である。つまり，同一ラベルに対する音声が，環境によっては異なるラベルとして人間によって識別されたり，音響的には全く異なる音声が，環境によっては同一ラベルとして人間によって識別されることが報告されている[4]。この音響的変動が考慮されるようになって以来音声認識の研究は，如何にこの "ゆらぎ" に対処すべきかが，一大テーマとして扱われるようになった[5]。間もなくこの "ゆらぎ" は時間方向にも周波数方向にも広く存在することが認識されるようになり，時間方向の変動に対しては，標準パターンとの間で非線形な伸縮を行なう DP(Dynamic Programming[3]，動的計画法)[6]が音声認識に導入され，更に周波数方向の変動 (音声スペクトルパターンの変動) に対しては HMM(Hidden Markov Model，隠れマルコフモデル)[7]による数理統計的手法が導入され

[2] 当然，各システムの制約はこれらの一部である。
[3] DTW(Dynamic Time Warping) とも言われる。

た。このように音声特有の音響的変動に対処すべく，種々の方法が考案されてきた訳だが，研究 B においては，これらの手法の更なる高精度化を研究目的として，人間の音声知覚過程への考察を含む，種々の観点から実験的検討を行なう。そして，その結果を踏まえた上で新たな手法を提案し，その有効性について論じる。

一方，人間に目を向けてみると，音声という絶えず変動している信号を何の支障もなく，しかも，かなりの悪条件の下でも知覚している。逆に言えば，あまりにも容易に行なっているが故に，我々人間は，その処理の難しさに気付いていない，と言うのが実情であろう[8]。このような人間に対して当然のことながら，「如何なる処理方式を用いているのだろうか？」という疑問が生じる。従来から機械処理における，人間の音声言語処理過程の重要性は認識され，医学的・生理学的・心理学的・認知科学的・哲学的・理学的・工学的立場から，人間を対象とした様々な研究が行なわれてきた。更に最近では，上述した現在の音声認識技術の限界を憂慮する声もあり[9]，音声処理の唯一のお手本である人間をもう一度見直そうという動きがある。それに伴って，音声知覚過程に関する研究もより広く，深く，注目されるようになった。また計算機技術の向上に伴い，次世代の情報処理として，「感性情報処理」と言うテーマの下，上記した幅広い分野の研究者が一同に介した研究も進められ，幾つかの成果も報告されている[10]-[14]。

音声知覚研究は人間が無意識的に行なっている処理過程を，心理実験 (知覚実験) の結果を基に "目に見える" 形へと分析・解明しようというものである[8]。そのため，実験計画には事前の入念な議論が必要であり，現在までに種々の知覚実験方法が考案されている[15]-[18]。しかし，一般に行なわれている知覚実験 (方法) を見ると，被験者に対して非日常的なタスクを課しているが故に，人間の知覚能力の一端を観測する形になってしまっているものが少なくない。我々の思いも寄らぬような，人間の持つ知覚能力の解明も非常に重要なテーマの 1 つではある。しかし，我々が日常的に行なっている言語活動との関係を考えた場合，上記のような実験結果のみが考慮されるならば，人間における音声言語処理の全体像が，逆に歪んでしまったり，或は，ぼやけてしまうことが危惧される。そこで研究 A では，人間の音声言語処理過程をグローバルな観点から分析することを研究目的として種々の知覚実験を行ない，最終的には，それら実験結果から得られた知見を基に，人間における音声知覚モデルの構築を試みる。

以下第 2 章で，本研究の背景と位置付けについて考察すると共に，その意義についても明確にする。第 3 章では，人間の音声知覚過程に関する研究，特にその "モデル化"，と言う観点からの先行研究について概観する。そして，種々のモデルを比較・考察すること

で，本研究で実施すべき知覚実験対象，及び目指すところの知覚モデルについて，その方向付けを行なう。本研究は，従来当研究室で行なわれてきた研究の延長線上に位置するものである。そこで，当研究室で行なわれた先行研究について，その後の議論で必要になると思われる実験例を紹介し，残された課題についても触れる。特に，重要なキーになると考えられる，内部辞書の検索過程については独立に節を設け，深く考察する。第 4 章において，実際に筆者が行なってきた一連の知覚実験を，着目する音声長の短い順 (単語→文) で紹介する[4]。主に，内部辞書過程に関する話題が多くなっている。この中では，当研究室の先行研究では全く触れられていない，音声が伝搬する言語情報に着目した実験も行なわれている。第 5 章では，第 3 章及び第 4 章において得られた，人間の音声知覚過程に関する知見をまとめ，人間の全体像を見渡すことのできる知覚モデルの構築を試みる。そして，その工学的応用の可能性についても考察する。第 6 章では，第 4 章における知覚実験で使用する音声試料作成用に構築された，高品質音声分析合成システムについて述べる。第 4 章の実験の一部は，本システムの存在により可能となったものである。

第 7 章以降では，研究対象を「人間による音声知覚」から「機械における音声認識」へと変更して議論を進める。第 7 章では，従来行なわれてきた音声認識研究において，重要な役割を果たしてきた手法である，DP と HMM について，歴史的背景をも含めて紹介し，両者の比較を行なう。特に本研究において使用する HMM については詳しく述べることにする。そして，従来の研究において十分な議論が行なわれていないと考えられる 2 つの点—音声の音響的特徴表現方式の動的変化の可能性及び必要性/HMM 継続時間長モデルによる音声の時間構造記述の問題点—について考察する。第 8 章及び第 9 章において，上記した 2 つの点について実験的に分析し，各々において認識精度向上を目的とした新しい手法を考案する。そしてその有効性について音素認識実験を通して検証する。

最後に第 10 章において，「人間による音声知覚」，「機械による音声認識」と言う 2 つの柱を持つ本研究を再度概観・総括し，将来的展望を述べることで，本論文を締めくくることにする。

[4] 必ずしも，実験の実施順にはなっていないことを断っておく。

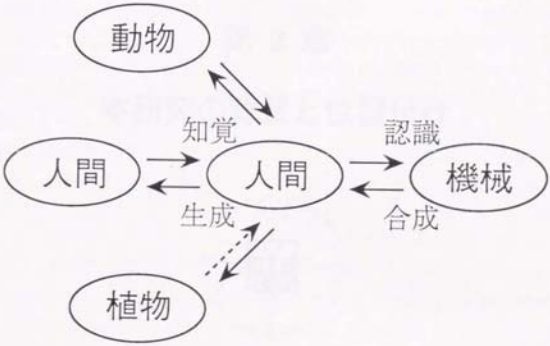


図 1.1. 音声を媒体とした情報伝達の概略図

第 2 章

本研究の背景と位置付け

第 1 章で述べたように本研究には，“人間による音声知覚過程の分析とそのモデル化” 及び “計算機による音声認識手法の高精度化” と言う大きな 2 つの柱がある。本章では，両研究を各々研究 **A** 及び研究 **B** と呼ぶことにする。さて，この研究 **A**，**B** は当然のことながら，各々，質の異なる背景を持っている。本研究の内容を詳説する前に本章において，両研究を取り巻く背景，及びその中における本研究の位置付け・意義について概説する。なお，各々の研究に直接関与する先行研究など，詳しい事柄については後ほど第 3 章及び第 7 章において述べることにし，ここではその概略を述べるにとどめる。

2.1 人間による音声知覚過程の分析とそのモデル化

一口に音声知覚過程の分析と言っても，その言葉が網羅する分野は非常に広く，様々な分野・観点からのアプローチが行なわれている。これらのアプローチを，対象とする人間の処理機構における，処理内容の複雑さ/高度さから大きく分類すると，次の 3 つに分けられると筆者は考えている。順に処理内容が低次から高次へと複雑化している。なお，数理学的にモデルを構築し，それを用いた計算機シミュレーション結果より，人間における処理過程を推察すると言う，言わば，間接的なアプローチも存在する。しかしここでは，人間に対して何らかの (音の) 刺激を与え，それに対する人間の反応を詳細に記録・観察し，得られた結果を基に人間における処理過程を分析・解明すると言う，より直接的なアプローチを考察対象とする。なお，以下の説明では，純音のような言語情報を含まない音を音響音，言語情報を含む音を言語音として区別して記述する。

1. 動物実験などを通して，聴覚における抹消神経系の挙動を観察し，種々の音響音，言語音に対する反応を見る医学的・生理学的なアプローチ。この分野の研究では，抹消神経系において，入力音に対して音響音/言語音の区別無く施される処理を詳細に観測することが頻繁に行なわれている[19]。そして，言語音が “音声言語としての” 処理を受ける以前の段階での処理内容を解明することをを目的の 1 つとしている[20]。
2. 種々の言語音を提示し，その言語的/辞書的意味・内容 (lexical meaning) の伝搬の様子に着目する認知科学的アプローチ。この分野で使われる刺激音は殆どの場合言語音である。しかし，必ずしも有意味な言語音のみを用いている訳ではない。即ち，音韻の知覚の様子[21]-[23]，或は，意味の有る音節列/無い音節列に対する知覚の様子の変化にも焦点が当てられている[24]。本アプローチにおける研究では，上記した有意味性 (内部辞書への登録の有無による影響) の他に，統語的整合性，意味的/談話的整合性の有無による音声知覚過程への影響などにも焦点が当てられ[25]-[27]，内部

辞書への検索処理過程[28]も一大テーマの一つとなっている。

3. 種々の言語音を提示し，その音声により喚起される感情やその音声が伝搬する発話者の意図に着目する心理学的アプローチ。この分野で使われる刺激音は多くの場合，有意味[1]な言語音である。このアプローチでは，刺激音を文字列に落した場合に，その文字列に含まれる情報以外の情報，即ち，音声であるが故に伝達可能となった情報 (超分節的情報) に焦点が当てられることが多い[29]-[31]。面白い研究例としては，まだ文字の読めないような幼児の，ある音声に対する感性と，同一音声に対して成人が示す感性とを分析・比較している研究例もある[32]。

筆者は人間対人間のコミュニケーションにおいて，メッセージの意味伝達に最も興味を持っており (上記 2.)，その結果，認知科学的な観点から人間を観察・分析することを研究 A における主な研究課題とした。音声を媒体としたこの認知科学的なアプローチは，注目する音声長の長さにより大きく，音素・音韻知覚/音節知覚/単語知覚/句・文知覚のように分類されている。

音声知覚過程に関する認知科学的な研究を国内で見てみると，音素・音韻/音節知覚に分類されるものが非常に多く，対象とする音声長が大きい場合でも，単語サイズの音声に関する研究に留まっているのが現状のようである。一方，海外に目を向けると，音素・音韻/音節はもとより，単語/句/文サイズの音声に着目した研究も盛んに行なわれている[25][26]。

当然のことながら着目する音声長が大きいほど，その処理過程は複雑となる。「小さな音声長の刺激に対する処理過程から順に分析/解明すべきだ」との議論もあるが，より大きな音声長での処理が小さな音声長での処理に影響を与えていると考えられる現象も報告されている[33]。更に，対象を小さな音声長での処理のみに限定してしまうと，その研究目的が，「人間における音声処理過程の解明」と言うより，「人間の音声処理能力の限界の解明」にあるようにも思われ，日常的に行なわれる音声言語活動とのズレが危惧される。筆者は，種々のサイズの音声長の研究は並行に行なわれ，相互の知見を組入れて進めるべきものであると考えている。そこで，国内での研究例の比較的少ない，単語以上の音声の知覚過程の分析を主目的とした (図 2.1 参照)。具体的には，

- 単語提示での単語知覚
- 文提示での文内単語知覚
- 文提示での文内単語知覚の相互作用 (文知覚)

[1] ここでは，ある感情が喚起されると考えられる音声を含めて，"有意味" と言う言葉を用いている。

に関する一連の知覚実験を行なった。そして，実験結果より得られた知見及び先行研究において得られている知見を基に，試験的にではあるが，人間による音声知覚モデルを構築する。詳しくは第 4 章及び第 5 章で述べる。

2.2 計算機による音声認識手法の高精度化

本研究のもう一つの柱である，機械への音声入力 (音声認識) 技術に関する研究 (研究 B) について，その背景と位置付け及び意義を概説する。音声認識の目的の一つは，音声信号に含まれる音韻性を正確に抽出し，“音声→テキスト” 変換の自動化を図ることである。音声認識に関する研究が開始された当初は，「同一シンボル (音素) に対応する音声の音響的特徴には，何らかの不変量があり，それを正確に抽出できれば音声認識は可能となる。」と考えられていた。しかし，研究が進むにつれ，同一シンボルに対応する音声でも，その音響的特徴[2]は時間方向及び周波数方向に広く分布する (揺らぐ) ことが明らかとなり，上記の「不変量の追求」から，「音響的特徴の “揺らぎ” に対する対処方法の追求」へと研究課題が変化して行った[5]。

音声認識手法はここ数年間で，DP から HMM へと，主流となる処理手法が変化してきたが，これは，上記した音声の “揺らぎ” のうち，どの “揺らぎ” に対処するのか，の変化であると考えることができる。両者の特徴を簡単に述べると，

DP　音声の時間的構造を直接反映した形で標準パターンを作成することができ，また，照合の際には入力音声と標準パターンを，条件付きで，非線形にマッチングすることとなり，時間方向の揺らぎに対しては柔軟に対応できる。しかし基本的には，標準パターンは少数 (あるいは 1 つの) 音声パターンから作成されるため，標準パターンの持つ周波数方向の揺らぎに対する記述力は，非常に貧弱なものとなる。

HMM　音声の時間方向の情報を圧縮した形で標準パターンが作成されるため，DP と比較して，時間方向の記述力が弱い。また，照合の際には，無制限での非線形マッチングを許すため，自由度が高くなり，時間方向での揺らぎに正確に追従することが困難となる。逆に，時間方向への圧縮の結果，周波数方向での分布の様子を統計的に表現することが可能となり，周波数方向への揺らぎに対しては柔軟に対応できる。

となり，どちらも一長一短である。しかし，HMM は DP と比較して認識時の計算量が少

[2] ここでは，単時間スペクトル包絡の時系列と考えてよい。

ない[3]にも関わらず，DP と同等或はそれ以上の認識結果を出している[34]。研究の方向性としては，DP に周波数方向の揺らぎの記述を組入れるか[35]，HMM に時間方向の揺らぎの記述を組入れるか[36]の 2 つの方向性がある。しかし，HMM で音響処理を実現した場合，言語処理も含めて，全てを数理統計的なパラダイムで記述できるようになり，実験結果の優位性の他に，この論理的な美しさから HMM の方が多く利用され，現在の主流を占めるようになった。HMM に対して時間方向の情報を導入させる方法としては，種々の方式が提案されてきたが，HMM と入力ベクトル系列を照合する際に得られる照合パスにおいて，その遷移が HMM の各状態にどの程度停留しているのかを予め学習しておき，それを認識時に利用する継続時間長制御の技術が最も広く使われるようになった[37]。

研究 B においても HMM を用いて，各音素別に音響モデルの作成を行なうが，以下の観点から，その高精度化を実験的に検討する (図 2.2 参照)。

- 音声の音響的特徴表現方式を動的に変化させた認識手法。
- 継続時間長モデルの時間構造記述力を向上させることを目的とした，学習データのクラスタリング手法。

前者は音声を工学的に扱う場合，どのような音響的表現方式を用いれば良いかという，HMM や DP と言った認識手法に拘らず存在する，基本的ではあるが非常に重要な問題に対する一つのアプローチである。研究 B においても，HMM を用いて各音素別に日本語音声の音響モデルを作成する訳だが，従来の方法では，如何なる音声事象 (イベント) に対しても，ある画一的な方法で種々の事象を捉え，記述していた。パターン認識の分野においては，入力信号を標準パターン P 及び Q と比較・照合する場合，入力信号の工学的記述方法は，異なる標準パターンに対して同一の方法を用いることが大前提となっている。また，標準パターン P との照合に着目した場合，その記述方法は時間に対して不変であることも前提となっている。しかし，HMM のような数理統計的は手法においては，照合スコア (尤度) は全て確率或は確率密度の形で算出される。すなわち，入力信号の記述方法を時間軸上で変動させた場合，或は，照合する標準パターンによって記述方法を変化させた場合でも，標準パターン P に対するスコア p と標準パターン Q に対するスコア q は直接比較可能な物理量となる。これはパターン認識を数理統計的な手法で実現する場合は，上記した前提を必ずしも必要としていないことを意味し，入力信号の工学的記述方法は時間軸上で動的に，或は，照合対象に依存して変化させることが可能であることを意味する。以上の考察の下，研究 B ではまず，音声の音響的表現方法を種々の観点から動

[3] 特に VQ を利用した離散 HMM の場合。

的に変化させて，自動認識することを試みる。そして，時間的に特性が変動する音声をより的確に記述すべく，より適切な音響的表現方法を実験的に検討する。なお，ここで提案する手法は，知覚モデルの知見を一部導入して考案された手法でもある[38]。

前述したように，HMM はその基本的構造だけでは，音声の時間方向の情報は圧縮した形でしか保有しておらず，時間方向の記述力に乏しいと言う欠点を持つ。これを改善すべく，"継続時間長制御" と呼ばれる技術が開発された。これは，HMM と入力音声との照合パスが，HMM の各状態にどの程度停留しているかを学習データを用いて予め学習・モデル化しておき，認識処理において，この継続時間長モデルを考慮しつつ，照合スコア (尤度) を計算 (修正) する手法を言う。この継続時間長モデルは非常に広く使用され，その有効性も数多く報告されている。しかし作成された継続時間長モデルと，対応する学習データの時間的構造との間にどの程度整合性があるのか，また，そこには音素依存性が観測されるのか，と言った観点からの議論が十分に行なわれていないように思われる。そこで研究 B における後者のテーマにおいては，まず，従来提案された継続時間長制御モデルによる，時間構造の記述の様子を音素別に詳細に観測する。そしてその結果を基に，継続時間長モデルの精度を向上させるべく，学習データの新しいクラスタリング手法を考案する[39]。音声認識の従来の研究においても，学習データのクラスタリング手法は数多く提案されている[40]。しかし，それらの多くは音声の周波数軸上での特徴 (スペクトル) に基付くクラスタリングであり，本研究で言う，音声の時間構造及びその標準パターン (継続時間長モデル) に基付くクラスタリングとは質的に異なるものである。

以上 2 つの観点から音声認識手法の高精度化を狙ったアプローチを第 8 章及び第 9 章において詳しく述べる。上述したように両手法は共に，従来行なわれてきた HMM による認識処理を実験的に詳細に観察し，得られた実験事実に基づいて考案した手法である。

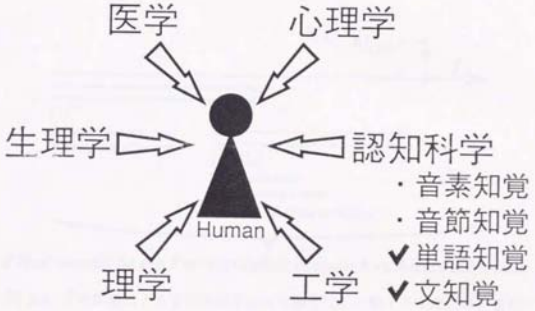


図 2.1. 『人間による音声知覚過程の分析とそのモデル化』に関する本研究の位置付け
図中，✓が本研究の該当箇所を表している。

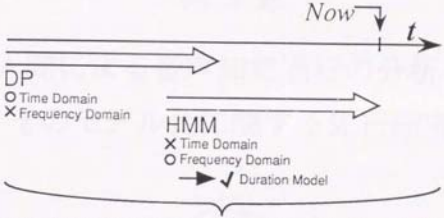


図 2.2. 『計算機による音声認識手法の高精度化』に関する本研究の位置付け
図中，✓が本研究の該当箇所を表している。

第 3 章

人間による音声知覚過程の分析とそのモデル化に関する先行研究

「人間は時々刻々と変化する音声信号からどのようにして，音響的・言語的情報など，第 1 章で述べたような種々の情報を，円滑にかつ正確に抽出しているのだろうか，そしてその処理方式はどのように表現され，モデル化されるべきなのか？ 一体我々人間は何をやっているのだろうか？」この問いに答えようと従来から多くの医学者，生理学者，心理学者，認知科学者，哲学者，理学者そして工学者たちによる様々な研究が行なわれてきた。本章ではまず，従来から音声知覚研究を行なう上で，重要な前提・仮定としての役割を果たしてきた (単語) 知覚モデルを，各モデル間の比較を含めて幾つか紹介する[41]。次に，各モデルに対して比較・検討を行ない，その中で，本研究の目指すところの音声知覚モデルの位置付けを行なうと共に，筆者が行なうべき知覚実験の方向付けについても考察する。

次に，筆者が行なった一連の知覚実験と関連が深いと思われる知覚実験を紹介する。第 2.1 節で述べたように，筆者が行なった実験は認知科学的なアプローチに分類されるものである。既述したように国内の研究においては，単語以下の単位の音声を扱うものが多く，ここで紹介する研究も，単語音声を対象とし，内部辞書への検索過程に着目した研究である。

筆者が実施した一連の知覚実験は，当研究室で藤崎教授 (現東京理科大学教授) 及び広瀬教授の指導の下に行なわれた先行実験の延長線上に当たるものである。そこで，第 5 章において構築する音声知覚モデルの導出・理解に必要となる先行知覚実験についても簡単に述べる。しかし当然のことながら，モデルの構築に際しては，先行知覚実験だけでは不十分であり，多くの課題を残す結果となっている。次の節では，この残された課題について触れ，その中でも特に，内部辞書検索過程及びそれに影響を与える諸要因に関しては，独立に一つの節を設けて詳しく考察し，本章を締めくくることにする。

3.1 提案されている単語知覚モデル・理論

3.1.1 Logogen モデル

Logogen モデルは，読語過程に関わる種々の現象を説明するための単語知覚モデル[1]として，心理学者 Morton によって 1969 年に発表された[42][43]。このモデルでは，全ての語は 3 種類の属性，音響的属性・視覚的属性・意味的属性を持っているとの仮定が行なわれている。語 W_i を知覚するために必要な，各属性に関する情報を A_i, V_i, S_i とすると，図 3.1 に示すように，A_i, V_i, S_i は各々，聴覚分析機構，視覚分析機構，及び認知系 (cognitive

[1] 音声言語の単語知覚モデルとしても，しばしば参照されている。

system) からの出力と定義される。認知系は，この系を含む全体をコントロールする機構や記憶機構を含む。更に，認知系において語の意味の理解がなされ，それに基付き，次入力に関する期待が生成される，としている。例えば，ある文脈が与えられると，その意味が理解され，その結果に基付いて種々の意味情報 (上で言う期待) が認知系からの出力情報として利用可能となる訳である。

ある語 W_iを理解する過程を考える。各種分析系から A_i, V_i, S_iが出力される訳だが，Morton はここで以下の 2 つの仮定を設けた。

- A_i, V_i, S_iの各情報全てを収集し，評価する機構が存在する。
- その機構は語毎に存在する。

この語毎の情報収集装置を Morton は Logogen と呼んだ。Logogen そのものは，入力に対する分析は一切行なわず，入力される情報を加算し，ある閾値を越えた場合に興奮 (発火) し，信号 P_iを出力する，としている。ここで P_iは，音韻符号列のようなものであると考えると分かり易い。つまり，Logogen はある語に関する低次から高次まで全ての情報を受け付ける "受動閾値素子" であると言える。そして，A_i, V_i, S_iによって活性化し，ある Logogen(L_i) の活性度が閾値を越えた場合 (興奮) に，語 W_iは L_iに該当する語として知覚される。ただし，どの Logogen の活性度も閾値に達しなかった場合は，各 Logogen の閾値を下げて再度試みる。また，A_i, V_i, S_iの相互作用についても Morton は，高次情報である S_iが大きい場合は，より小さい低次情報の A_i, V_iで十分に活性化され，その逆も成立する，としている。このように Logogen モデルは，高次情報と低次情報とが相互作用を及ぼし合う interactive なモデルであると言える。また，図 3.1 にあるように，出力バッファー及びリハーサル・ループを仮定し，知覚された語 (Logogen) 自身の興奮を直接持続させる機構も備えている。更に，A_i, V_i, S_iによる興奮性刺激の時定数的性質にまで，定性的にではあるが議論されている。つまり，感覚器官を通して入力される低次情報である A_i,V_iは一過性であるが，高次情報である S_iは長時間 Logogen を興奮させ続けるとしている。

このモデルによって，文脈効果[44]やプライミング効果[45][46]の他に，使用頻度による効果 (Frequency Effect)[47]も定性的に議論できる。例えば，長期的に使用頻度が高くなった語の Logogen は，高頻度の活性化により，閾値が低く shift しており，そのため，弱い刺激に対しても十分に活性度が閾値を越え，興奮状態となり，P_iを出力できると考えれば良い。なお，Logogen モデルは本来，読語過程における単語知覚に対して提案されたものだが，以下に述べるモデル・理論は全て，音声言語における単語知覚に対して提案され

たものである。

3.1.2 Cole と Jakimik による単語知覚モデルの仮説

Cole と Jakimik は，その後の単語知覚研究及び単語知覚モデル構成に対して大きな影響を与えることになる，以下の仮説を立てている[48]。これらの仮説は全て，音声知覚実験結果に基付いて考えられたものである。

1. 単語は，抽出される音響的特徴と予め保有されている知識の相互作用によって知覚される (interactive 処理)。
2. 音声は単語単位に連続的に処理され，知覚された単語は，その直後の単語の開始時点とその単語に対する構文的および意味的制約を与える。
3. 知識として LTM(Long Term Memory，長期記憶) に記憶されている内部辞書 (Mental Lexicon) 中の単語は，語頭音によってアクセスされる。
4. 単語の音響的構造の分析が進むにつれ，候補単語が 1 つに絞られた段階で，知覚は完了する。

3.1.3 Cohort モデル

上記した，Cole と Jakimik の 4 つの仮説に基付いて Marslen-Wilson らによって 1978 年に提案された単語知覚モデルであり，Logogen モデルと同様にボトムアップ処理とトップダウン処理の相互作用を考慮した interactive なモデルである[49][50]。しかし，読語過程のモデルとして提案され，完全に単語を単位として処理が行なわれる[2]Logogen モデルと異なり，Cohort モデルでは，時間の推移と共に，その時点までの照合処理結果を基に単語候補群を絞りこむ，と言う処理を基本としている。具体的には以下のようになる。まず，単語先頭音の音響的特徴により，その先頭音を持つ単語の集合 (Cohort) が形成される。例えば，入力音声が/sling/であれば，/sight/, /slight/,/slave/など，先頭に/s/を持つ内部辞書中の全ての単語が活性化され，その集合体 Cohort が形成される。この段階の処理は，完全なボトムアップ処理のみである。次に，時間軸に沿って，入力音声と Cohort 中の全単語候補との照合処理が並列的に行なわれ，その結果，/sl/で始まる単語以外の語が Cohort から消去される。この時同時に，文法的或は意味・談話的に矛盾を生じる候補も消去される (トップダウン処理)。やがて，Cohort 内の単語が/sling/1 つになった時点 (ユニークネスポイント) で単語知覚は完了し，それ以上の分析は行なわれないとしている。図 3.2 に以上の処理過程の様子を模式的に示す。このように Cohort モ

[2] それ以下の処理単位への言及や，単語入力途中の段階への考慮が乏しい。

デルでは，時間軸に沿った left-to-right 処理を基本とし，ボトムアップ処理とトップダウン処理を組み合わせて (interactive 処理)all-or-nothing の単語選択 (消去) を行なう処理を仮定することで，単語知覚を説明している。その結果，最適効率で単語が知覚されることになる。このモデルによって説明される知覚現象も多く，また，上記したように処理の時間的側面を明確に示しており，これまでに多くの研究者の注目を集め，現在でもこのモデルを支持する研究者は多い。

しかしこのモデルは，以下のように大きな欠点を 2 つ持つ。まず第一の欠点は，このモデルが left-to-right の処理に基付くモデルである点に起因する。つまり，先頭音によって初期 Cohort が形成される訳だが，この時点での知覚誤りを，その後修正する手段が無いと言う点である。初期 Cohort に正解となる単語が含まれていない場合には，このモデルでは破綻を来たしてしまう。逆に言えば，語頭音に対しては常に精度の高い音響的特徴を要求し，語頭音が正確に知覚されることが前提となっているモデルであるとも言える。第二の欠点は，このモデルはあくまでも単語音声に対する処理，即ち，単語の始端，終端の検出を前提としている点である。実際の音声波形を分析すると，音響的な単語境界は必ずしも明確でない場合が多く，この前提に固執したままでは，人間の音声処理過程の一部のみを記述するモデルに留まってしまうことになる。また，単語知覚に及ぼす出現頻度の影響や単語尾付近の音声のみによる単語同定[26][51] など，十分に説明できない単語知覚現象もある。

これらの問題を解決するために，Cohort モデルの新バージョンが提案され[52]，単語候補に活性度を導入している[3]。しかしその結果，従来本モデルが持っていた単語知覚の時間的側面への明解性を欠く結果になっているとの批評もある[8]。

3.1.4 Phonetic Refinement 理論

Logogen，Cohort モデルでは，入力単語の全体，あるいは単語中の一セグメントにおいて，“音響的特徴↔(ある種の) 詳細な音声学的符号” 間の対応付けを行なっている。そして，その結果を基に候補を絞るなどして，最終的な識別結果を求めている。これに対して Pisoni らは，詳細な音声学的符号への対応付けと言う観点からではなく，音響的特徴から得られる概略的な/緩やかな制約条件 (constraint)，及び，その蓄積によって結果的に得られる詳細な制約条件に基付いて単語候補を絞り込む，と言う観点から単語知覚を説明している[28][53][54]。

まず Pisoni らは，従来のモデルでは，内部辞書の構造と単語知覚過程との関係，及び，

[3] Cohort モデルに限らずこれらのモデルは，日々改良されている。

単語自身の構造についての検討が十分に行なわれていないことを指摘した。即ち，内部辞書内に一項目として存在する単語は，どのような環境下に存在しているのか (疎な空間に存在するのか，密な空間に存在するのか，など)，更に，その単語を構成する各セグメントは単語内で，各々どのように影響を及ぼしあっているのか，などに言及している。ある実験結果によれば，20,000 単語辞書の各単語内の全音素を，6 種類のカテゴリ[4]に分類して記述し直した場合，ある単語と全く同一のカテゴリ列を持つ単語は平均して，もう 1 単語しか存在しないことが明らかになった[55][56]。つまり音素空間を粗く 6 分割し，入力音声の各セグメントがどの空間にマッピングされるかを見るだけでも，20,000 単語候補は 2 単語へと縮小される訳である。更に Pisoni らによれば，単語長 (セグメント数) も単語候補削減に非常に有益であるとしている。彼らの提案する Phonetic Refinement 理論は，上記等の実験事実に基付き，入力音声中のあるセグメントが "Vowel である/ない"，或は，"入力単語長は n セグメント" である，と言った (概略的な) 制約条件を入力単語に課すことで，候補となるカテゴリ系列を活性化する[5]ことを基本にしている。そして，活性化した候補カテゴリ系列より，入力単語中の他のセグメントへの制約条件も生まれてくる。当然のことながら，時間の経過と共に，あるセグメントが満たすべき制約条件は増加するため (制約条件の詳細化)，そのセグメントは，より詳細な音声学的符号 (カテゴリ) への変換が可能となる。即ち，より詳細なカテゴリを用いた系列が活性化されるようになる訳である。また，あるセグメントが強く (明確に) 発声された場合，他のセグメントからの制約条件を待たずして，より詳細な音声学的符号へと変換可能である，としている。最終的には，詳細なカテゴリ (音素) 空間におけるある単語の活性度が，同一空間の他の単語の活性度及び，他の (より概略的な) カテゴリ空間において，最大活性度を有するカテゴリ系列の活性度より高くなった時に単語知覚が終了するとしている。

この理論によれば，単語先頭音の重要性は，以下のように説明される。『単語先頭音は当然のことながら，最も多く単語中の他のセグメントからの制約条件を受けることになる。その結果，最も詳細 (refine) な音声学的符号へと変換可能なセグメントとなる。単語の知覚が，この最も詳細に記述されるセグメント (語頭音) に大きく影響されるのは十分妥当である。』即ち音声の時系列性と言う基本的性質から，単語先頭音の重要性を直接的に説明した。また，Cohort モデルでは分析は単語途中でも終わるとしているが，Pisoni らは単語尾に同一の音響的特徴を持つ内部辞書内項目群は，その単語尾の音によっても互

[4] stop consonant, strong fricative, weak fricative, nasal, liquid/glide, vowel の 6 種類。
[5] 制約条件を満たした場合，活性化する。

いに活性化されること (つまり単語尾部分の処理も行なわれている)，更に，語尾部分の音響的特徴のみでも単語の識別が可能となること[26]を実験的に示した。Cohort モデルは先頭音の正確な識別に非常に大きく依存するモデルであったが，これらの実験結果は，先頭音の正確な識別が必ずしも必要でないことを示しており[6]，更に，先頭音の誤った識別後，或は，先頭音の識別無しにも，単語は正しく知覚され得ることが，知覚モデルによって説明されるべきであることを示している。第 3.1.3 節に述べたように Cohort モデルではこれらの現象を説明することは基本的には出来ない。一方 Pisoni らの理論では，時間と共に増加する他セグメントからの制約条件により，最も refine された部分を基にした単語知覚が行なわれるため，その基になる部分は必ずしも先頭に存在する必要性は無い。そのため，先頭音が noisy で，その後の音声が明瞭であった場合，いくら分析時間が短くても後者の方が相対的にはより詳細な音声学的符号へと変換されることになる。そして，より詳細な符号へと変換された部分を基に知覚が行なわれるとしている。この他にも，出現頻度による効果は，Cohort モデルでは説明が困難な現象も，内部辞書の構成へ言及することで説明し，プライミング効果に代表される文脈効果は，制約条件の中に言語的な条件を付加することで簡単に説明できるとしている[7]。

3.1.5 LAFS(Lexical Access From Spectra) モデル

このモデルは，Klatt によって 1979 年に提案された[37]。計算機上での音声認識の影響を強く受けたモデルである。内部辞書中の各単語は，2000 個程の diphone(2 音素結合) スペクトルと音韻学的規則を用いてネットワークで表現されている。図 3.3 にその様子を示す。図では，/limb/, /list/, /summer/, /sum/の 4 単語に対して，音素記号 (phonemic) ネットワーク→音声記号 (phonetic) ネットワーク→diphone ネットワークを構成する様子を示している。diphone ネットワークでは，各 diphone に割り振られた番号で示してある。音素記号ネットワークから音声記号ネットワークへの変換の様子を見ると，この時点で音韻学的規則を組み込んでいることが分かる (/list/+/summer/と連続発声された場合の音声記号列等)。これにより，調音結合など，コンテキストに依存した変動をある程度回避している。実際の認識に際しては，まず，入力単語に対するスペクトル系列が求められる。そして，各スペクトルは辞書中の diphone スペクトル表現との照合が行なわれ (その結果，番号が割り振られ，入力は番号系列となる[8])，その結果を基に，ネットワーク内

[6] あくまでも，必ずしも，と言うことである。重要であることには変わりない。
[7] 但し，この理論を発表した時は，孤立単語を対象としていた。
[8] 現在で言うところの，VQ(Vector Quantization) と同様の操作であろう。

で最も適合するパスが求められる。このモデルの特徴は，入力音声に対して音素や音節と言った小さな単位へのセグメンテーションを必要とせず，音声スペクトル系列と内部辞書項目とが直接比較される点にある。実際の計算機上でのインプリメントに際しては，音声波形を音素列に変換する SCRIBER モデル[57]を並列に機能させ，未知語に対する処理も行なわせている。

3.1.6 TRACE モデル

このモデルはコネクショニストの考え方を基本とし，PDP(Parallel Distributed Processing) の研究の中で，1981 年，Elman と McClelland によって提案された[58][59]。音響特徴抽出レベル，音韻検出レベル，単語検出レベルの 3 層で構成され，各レベル内には，音響的特徴ユニット，音韻ユニット，単語ユニットが存在する (図 3.4 参照)。各ユニットは活性度，閾値，resting value を持つ。音韻ユニット，単語ユニットには対象とする音声言語の音韻数，単語数だけのユニットがあるが，音響的特徴ユニットには，母音性・拡張性など，弁別的特徴 (distinctive feature) に似た 7 種類の特徴が割り振られている。また，同一レベル，或は，隣り合うレベルの各ユニットは双方向性の結合で結ばれている。同一レベルのユニット間結合は抑制性であり，隣り合うレベルのユニット間結合は興奮性である。即ちユニットは，直上/直下のレベルのユニットからの興奮性入力を受けて活性化すると共に，自分自身の活性度に応じて，同一レベルの他ユニットの活性化を妨げようとする。この相互作用によって，各層において，最も適したユニットのみが最終的に生き残ることになる。そして，単語レベルのあるユニットが生き残り，かつその活性度から計算される反応確率がある閾値を越えた場合に，単語知覚が終了するとしている。

TRACE モデルの各ユニットの入力は，時間幅を持っており，上位レベルほどその幅は大きくなる。例えば，単語レベルの各ユニットはその単語長に応じた時間幅を持っており，語尾の部分が入力される前に，(その時点までに入力された音声によって活性化された音韻ユニットにより) 活性化された単語ユニット群は，逆に語尾に存在する音韻ユニットに対して作用する。当然のことながら，入力単語途中のある部分の音声により (音韻レベルを通して) 活性化される単語ユニットもあり，その単語ユニットが語頭の音韻ユニットを活性化するバックワードの処理も実現される。それ故，Cohort モデルで扱えなかった語頭部の音韻知覚誤りに対しても柔軟に処理できる。

Phonetic Refinement 理論でも，制約条件の生成がセグメント→単語→セグメントへと伝搬する (異なる処理レベル間の相互作用) ことは述べられており，その点では類似したモデルであると言える。しかし，TRACE モデルは計算機上でのシミュレーションも行

なわれており，音響特徴レベル，音韻レベル，単語レベルと異なるレベル間，或は，同一レベル間の相互作用の時間的側面を明確に示した点で他のモデルと異なっていると言える。それ故，工学のみならず，心理学・認知科学の分野でも，このモデルを支持する研究者は多い。しかしながら，TRACE モデルの範囲内では説明ができない知覚現象も多い。例えば，プライミング効果に代表される文脈効果は説明が困難である。これは，Cohort モデル同様，TRACE モデルの扱う対象が (現在のところ) 単語に留まっていることに起因する。しかし，プライミング効果は音韻レベルでも観測されており[60]，中間層として存在する音韻レベルでの現象について説明困難な現象が存在する事実は，TRACE モデルが単語知覚モデルとしてもまだまだ改良されるべきモデルであることを示している。更に，韻律的特徴を扱う機構が無いことや，TRACE モデルの根本的問題として，音響的特徴，音韻，単語と言うユニットの設定が，知覚ユニットとして，心理学的に適当であると言う保証が無いと批評する研究者もいる[9]。

[9] しかし，このような議論を更に押し進めれば，「内部辞書項目として当然のごとく設定される，音節，単語などと言う単位が，心理学的に本当に適当であるのか？」と言う議論にまで発展しかねない。

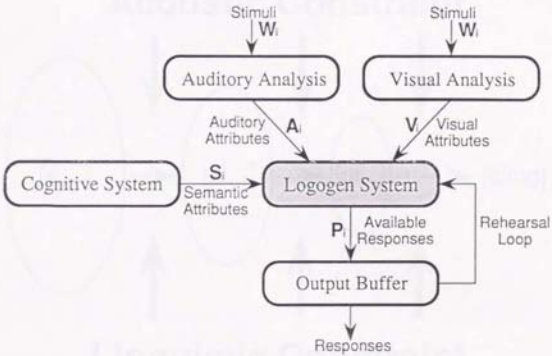


図 3.1. Logogen モデルの概念図

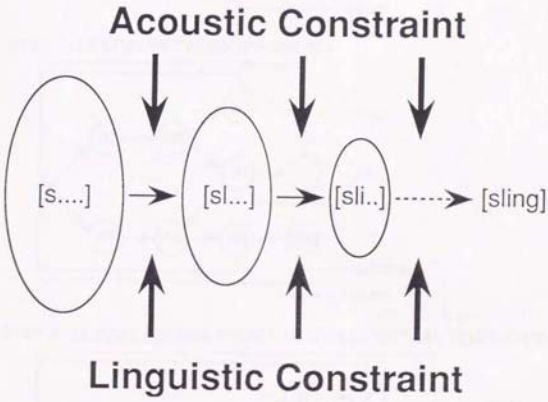


図 3.2. Cohort モデルの概念図

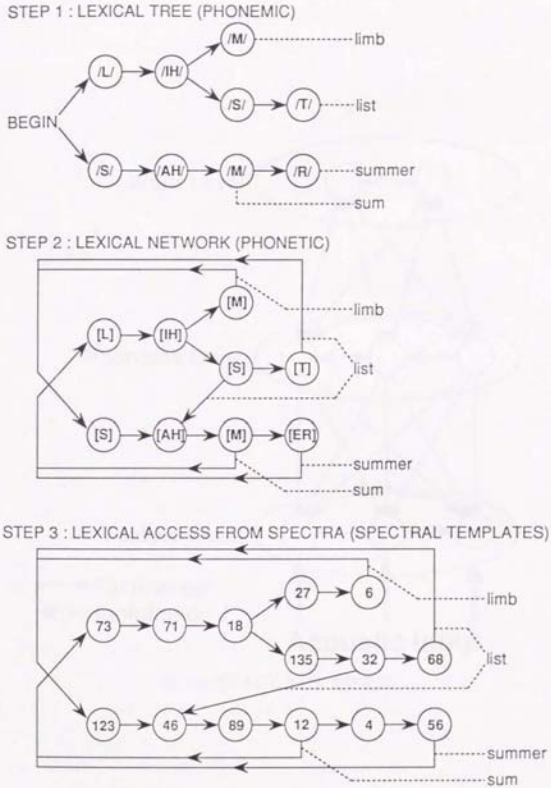


図 3.3. LAFS モデルにおける，音韻記号・音声記号・diphone ネットワーク

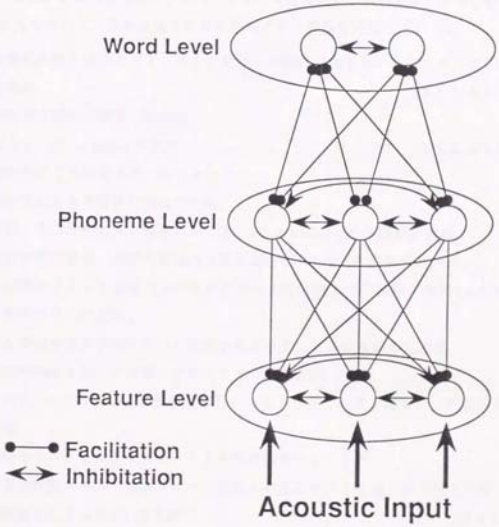


図 3.4. TRACE モデルの概念図

3.2 本研究の目指す音声知覚モデル

第 3.1 節において，従来の音声知覚研究において提案され，前提・仮定として重要な役割を果たしてきた幾つかの単語知覚モデル・理論を紹介した。これらのモデル・理論の構築の際に着目された単語知覚の特性を (是非を問わず) まとめると以下のようになる。右に示してある数字は，各モデル・理論が紹介された節番号である。当然のことながら，各モデル・理論が着目した特性は，相反するものも含まれる。本節では，これらの特性を踏まえ，私見も含めて，従来提案された知覚モデル・理論を考察していく。

1. 音響的特徴を用いたボトムアップ処理と言語的特徴を用いたトップダウン処理の相互作用 (3.1.1, 3.1.2, 3.1.3)
2. 内部辞書項目の奥行の持続性 (3.1.1)
3. 完全な Left-to-Right の処理 (3.1.2, 3.1.3, 3.1.5)
4. 語頭音による辞書項目へのアクセス (3.1.2)
5. 語頭音による単語候補集合の生成 (3.1.3)
6. 単語入力の途中のある段階における，all-or-nothing の単語選択過程 (3.1.3)
7. 内部辞書の構造・単語内構造の情報に基付いた処理 (制約条件の生成) (3.1.4)
8. 各音声セグメントが満たすべき制約条件の時間軸上での蓄積・増加 (より詳細な音声学的符号への変換) (3.1.4)
9. 最も詳細な音声学的符号へと変換されるセグメントを基にした知覚 (3.1.4)
10. 複数の精度を用いた単語・音声セグメントの表現 (3.1.4)
11. スペクトルパターン+音韻結合規則によるネットワークを用いた，内部辞書項目の表現 (3.1.5)
12. 純粋なスペクトルテンプレートとの受動的なマッチング (3.1.5)
13. 音響的特徴レベル・音韻レベル・単語レベルにおける知覚の間の相互作用 (3.1.6)
14. 閾値素子による判別=知覚終了 (3.1.1, 3.1.6)
15. 1 候補への候補限定時=知覚終了 (3.1.2, 3.1.3, 3.1.6)
16. 全精度，全レベル (音素/音節/単語) を考慮した上での，最大活性度を持つ候補の出現=知覚の終了 (3.1.4)
17. 最大活性度による識別=知覚終了 (3.1.5)

これらのモデルのうち，LAFS モデルは計算機上での音声認識に非常に強く影響を受けたモデルである。内部辞書項目の表現方法も diphone 単位のスペクトルパターンのネット

ワークで行ない，また，入力音声と辞書項目 (系列) の照合処理も，入力を "スペクトルパターン→対応する diphone の ID" に変換して行なっている。辞書項目の表現に音韻学的規則を採用することで，言語的情報の一部を汲み入れてはいるが，処理方法そのものは，現在の音声認識処理と何ら変わりが無い。即ち，従来の音声認識のパラダイムで，音声知覚を説明しようとした感が拭えない。結局，人間における音声知覚モデルと言うよりも，計算機による認識モデルと言うべきものであり，知覚モデルとしては，不十分な点が多いように筆者は考える。

一方，心理学者らによる Logogen モデルは内部辞書の項目を受動的な閾値素子 Logogen としてモデル化している。即ち，Logogen そのものは，知覚過程 (処理) のモデル化ではなく，内部辞書項目のモデル化であると考えられる。興奮の持続性をバッファを設けてモデル化するなど，Logogen 周辺の処理部に対するモデル化も行なわれてはいるが，肝心の音響的処理[10]及び言語的処理に関しては極端なブラックボックスのままであり，人間の知覚過程のモデルとしては，これも不十分な点が多いと筆者は考える。これに対して，Cohort モデルは逆に，処理過程の方を重視しすぎているように思える。即ち，音響的特徴に対するボトムアップ処理，及び，言語的特徴に対するトップダウン処理による候補削減だけでは，内部辞書の存在が見えてこない。先頭音によって候補単語集合 (Cohort) が生成されるとあるが，各候補単語への検索の "し易さ/難さ"，即ち，項目固有の特徴 (内部辞書の構造とも関係してくると考えられる) を考慮する必要は無いのだろうか？また，処理が進むにつれて all-or-nothing の削除が行なわれていくとあるが，ここでも，その項目固有の性質として削除の "され易さ/難さ" があるはずである。更に，Cohort モデルの唱える全面的 Left-to-Right 処理では，説明できない知覚現象があまりにも多いように思う。原則としての Left-to-Right 処理と言う立場を採るべきであり，全てを Left-to-Right で語るのは無理があると筆者は考える。加えてこれら両者のモデルは，TRACE モデルの言う，音韻レベルでの知覚と単語レベルでの知覚と言った，異なる処理単位による知覚過程間の相互作用についての記述も乏しい。これら両モデルは音響的情報処理と言語的情報処理の相互作用によって知覚が行なわれる interactive なモデルと言われる。言語的情報は大きく，談話的情報・意味的情報・統語的情報に分れるが，夫々の相互依存性や処理における優先度などの記述も見当たらない。また，音響的情報に関しても，音韻情報を伝搬する分節的特徴 (segmental feature) と韻律的情報を伝搬する韻律的特徴 (prosodic feature) との関係の記述も無く，音響的・言語的処理の両側面に渡って，処理過程が抽象

[10] 統語過程に即して言うならば，入力刺激 (文字) に対する (低次) 特徴抽出処理。

化され過ぎていると筆者は考える。

認知科学者 Pisoni らの提案した Phonetic Refinement 理論は，Logogen モデルや Cohort モデルにあった，"内部辞書か処理過程か" と言った偏りが少ない。即ち，内部辞書の構造，単語内部の統計的構造，複数の精度を用いた単語の表現と言った，内部辞書への考察を行なうと同時に，処理過程のモデル化と言う観点から眺めた場合も，内部辞書の構造及び単語の統計的内部構造を利用した制約条件の生成，それを用いた活性化，異なる処理単位間での活性化の伝搬，と 2 つの側面に対してバランスのとれた知覚理論であり，非常に見るべきところが多い理論であると筆者は考える。特に，複数の精度[11]による辞書項目表現に関しては，他の知覚モデルでは明示的に論じられてはいない点である。この理論において唱えられている，単語内セグメントが生成する制約条件による候補単語の活性化，及び，候補単語が生成する制約条件による単語内セグメントの活性化，と言った "セグメント↔単語" 間の相互作用は，TRACE モデルの唱える，異なるレベル (音韻/単語) の知覚過程間の相互作用とほぼ同じであると考えることができる。しかし上記したように，制約条件の蓄積による，より詳細な分析結果の導出 (即ち複数の精度による音声学的符号の事前の登録) がこの理論の独自性を生み出している。このように優れた理論であるが，言語処理における統語解析・意味解析・談話解析の相互作用や優位性，或は，音響的処理における音韻情報処理と韻律的情報処理の役割については，まだ述べられていない。更に，これはどのモデルに対しても言えることであるが，単語以上の入力音声への考慮も不十分である[12]。

TRACE モデルは計算機上でのインプリメントを考慮しつつ主に工学者らによって構築されたモデルである。異なるレベル間での相互作用をシミュレーションを通して明確に提示する[58][59]など，見るべきところの多いモデルである。しかし，計算機上で十分に実現されていない言語的情報の利用 (知識処理等) に関しては，音響的特徴の処理手法ほど詳細に論じられていないのも事実である。しかし既述した特性 1. の正当性は自明であり，知覚モデルに 1. を明記しないのはモデルとして不完全であると筆者は考える。また，各層間の相互作用についても，音韻単位の知覚結果を基にして単語単位の知覚が行なわれ，そのフィードバックが音韻単位の知覚に影響する，と言うように基本的には音韻単位での知覚が先行している (音響特徴レベルから単語レベルへ直接の作用は無い)。しかし人間は，雑音の多い環境でも苦もなく会話できたり，不明瞭に発声された音声に対しても支障

[11] 複数の処理単位・レベルではない。
[12] どのモデルも，単語知覚モデルとして提案されているので，当然と言えば当然である。

なく知覚できる。これは入力音声に対して，局所的な各音韻性への注意よりも，まず辞書項目全体での類似度を基にした知覚が行なわれ，その後に，知覚結果となった項目を構成する個々の音韻と入力音声との対応付けが行なわれると言った，より大きな単位による処理が優先する場合も多いと考えるべきではないだろうか。更に，何度も記すようだが，単語モデルと言う枠を越えた知覚モデルの構築も望まれるところである。

以上は，従来提案された知覚モデル・理論を通しての考察であるが，従来のモデル・理論に欠けていると思われる部分について記す。単語知覚の枠を越えたモデルの構築は勿論であるが，それ以外に以下のことが考えられる。音声知覚とは未知入力を何らかのカテゴリに属させる範疇化であり，“物理的な音” として入力された音声は，照合処理を経ることによって “シンボル列” へと変換される。この時，照合前後によって音声の持つ情報の質は大きく変化している。つまり，照合前は純粋に音響的情報としての連続量であるが，照合後はあるシンボルの並びとしての離散量となる[61]。この音声の情報としての質的変化について言及しているモデルは上記のモデルには無い。また，単語を越えたモデルを考える場合，単語以上の句や文といった処理単位も考えられ，より大きな処理単位への実験的考察も必要である[62]。

このように各モデル・理論はその構成において，背景となる分野が影響を及ぼしていること，単語知覚と言う枠によって十分に表現できていない知覚現象があること，異なる処理単位による処理間の相互作用を議論する場合でも，より小さな単位を用いた処理が先行することが前提となっていること，などの種々の問題点を含んでいる。そこで本研究では，まず単語と言う枠を越えた，人間の音声言語処理過程をグローバルな観点からモデル化することを目的とした知覚実験を行なう。そのためには，上記の考察より以下の項目を対象とした実験計画を行なうべきであろう。

- 単語以上の処理単位の存在の是非。
- 存在する場合，その処理単位を用いた処理の特性。
- 複数精度による (照合) 処理と複数単位による (照合) 処理との関係。
- 内部辞書の構成・構造。
- 辞書検索過程へ影響を及ぼす諸要因。項目固有の特性がもたらす作用 (単語知覚) と複数の単語が一文中に存在することにより生じる単語間の作用 (文知覚)。
- 韻律的情報処理と音韻情報処理の単語内/文内における相互作用。
- 言語的情報処理が音声知覚過程に及ぼす影響。特に統語的・意味的・談話的情報処理による影響の差異。

3.3 本研究に関連する研究

ここでは，本研究に関連する研究を 2 つ紹介する。どちらの実験も第 4 章で述べる知覚実験で参照している実験である。

3.3.1 刺激音声の有意味性が単語内音韻知覚に与える影響

背景と目的

音声知覚は音響特徴/音韻/単語/文 (構文) レベル … と言うように複数のレベルの処理過程が存在し，それらが相互に作用を及ぼしながら全体としての処理が進むものと考えられる。この研究では焦点を単語以下に絞り，単語と音韻レベルの知覚と両者の相互作用を観測することを目的とする[63]。

実験方法

出現頻度が十分高いと考えられる単語音声 (A) 及び非単語音声 (B) を数十種類用意する。更に，日本語 100 音節音声 (C) も加えて用意する。そして，C を連結することにより，(同一音韻情報を持ち) 調音情報が欠落した単語セット (A') 及び非単語セット (B') を作成する。更に，ターゲット音韻を/k/に定め，各セット A, A', B, B' の半数が語中に/k/を一つ含むようにする。これらの 4 セットの音声を混合してランダムに並び変え，被験者にヘッドフォンを通して提示する。被験者には以下の 2 通りのタスクを課す。

- 提示音声中に/k/が発見された場合，出来るだけ早く反応キーを押す。
- 提示音声の "単語/非単語" の判断を行ない，出来るだけ早く反応キーを押す。

当然のことながら，上記 2 つのタスクは各々異なるセッションで行なわせる。

結果と考察

(非) 単語の判断によるキーインが行なわれた時点から 100[msec](単純反応時間) 前の時点を単語認識時点と呼ぶことにする。この単語認識時点を 0[msec] として，/k/が存在する時点を横軸に，/k/に対する反応から 100[msec] 差し引いた時間 (RT) を縦軸にとって示したのが，図 3.5，図 3.6 である。但し，前者が自然音声に対するもの，後者が連結音声に対する結果である。図より，単語認識時点よりも 300[msec] ほど前から，/k/に対する反応が，自然音声/連結音声によらず，単語音声の方が短くなっていることが分かる。即ち，本実験で得られた効果は，調音結合の情報によるものではなく，純粋に入力単語が内部辞書に登録されているか否か，に起因するものであると結論できる (Lexical Effect)。筆者は更に，本実験で得られた結果を TRACE モデルを用いて説明している。

3.3.2 大規模データベースを用いた単語親密度の測定

背景と目的

単語知覚の難易度に影響を与える要因として，1) 出現頻度 (Frequency)，2) 獲得年齢 (AOA:Age of Acquisition)，3) 心像性 (Imagery)，4) 具象性 (Concreteness)，5) 連想価 (Association Value)，6) 有意味度 (Meaningfulness)，7) 連想基準 (Norm of Association)，8) 親密度 (Familiarity) などが考えられる。知覚実験において使用する刺激は，上記の指標について統制のとれたものが望ましい。しかし日本語を用いて，これらの指標を大規模データベースに対して求めた実験は行なわれていない。英語の場合，上記の指標を求めたデータベースは存在するが，多くの場合視覚提示して求めたものである。そこで，延べ 62,000 語の単語データを用いて「視覚提示」,「聴覚提示」における親密度の測定を行ない，両者の相違を見ることを目的とする[64][65]。

実験方法

実験 A 音声刺激をランダムにヘッドフォンより提示。聴取後なじみの程度を主観的に評価させ，PC 上の 7 段階のスケールに，マウスでクリックさせる。

実験 B 漢字かな交じり文字刺激をランダムに PC のモニタ上に表示。その後なじみの程度を評価させ，PC 上の 7 段階のスケールに，マウスでクリックさせる。

結果と考察

延べ 62,000 単語に対する結果から，同音異字語，同字異音語の結果を取り除いて，視聴覚間の相関をとってみたところ 0.734 となり，また，全単語に対して相関を求めると，0.700 となった。同様な実験は英語において実施されており，英語の場合は，相関が 0.9 以上と高い値が得られている。これは，日本語の表記が表音/表意文字である，かな/漢字の混合であることに依る，と考察している。

なお，本実験を紹介したのは，聴覚/視覚間での情報の表出/受容の "ずれ" への興味以上に，本実験が日本における音声知覚研究の実態を，ある意味で示していると思われるからである。従来の音声知覚実験においては，音声試料の難易度を揃える場合に「小学校の国語の教科書から選出する」などが行なわれていたが，選ばれた単語セットに内在する難易度の "揺れ"(分散) の正確な測定などは不可能であった。これは，各単語に対して各指標の値を定めた，基準となるべきデータベースが存在しなかったからである。今後も知覚実験環境を整備すべく，上記したような研究が充実していくことを期待したい。

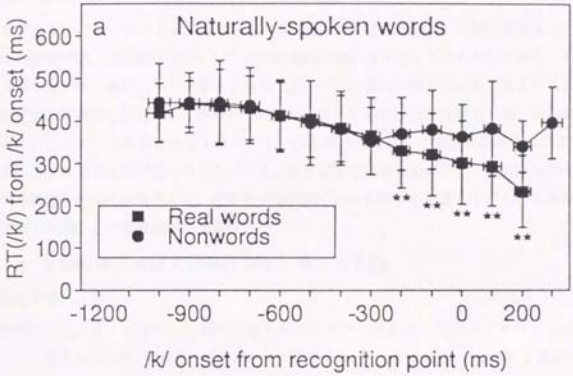


図 3.5. 自然単語音声内の/k/に対する反応時間 (RT)

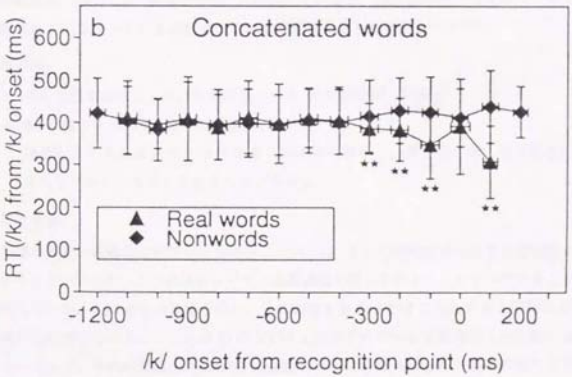


図 3.6. 連結単語音声内の/k/に対する反応時間 (RT)

3.4 当研究室で実施された先行知覚実験

本研究は，過去において東京大学名誉教授 (現東京理科大学教授) の藤崎教授及び広瀬教授の指導の下，当研究室で行なわれた知覚実験の延長線上にも当たるものである。そこで本節では，第 5 章において構築する音声知覚モデルの導出・理解に必要と考えられる先行知覚実験を簡単に紹介する。当然のことながらこれらの先行知覚実験から得られる知見だけでは，モデルの構築には不十分である。即ち種々の残された課題が存在する訳だが，それらについては次節で述べることにする。なお以下ではまとめの欄に，各実験の現象としての結果を直接的結果として，考察及び他実験における結果を考慮して得られる知見を考察及び知見として箇条書きして記す。

3.4.1 音韻知覚における範疇化効果に関する実験

背景と目的

音韻の知覚とは，何等かの音響的特性を帯びた音声を有限個の言語的カテゴリ (カナ) の一つに変換する作業 (範疇化) である。一方従来から，2 つのカテゴリ (A・B とする) 間に存在する音韻をその音響的性質が A⇒B となるように連続的に変化させて提示すると，ある点を境界として知覚結果が A⇒B へと急速に変化する (範疇的) ことが知られている。この現象は一般的には "範疇的知覚" と呼ばれているが，何故範疇的になるかについては解明されていなかった。本実験ではこれを解明することを目的とする[61]。

実験方法

1. ある音韻を提示し，それを範疇化させる (絶対的判断，同定)。
2. ABX 法による弁別 (相対的判断) 能を測定する。
   (ABX 法とは A-B-A または A-B-B と刺激音を提示し，第 3 音が第 1 音と第 2 音のどちらであるかを答えさせるものである)。

結果と考察

言語音声及び範疇化が可能な非言語音については，2 つの範疇境界付近での弁別能が上昇することが分った。この効果について，処理過程を図 3.7 のようにモデル化することで説明している。A は音色の知覚を行い，その結果を B の STM に保持すると同時に C で音素の同定が行なわれる。これは D の LTM に保持されている音素境界との比較に基づくものであり，その結果は E の STM に保持される。F は A, B, X の各々に関する知覚に基づいて弁別を行なう過程であるが，A, B を異なる音素と判断した場合には X の判定

に E の STM 内の情報が用いられ，同一の音素と判断した場合は B の STM 内の情報が用いられる。ここで注意すべきは B と E の STM の性質の差である。B に保持される情報は連続量であり，記憶・再生の過程での変動の影響を被り易いのに対して E に保持される音素は離散量であり，数秒程度の実験内では極めて安定であり，この差が実験結果に反映しているとしている。つまり，従来から範疇的知覚と呼ばれていた現象は，知覚そのものが範疇的なのではなく，言語音であるが故に必然的に受ける範疇化作業及び，その後の識別に用いられる情報の質の差による効果であると説明している。

まとめ

考察及び知見

- 入力音が言語音として知覚された場合，それは必然的に，情報の形態が連続量から離散量へと変換したことを意味する。
- 離散化されることで，記憶の中にも安定して保持されることとなり，その結果次入力の音の知覚へ及ぼす影響もより大きなものとなる (→範疇化効果)。

3.4.2 音声処理単位の多重性に関する実験

背景と目的

人間による音声知覚には，その処理の一部として，内部辞書項目との照合処理が存在すると考えられている。そして，未知語に対する照合処理においては，音韻あるいは音節レベルの辞書項目を用いた処理が行なわれていると考えられる。しかし，音韻・音節と言った小さな単位における音声信号は，前後の環境により，その特性が大きく変動する。連続音声中の既知単語に対しても，このように特徴変動の大きい音韻，音節を単位とした照合処理を常に行なっているとは考えにくい。そこで人間の音声処理単位の大きさと，その特徴を分析することを目的とする[66]。

実験方法

同一文章音声から以下に示すコンテキスト長の異なる 4 つの TYPE の音声を作成する。

TYPE 1　文章音声を単語単位で区分し，ランダムに並べかえたもの。

TYPE 2　文章音声を文節単位で区分し，ランダムに並べかえたもの。

TYPE 3　文章音声を句単位で区分し，ランダムに並べかえたもの。

TYPE 4　区分も並びかえも行なわないもの (元の文章音声)。

その後，図 3.8 に示すように，いくつかの音節を無音で置換して，被験者に提示する。タスクとしては，無音置換の個数を数えさせる。指標としては，TYPE 別の無音置換検出率を見る。なお，提示する音声は全て自然音声である。

結果と考察

図 3.9 に示すように TYPE 1, 2 における結果はほぼ同じで 70[%]，TYPE 3, 4 についてもほぼ同じで 40[%] となった。TYPE 3, 4 の場合でも音節単位で処理しているのならば，検出率は高いはずである。また，TYPE によらず同一の単位を用いているのなら検出率は全 TYPE を通して一定のはずである。実験結果より，人間には複数の処理単位が存在し，前後のコンテキスト長を一つの要因として，用いられる処理単位長は変化していると言える。つまり，コンテキストが単語の場合，音節単位で知覚を行ない，コンテキストが句，文になると単語・文節を単位とした処理が行なわれ，無音置換による音響的特徴の変化は吸収され検出率が低くなった，と説明している。

まとめ

直接的結果

- コンテキスト長の違いが無音部の検出率に大きく影響している。
- コンテキスト長が単語サイズの場合は無音部の検出率は高いが，コンテキスト長が文，句サイズになるにつれて，検出率は極端に低くなる。

考察及び知見

- 音声処理単位は複数の大きさのものが存在する。
- 処理単位長は，入力音声のコンテキスト長を一つのパラメータとして変化する。
- 文章音声の場合 (一般の言語活動がこれに相当する)，単語・文節ほどの大きさの (少なくとも音節よりも長い) 音声長を単位にした処理が主に行なわれる。

3.4.3 処理単位長の違いが知覚の早さに与える影響に関する実験

背景と目的

第 3.4.2 節において，音声処理単位の複数性が実証された。さて，同一音声に対して，複数の異なるサイズの単位による処理が行なわれた場合，その処理間にはどのような差が生じるのだろうか？そこで本節では，処理単位長と，最終的に知覚結果を出すまでに必要とされる時間 (知覚の早さ) との関係について実験的に検討する[67]。

実験方法

単語提示/口頭再生などの様に，高次の言語処理を必要としないタスクに対する (単語) 知覚の "早さ/遅さ" を議論する場合，以下の 3 つの処理にかかる時間の合計を考慮すべきであると考えられる。それらは，"音響的特徴抽出処理"+"内部辞書検索処理"+"音響的照合処理" である。また，これら 3 つの処理はそれぞれ同時に，並行して行なわれていると考えられる。さて，本実験では，刺激となるべき単語を前もって被験者に示しておく方法を考える。こうすることで，辞書検索処理にかかる時間を各提示音声で一定 (≒ 0.0) にすることが可能となる。そして，単語内の一音韻を他の音韻と置換したものを提示し，その単語が，前もって示した単語か否かを判断させ，Y/N の判断を下す (Y/N のキーイン) までの時間を測定する。刺激音声には/tokuʃima/が選ばれた。

結果と考察

図 3.10 に実験結果を示す。単語後部の置換部に気付くまでの時間よりも "単語全体が" 正しいと判断するまでの時間が短い。これは，置換部に気付くために働いている音韻単位の処理過程とは異なった別の処理過程があることを示唆する。そして，音韻単位の処理過程が最後の音韻が正しいと結果を出す前に，その別の処理過程が，単語全体が正しいという答えを出していると説明している。第 3.4.2 節の実験結果を考慮すると，この過程が単語・文節単位の照合処理であると推測される。

実験方法を考慮すると上述したように，検索過程に必要な時間はどの刺激音声に対しても≒ 0.0 としてよい。故に反応時間の差は，"音響的特徴抽出処理"+"音響的照合処理"(但し並行して行なわれている可能性大) に必要とされる時間の差と言うことになる。文献ではまず図 3.11 に示すように，「同一の入力音声から複数の精度の音響的特徴が並列に抽出され，時間的にはより低精度の特徴から分析・抽出が完了し，いち早く次処理部へ出力される」と言う音響的特徴抽出部，及び各々の精度に対応する音響照合処理部を想定している。そして，「使用する特徴量の精度に拘らず，音響的照合処理に必要とされる時間は等しい」と言う暗黙の前提を置くことで，本実験の結果から，"より大きな処理単位"→"より早期に抽出が完了する低精度の特徴による正しい照合"→"より早期の段階での照合結果の出力" と言う関係を導いている。

しかし，精度の異なる複数の特徴量を用いた，複数の照合過程を考えると，同一長の音声に対しては，粗い特徴量を用いる方が必然的に単位時間当りに処理する情報量は減り，照合処理そのものに要する時間も短くなることが考えられる。つまり，音響的特徴抽出部及び音響的照合処理部の両者における時間差が実験結果として現れていると再考察する

こともできる。また上記の考察では，音韻置換が検出された場合に，使用された処理単位は必ず音韻サイズであり，単語・文節サイズの処理単位ではその検出ができないとの前提も置いていることが分かる。

まとめ

直接的結果

- 単語尾付近の音韻置換の検出に要する語頭からの提示長は，単語全体が正しいと判断するために必要な提示長よりも長い。

考察及び知見

- 同一音声を異なるサイズの単位を用いて処理させた場合，より大きな単位における処理ほど，早期に抽出が完了する低精度の音響的特徴量で正しい照合が可能であり，最終的には，正しい知覚に必要な処理時間も短縮される。

なお，以降の章・節で，高/低精度の音響的特徴と言う言葉がしばしば登場するが，ここで“緩やかな”定義を行なっておく。“緩やかな”と言うのは，内部辞書項目における音響的特徴の表現方式 (representation) が明らかにされていない現段階で，厳密な定義を行なうのは困難であり，また無意味でもあるからである。

“高/低精度の音響的特徴量”の定義

高精度　音節サイズ以下の単位で処理を行なう為に必要な音響的特徴量 (高情報量)
低精度　単語サイズ以上の単位で処理を行なう為に必要な音響的特徴量 (低情報量)

一般に聴取妨害等により，音節単位での知覚・同定が不可能な音声でも，単語単位での知覚・同定が可能となることは多い。これは，音節単位での処理が単語単位での処理よりもより精密/高精度の特徴量を要求しているからであると考えられる。上記の“高/低精度”は“より小さな/大きな”処理サイズでの処理を行なう為に必要な音響的特徴量と定義することも出来るが，第 4 章に述べる知覚実験との関連から，音節と単語の間に境を設けて上記のように定義しておく。

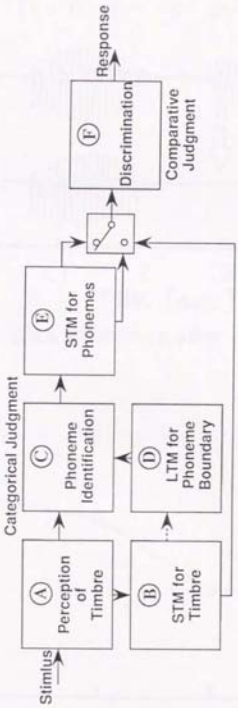


図 3.7. 音韻知覚における範疇化効果を説明するモデル

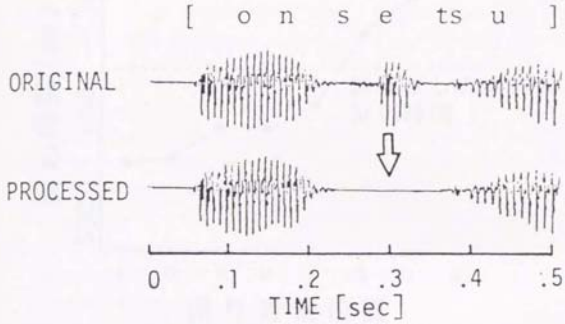


図 3.8. 音節単位での無音置換

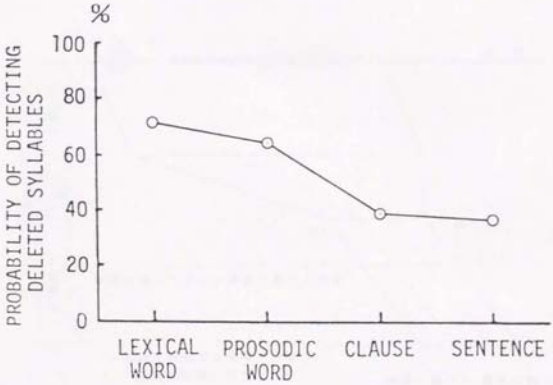


図 3.9. 音声処理単位の多重性に関する実験結果

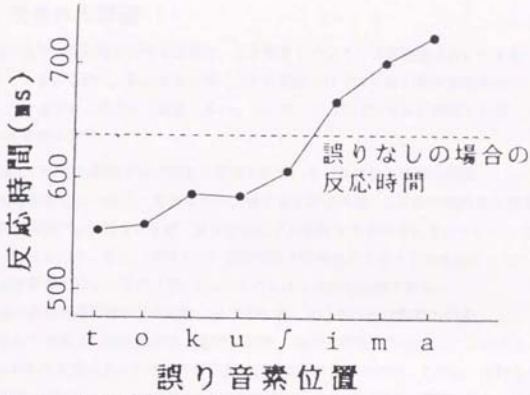


図 3.10. 処理単位長の違いが知覚の早さに与える影響に関する実験結果

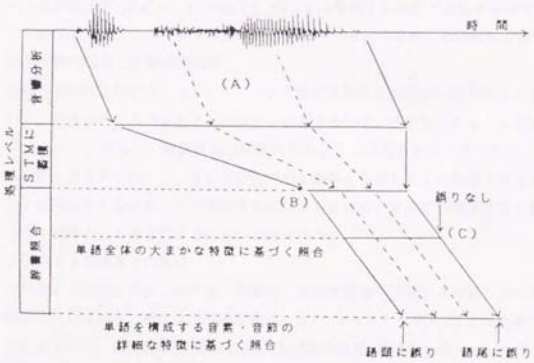


図 3.11. 処理単位長と知覚の早さとの関係を説明するモデル

3.5 残された課題

人間の音声知覚過程に対する研究は，以上報告したように本研究室においても幾つか行なわれている。しかし，第 3.4 節で述べた先行実験だけでは人間の音声知覚過程の一部を観測したに過ぎず，残された課題も多い。そこで，この節ではそれら課題を列挙し，各々について考察する[68]。

1. 種々の音響的特徴が音声知覚に及ぼす影響，及び各特徴の影響の相違

   音響的特徴は一般に，音韻情報を伝搬する分節的特徴[13]と韻律的情報を伝搬する韻律的特徴[14]に区別されるが，夫々どのような影響を音声知覚に及ぼすのか。その違いは何なのか。特に，機械上での音声認識では現在のところ十分考慮されていない，韻律的特徴についての単語/文レベルにおける検討が必要である。

2. 種々の言語的特徴が音声知覚に及ぼす影響，及び各特徴の影響の相違

   言語的特徴は，統語的特徴，意味的特徴，談話的特徴に大別することができるが，これらの特徴は各々どのような影響を音声知覚に及ぼすのか。その違いは何なのか。音声生成において，1. で述べた韻律的特徴は，文音声の統語的構造・話者の意図/感情との関係が従来より報告されているが[69]-[71]，知覚面における影響はどのようなものなのか。更に，言語的情報を定量的に表現する場合，従来より確率論的情報論による表現方式 ($I(p) = -p \times \log(p)$，但し p は着目する事象が生起する確率) が用いられてきたが，この方式を踏襲すべきか否か，と言う根本的な問題も存在する。

3. 内部辞書の構造と辞書検索過程

   内部辞書の構成はどうなっているのか。辞書項目単位と知覚 (処理) 単位とは等しいのか。各項目のフィールドとしては何が存在するのか，それはどのように表現されているのか。そして，各項目間の関係はどのように表記されるべきなのか。また，人間は入力音声に対して，常に全辞書項目を検索，照合した上で結果を出力しているとは考えられないが，その検索法は如何なるものなのか。その検索方法に影響を与える要因としてはどのようなものがあるのか。

4. 文脈による冗長度との関係

   大局的，局所的な文脈における，音響的・言語的情報の冗長性は知覚とどのように関係しているのか。それは，円滑な音声知覚 (コミュニケーション) には必要不可欠な要素なのか。低次処理 (音響的処理)，及び高次処理 (言語的処理) の各々に対して，

[13] 物理的には，スペクトル包絡の時系列で表現される。
[14] 物理的には，ピッチ，パワー，ポーズなどで表現される。

どのような影響を与えているのか。

5. island-driven 的な処理

第 3.1.4 節で述べた Phonetic Refinement 理論では，時間と共に蓄積される制約条件により，最も refine されたセグメントを基にした知覚が行なわれると説明している。これは，一般に言われる island-driven 的処理における，island の決定手段を論じていると見ることができる。Phonetic Refinement 理論の是非は別にして，この island-driven 的な処理の結果を基に，全体の特徴が知覚され，最終的な結果を出力すると言う方向性を積極的に否定することは出来ない。但し，この island として，音韻・音節レベルのセグメントが適当であるのか，単語・文節レベルのセグメントが適当であるのか，或は両者が使用される場合でもその使い分けは何に依存して行なわれるのかなど，まだまだ議論する余地は残されている。

6. repeated-listening の効果

一般的な自動音声認識では，同一音声を繰り返し提示しても毎回同じ処理を行なうだけである (同じ結果しか生れない) が，人間の場合，前回聞き逃したところのみを聞く等，その処理は自ずと異なってくる。このように短期的に頻度が高くなった文脈が音声知覚に与える影響はどのようなものなのか。一般に frequency effect と呼ばれるのは，長期的な高頻度を問題としており，ここで言う repetition(短期的な頻度の向上) とは性質を異にすると考えられる。

7. 処理能力の最適配分

音声信号は時間とともに消滅するものである。即ち，円滑なコミュニケーションを行なうためには，入力音声に対する処理時間の上限と言うものが存在する。一方，人間の情報処理能力にも限界は存在するであろう。ここで，膨大な情報量が入力として与えられた場合，人間は処理能力の最適 (あるいはそれに近い) 配分を行ないながら，円滑なコミュニケーションを続けようとすると考えられる。この場合の最適配分が実験的に考察できれば，通常の入力音声に対する処理能力配分をも，近似的に観測・考察できると考えられる。

以上，ややオーバーラップするものもあるが，人間の音声言語処理過程に関する研究における残された課題について，筆者が考えるところを簡単にまとめてみた。特に，言語処理部についての解明が待たれるところである。そこで次の節で，その中の 1 つである内部辞書項目への検索方法について詳しく考察することにする。

3.6 内部辞書検索過程についての考察

3.6.1 内部辞書の検索処理

本来，音響的情報処理と言語的情報処理は切り離して考えるべきものではなく，相互に助け合って行なわれるべきものである。即ち，ある音声を認識するために人間は，前後の文脈及び着目する音声から得られる音響的特徴・言語的特徴を用いて，音響的/統語的/意味的/談話的整合性を総合的に評価し，最終的な判定を下していると言える。従って，その音声に対応する内部辞書項目には必然的に，分節的特徴，韻律的特徴，統語的役割，意味的属性，他項目との統語的/意味的/談話的関係等を兼ね備えている必要がある。そして照合処理時には，まず内部辞書項目への検索処理が照合に先行して行なわれる訳だが，当然のことながらこの辞書のサイズは膨大なものとなる。一般に計算機上での自動音声認識では，入力音声から抽出される音響的特徴パターンと辞書内の全項目[15](有限個)とを照合し，最大類似度を示す項目を結果として出力する。しかし人間は，入力単語 (項目) の各々に対して，数万，十数万とも言われる内部辞書内全項目との検索・照合を常に行なって最終的な結果を出力しているとは考え難い。次の例を考えてみる。

あの青い空に浮かぶ白い /kumo/ にのってどこかへ行きたい。 (文 1・同音異義語)

この文音声中の音声/kumo/に対して，あの 8 本足の「蜘蛛」を想起することはまず有り得ない。この現象に対して 2 つの考え方が可能である。1 つは，音声/kumo/に対する検索・照合が辞書項目「雲」に対していち早く行なわれ,「蜘蛛」に対してまでは行なわれなかったとするものである。この場合，先行の文脈により適切な辞書検索の動的制御が行なわれ，「蜘蛛」に先行して「雲」との照合が完了し，十分に音響的/統語的/意味的/談話的整合性がある (活性度がある閾値を越えた) と判断されて，その結果，/kumo/が「雲」であると知覚されるとする考え方である (閾値操作)。もう一つは，音声/kumo/に対する照合は「蜘蛛」に対しても「雲」とは独立して行なわれ[16]，両者の音響的/統語的/意味的/談話的整合性を比較した結果，注目する音声は「雲」であると判断するとの考え方である (最大類似度候補[17])。即ち両者の相違点は,「蜘蛛」への辞書検索及び照合処理が行なわれているか否か，換言すれば,「雲」の結果と「蜘蛛」の結果が比較されるか否か，の違いである。しかし，前者の考え方を採用した場合でも，Pisoni らの言う辞書構造を考慮するならば，辞書項目「雲」とその同音異義語「蜘蛛」とは内部辞書内で，比較的近い位置に配置

[15] 照合前に候補を限定する手法も考案されているが，動的かつ柔軟な方法はまだ難しいようである。
[16] どちらが先かは不明。あるいは並列的に行なわれるとも考えられる。
[17] 即ち，現在の自動音声認識において採用されている方針である。

されている，或は，何らかのポインタで関係付けられていると考えられる。その結果「蜘蛛」に対する検索 (活性化) も，照合結果「雲」を通して間接的にではあるが，行なわれていると考えられる。単語尾に同一音を持つ単語ですら，その知覚が容易に/早くなるとの報告[72]を考慮すると，同音異義語に対する活性化はかなり強いものであり，上記の間接的活性化は，十分に受け入れられるものである。即ち筆者は，上記の 2 つの考え方に依らず，結果的には項目「蜘蛛」への検索も行なわれ，活性化も行なわれると考えている。直接的か間接的かの違いだけである。次の例文を考える。

/kumo/ を見た。 (文 2・同音異義語)

この文音声の/kumo/はどちらの "くも" であろうか？当然のことながら，前後の文脈が無ければ，断定は不可能である。今にも「雲なの蜘蛛なの？」と言う応答文が聞こえてきそうである。さて，ここで上記の議論を繰り返してみる。「雲」か「蜘蛛」か，最終的な判断が出来ないとしても，まずどちらかの「くも」が先行して，検索・照合されるのか (そしてポインタにより他方の「くも」も活性化される)，あるいは，両者が各々独立に検索・照合され，比較されるのか，である。ここで，注目したいのは，内部辞書項目が各々，固有の性質として "検索のされ易さ" を持つと考えられることである。例えば/hiʃo/と言う言葉に対して，何も文脈が与えられなかった場合，筆者の場合は，まず「秘書」が想起される[18]。「避暑」と言う言葉は現在の筆者にとって比較的縁遠い言葉である[19]。このように，各辞書項目には，"検索のされ易さ" と言うパラメータが存在する。即ち，ある入力音声に対する種々の辞書項目の検索順序と言うのは，文脈の有無に拘らず存在する。話を文 2 に戻す。上記の/kumo/にしても，聴取者にとって検索され易い "くも" がまず先行して照合されると考える。そしてこの場合当然のことながら，高い音響的類似度を示すはずである (高い活性度を示す，と表現することもできる)。ここで，活性化された項目の周辺に存在すると考えられる，他方の「くも」に対して，全く独立の検索が行なわれると考えるのは無理があると筆者は考えている。と言うのも，検索の順序が文脈の有無に依らず存在する事実は，ある項目 **A** が検索・照合され，その結果活性化された場合，関連項目 **Â** への検索は **A** を通して行なわれることを十分に示唆するからである。以上の考察により，筆者は各々の「くも」が独立に検索・照合されるのではなく，どちらかをキーにして他方が検索・照合され，最終的には (この場合は) 両方の項目が検索されると考えている。

さて，再度話を文 1 に戻す。この場合，先行文脈よりまず「雲」への検索が行なわれる

[18] 特に，これと言う理由は無い。以前，「秘書」の同音異義語が暫く思い浮かばなかったことがある。
[19] /hiʃo/の後に/chi/が付くと，状況は一変するが。

と考えられる。そして活性化された「雲」を通して「蜘蛛」への活性化も行なわれることになる。しかし,「雲」によって必ず「蜘蛛」は活性化されるのだろうか?「雲/蜘蛛」のような同音異義語の場合と,「大学/キャンパス」のような関連語とでは間接的活性化の生起の様子,或は活性化の度合に差は生じないのだろうか?次の例を考えてみる。

大雪のため,列車のダイヤが一日中,/koNdaN/ した。 (文 3・類似語)

という文音声を被験者に聞かせたところ,約 50[%] の人が/koNdaN/を「懇談」ではなく,「混乱」とみなすという結果がある[73]。この場合も,各項目が固有の特性として持つ検索のされ易さ,及び文脈によって生じる検索の方向性によって決定される,検索の順序に従って,「懇談」と「混乱」のどちらかが先に検索・照合されるはずである。しかし「懇談」の場合は音響的には十分な整合性を示すが,意味的整合性が低く,逆に「混乱」の場合は意味的整合性は十分であるが,音響的整合性が低いとの結果を示すはずである。さて,ここで問題となるのは,先行して検索された,音響的/意味的のいずれかで高い整合性を示す項目との検索・照合が終了した時点で/koNdaN/に対する処理が全て終了するのか,或はそれをキーにした間接的活性化が更に関連語に対しても行なわれるのか,である。単語全体の音響的類似度は比較的高いと考えられる両単語であるが,上記した間接的検索が行なわれる可能性は,同音異義語に比較すると,当然低くなっているはずである。筆者が考えるに,これは更に上位の処理系からの作用に依存すると推察している。即ち被験者が分析的な態度でこの音声を聴取した場合は,どちらか一方の項目が検索・照合された後に,他の (関連) 項目への間接的検索も行なわれ,その結果「懇談か,混乱か」との反応を示すであろうし,非分析的な態度でこの音声を聴取した場合は,間接的検索を行なう可能性は大幅に低減される[20]と筆者は考える。更に,人間が行なっている円滑なコミュニケーションには,この検索の "打ちきり" とも言うべき巧妙な操作が非常に重要な役割を果たしていると筆者は考えている。即ち,注目音声に対して,検索の幅を大きくした処理を常に行なっているならば,音響的には「懇談」,意味的には「混乱」,XY 的には「xy」と言ったような事態が生じ,円滑な音声処理は不可能なものとなってしまうからである。

3.6.2 辞書検索に影響を及ぼす要因

上述したように,人間は入力音声と内部辞書内の項目を照合する際,先行文脈の音響的・言語的情報から,予備的に辞書項目を選択・限定して照合を行なっている (検索・照合の順序を動的に変化させる) と考えられる。この場合当然のことながら,限定の度合は

[20] 但し上記したように,低減の度合も,両単語の音響的類似度をパラメータとして変化するであろう。

その文脈に依存し，ある特定の項目が明確に予測される場合もあれば，ある単語群が興奮の閾値に達しないまでも，緩やかに活性化される場合もある。この選択・限定処理が柔軟かつ適切に行なわれるため，辞書項目群の一部と照合しただけで適切な知覚が行なわれ，その時間的効率，信頼性も機械の認識に比べ格段によいものとなっていると言える。そこでこの節では，辞書検索順序の動的変化に対して影響を及ぼしていると考えられる要因を，以下，仮説として考察する[68]。

1. 音響/音声学的要因

   入力音声の一部，または全体的な音響/音声学的特徴が知覚されると，それを基に辞書検索範囲を制限することができる。但しこの場合必ずしも TRACE モデルが言うような，先行する音韻レベルでの知覚と，後続する語レベルでの知覚の相互作用を意味しているのではない。即ち，語を構成する音韻全てに対する "音韻レベルの" 知覚が完了する前，或は，音韻レベルでの知覚が困難な場合でも，語レベルの知覚が適切に行なわれ，その結果を基に音韻レベルでの知覚が行なわれるプロセスの可能性も含めているものである。更に，分節的特徴による音韻情報の伝搬が妨げられている場合でも，韻律的情報は正しく知覚されることがある (韻律的特徴の頑強性)。当然のことながら，韻律的情報のみでは，候補をある特定の一項目に限定することは出来ない。しかしながら，辞書検索範囲を絞ることは十分可能であり，その効果により，最終的にある特定項目へと限定するまでの処理時間が短縮されたり ("早い" 知覚)，より少ない音韻情報が提示される環境下においても，正確な知覚が可能となる ("容易な" 知覚) ことは十分に考えられる。

2. 辞書構造的要因

   Phonetic Refinement 理論では，内部辞書内の項目が，項目の存在する空間 (ここでは，項目空間と呼ぶことにする[21]。) 中にどのように配置されているか，についての考察が行なわれ，項目空間の疎/密と単語知覚との関連について述べられていた。このように，内部辞書の構造がある項目の検索に関する "容易さ" 及び "早さ/順序" に影響を及ぼしていることは十分考えられる。本研究では，この辞書構造的な要因として，以下の 2 点について実験的に考察する。

   長期的頻度 長期的使用頻度 (出現頻度) が高いものほど，その項目への検索はより早く行なわれると推測される。これに対して，使用頻度を一パラメータとして，内部辞書に対する動作が決定される処理部を仮定することもで

[21] Pisoni らによれば，"multi-dimensional acoustic-phonetic space" と呼ばれている。

きる。しかし内部辞書 (特に単語辞書) の静的構造が長期間に渡る学習を経て決定されることを考慮すると，各項目の長期的頻度は，内部辞書構造そのものに直接影響を与えるパラメータの一つとして捉える方が，より適切であると考えられる。

短期的頻度 短期的使用頻度 (出現頻度) が高くなった項目も同様に，その検索が早く行なわれると推測される。但し，内部辞書は LTM(Long Term Memory, 長期記憶) であるため，短期的頻度の一時的な上昇が辞書の構造そのものを変化させるとは考え難い。つまり，前者は純粋に辞書の構造に起因するものと考察できるが，後者は正確に言うならば，知覚結果を保持する STM(Short Term Memory, 短期記憶) が検索処理過程に対して，cache 的な役割を果たしているものと推測される。

3. 統語的要因

文音声中では先行文脈から次に発声される音声の統語的役割が限定されてくる。その結果，その役割を担える項目への検索が他項目より先行されると思われる。但し，日本語と英語を較べた場合，後者の方が一般に，より明確な統語的規則に従った言語であると言われるように，統語的要因による後続項目の制限に関しては，各言語による差があることも予想される。また，統語的要因による辞書検索限定の作用を観測する場合，ある文が対象とする言語の文法規則に対して持つ統語的整合性を，"有/無" の 2 値ではなく，連続量として定量的に定義できるのか？と言う課題も興味深いところである。

4. 意味的要因

先行文脈からの意味的係り受けによって辞書検索範囲を限定することができる。文 1, 2, 3 がその例でもある。従来行なわれてきた意味的要因に関する知覚実験では，その意味的整合性として上記の統語的要因と同様に，"有/無" の 2 値化を行なって音声試料が作成されている例が多い。このような試料を用いた場合，その定性的な特徴は得られるが，(心理学的に) 定量的な測定は不可能である。意味的整合性の "定量化" も課題の一つである。

5. 談話的要因

上記の意味的要因は，2 つの語 (主語+述語，述語+目的語，修飾語+名詞など) の局所的な意味的関係，あるいは文としての意味の "有/無" に対して議論される場合が多い。しかし，人間の音声言語活動を眺めた場合，このような局所的な言葉の繋が

りだけでなく，より広い大局的な言語的文脈の意味的整合性にも注目すべきである。即ち，ある文章音声において，各文は統語的/意味的にはどれも正しい文であっても，各文の間に何らの言語的関係を見い出すことができない場合，その文章音声の知覚は困難なものとなるであろう。このように，内部辞書の検索処理において，まず現在の "話題・焦点" に関連した項目に対して検索範囲を絞り，照合を行なっていると考えられる。そして，十分な整合性を示すものが得られなかった場合に，その検索範囲を広げていくと思われる。この談話的整合性に関しても，上記した "有/無" による定性的な実験が多く，"定量化" が必要とされるところである。

以上，辞書検索に影響を与えると考えられる要因を，仮説として考察してきた。次章において，上述した要因の幾つかに対して実験的にその効果・作用の様子を明確にしていく。なお，従来本研究室で行なわれた知覚実験を見ると，音声試料に対して行なわれる操作の対象は，殆ど音響的特徴に限られている。これでは音声言語の片面のみを分析しているに過ぎない。即ち，言語的特性を操作した上で音声試料を作成し，言語的特性の違いによって，音声知覚がどの様に影響を受けるのかを観測する必要がある。また，音響的操作と言う観点から第 3.4.2 節及び第 3.4.3 節を眺めた場合，どちらの実験も，ある音声区間の音韻情報を削除する，或は，異なる音韻情報を持った音声と置換することが行なわれており，これは分節的特徴を操作対象としていることに他ならない。一方，韻律的特徴に着目して本研究室で行なわれた実験を見ると，F_0 の外挿効果など，言語音としての性質が薄い音を対象としており，人間が音声言語を処理する場合に，韻律的特徴がどのような役割を持つのか？と言う問いに答えるには至っていない。以上の考察の下，次章で述べられる実験では，「言語としての音声」，「韻律的特徴 (特に F_0) の果たす役割」と言う観点からの分析も数多く行なわれている。

第 4 章

知覚実験による人間の音声知覚過程の分析

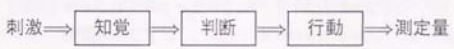
本章ではまず，知覚実験の構成について簡単に触れた後，第 3.4 節で概説した先行実験に引き続いて，筆者が中心となって行なった一連の知覚実験について説明する。これらの実験は，第 3.2 節で述べたように，人間の全体像を記述することができる知覚モデルの構築を主目的としており，第 3.6 節で考察した「内部辞書への検索過程，及び検索処理を制御する音響的/言語的要因」の分析が中心課題となっている。特に音響的要因に関しては，第 4 節最後の考察にあるように，韻律的特徴を考慮した知覚実験も行なっている。なお，本章でも第 3.4 節と同様に，「まとめ」の欄において，各実験の現象としての結果を "直接的結果" として，考察及び他実験における結果を考慮して間接的に得られる知見を "考察及び知見" として箇条書きして記すことにする。

第 1 章でも述べたように，音声知覚実験は対象とする音声長により，音韻/単語/文知覚実験のように分類される。筆者が行なった実験はこの中で，単語以上の音声を扱ったものである。当然のことながら扱う音声長が長くなるほど，注目すべき要因/現象も増える。そこで，実際に実施された時間的順序とは異なってくるが，対象とする音声長の順に (単語→句→文) 各実験を紹介していくことにする。

4.1 知覚実験の構成

4.1.1 知覚実験のモデル化

一般に行なわれている知覚実験をモデル化すると次のようになる。



このモデルを音声を刺激とした知覚実験に即して考える。まず，刺激音声が被験者に提示され，何らかの言語的符号へと変換される (知覚)。本来ならば，この段階までの処理を詳細に分析するのが，知覚実験に課された課題である。しかし，この知覚は受動的行為であり，この行為そのものを観測しても何ら測定量は得られない。即ち被験者に，知覚結果に基付き何らかの判断を行なってもらい，そしてその判断に基付いて，予め指定しておいた能動的な行動を行なってもらう必要がある。この，能動的な行動の段階まで来て初めて，実験者にとって測定可能な物理量が得られることになる。第 3.4.3 節の実験をこのモデルに当てはめるならば，/toku?ima/と言う刺激音声が提示され，その刺激 (特に/?/の部分) が何と言う音韻であるかが知覚される。その結果に基付き[1]，刺激が/tokuʃima/であったか否かを判断する。そして，判断結果に基付き，指定されたキーを押すと言う行動を行な

[1] あるいは "単語全体の特徴に基付き" となる。

う。このように非常に簡単なモデルを通して知覚実験を考えることができるが，以下の点に十分に注意すべである。即ち,「知覚結果に基付く判断および行動は出来る限り，容易なものとする。」と言うことである。知覚されてから測定量が観測される間の処理が複雑なものとなると，測定量に変動が生じた場合，それが知覚までの段階における変化に依るものなのか，その後の段階における変化が原因であるのかが特定できなくなるからである。被験者に課す行動の種類によっては判断が必要無いものもあり[2]，そちらの方が，知覚までの段階をより的確に反映した結果が得られると言える。なお第 3.4.3 節では，知覚のプロセスを更に「音響的特徴抽出処理」,「内部辞書検索処理」,「音響的照合処理」の 3 つに細分化して考察が行なわれている。

4.1.2 音声知覚実験における音声試料への操作

次に，音声試料 (刺激音声) に対して行なう操作について考える。当然のことながら，次節で考える被験者に課すタスク (モデルにおける "判断"+"行動") と関連してくるが，ここでは次のように大きく 2 つに分類して考える。

音響的操作 雑音重畳，帯域制限による音韻情報の妨害，フォルマントの推移による音韻情報の変形と言った分節的特徴の操作や，基本周波数 (以下 F_0 と記す) パターンのの変形，パワー (パターン) の変形，音韻長の変形などによる韻律的特徴の操作に区分される。

言語的操作 統語的/意味的/談話的正当性の有/無と言った定性的な操作の他に，ある定義の下に，談話的属性 ξ が，$\xi_- < \xi < \xi_+$ を満たす文セット，と言った定量的な操作も可能である。

以上の分類の他にも，提示音声の全体を操作するのか，特定の一部を操作するのか，などによる分類も可能である。

4.1.3 音声知覚実験で用いられる指標 (タスク)

本実験で用いる指標は，大きく次の 2 つに分かれる。1 つは，種々の音響的/言語的聴取妨害の下での音声提示に対する正答率に基付くもので，その結果測定量は，"知覚の容易さ" を表すものとなる。もう 1 つは，提示音声に課した判断/行動が完了するまでの時間を測定するものである。その結果測定量は，"知覚の早さ" を表すものとなる。この場合も被験者に要求する行動を工夫することで，判断に必要とされる時間を十分抑えることが可能である。

[2] 口頭再生などは，知覚がそのまま行動に繋がる。

4.2 長期的頻度が単語音声知覚過程に及ぼす影響に関する実験

4.2.1 背景と目的

まず，単語固有の性質が単語音声知覚に及ぼす影響を考える。人間は，ある (単語) 音声に対してそれが既知であるか未知であるか，即ち内部辞書内に存在する項目か否かの判断ができる。しかし既知の項目全てに対して等しい処理方式を用いているとは考えにくい。そこには，日常会話を長いスパンで観測した場合の出現頻度，即ち長期的頻度によって，処理方式に差が生まれてくることが予想される。また，この長期的頻度は長い時間をかけて生成されるものであり，LTM の中に，該当する項目の静的な特徴として，何らかの表現形式で記載されていると考えられる。この長期的頻度の及ぼす影響を実験的に実証することを本実験の目的とする[62][74]。

なお frequency effect として知られる，出現頻度が音声知覚に及ぼす影響に関しては多くの先行実験がある[47]。しかし，これらの実験の多くは，実験中に，ある特定の音声試料の出現頻度を操作したり，新聞記事などのデータベースより計算される頻度を用いて音声試料を作成しているものが多い。前者のように人工的に (動的に) 生み出された頻度は，一時的な出現頻度の変動を意味しており，LTM で構成される内部辞書内に項目固有の性質として記述されてあるとは考え難い。また後者は，対象とするデータベースの量を増やせば長期的頻度を定義することは可能である。しかし，算出される頻度は，経済・政治用語に対して極めて高いスコアを示すなど，新聞の特性から来る影響が不可避的に付随してしまう。また，日本語の場合第 3.3.2 節にあるように，文字提示された場合と音声提示された場合とで，親密度[3]にずれがあることも懸念される。上記の考察の下，本実験は，内部辞書内に静的に[4]，項目固有の性質として記載されてあると考えられる長期的頻度の影響を，特定の分野への偏りが存在しないと判断できる音声試料を用いて評価しようとするものである。なお，実験中の音声試料を用いて行なう出現頻度の制御 (短期的頻度) については，次節で扱うことにする。

4.2.2 実験方法

音声試料

本実験を定量的に評価するためには，特定の分野への偏りが無く，かつ長期的頻度が客観的に評価できると考えられる音声試料を用意する必要がある。これらの条件を満足させ

[3] 厳密な意味では長期的頻度と異なる指標であるが，両者の間には高い相関があると考えられる。
[4] 即ち，文脈などに依存しない。

る単語として，名字音声に着目した。名字音声は，音声試料としてはやや特異な試料であるが，名字音声であることを予め明示しておくことで，被験者に，内部辞書中で名字(単語)が記載されている部分のみを対象とした検索処理を行なわせることができる。逆に，「ある事象(経済・政治など)に関連する単語」などような分野の限定を行なった場合，内部辞書中のどの部分までを検索対象とすべきかが非常に曖昧なものとなる。一方，名字音声の場合は，そのような曖昧性が極端に低く，この意味でも名字音声は本実験に都合の良い音声試料であると言える。

日本人の名字のランキング表(約 6,000 位まで)[75]から，3 モーラ名字群を 7 グループ(**A**~**G**，1 グループ 10 個で計 70 個)を，グループに対応する人口が **A** から **G** の順に約 1/2 ずつ減少するように選ぶ。選出された名前のリストを，各グループの平均人口と共に表 4.1 に示す。この名字単語を成人男性 1 名に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。その後，S/N=-7.5, -5.0, -2.5, 0.0[dB] となるようにランダムノイズ[5]を重畳し，1 名字音声に対して，4 種類の雑音重畳音声を作成し(合計 70×4=280 個)，これを音声試料として使用する。但し S/N 比は，着目する名字音声全体の平均パワーとランダムノイズのパワーとの比として定義している。

被験者

東京大学工学部学生 7 名，同大学院生 3 名，計 10 名

実験手順

同一 S/N 比の 70 種類の名字音声をランダムに並び替え，70 個全てを 1 セッションとしてヘッドフォンを通して両耳から提示する。これを，S/N 比の低い方から(明瞭度の悪い方から)順に計 4 セッション(1 日 1 回，計 4 日)行なう。被験者には，聴取用意が出来次第，パソコンのリターンキーを押すよう指示しておく。その 4 秒後に雑音が重畳した名字音声が提示される。下に示すタスクを行なった後，リターンキーを押すことで次の音声の入力待ちとなる。以上の操作を刺激音声数だけ繰り返す。また，実験後アンケートを行ない，使用した名字が既知であったか未知であったかを調べた。なお，追加実験の結果，実験で用いた音声試料は無雑音下では 100[%] の正答率であった。

タスクと指標

インストラクションに示すように，提示音声聴取後，即座に口頭再生させる。指標としては，グループ間の正答率の相違及び S/N 比の変化による正答率の相違を見る。

[5] ホワイトノイズの近似として用いた。

インストラクション

これから，日本人の名字の音声を聞いてもらいます。音声はリターンキーを押すことで提示されます。但し，音声にはノイズが重畳してあります。まずノイズが約 1 秒間聞こえ，その後ノイズに埋れた形で音声が提示されます。音声を聞き取ったら何と聞こえたかをマイクに向かって発声して下さい (口頭再生)。分からなければ発声する必要はありません。音声はヘッドフォンを通して両耳から提示されます。口頭再生後，用意が出来たら再度リターンキーを押して，次の音声の聴取に備えて下さい。それでは宜しくお願いします。

4.2.3 実験結果

図 4.1 にグループ別，S/N 比別の実験結果を示す。なお，横軸はグループの平均人口 ([万人]) を log 軸で示したものである。

4.2.4 考察と検討

今回使用した名字音声試料は人口分布に基付いてグループ分けされているので，各グループの長期的頻度は，A,...,G の順番に確実に低くなっている。実験後のアンケートの結果から，A, B, C については 100[%]，D, E についても 90[%] 以上の名字が既知であり (以下，ここまでを既知音声とする)，F は 68[%]，G は 28[%](ほぼ未知に近い) であった。これは，G において正しく知覚された名字が，音節を単位とした処理結果であることを意味する。図 4.1 より，G に対する正答率は非常に低く，特に S/N=-5, -7.5[dB] の実験条件下では，音節単位での処理が十分に機能していないことが分かる。当然のことながら，G 以外の刺激音声に対しても音節単位の処理は正しく機能せず，従って本実験 (特に S/N=-5, -7.5[dB]) の結果は，"語" 全体を単位とした処理結果であると考えて良い。このように考えると，正しく識別された名字音声に対しては，「まず (名字) 単語辞書に対して該当項目への検索が行なわれ，検索された項目と入力音声を項目全体の特徴を利用して照合した結果，十分な音響的整合性を示した。」と言うことができる。逆に，誤って識別した名字音声に対しては，「検索はされたものの，項目全体の特徴を利用して照合を行なっても，十分な音響的整合性が得られなかった (かつ，後続して検索された項目で十分な整合性を示すものがあった)。」のか，或は「先行して検索された項目に十分な音響的整合性を示すものがあり，そこで入力音声に対する処理が終了した。」のどちらかであると解釈できる。両者のどちらであるかについての議論は，第 3.6.1 節を参照して頂きたい。

既知音声 (A~E) の場合，全ての環境下で長期的頻度が高いほど正答率も高くなる結果

が得られた。これらより，同一処理単位 (この場合，“語”) が用いられている場合でも，長期的頻度が高い項目は，より少ない音響的情報量 (第 3.4.3 節における低精度の音響的特徴に対応) で正しく知覚される，即ち知覚が容易な項目であることが分かる。これは，

- 同一処理単位においても複数の精度の音響的特徴を扱う機構が備っていること。
- 項目固有の性質である長期的頻度を一パラメータとして，「適切な辞書検索・正しい音響的照合は，どの程度の低精度特徴まで可能であるのか」が決定されること。

を示している。第 3.4.3 節の考察において，低精度の音響的特徴はより早期の段階で抽出が完了するとの処理部を仮定した。即ち，長期的頻度の高い項目はその項目固有の性質として，「音声処理におけるより早期の段階で，正しい音響的照合が可能な項目」と言える。但し，辞書検索処理において該当項目が検索されなければ，いくら上記の性質を有していても，照合処理は行なわれない。人間における情報処理が自然淘汰の下最適化されていると考えるならば，上記の性質が十分生かされた処理系が実現されているはずである。即ち長期的頻度の高い項目は，その固有の性質として，「早期の段階で辞書検索され」かつ「早期の段階で抽出される低精度の特徴量と，正しい照合が開始される」項目であると推測される。更に情報量の側面から考えた場合，低精度の特徴を用いた音響的照合に必要な処理時間は，高精度の特徴の場合より，長くはないと容易に仮定することが出来る[6]。その結果長期的頻度の高い項目は，最終的に，より早く知覚されることが十分に予測される。

また，誤識別の結果を見ると長期的頻度の高いものとして知覚しているものが殆どであった[7]。これに対して上記した 2 つの観点からの考察が可能であるが，本実験だけで結論を出すことは出来ない。しかし，第 3.6.1 節での考察を考慮すると，いわゆる “打ち切り” による後者の説を採択すべきであると筆者は考えている。本実験の場合は，第 3.6.1 節での考察とは異なり，孤立発声された単語が対象であり，“各項目固有の性質の差” が “実験結果における差” を生み出している。「優先的に検索・照合される」と言う性質は第 3.1.1 節，第 3.1.3 節の言葉を借りるならば，“無文脈で既に活性化されている” 状態にある項目である。そして，聴取時の態度が分析的では無い場合，活性化済みの項目に十分音響的整合性を満たすものが存在すると，そこで処理が終了する，と言う訳である。当然のことながら雑音重畳している本実験の場合，被験者は分析的態度をとること (正確には，分析的態度による効果を望むこと) は困難である。

なお本実験では，第 4.1 節で述べたような，単語音声知覚を「音響的特徴抽出処理」，

[6] 第 3.4.3 節 (先行研究) では等しいとの仮定を置いている。
[7] 但しこの点に関しては，正確なデータを残していないことを断っておく。

「音響的照合処理」,「辞書検索処理」に明確に区分した形で観測しておらず，上記考察においても特性の一処理部へ限定しては行なわれていない。例えば，第 3.4.3 節で行なわれたように，提示音声を予め被験者に提示することで，"検索された/されない" に対する議論は必要なくなる。即ち音響的照合処理に絞って議論を進めることが可能となる。今後の課題の一つである。また，今回の実験ではデータ数との兼ね合いから，図 4.1 の定式化は行なわなかった。人口と言う定量的指標が与えられていることを考えると，1 グループの平均人口を更に細かく設定し，被験者を増やした上で実験を行なうことも残された課題であろう。

4.2.5 まとめ

直接的結果

- ある辞書項目を正しく知覚・同定するために必要な音響的情報量は，同一処理単位が用いられている場合でも，項目固有の特徴である長期的頻度を一つのパラメータとして変動する (同一単位による処理においても，複数の精度の特徴量を扱う機構が存在する)。
- 長期的頻度の高い項目ほど，少ない音響的情報量 (低精度の特徴) で正しい照合が可能と (知覚が容易に) なる

考察及び知見 (予測)

- 長期的頻度の高い辞書項目は，早期に検索され，また，早期に抽出が完了する低精度の音響的特徴との正しい照合が行なわれ，その結果，いち早く知覚が終了する。

表 4.1. 実験で使用した日本人名字リスト

数字は各グループの平均人口 [万人] であり，A から G の順に，対応する平均人口が約 1/2 倍されるよう名字単語を選んでいる。

グループ A	スズキ	サトウ	タナカ	イトウ	カトウ
(86.5)	ヤマダ	ヨシダ	ササキ	キムラ	シミズ
グループ B	ハヤシ	オガワ	イケダ	ウチダ	オカダ
(25.5)	アオキ	カネコ	オオタ	コジマ	シマダ
グループ C	マエダ	イシイ	ヨコタ	ハラダ	ノムラ
(11.1)	タナベ	イシダ	マツダ	クロダ	イマイ
グループ D	ヤジマ	イナバ	イグチ	オカノ	ニシオ
(5.0)	オオキ	ノザキ	ヤスイ	エグチ	オクノ
グループ E	エハラ	タカミ	アリガ	フカイ	アリタ
(2.2)	ヨコオ	シマノ	シオノ	タザキ	テライ
グループ F	セキノ	ツジタ	ニシデ	エトウ	ヤナセ
(1.0)	ムラキ	カタノ	マエノ	イソダ	マブチ
グループ G	ツザキ	モトダ	テライ	イクノ	アラオ
(0.5)	アナミ	ヌマノ	ワケベ	シブエ	キヨセ

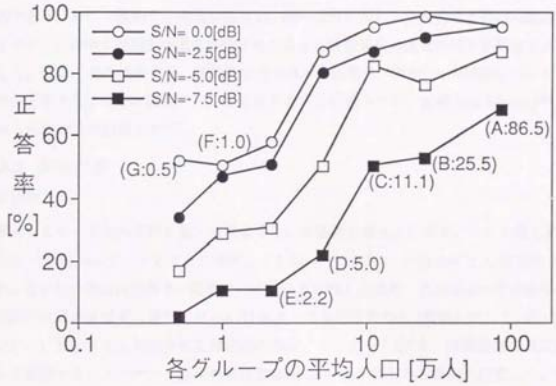


図 4.1. 長期的頻度が単語音声知覚過程に及ぼす影響に関する実験結果
横軸は log軸で示されており，括弧の中の数字は各グループの平均人口［万人］を表す。

4.3 短期的頻度が単語音声知覚過程に及ぼす影響に関する実験

4.3.1 背景と目的

第 4.2 節の最後に，長期的頻度の高い項目は時間的に早く知覚されることを予測した。長期的頻度の異なる 2 つの項目を比較する場合，異なる項目を用いると，項目の意味的容易さ・親密度や，構成する音韻の音響的特徴の違いに等による影響が生じるため，同一項目を異なる頻度を用いて提示することが望ましい。しかし，長期的頻度は，辞書項目に固有の性質と考えられるため，同一項目に対して異なる長期的頻度を付加することは非常に困難である。そこで観測する期間を短縮し，同一項目に対し，実験内でその出現頻度を変化させる，いわゆる短期的頻度を変化させた場合の知覚過程の変化の様子を観測することとした。そしてその結果より，長期的頻度の違いと知覚の "早さ" との関係についても間接的に考察する。なお本実験では，知覚の早さを直接扱うため，追唱 (shadowing)[76] と呼ばれる実験方法を採用した[77]。

4.3.2 実験方法

音声試料

平易な 4 モーラ名詞単語を数十個用意する。単語例を表 4.2 に示す。これを成人男性 1 人に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。なお，各単語の意味的容易さ・親密度のばらつきを抑えるため，各単語は小学校低学年用の国語の教科書を参考に選出した。A/D 変換した単語音声から (重複を許して) 約 150 個ランダムに抽出する。作成された単語列の中に，ターゲット語[a]を，非関連語を間に数個挟んで配置する。ターゲット語間の単語数は 0, 2, 4, 6, 8 以上の 5 段階を用意した。ターゲット語の配置された単語列リストを参照して，刺激間隔 4 秒となるように D/A 変換し，DAT に録音したものを音声試料として使用する。このように単語音声は全て一度 A/D 変換したものを使用するため，提示音声中の同一単語は同一波形をもった音声となる。ターゲット語には，波形による視察を容易にするため語頭音が/t/である単語を用いた。具体的には/takuaN/ 及び/tamaʃii/の 2 つである。なお，刺激間隔の 4 秒と言うのは，前置単語の音響的情報が STM から消滅し，言語的情報のみが残った状態で次単語が提示されるよう，決定されたものである。

被験者

東京大学工学部生 1 名，同大学院生 1 名，技官 1 名，計 3 名

[a] 実験者側が着目する音声試料。

実験手順

刺激間隔 4 秒で作成された単語音声列 (約 150 個) を被験者に提示し，以下のタスクを行なわせる。

タスクと指標

以下に示すインストラクションの下，各提示単語音声に対して追唱を行なわせる。追唱とは図 4.2 に示すように，提示された音声を，即座にそのまま口頭で再現する操作のことを言う。但し，当然のことながら異なる音韻に対しては異なるタスクを要求することになる (異なる音韻を発声させる)。その結果，音韻別の発声し易さの違いの影響を受けてしまう。そこで語頭音韻を揃え (/t/)，指標としては，追唱開始遅れ時間を波形視察より求めることとした。また，追唱は個人差が大きい実験方法でもある。そこで，予備実験としてまず音節追唱を行ない，被験者を，比較的安定して追唱を行ない，かつ，個人差の少ない 3 人に絞った。

インストラクション

これから，刺激間隔約 4 秒で，単語音声群を聞いてもらいます。但し，耳に入力された音声を即座に発声するように努めて下さい。特に語頭の部分では，意味をとってから，あるいは考えてから提示音声を発声することの無いようにして下さい。聞こえた音声をそのまま，口に出して頂ければ結構です。語頭から語尾にかけて (意味が分かり始めると)，発声速度が上がることと思いますが，それは差し支えありません。なお，実験開始に当たって「追唱実験を始めます」というアナウンスをして下さい。その後，音声提示が始まります。

4.3.3 実験結果

図 4.3 に示す。ターゲット語間の単語数が多くなるほど，追唱開始遅れ時間が大きくなっていることが分かる。

4.3.4 考察と検討

第 4.1 節で考察したように，孤立単語音声に対する知覚の早さは "音響的特徴抽出処理"+"辞書検索処理"+"音響的照合処理" に要する時間の大小によって決定されると考えられる。本実験では，提示音声に自然音声を用い，また，雑音重畳や帯域制限などの聴取妨害も施していない。また，追唱と言うタスクの性質上，被験者は実験中，非常に分析的な態度になるため，本実験条件下では音韻/音節単位での高精度の特徴までをも抽出する

処理も働いていることは十分考えられる。かつ，その抽出結果を受けて，高精度特徴での照合処理が行なわれていることは十分考えられる。このように，高精度処理まで行なわれている特徴抽出部や照合処理部に対して，反応時間縮小の主要因を考えることは困難である。更に，ターゲット語間の時間間隔は 4 秒以上に設定しているため，STM に一時的に保存される音響的特徴を使用したがために追唱開始が早くなった，との議論も困難である。また，本実験において注目する短期的頻度の変化は，同一音響音の繰り返しとは異なり，同一オブジェクトを指す (意味を持った) 言語音の繰り返しである。その結果，音声知覚過程における言語処理部に対して何らかの動的な作用を及ぼすことが考えられる。一方，特徴抽出部や照合処理部は純粋に音響的な低次の処理であり，この点からも上記 2 つの処理部に対して，直接的に，本実験の主要因を考えるのは困難と言える。そこで，本実験結果を "辞書検索処理" の短縮化と考える。こうすると非常にスムーズに考察できる。内部辞書は莫大な情報を抱えたデータベースであり，その検索順序を誤れば正しい項目への検索に必要とされる時間に与える影響も非常に大きいと考察される。この検索処理において，最近検索された項目が優先的に (より早期の段階で) 検索されるようになると言うのは非常に自然な考えであり，また本実験の現象を説明するに十分な妥当性もある。つまり，提示音声に対してより早く正しい追唱が行なえる場合，それはより短い提示長で該当項目への検索が可能な状態 (活性化，早い知覚へと繋がる) となっていると考えられる。また，この検索処理部を制御する一処理部として言語処理部を考えれば，上記の動的な言語的作用とも合致する。

ここで，短い提示長中に含まれる特徴量と，第 4.2 節の考察で述べた，早期の段階で抽出が完了する特徴量とは厳密な意味では異るが，情報量と言う観点から眺める限り，両者を等しく扱ってもよいであろう。そして，第 3.4.3 節，第 4.2 節を考慮すると次のことが導かれる。短期的頻度の上昇した項目は，優先的に辞書検索されるようになり (活性化)，照合処理部へ，より早期に送られることになる。しかし，語頭からの短い提示長中に含まれる低情報量の特徴のみでは，正しい照合が行なえない (該当する項目との照合の結果，十分な整合性が得られない) 状態にあるとするならば，本実験のような結果は望めない。故に，早期に検索されるように活性化された項目は，早期の段階で抽出される低情報量の特徴量を用いた照合処理においても正しい照合が可能となっている項目であると言える。

以上の結果を，内部辞書の構造的変化と捉え，短期的な出現頻度の上昇により，辞書内項目の一特性が変化し，その結果，より早期の段階で検索されるようになった，と考察するには無理がある。と言うのも，内部辞書は LTM で構成されており，その中にある各項

目の特性も，長期に渡る学習/リハーサルによって獲得される，言わば静的なものだからである。これらの特性を変動させるためには，十分な時間を費やして，再度学習/リハーサルする必要がある。さて，それでは上記の短期的頻度の向上による影響はどのようにモデル化されるべきなのだろうか？次のように考えてみる。一度知覚・同定された項目が cache 的な STM に保存され，その STM が LTM(内部辞書) に先立って優先的に検索される。この STM 内の項目は，LTM において長期的頻度の高い項目が持つ特性と似た特性 (上記で言う "活性化" された状態,「早期に検索」・「低情報量での正しい照合可能」) を持っていると考察することができる。また，図 4.3 より，項目の出現頻度が落ちてくる (ターゲット語間の単語数が増えると) と検索の優先度も低くなる様子が伺える。しかし，今回の実験では追唱の特徴上，被験者を前もって絞るということを行なったが，実験方法を工夫して，多くのターゲット語と多くの被験者の下で実験を繰り返す必要があるだろう。

第 4.2 節より，長期的に高頻度の項目は，低精度の特徴で正しい照合が可能であることを直接的に見た。一方本節では，短期的に高頻度の項目は辞書検索が優先される (早くなる) ことを直接的に見た。そして，各々において両者を結び付ける知見が導出された。このように，"低精度の特徴による正しい照合の可否 (容易な知覚)" と "早期の段階での辞書検索の可否 (早い知覚)" とは非常に高い相関があることが分かる。

4.3.5 まとめ

直接的結果

- 同一項目の場合でも，短期的にその出現頻度が高くなると，正しい追唱を開始するに致るまでの時間 (反応時間) は短縮される。

考察及び知見

- 短期的に出現頻度の高くなった項目は，cache 的 STM に保存され，優先的に検索されるようになる (活性化)。
- 優先的に検索される (活性化されている) 辞書項目は，低情報量の特徴で正しい照合が可能な状態となる。そして，早期の段階で抽出される，低情報量の特徴との照合により，正しい知覚がいち早く完了する。

表 4.2. 本実験で用いた単語リスト例

ターゲット語は枠で囲んで表示してある。再左列から順に，ターゲット語間の単語数が，0, 2, 4, 6 となっている。

ライオン	れんこん	カシミヤ	にっさん
えんぶつ	ベクトル	ライオン	ながさき
たいふう	タバスコ	にっさん	タレント
いのしし	にっさん	チーター	もっきん
ニッパー	へいめん	たいすう	くつした
じゃがいも	ちからこぶ	おわかれ	たくあん
タイトル	おわかれ	かきぞめ	たいふう
こんちゅう	じゃがいも	ベクトル	べきじょう
あしくび	たつまき	たくあん	もっきん
こんちゅう	たくあん	にくたい	れんこん
ぼうれい	じゃがいも	カステラ	ぶっしつ
じゃがいも	あしくび	チーター	オランダ
クレヨン	たくあん	あしくび	たくあん
たくあん	カシミヤ	たくあん	カシミヤ
たくあん	はねつき	いのしし	タバスコ
どしゃぶり	へいめん	おやゆび	おわかれ
もっきん	そつろん	だいぶつ	きかがく
：	：	：	：

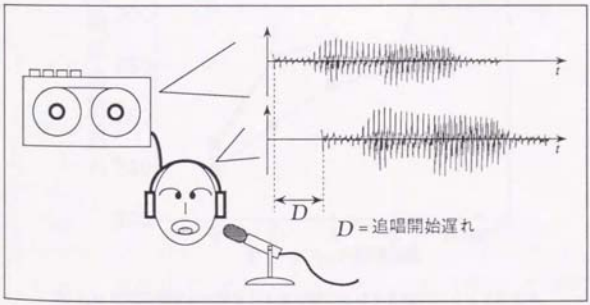
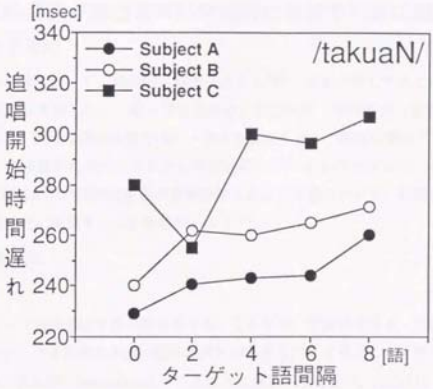


図 4.2. 追唱 (shadowing)



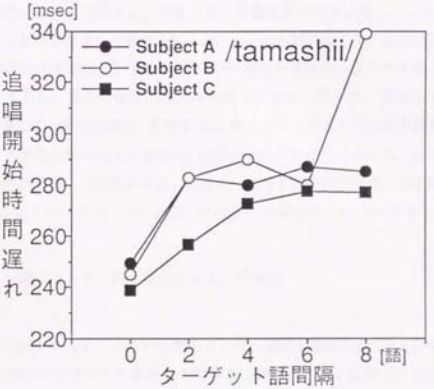


図 4.3. 短期的頻度が単語音声知覚過程に及ぼす影響に関する実験結果
上図がターゲット語/takuaN/に対する結果であり，下図がターゲット語/tamaʃii/に対する結果である。

4.4 意味的要因が単語音声知覚過程に及ぼす影響に関する実験

4.4.1 背景と目的

第 4.3 節の実験において，短期的に繰り返される[9]同一項目に対して生じる効果を言語処理との関連から考察した。一般に言語処理は，統語解析・意味解析・談話解析の 3 つに分類されるが，第 4.3 節の実験では同一項目の反復と言う，特殊な環境下での実験であり，上記の 3 つの解析処理のいずれかに直接対応しているものではない。そこで本節では，第 3.6.2 節に示した意味的要因が辞書検索に及ぼす影響について，単語音声提示のパラダイムで実験的に実証することを目的とする[77]。

4.4.2 実験方法

音声試料

平易な 3 モーラ名詞単語を数十個用意する。これらは，意味的容易さ・親密度のばらつきを抑えるため，小学校低学年用の国語の教科書を参考にして選出したものである。これらを成人男性 1 人に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換したものを音声試料として使用する。この単語音声群から (重複を許して) 約 70 個を選び，単語リストを作成する。単語リストを表 4.3 に示す。但しリストの中には，4 種類のターゲット語が含まれており，各々のターゲット語に対して，先行単語列に関連性のある単語が存在する場合 (**TYPE 1**) と，非関連性の単語のみ存在する場合 (**TYPE 2**) を両方用意しておく。後者の場合，非関連単語 1 つを間に置いて，更にターゲット語を繰り返して配置した (短期的頻度，**TYPE 3**)。表 4.3 のリストの順に音声試料は提示される。このように単語音声は全て一度 A/D 変換したものを使用するため，提示音声中の同一単語は同一波形を持った音声となる。なおターゲット語の語頭音は，波形視察を容易にするため/t/に揃えた。各々，/tatami/, /taNbo/, /taNgo/, /taiho/である。

被験者

東京大学工学部学生 5 名，同大学院生 3 名，計 8 名

実験手順

被験者がパソコンのリターンキーを押すことで，実験が開始される。4 秒後に単語リストの最初の単語がヘッドフォンを通して両耳より提示される。被験者には以下に示すタスクを課し，その結果に基付いて Y/N のキーインをするよう指示しておく。音声提示は，

[9] 但し，音響的情報が十分に STM から消滅するだけの間隔は置いている。

キーインと同時に停止する。その 4 秒後に次単語が自動的に提示される。以上の操作を音声試料数だけ繰り返す。なお，提示間隔 4 秒と言うのは，第 4.3 節における実験と同じように，STM から音響的情報が消滅し，言語的情報のみが残った状態で次の音声試料が提示されるようにとの考慮からである。

タスクと指標

以下のインストラクションに示すように，入力音声に対してその単語が「動物か否か」を判断させ，パソコン上に設定した Y/N に該当するキーを押させる。指標としては語頭からキーインするまでの反応時間を測定する。また実験前に，Y/N に該当するキーをビープ音に反応して押す，という形で "単純反応時間" を測定した。なお，表 4.3 を見て分るように，全ターゲット語に対して "NO" の判断が下されるはずである。

インストラクション

これから，単語音声を聞いてもらいます。但し，音声を聞きながら，それが動物か否かの判断をして下さい。そして動物であったならば右手の人差し指でカナの「ロ」を，そうでなければ左手の人差し指でカナの「ツ」を押して下さい。音声はそこで止ります。このキーイン後，4 秒して次の単語が自動的に提示されます。なお，音声はヘッドフォンを通して両耳より提示されます。また，各キーの上に軽く指を載せた状態で提示音声を聴取するようにして下さい。それでは宜しくお願い致します。

4.4.3 実験結果

語頭からキーインまでの時間から，単純反応時間を引いたものを "純粋反応時間" と呼ぶことにする。図 4.4 に TYPE 2 の純粋反応時間に対する TYPE 1 の純粋反応時間の比を各単語毎に示す。図 4.5 に TYPE 2 の純粋反応時間に対する TYPE 3 の純粋反応時間の比を各単語毎に示す。

4.4.4 考察と検討

ある単語を「知覚」してから，"動物であるか否か" の「判断」に必要とされる処理量は，同一単語であれば，有意な差は生じない[10]と仮定できる。逆に，異なる単語間では「判断」に要する処理量/時間が異なってくると考えられ，これらを直接比較することは望ましくない。図 4.4, 4.5 を各単語毎に示したのは以上の理由からである。今回の実験では，1 被験者に対して，ターゲット語毎に 3 回同一の判断を行なわせているが，同一ターゲット語

[10] 特に単語音声提示の実験パラダイムでは。

間の距離は十分離れているので，この「判断」についての "慣れ"，及びそれによる処理時間の短縮，というのも考えにくい。以上の考察より，図 4.4 に示されている TYPE 2 に対する純粋反応時間の減少は「知覚」の段階での反応の (即ち，知覚の "早さ" における) 差であり，その差をもたらしたのは，提示単語音声を取り囲むコンテキストの変化であると言える。

本実験で観測された「知覚の早さ」の差も第 4.2 節や第 4.3 節と同様に考察することができる。つまり，関連性のある項目が先行コンテキストにおいて検索 (及び活性化・興奮) された結果，ターゲット語の提示以前に，第 3.6 節で言う間接的活性化がターゲット語に対して行なわれる。即ち，一時的に，かつ緩やかに活性化された状態となる。この状態は第 4.3 節での考察と同様に「正しい照合に必要な音響的情報量は少なく」かつ「早期の段階で辞書検索が行なわれる」と言う特性を持つ[11]と考えられる。その結果，先行コンテキストに意味的関連性のある単語が存在する場合も，より早く「知覚」が行なわれるようになる，と言える。

ここで，本実験と第 4.3 節の実験における相違について更に深く考察する。但し，ここでは被験者に課したタスクではなく，被験者に提示した刺激音声の質的違いに注目する。第 4.3 節においては，先行コンテキスト中にターゲット語と全く同一の単語音声を挿入した。本実験では，十分に関連があると判断される単語を挿入している。つまり前者の場合，ターゲット音声に対応する辞書項目は，聴取前に，確実に直接的検索・照合処理を経ていると考えられるが，後者の場合，その辞書項目は必ずしも，直接的検索・照合処理を経ているとは言えない。第 3.6 節での議論を考慮すると，後者の場合においても，間接的な検索は十分に考えられる。ここで問題となるのが，同一項目の繰り返しによる効果に対しては cache 的 STM の存在を考察したが，本実験においても，先行コンテキストに関連する項目は常に cache 的 STM にその情報が逐次保存されるか否かである。もし，先行コンテキスト中の単語に関連する項目が常に cache 的 STM に入力されているとした場合，その容量は莫大なものとなるか，あるいは，cache 的 STM 内の情報は非常に短期間に消滅してしまうことになる。まず，後者に対して第 4.3 節の結果を基に考察する。第 4.3 節において，一度 (検索・照合処理を経て) 知覚された辞書項目は，全く関連性の無い項目を更に数個を知覚する間は，cache 的 STM 内に保持されていることが示された。当然のことながら，一般の言語活動においては，関連性を持つ辞書項目が連続しており，cache 的

[11] 但し，その特性を実現する手段が同一かどうか (即ち cache 的 STM か否か) については更なる議論が必要である。

STM 内に保持される時間も第 4.3 節程度，あるいはそれ以上と考えられる。即ち，cache 的STM 内の情報は，非常に短期間の時間経過では消滅しないことが分かる。即ち，cache 的STM の存在で本実験が説明されるためには，容量の大きい STM(即ち前者の説) が必要となってくる。しかし，7 chunk と呼ばれる STM の容量を考えるならば，これも非常に困難な仮説である。以上の考察より，先行コンテキスト中の項目と意味的関連のある項目に対する検索における優先性と言うのは，cache 的 STM のような辞書構造的な要因ではなく，LTM 内にある内部辞書項目への検索順序を管理する辞書検索処理部が存在し，その処理部の特性における動的変化によるものであると考察される。

4.4.5 まとめ

直接的結果

- 提示音声がある属性を持つか否かの判断を行なわせた場合，先行コンテキストに提示音声と意味的関連性のある音声が含まれると，その反応時間は短くなる。

考察及び知見

- 先行コンテキスト中に，項目 A に対して意味的関連性の有る項目が存在していた場合，辞書検索処理部における，検索方法の動的変化によって，項目 A への検索処理が優先されるようになる。

表 4.3. 本実験で用いた単語リスト

ターゲット語は枠で囲って示してある。また，ターゲット語に対して関連性が有るとして配置された単語は先頭に• を付けて示してある。

ヨット	キリン	たたみ	サラダ	テレビ
さくら	• そうさ	やもり	たいほ	たぬき
ひつじ	たいほ	あたま	とかげ	マイク
タイヤ	たぬき	こねこ	たいほ	すずめ
きつね	たんぽ	こいぬ	ことり	たんば
きつね	たぬき	うさぎ	ひぐま	• しゅしょく
ぽると	たんぽ	• ダンス	やもり	• ごはん
りんご	すみれ	• タップ	かもめ	マイク
タンゴ	さんそ	こねこ	• しょうじ	• おこめ
さんま	ちっそ	• おどり	• わしつ	たんぽ
タンゴ	ねずみ	タンゴ	ひぐま	
キリン	たんそ	パンダ	• ふすま	
• じゅんさ	たたみ	コアラ	たたみ	
• けいじ	すみれ	インコ	ひつじ	

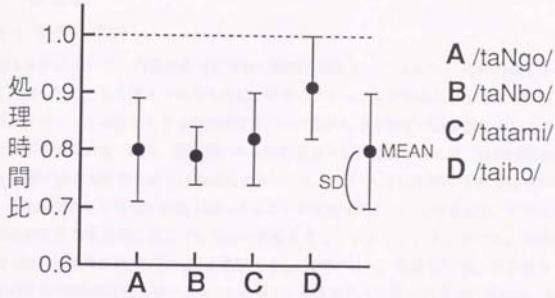


図 4.4. 意味的要因が単語音声知覚過程に及ぼす影響に関する実験結果 (TYPE 1)

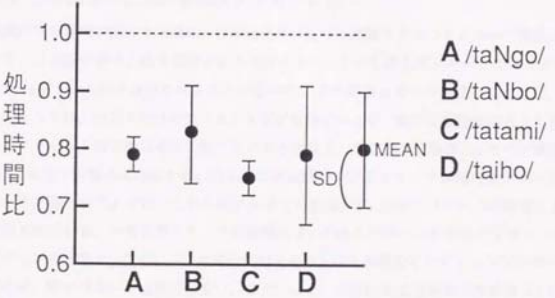


図 4.5. 意味的要因が単語音声知覚過程に及ぼす影響に関する実験結果 (TYPE 3)

4.5 単語アクセントが単語音声知覚過程に及ぼす影響に関する実験

4.5.1 背景と目的

第 4.2 節において，内部辞書項目が持つ固有な特徴の一つとして，長期的頻度を考えた。同様な特徴として考えられるものに，単語のアクセント型がある。もちろん，入力音声のアクセント型が既知となった場合でも，その音声をある特定の辞書項目に対応付けることは不可能である。即ち，最終的にある辞書項目として識別するには，音韻情報を伝搬する分節的特徴の分析を経なければならない。しかし，人間は音声の音韻情報のみならず，韻律的情報をも容易に知覚・識別することが可能であり[12]，この事実は，アクセントの情報が音声知覚過程に対して何らかの影響を与えていることを予測させる。韻律的情報は一般に基本周波数 (以下，F_0 と略記する)，音源パワー，音素 (節) 長，休止長などの音響的特徴 (韻律的特徴と呼ばれる) によって伝搬されるが，特に日本語の場合は，F_0 によって伝搬される情報が重要な位置を占めていると言われる[69][70]。この点からも，単語音声の F_0 パターンのを分類することにより定義される，アクセント型の及ぼす影響は興味深い。以上の考察より，アクセントが音声知覚過程に及ぼす影響を実験的に検証することを目的として，単語提示のパラダイムで知覚実験を行なうこととした[78]-[81]。

4.5.2 日本語音声における単語アクセント型

実験内容を詳細に記述する前に，日本語音声における単語アクセントについて簡単に説明する。日本語の場合，孤立発声された単語アクセントはしばしば，モーラ数分の F_0 の "High/Low" 2 値系列で表現されることが多い[82]。その結果 n モーラ単語に対しては，パターンとしては，2^n 通りだけのアクセント型が可能であるが，実際に単語音声として使用されるアクセント型の数はそれに較べてかなり少なく，東京方言の場合，n モーラ単語に対して，$n+1$ 種類のみ存在する。図 4.6 に日本語における 4 モーラ単語の全アクセント型を示す。図に示すように，これらのアクセント型は，F_0 の降下パターンの位置によって分類されている。一般に第 i モーラの直後に F_0 の降下パターンを持つアクセントを i 型アクセントと呼ぶ。但し，i=0 の時は例外であり，i=0 の場合そのアクセントパターンの中には，降下パターンは存在しない。また，n モーラ単語を孤立発声した場合，0 型単語と n 型単語では，全く同一の F_0 パターンを示す。両者の区別は，単語の後に助詞等が接続された時の，その助詞のアクセントによって行なわれる。即ち，0 型単語の場合は後

[12] だからこそ，音声の存在意義があるのかもしれない。

続助詞の F_0 も High のままであるが，n 型の場合，後続助詞の F_0 は降下パターンを描く(図 4.6 参照)。本実験では，4 モーラの孤立単語音声のみを扱うため，音声試料としては，0~3 型アクセント単語のみを使用することとした。

4.5.3 実験方法

音声試料

表 4.4 に示すように，0~3 型アクセント単語を，各型につき 12 個ずつ，計 48 個用意する。これらの単語を東京方言話者の成人男性一人に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。これらの音声データを基に，PARCOR 分析合成音[83]を作成する。但し，分析後の再合成の前に以下に示す 3 通りの操作を F_0 に対して施す。

CASE 1 F_0 を 110[Hz] 一定に変換した後に再合成。

CASE 2 F_0 を操作することで，アクセント型を他の型へと変換した後に再合成。

CASE 3 F_0 へは何も操作せず，再合成。

上記の操作の中で，**CASE 2** の操作は F_0 パターン生成モデル[84](以下，F_0 モデルと略記する) に基付いて行なわれた。F_0 モデルにおいて，アクセント型はアクセント指令の立ち上がり/立ち下がり時刻，及びその大きさ等によってモデル化されているが，本実験では，立ち上がり/立ち下がり時刻を操作し，High あるいは Low のモーラの位置をずらすことで，アクセント型の変換を実現している。本来ならば，アクセントの型によってアクセント指令の大きさも異なってくるはずであるが，ここでは制御を簡単にするため変換前後におけるアクセント指令の大きさまでは操作していない。表 4.4 は，**CASE 2** における型変換と刺激音声との対応も示している。表にあるように，一つの単語に対して，可能な型変換を全て行なっている訳ではなく，任意の一単語に対して施す型変換は一種類のみである。表の左側に $i \rightarrow j$ と示してある単語群は，本来 i 型アクセントを持つ単語を **CASE 2** において，j 型に変換していることを示し，以後 "単語群 $i \rightarrow j$" と呼ぶことにする。以上のようにして得られた合成音声に対して，0.5[kHz]~3.0[kHz] の帯域制限 (BEF) を更に施したものを，音声試料として使用する。この帯域制限は，音韻情報 (分節的特徴) の伝搬を制限することで，提示音声に対する音節単位の left-to-right 処理を困難にすることを目的としている。先行研究によれば，通常の音声言語活動では，語レベルの音声知覚単位が優先的に使用されるとの知見がある[66]。しかし実験下においては，被験者がより分析的な態度になることは容易に推測され，その結果，音節レベルを知覚単位とした処理が行

なわれることが予想される。入力音声を音節系列として容易に捉えることが可能な環境では，提示音声のアクセント形態への依存性は低くなると考えられる。しかし上述したように，これは通常の音声言語活動とは異なる環境下での観測となる恐れがある。そのため，人工的に語レベル以上での知覚のみが可能となるよう帯域制限を施した。即ち，単語全体の特徴で知覚を行なわせることを目的として行なったものである。

被験者

成人男性 10 名

実験手順

F_0 パターンに対して上記の操作を行なって作成した，各 CASE48 個の分析合成単語音声を，成人男性 10 人に提示間隔 4 秒でヘッドフォンを通して両耳より提示し，これを 1 セッションとする。被験者には以下に示すタスクを課す。セッション終了後数時間して，同一被験者に対して，異なる CASE の合成音声 48 個を再度提示する。なお，提示順序は CASE 2→CASE 1→CASE 3 であり，合計 3 セッションで実験は終了する。なお，全セッションに先立って，上記した 48 個とは別個に作成された，F_0 操作，帯域制限後の単語音声を使って，聴取の訓練を行なっている。

タスクと指標

インストラクションに示すように，被験者には予め，聴取後即座に提示音声を同定し，その結果を単語本来のアクセント型で口頭再生するよう指示しておく。

インストラクション

これから，単語音声を聞いてもらいます。但し，提示音声はある操作が施されており，聞き難くなっています。単語音声の提示が終了したら即座に，何と言う単語であったかを正しいアクセントで発声して下さい (口頭再生)。単語提示の終了後 4 秒して，次の単語が提示されますので，次単語の聴取の妨げにならぬよう，口頭再生は聴取後即座に行なって下さい。単語は 48 個，ヘッドフォンを通して両耳より提示されます。それでは宜しくお願い致します。

指標としては，アクセント型別の正答率を以下に示す 2 通りの観点より求める。まず，表 4.4 の単語音声データを以下 2 通りの方法で，各々4 グループに分類する。いずれの方法も CASE 2 の型変換を基に行なった分類である。そして，得られた結果から各グループ毎の正答率を求める。

- 単語群 $i \to j (0 \leq i, j \leq 3)$ を i を基に分類。単語群 $i \to *$ が 1 グループを形成することになる。これは提示単語本来のアクセント型による分類である。以降分類 A と呼び，$i \to *$ を分類 A の i 型と呼ぶ。

	0 型	1 型	2 型	3 型
分	0→1	1→0	2→0	3→0
類	0→2	1→2	2→1	3→1
A	0→3	1→3	2→3	3→2

- 単語群 $i \to j (0 \leq i, j \leq 3)$ を j を基に分類。単語群 $* \to j$ が 1 グループを形成することになる。これは型変換後 (CASE 2) のアクセント型による分類である。以降分類 B と呼び，$* \to j$ を分類 B の j 型と呼ぶ。

	0 型	1 型	2 型	3 型
分	1→0	0→1	0→2	0→3
類	2→0	2→1	1→2	1→3
B	3→0	3→1	3→2	2→3

4.5.4 実験結果

分類 A による結果を各型 (単語群) 毎に図 4.7 に，分類 B による結果を各型 (単語群) 毎に図 4.8 に示す。各々横軸が F_0 パターンに対して行なった操作であり，縦軸が正答率である。図中 CASE 1~3 では，提示音声と被験者からの応答音声とが "単語として" 一致した場合のみを正解とした正答率を示している。一方 CASE 2′では，CASE 2 の実験結果を，提示音声の提示時のアクセント型と，被験者からの応答音声のアクセント型が一致した場合[13]のみを正解とした正答率 (以後アクセント型正答率と呼ぶ) を示している。また，図 4.7, 4.8 の TYPE n とは上述した分類 A/B の n 型に属する単語群を意味する。

4.5.5 考察と検討

図 4.7, 4.8 の CASE 3 の結果より，A/B の分類によって定義される各型間に，同定の難度に差が無いことが分かる。これは間接的にではあるが，各型間に合成音としての品質に差が無く，両分類における，各型間の直接的な比較が可能であることを意味している。但し，合成音の絶対的品質，及び帯域制限から来る同定の難度より，その正答率は 80[%] 弱であり，語サイズの知覚単位が使用された場合でも，その 8 割しか正しく同定できない環境下での実験であることをも意味している。

図 4.7 における，CASE 3 と CASE 1 との結果を比較してみると，F_0 を平坦化 (未知アクセント型) することによる正答率の大幅な低減が，0 型以外の音声に対して観測され

[13] 単語として見た場合は，当然誤認識である。

ている。これは，1, 2, 3 型アクセント単語音声の処理過程において，韻律的特徴が有効に寄与していることを示している。0 型に単語に対しては正答率の低減が観測されていないが，これは 0 型の F_0 パターンはアクセント核 (F_0 降下パターン) を持たず，かつ **CASE 1** によって変換された F_0 パターンとの間に他型ほど差が生じなかったからであろう。次に **CASE 2** の結果を見てみると，この場合は，0 型においても大きな正答率低減の様子が観測されている。しかし，この実験結果より，「0 型単語においても韻律的特徴が有効に作用している」との結論を導くことはできない。即ち，**CASE 2** は既知の他型への変換時の正答率であり，「提示単語は本来 0 型である」と言う情報と「提示アクセントは既知の $i(\neq 0)$ 型である」と言う情報のどちらが作用してこの正答率低減を招いたかは判断出来ない[14]からである。この問題については，F_0 上昇/降下パターンを持つ未知アクセント型を用いた実験などが更に必要である。

1, 2, 3 型において，韻律的特徴の音声知覚に対する有効的寄与が観察されたと述べたが，この様子について更に詳しく見ることにする。図 4.7 の **CASE 1, 2** を見ると，1 型の単語群の正答率は他型のものに較べて極端に低いことが分かる。これは，本来 1 型アクセントを持つ単語音声を F_0 一定 (**CASE 1**)，或は他アクセント型へ変換 (**CASE 2**) すると，途端にその同定が困難になることを意味する。また，図 4.8 の **CASE 2** を見ると，1 型の単語群の正答率は図 4.7 ほど顕著ではないものの，他の型より低くなっている。これは，非 1 型アクセントの単語を 1 型に変形して提示すると，その同定が他型と比較して相対的に難しくなることを意味する。更に **CASE 2′**を見ると，1 型の単語群の正答率が他型より相対的に高くなっている。これは非 1 型アクセントの単語を 1 型に変形すると，提示時のアクセントにつられて，本来 1 型アクセントを持つ単語として同定される (誤認識) 傾向がより高いことを意味する。以上をまとめると，1 型アクセントは他型とは異なって形態で知覚され，その結果，1 型アクセント単語は提示時の F_0 パターンに大きく依存して (F_0 パターンを大きな手がかりとして) 知覚されることになる，と言える。上記の現象は次のように説明される。1 型アクセントは第 1 モーラ直後にアクセント核 (F_0 の降下パターン) を有しており，これが単語音声知覚の早期の段階で検出され，“提示音声が 1 型である” 或は “1 型ではない” と言う情報がいち早く使用可能となる。その結果 1 型単語は，この F_0 パターンから得られる情報により依存した形で知覚され，また，提示音声を 1 型に変形してしまうと，内部辞書 (LTM) 照合処理において検索範囲が限定され，図 4.8 の **CASE 2′**の様な誤認識も引き起こされることになる。

[14] 即ち，0 型で無くなったことが効いているのか，既知の $i(\neq 0)$ 型になっていることが効いているのか。

しかし，アクセント核が後方にずれる 2,3 型アクセントの順に，単語音声知覚に F_0 パターンが利用される度合が減少している様子は特に観測されていない。これは単語尾まで入力される前の段階で，その時点までの分節的特徴を用いた単語全体での知覚が可能となっていることを示唆する[67]。

本実験は，「帯域制限をした音声に対する，聴取直後の口頭再生における正答率」と言う形での評価であり，上記した 1 型アクセントの同定の "早さ"，及びそれによる 1 型アクセント単語の知覚の "早さ" を直接的に見たものではない。そこで，Gating Paradigm (第 4.8 節を参照) の技法を用いた異なる実験[15]の結果をアクセント型別に見たところ，1 型と非 1 型との間で，「該当辞書項目へ確実に検索が行なわれるために必要な語頭からの音声提示長」に，有意差が見られた。具体的な分析結果などについては，第 4.8 節で述べることにする。

4.5.6 まとめ

直接的結果

- 音韻情報の伝搬への妨害を行なった上で，辞書項目固有のアクセント型とは異なるアクセント型で提示した場合，その正答率は大きく減少する。
- 孤立単語音声知覚のアクセントへの依存性は，その型によって異なる。特に 1 型の場合，その影響力は大きくなる ("提示音声が 1 型である" と言う情報が，知覚の "容易さ" を増大する)。
- 提示時のアクセントにつられて誤認識する傾向が最も大きいのは，非 1 型を 1 型への変換した場合である。

考察及び知見

- "アクセント核が語頭に存在する" と言う性質が，1 型アクセントの同定を早める。
- その結果 1 型アクセント単語は，より早く知覚されることになる。この知見は，異なる目的で行なわれた過去の実験結果を再分析することで，部分的に実証された。

[15] 第 4.8 節で紹介する実験である。但し，BEF 等の後処理は無く，音節単位での知覚を許している。

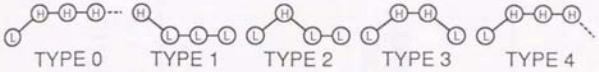


図 4.6. 日本語 4 モーラ単語におけるアクセント型

表 4.4. 音声試料と CASE 2 における型変換との対応

i→j = CASE 2 において本来 i 型アクセントの単語を j 型へ変換

0→1	"raioN"	"akabou"	"niNjiN"	"naiyou"
0→2	"shiNgou"	"omatsuri"	"yokujitsu"	"amerika"
0→3	"hiroshima"	"aimai"	"orugaN"	"raihiN"
1→0	"nekutai"	"koumori"	"wakuchiN"	"raNdamu"
1→2	"naitaa"	"monoraru"	"amazoN"	"uNsei"
1→3	"kamakiri"	"uNmei"	"ookami"	"eNbuN"
2→0	"imomushi"	"norimaki"	"omusubi"	"yononaka"
2→1	"mimizuku"	"katakori"	"onigiri"	"nodoame"
2→3	"aomori"	"toraburu"	"murasaki"	"origami"
3→0	"kamisori"	"tamanegi"	"nokogiri"	"machigai"
3→1	"kaminari"	"nissuu"	"neNryou"	"noNbiri"
3→2	"kaminoke"	"nakigoe"	"noumiso"	"teNkizu"

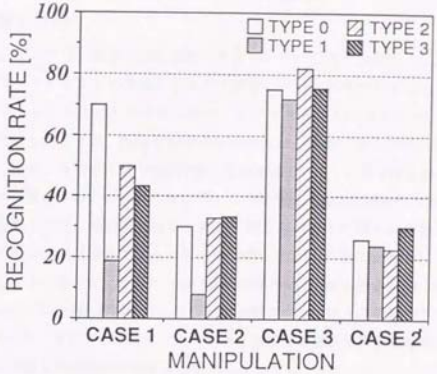


図 4.7. 単語アクセントが単語音声知覚過程に及ぼす影響に関する実験結果 (分類 A)

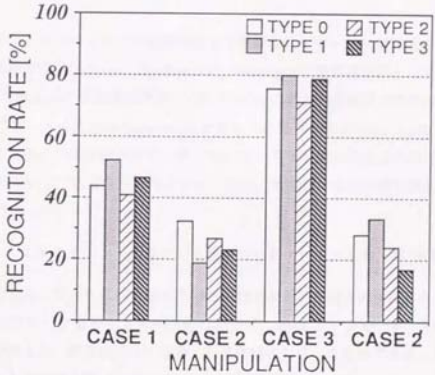


図 4.8. 単語アクセントが単語音声知覚過程に及ぼす影響に関する実験結果 (分類 B)

4.6 単語アクセントの知覚に関する実験

4.6.1 背景と目的

第 4.5 節においては，孤立単語聴取時のアクセントの果たす役割について実験的に考察した。即ち，アクセントの情報がもたらす作用についての考察であった。本節では更に進んで，アクセントそのものが持つ特性，或はアクセントの知覚について検討する。第 3.3.1 節に示したように，単語音声中の音韻知覚は，無意味語中の同一音韻の知覚より早くなる。即ち，提示音声が "内部辞書に登録されている" と言う情報がその単語を構成する音韻の知覚を早めている訳である[63]。この効果は "lexical effect" と呼ばれている。第 4.5 節で見たように，単語アクセントの数と言うのは非常に限られた数であり，単語と同様に，何らかの表現形態でアクセントの情報も LTM の中に保有されていると考えられる。そしてこの場合も，アクセントパターン中のある F_0 の動きに対して，上記した "lexical effect" に似た現象が観測されることが予想される。以上の考察の下，第 4.5 節では使用されていない未知アクセントパターンへと変換した単語音声をも使用し，アクセントそのものの知覚を実験的に考察する[78][79]。

4.6.2 実験方法

音声試料

3 型アクセントを持つ 4 モーラ名詞単語を 3 種類及び 4 モーラ無意味語を 3 種類用意する。今回の実験では，"aozora", "geNshiro", "korigori" を名詞単語として，"imeyuro", "nemeira", "ominere" を無意味語として使用した。これらの単語及び無意味語を成人男性の東京方言話者一人に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。これらの音声データに対して，第 4.5 節と同じように PARCOR 分析合成方式を用いて，合成音声を作成する。但し，本実験では各単語音声に対する F_0 パターンの操作として，

- 0, 1, 2, 3 アクセント型及びの 6 種類の未知アクセント型，計 10 種類。

の型変換を行なう。図 4.9 に本実験で用いた未知アクセント型を示す。これらの未知アクセント型には 3 モーラ目から 4 モーラ目にかけて，F_0 パターンの上昇/下降のあるものと無いものがあるが，後者はダミー音声として作成されているものである。本実験における，アクセント型変換も第 4.5 節と同様に，基本的には F_0 モデルにおけるアクセント指令の立ち上がり/立ち下がり時間を操作することで行なっているが，第 3, 4 モーラにおけ

る F_0 の動きに関しては，更に以下に述べるような操作を行ない，各刺激音声間での統一を図っている。図 4.10 に模式的に示すように，3 モーラの後半から 4 モーラの頭にかけて，少なくとも 100[msec] は F_0 = 一定となるようにし，4 モーラ目の先頭部分で F_0 パターンの上昇/下降を行ない，その後 4 モーラの F_0 も一定となるように変形した。更に上昇/下降は，75[Hz]→110[Hz]/110[Hz]→75[Hz] で行なわれるようにし，4 モーラの F_0 もその値を使用した。このような制御の下作成された 6×10=60 個の音声データに対して，以下の 2 種類の方法で無音置換 (図 4.11 参照) を行なったものを音声試料として使用する。この手法は，Gating Paradigm[85] と呼ばれる手法である。

CASE A 語頭から x[msec] を残し，それ以降を無音置換する。

CASE B **CASE A** の操作に加えて，第一，二モーラをも無音で置換する。

CASE A において，x[msec] は第 3 モーラまでの音声長を初期値として，5[msec] ずつ増やしていきながら無音置換を行ない音声試料を作成する。一方，**CASE B** の方の音声試料は，第 3, 4 モーラの一部のみの提示であり，F_0 の上昇/下降そのものの知覚の様子を観測するために作成されたものである。以下 x のことを Gating Period と呼ぶ。

被験者

成人男性 1 名，成人女性 3 名，計 4 人。

実験手順

有/無意味単語 6 語に対して各単語別に，10 通りのアクセントパターン及び，ある Gating Period x に対応する，10 種類の音声試料をヘッドフォンを通して両耳より提示し，これを 1 セッションとする。但し，まず **CASE B** によって作成される音声試料の提示から行なう。**CASE B** の音声試料の提示では提示間隔を 2 秒に設定する。セッションを繰り返す毎に x を 5[msec] ずつ増やしていき (x の初期値は上記した通り)，セッションの繰り返しは確実に F_0 の上昇/下降が知覚できるようになるまで行なわれる。なお，音声提示のアクセントパターンによる順序は，セッション間でランダムに変更する。**CASE B** の実験終了後，**CASE A** の音声試料を用いた実験をほぼ同様な手順で行なう。正し，提示間隔は 6 秒に設定し，セッションの繰り返しは，既知/未知に拘らずアクセントパターンの同定が正確に行なわれるようになるまで行なう。

タスクと指標

インストラクションに示すように **CASE B** の実験においては，被験者に提示音声のアクセントについて，上昇/下降/平坦の 3 通りから選択させ，所定の用紙に記述するよう指

示した。一方 CASE A の実験では，提示音声のアクセントパターンを "High/Low" の 2 値時系列で，所定の用紙に模式的に記述させた (図 4.6 参照)。なお，CASE A の音声試料提示に対してアクセントパターンが同定できない場合は，その旨を用紙に記述させた。指標としては，F_0 の上昇/下降のみを提示した場合の (CASE B)"F_0 の動き" の知覚に必要な提示長，及び，同一波形が単語内に存在した場合の (CASE A)"F_0 の動き" の知覚[16]に必要な提示長を実験結果より求める。

インストラクション (CASE B)

これから，音声を聞いてもらいます。但しその音声長は 1 音節以上，2 音節以下です。音声の聴取後，第一音節から第二音節にかけて，アクセントが上昇しているか，下降しているか，あるいは平坦かの判断をし，その結果を所定の用紙に書いて下さい。上昇の場合は右上がり，下降の場合は右下がり，平坦の場合は水平の線分を書いて頂ければ結構です。なお，音声提示は 2 秒間隔で行なわれますので，次提示の音声の聴取を妨げることの無いよう，記述は素早く行なって下さい。また，全体の音声提示の終了はこちらから指示致します。それでは，宜しくお願い致します。

インストラクション (CASE A)

これから，4 モーラの単語音声を聞いてもらいます。但し，音声の後半部分は無音で置換され，不完全な状態となっています。音声の聴取後，第一モーラから第四モーラにかけてのアクセントの動きを，所定の用紙に，例にならって模式的に記述して下さい。なお，第三モーラから第四モーラにかけて，アクセントの動きが聞きとれなかった場合は，その旨書いて頂ければ結構です。音声提示は 6 秒間隔で行なわれますので，次提示の音声の聴取を妨げることの無いよう，記述は素早く行なって下さい。また，全体の音声提示の終了はこちらから指示致します。それでは宜しくお願い致します。

なお，以下では男性被験者 (1 人) の結果の分析を行なうが，他の女性被験者 (3 人) の結果も同様の傾向を示していることを予め断っておく。

4.6.3 実験結果

図 4.12 に有意味語に対する CAES A, B の結果を示し，図 4.13 に無意味語に対する CASE A, B の結果を示す。両方の図において，縦軸は F_0 の上昇/下降を確実に知覚するために (CASE B)，或は F_0 のパターンとしての正確な知覚に (CASE A) 必要な，第四モーラ先頭からの提示長を意味する。また横軸には，本実験で使用した 10 通りのアクセ

[16] 但し，こちらの場合は "語全体のアクセントのパターンの知覚" とも言える。

ントパターンの内，第三モーラから第四モーラにかけて，F_0 の上昇/下降が存在する 5 種類のパターンを記載している。これらのパターンは，既知パターン 1 種類 (3 型)，未知パターン 4 種類である。

4.6.4 考察と検討

まず，CASE B の結果について考える。図 4.12，図 4.13 より有意味語/無意味語どちらのグループに対しても，F_0 の上昇に対する知覚がより早く行なわれることが分かる。但し CASE B の場合は，後半の 2 音節に対して Gate をかけながら提示するため，無音置換前の単語の有意味性は殆ど関係しない。しかし，F_0 の上昇/下降部のみの知覚に対して，この様な差が生じたことは興味深い。一方 CASE A の結果を見ると，提示単語の有意味性に依らず，既知パターンである 3 型の反応が，未知パターンのそれより，いち早く行なわれることが分かる。ここで，3 型の第 3→4 モーラは下降パターンであり，未知パターンの内 3 つは上昇パターンを有していることにも着目したい。更に，CASE A と CASE B の結果を比較すると，3 型アクセントにおいて，CASE A での反応が CASE B より早く行なわれていることが分かる。そして，この傾向も有意味性に依らず観測されている。しかし，観測されているのは 3 型においてのみであり，未知アクセント型においては，上記の傾向が明確に現れているものは無い。ここで仮に，アクセントパターンの情報が単語の意味，品詞などを記述する辞書 (以下，単語辞書と呼ぶ) 中に記述してあると仮定してみる。当然のことながら，無意味語はこの単語辞書に登録されてはいない。その結果無意味語に対しては，提示音声のアクセントパターンが既知である/ないによる差は生じ得ないことになる。これは明らかに実験事実と相反するものである。以上の考察より，アクセント型の情報は単語辞書とは別個の辞書として，LTM 内に存在していると言える。そして，アクセント辞書と単語辞書との各項目間には，ポインタによって結合されていると考えられる。更に，入力音声の有意味性によらず，既知アクセント型の反応が早いと言う実験事実は，人間の音声知覚過程には，入力音声を有限個のアクセント型のいずれかであると仮定して処理を行なう機構が存在することを示唆する。上記したアクセント型辞書であるが，当然のことながら，単語辞書に較べて，そのサイズは小さいものとなり，検索に要する時間，処理量はかなり低くなる。故に，単語辞書の検索に先だってアクセント辞書による照合が行なわれ，一早く得られるアクセント型に関する情報を利用して，単語辞書の検索が行なわれると考察される。この考察の妥当性は，第 4.5 節に述べた，一型アクセントの単語知覚実験結果が十分に示している。

4.6.5 まとめ

直接的結果

- F_0 の上昇/下降パターンを単独で提示した場合，上昇の知覚に要する刺激長の方が短い。
- 一方，既知&未知アクセント型語尾に，F_0 の上昇/下降パターンを挿入して提示すると，既知アクセント (F_0 の下降に相当する) に対する反応が早い。これは，提示音声の有意味性によらず，観測された。
- F_0 の下降を単独提示と単語内提示で比較した場合，アクセント型が既知パターンである刺激に対してのみ，単語内提示の反応時間の方が確実に短くなる。この現象も，有意味性によらず観測された。

考察及び知見

- アクセント型の辞書が各単語の意味・統語的役割などを記載する辞書とは独立して存在する。
- 入力音声の有意味性によらず，既知アクセントパターンで音声を捉えようとする機構が存在する。

以上述べてきた実験は，基本的には単語音声知覚と呼ばれる領域の知覚実験である。しかし本研究の目的の一つは，「人間の音声知覚過程全体を見渡すことのできる知覚モデルの構築」であり，そのためには日常の音声言語活動の観測・分析，即ち文 (章) 音声を用いた実験が不可欠である。唯一第 3.4.2 節の実験において，提示音声の音響的・言語的環境の 1 つであるコンテキスト長に目が向けられ，単語以上のサイズの音声試料が用いられていたが，上記の目的を達成するためには，より深い考察の下，更なる実験が必要であることは明らかである。そこで以下の実験では，文の統語構造あるいは，文の記述する意味的内容にまで目を向け，人間の音声知覚過程の全貌に迫ることにする。

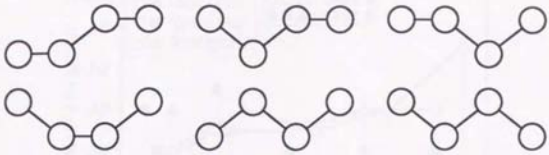


図 4.9. 本実験で使用した未知アクセント型

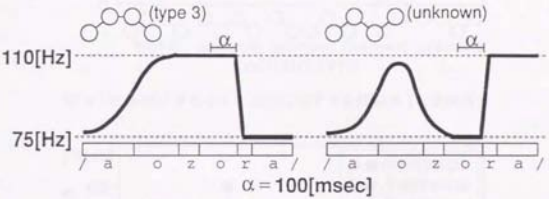


図 4.10. 第 3, 4 モーラに対する F_0 の制御

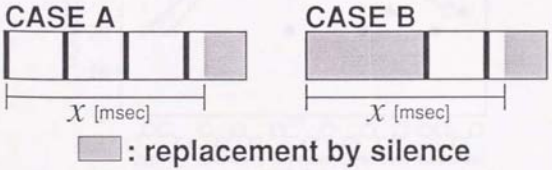


図 4.11. 2 種類の無音置換

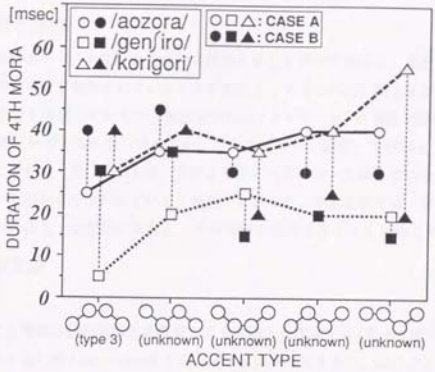


図 4.12. 単語アクセントの知覚に関する実験結果 (有意味語)

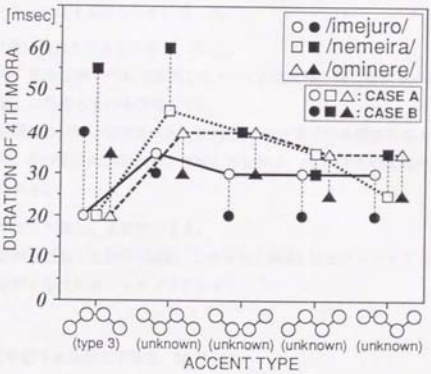


図 4.13. 単語アクセントの知覚に関する実験結果 (無意味語)

4.7 文節以上の音声処理単位に関する実験

4.7.1 背景と目的

第 3.4.2 節において，人間は入力音声を複数の単位を用いて処理し，連続文章音声では主に文節程度の単位が使用されていることが知見として得られた。第 3.4.2 節では単位長として文節までを考慮していたが，提示音声内容 (メッセージ) の言語的特性によっては更に大きな単位が考えられないだろうか。つまり慣用句・諺等，文字列として覚えてしまっているような句・文においては，文節より更に大きな句・文単位での全体的・大局的な特徴に基付く照合も行なわれていると推測できる。そこで本実験では，句・文レベルの照合処理単位の存在を実験的に実証し，その特徴を明示することを目的とする[62][74]。

4.7.2 実験方法

音声試料

以下に示す 3 種類の言語的特性を考慮して 4 文節文を作成する (表 4.5 参照)。これらの文を成人男性 1 名に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。その後，3 種類の方法で音声をセグメントし (**CASE A, B, C**)，S/N=-5[dB] でランダムノイズを重畳する。但し S/N の値は，各 **CASE** において，着目する音声単位全体の平均パワーに対する値として計算した。

TYPE 1 諺 (使い慣れた文として使用，原文)。

TYPE 2 諺の一部を文節レベルで置換したもの (文節置換，但し置換前の諺は **TYPE 1** の諺とは異なるものを使用する)。

TYPE 3 文節間に特別強い関連性の無い文 (文節程度の単位で処理されると考えられる)。なお，**TYPE 2** の文節置換で使用した文節は，本 **TYPE** の文節の一部として使用される。

CASE A セグメント無し。文音声のまま。

CASE B 文音声を前後 2 文節毎 (以降，これを句と呼ぶ) にセグメントする。

CASE C 文音声を各文節毎にセグメントする。

被験者

東京大学工学部学生及び大学院生，計 14 名

実験手順

上記被験者を各々7人の2グループに分け，一方に CASE A を，他方に CASE B 及び C をヘッドフォンを通して両耳より提示する。パソコンのリターンキーを押すと4秒後にノイズが2秒提示され，続いてノイズが重畳した音声試料が提示される。音声提示が終わって1秒ノイズが続く。このノイズが終わると同時に被験者には以下に示すタスクを行なわせる。その後，再度リターンキーを押すことで次の音声の入力待ちとなる。以上の操作を音声試料の数だけ繰り返す。音声試料の提示順序は表 4.5 の文音声をランダムに並び変えたものである。なお本実験は，音声試料の言語的特徴による差に着目しているため，同一内容の音声を同一被験者に繰り返し提示することは望ましくない。そこで，提示回数は各 CASE の各音声につき，各々1 回ずつである。

タスクと指標

インストラクションに示すように，音声提示終了後，即座に「聞こえた通りに」再現させる (口頭再生)。「聞こえた通りに」というのは被験者に入力音声に対する音響的照合のみを行なわせ，曖昧に聞こえたところに対して高次の処理 (知識による埋め合せ等) を極力行なわせないようにするためである。指標としては，提示した音声試料 (文/句/文節) と正答率との関係，及び TYPE と正答率の関係を見る。

インストラクション

これから文/句/文節音声を聞いてもらいます。ただし，音声はノイズに埋もれた形で提示されます。音声 (ノイズ) が終了したら即座に，聞こえた通りにマイクに向かって発声して下さい (口頭再生)。分からない部分は再生する必要はありません。音声はリターンキーを押すことで，ヘッドフォンを通して両耳から提示されます。口頭再生後，用意が出来たら再度リターンキーを押して，次の音声の聴取を行なって下さい。

4.7.3 実験結果

文節単位で提示した場合の正答率を文節正答率，句の場合を句正答率，文の場合を文正答率と呼ぶことにする。TYPE 別の平均正答率は表 4.6 のようになる。また，実験結果例を表 4.7 に示す。枠内の数字はその部分を単独で提示した際の正答率で，括弧内の数字は TYPE 2 の文を置換前の語として口頭再生した率である。

4.7.4 考察と検討

文正答率を見ると TYPE 1 は高く，TYPE 2・3 は非常に低い。但し，TYPE 2 の場合，その半数以上が元の諺として同定され (文節修復，57.1[%])，正しく正答した割合と合計すると約 70[%] になる。

TYPE に依らず非常に低い文節正答率に対して，文提示した場合における，文中の文節毎の正答率は TYPE 1 で 86.4[%]，TYPE 3 で 33.6[%] であった。文節単位での知覚が行なわれていると考えられる TYPE 3 における正答率の上昇は，音声試料の形態が文節→文へとなったことにより，統語的・意味的情報が使用可能となった結果であると考えられる。インストラクションとして "聞えた通りに" 再生するよう指示してはいても，完全には高次処理と低次処理を分離して知覚することが困難であること (不可分性) を示している。しかし，TYPE 1 における正答率の上昇は TYPE 3 と比較して大きく異なっている。この TYPE 1 における大幅な上昇に対しても，不可分な高次処理による影響のみで説明するのは非常に無理がある。つまり，文節単位で処理が行なわれている TYPE 3 とは性質の異なる処理過程 (†) が TYPE 1 の音声試料に対して作用していると推測される。

TYPE 2 における文節修復であるが，正しく知覚された周辺部分と知識からの埋め合せで合成したのでは，という議論がある。しかし，これに対しては文節正答率の大幅な減少から否定することができる。つまり，置換していない周辺部分においても文節単位での知覚が困難なのだから，穴埋め的な知覚・再生も当然のことながら，非常に困難なものとなる。そこでこの文節修復の現象に対して，文・句サイズの処理単位 (辞書項目) の存在を仮定して議論を進めてみる。第 3.4.3 節において，音節単位での無音置換は，単語・文節サイズでの処理単位によって処理単位内に吸収されてしまったとの結論を出しているが，本実験においても同様な考え方をすることができる。即ち，文・句サイズの処理単位が用いられたことで，それよりも小さな文節サイズでの特徴変動は吸収され，辞書項目として存在する現文 (諺) として知覚された，と考察することができる。

第 3.4.3 節において，処理単位長と正しい照合処理に必要な音響的特徴精度との知見が得られている。この知見を適用すると，文・句サイズを単位とした処理は更に低精度の特徴量で正しい照合 (知見) が行なわれることになり，これは TYPE 1 と TYPE 2, 3 における文正答率の差を生じさせた原因として十分妥当性がある。当然のことながら，諺である TYPE 1 において文・句サイズ単位の処理が行なわれること[17]は容易に考えられる。更に，上記した TYPE 3 に対する処理とは異なる処理過程 (†) として，文・句サイズの処理

[17] もちろん，文・句サイズの存在を仮定した場合，である。

単位を考えてみる。辞書項目は全て学習/リハーサルの繰り返しにより，後天的に構築されるものであることを考えると，TYPE 3 における，意味的・統語的情報の利用による単語間の結合性よりも，辞書項目として存在する TYPE 1 における単語間の結合性[18]が極めて強いことは容易に考えられる。そしてこれが TYPE 1 と TYPE 3 における文節正答率上昇の差を生じていると考察される。

以上考察してきたように，本実験の結果は，句・文サイズの処理単位 (辞書項目) を仮定することで，全て説明される。これは，この仮定に十分な妥当性を与えるものである。但し本実験は，S/N=-5[dB] と言う悪条件及び，諺と言う特殊な音声試料を用いて行なわれている。即ち，文・句単位での処理が駆動される様子を，非常に限られた状況で観測したものであり，文・句単位での処理過程の存在 (及びその特性の一部) を明らかにしたに過ぎない。また，本実験のような悪条件及び，諺のような音声に対してのみ，文・句単位での処理が正しく機能するのであるならば，この処理を人間の音声処理過程の中心に据えることは問題があると筆者は考える。

4.7.5 まとめ

直接的結果

- 文節音声の知覚が非常に困難な環境では，当然のことながら，通常の句・文音声の句・文単位での正確な知覚も困難なものとなる。しかし，諺音声の句・文単位での知覚は正確に行なわれる。
- 文節正答率を，文節提示及び (文節間に強い関連性の無い) 文提示の間で比較すると，後者の方が正答率は高い (知覚が "容易に" なる)。

考察及び知見

- 人間は連続音声を複数の単位で処理しており，そのサイズは句・文にまで及ぶ。
- 単語・文節サイズと句・文サイズの単位を用いた処理を比較した場合，正しい知覚に必要となる単位時間当たりの音響的情報量には大きな差があり，より大きな単位ほど少なくなる。

[18] 項目として存在しているのであるから，このような表現は不適切であるかもしれない。

表 4.5. 本実験で使用した文リスト例

TYPE 1 諺 (使い慣れた文として使用，原文)。

TYPE 2 諺の一部を文節レベルで置換したもの (文節置換，但し置換前の諺は TYPE 1 の諺とは異なるものを使用する)。

TYPE 3 文節間に特別強い関連性の無い文 (文節程度の単位で処理されると考えられる)。なお，TYPE 2 の文節置換で使用した文節は，本 TYPE の文節の一部として使用される。

TYPE 1	かわいい	こには	たびを	させよ
	じゅう	よく	ごうを	せいす
	たびは	みちずれ	よは	なさけ
	いちを	きいて	じゅうを	しる
	めは	くちほどに	ものを	いう
TYPE 2	トラハ	なくても	こは	そだつ
	カタナ	かくして	しり	かくさず
	いぬも	ナツケバ	ぼうに	あたる
	のうある	サラハ	つめを	かくす
TYPE 3	トラハ	あるいて	めを	こする
	カタナ	こわして	めが	みえない
	とらも	ナツケバ	めしが	うまい
	つめたい	サラハ	いろが	わるい
	あの	あしほどに	しろい	かみ
	げんきな	とりは	くちが	くろい
	あの	しょうねんは	しおを	のむ

表 4.6. 文節以上の音声処理単位に関する実験結果 ([%])

	文節	句	文
TYPE 1	12.9	39.7	88.6
TYPE 2	3.6	14.3	10.7
TYPE 3	15.1	20.0	10.7

表 4.7. 実験結果例

各枠内の数字は該当するサイズにおける正答率 [%] を示す。また括弧内の数字は，諺から文節置換した文音声を元の諺として口頭再生した割合 [%] を示す。

かわいい	こには	たびを	させよ
42.9	0.0	42.9	0.0
71.4		28.6	
100.0			

いぬも	なつけば	ぼうに	あたる
0.0	0.0	0.0	0.0
28.6		57.1	
0.0 (57.1)			

あの	しょうねんは	しおを	のむ
28.6	14.3	0.0	0.0
28.6		0.0	
0.0			

4.8 種々の言語的情報が文音声知覚過程に及ぼす影響に関する実験

4.8.1 背景と目的

第 4.7 節で述べた実験は，筆者にとって提示文音声に含まれる言語的情報 (言語的内容) に着目した最初の実験であった．提示音声は 1) 諺，2) 諺中の一文節を異なる文節で置換した有意味文，3) 有意味文の 3 タイプに分類され，音響的提示条件 (提示音声に含まれる音響的情報量) は S/N=-5[dB] と一定条件の下で実験が行なわれた。しかし，以上の分類は非常に粗いものであり，提示音声に含まれる言語的情報量の及ぼす影響を調べるためには，より系統立った分類・制御に基付いた実験が必要である。また，音響的提示条件も可変化する方が実験の枠組みとしては望ましい。そこで本節では，言語的情報量による分類を，有/無意味を表わす意味的情報による分類と，統語構造の有/無を表わす統語的情報による分類とを組み合わせることにより，定性的に構成し，意味的情報と統語的情報が果たす役割を明確化することを試みた。更に辞書項目へのアクセスに着目し，音響的提示条件も Gating Paradigm の技法を用いて定量的に可変化し，提示音声のどの部分までが入力されれば該当する項目へ確実に検索が行なわれるか，つまり確実な検索に必要な提示時間長 (知覚の早さ) の違いを，提示音声に含まれる言語情報量の定性的な違いをパラメータとして調べる。その結果を基に，文音声の知覚における音響的情報と意味的・統語的情報の及ぼす影響の相互関係を考察する[86]-[88]。

4.8.2 実験方法

音声試料

以下に示すように，有意味性及び統語構造の有無に基付いて定性的に定義された 5 つの TYPE の 4 文節文を各 TYPE10 文ずつ，計 50 文用意する。この場合，単語間の熟知度及び意味的難易度のばらつきを抑えることを目的として，小学校低学年の国語の教科書を参考にして単語を選んだ。また，各文節のモーラ数を考慮して，各 TYPE 毎の文節長の平均及び分散がほぼ等しくなるよう作成した。実際に使用された文のリストを表 4.8 に示す。この文リストを成人男性一人に約 7[mora/sec] となるように発声してもらい，12[bit]・10[kHz] で A/D 変換する。得られた音声データを波形視察により文節毎にセグメントし，各 TYPE 毎の文節長の平均・分散を算出した結果を図 4.14 に示す。TYPE 間では文節長に有意な差は認められていない。この音声データを提示形態の違いにより，以下 2 種類の音声試料を作成する。

• 音声試料 1 —文音声—

刺激音声に含まれる言語情報量の定性的な違いにより定義された 5 TYPE の 4 文節文音声。各 TYPE10 文ずつ，合計 50 文。以下では意味的及び統語的情報の有無を○，×を用いて示している。

TYPE 1 諺など，単語系列として覚えてしまっていると考えられる文 (意○, 統○)。
TYPE 2 日常的な事象を記述した有意味文 (意○, 統○)。
TYPE 3 文節間に意味的矛盾は無いが，非日常的な事象を記述した有意味文 (意○, 統○)。
TYPE 4 文節間に意味的矛盾を持つが，統語的には正しい文 (意×, 統○)。
TYPE 5 非文 (意×, 統×)。

• 音声試料 2 —文節音声—

音声試料 1 の全文音声を波形視察により，文節単位で切出したもの。

これらの試料に対して図 4.15 の方法で無音置換したものを音声試料として使用する。文音声の場合，各文節頭から d[msec] までを保存し，それ以降文節尾までを，各文節毎に無音置換する。文節音声の場合，始めの d[msec] までを保存し，それ以降を無音置換する。なお，切り出し部にはクリック音の発生を防ぐため，10[msec] の三角窓をかけてある。以下，無音置換された文/文節音声を簡単のため，単純に文/文節音声と記す。

被験者

東京大学工学部学生及び大学院生，計 11 名

実験手順

11 人の被験者を 6 人と 5 人の 2 グループに分け，前者に文音声を後者に文節音声を提示する。文音声を提示する際には，無音置換を施した 50 文を 1 セッションとし，提示間隔 "提示文長+7"[sec] でヘッドフォンを通して両耳よりランダムに被験者に提示する。各文節の提示音声長 d は 50[msec] からセッション毎に 50[msec] ずつ増やし，450[msec] まで，計 9 セッション行なう。一方，文節音声を提示する場合は，50×4=200 個ある文節音声を 2 つに分け (100×2)，100 個を 1 セッションとし，提示間隔 5[sec] でヘッドフォンを通して両耳よりランダムに提示する。提示音声長 d は文音声と同様に，50[msec] からセッション毎に 50[msec] ずつ増やし，450[msec] まで計 9 セッション行なう。文提示/文節提示各々の場合において，被験者には，以下に示すタスクを行なわせる。

タスクと指標

インストラクションに示すように，被験者には聴取後即座に無音部を推定し，文あるいは文節音声全体を口頭再生するよう予め指示しておく。但し，音声終了の合図として，各文或は各文節尾にクリック音を挿入した。口頭再生が非常に困難な場合は，“分からない”旨を答えるよう指示した。文/文節音声何れにおいても，第 1 セッションの始め 20 回を練習用に設け，以後各セッションの始め 5 回を練習用として設けた。指標としては，提示長と正答率 (正確には，本来そこに存在していた項目として再生された率) の関係を各 TYPE 別に見る。なお，口頭再生された文節の正否を判定する場合，助詞の違い，用言の活用の違い，時制の違いは無視した。

インストラクション (文音声)

(連続) 音声がヘッドフォンを通して両耳より提示されます。音声の提示が終了するとクリック音が 2 度提示されます。始めのクリック音を合図にして，提示された音声を復唱して下さい。2 度目のクリック音は提示音声長+1 秒後に提示されます。この 2 つのクリック音の間に，提示音声の復唱を行なって下さい。但し，音声は数箇所無音置換されており，その無音部を推定して埋める形で復唱して下さい。復唱時間が限られていますので，無理に考え込む必要はありません。何かしら想起される単語があればそれで埋めて頂ければ結構です。分からない場合は復唱しなくても結構です。第二のクリックの後，2, 3 秒して次の音声の提示が開始されます。以上のことを繰り返して下さい。それでは宜しくお願い致します。

インストラクション (文節音声)

文音声から文節単位で切り出した音声がヘッドフォンを通して両耳より提示されます。但し，文節の一部は無音置換されています。聴取後，無音部を埋める形で復唱して下さい。文節音声の提示間隔は 4 秒となっておりますので，聴取後できるだけ即座に復唱を開始し，次提示の音声の聴取が妨げられないよう，注意して下さい。復唱時間が限られていますので，無理に考え込む必要はありません。何かしら想起される単語があれば，それで埋めて頂ければ結構です。分からない場合は復唱する必要はありません。なお，1 日 10 分を 2 回行ない，それを 8 日間行ないます。宜しくお願い致します。

4.8.3 実験結果

図 4.16 に文節音声提示における，提示長と文節正答率との関係を TYPE 別に示す。図 4.17 に文音声提示における，提示長と文中の文節正答率との関係を TYPE 別に示す。

何れの図においても，プロットされてあるデータは，各提示長に対する正答率を示し，曲線は最尤推定法[89]を用いて累積正規分布近似を行なった結果である。累積正規分布近似より平均値μ及び標準偏差σが得られるが，μは正答率 50[%] に相当する提示長であり，以後これを同定の閾値 (identification threshold) と呼ぶ。表 4.9 に図 4.16, 図 4.17 における同定の閾値を各々示す。

4.8.4 考察と検討

図 4.16 より，文節提示における提示長と文節正答率との関係はタイプ間で殆ど差が無いことが分かる。但し同定の閾値を見ると，TYPE 1 の閾値が他より約 20[msec] 大きいことが分かる。これは，TYPE 1(諺) を構成する単語の中には，その熟知度が低い (単語としての使用頻度が非常に低い) と考えられる単語も見受けられ，その分閾値がシフトしたと考えられる。しかし，シフト幅は提示長のステップ幅 (50[msec]) の 1/2 以下であり，本実験においては十分小さいと考えてよかろう。一方図 4.17 より，文提示における提示長と文中の文節正答率との関係は，TYPE 間で大きな差があり，提示文中に含まれる言語情報量の違いが，文音声を構成する単語の同定に対して大きく影響していることが分かる。同定の閾値もタイプ番号と共に増加している。しかし，TYPE 4 と 5 の差は他と比べて非常に小さいものとなっている。TYPE 3 と 4 の違いが意味的情報の有無であり，TYPE 4 と 5 の違いが統語的情報の有無であることを考えると，これは意味的情報の有無が統語的情報の有無に優先して影響を与えていることを示唆している。しかし，本実験では統語的構造は崩れているが意味的には正しい文は，その定義が困難であったため実験を行なっていない。そのため，意味的情報と統語的情報の交互作用が完全には考慮できておらず，この点で更に検討が必要であろう。

次に，文・文節提示の 2 つの実験結果を比べると，TYPE 1, 2 においてのみ文提示の場合の閾値がより小さくなっていることが分かる。これは，統語的構造が保たれている有意味文であっても，記述する事象の通常性によっては，その文脈が内部辞書項目へのアクセスを妨げる方向に作用することを意味しており，この結果からも意味的整合性の統語的整合性に対する優位性が伺われる。

図 4.18 に文音声提示において，文節正答率に対する「文正答率/文節正答率」の関係を示す。この図において y 軸の値が 1.0 である場合，それは文が不完全に同定される (4 文節中 n(<4) 文節が正しく同定される) ことが全く無い状態を意味する。つまり，正しく同定された文節は必ず正しく (4 文節全て) 同定された文中の文節となる。このことに考慮して図 4.18 を見ると，TYPE 1 では文節正答率が非常に低い時から「文正答率/文節正答

率」ははぼ 1.0 を示している。これは，TYPE 1 の文が同定される場合は，必ず文として知覚され，部分的に正しく同定されることが無いことを意味し，第 4.7 節で言う，文単位の辞書項目の存在を支持する結果であると言える。また，この図においても統語的情報の有無による差が非常に小さいことが分かる。

第 3.4.3, 4.7 節，第 4.2 節，第 4.3 節などの実験より，「処理単位のサイズと使用される音響的特徴量との関係」，「辞書項目固有の性質と使用される音響的特徴量との関係」，「cache 的 STM に対する辞書検索の優位性」などが知見として得られている。本実験結果 (及び第 4.7 節における，文節正答率の文節提示→文提示における変化) より，熟知度の高い文脈中では，より低情報量の音響的特徴で，無音置換前に存在していた項目への "確実な辞書検索 (正しい照合)" が可能となることが示された。この文脈中における各辞書項目に対する処理を，「処理単位サイズ」及び「辞書項目固有の性質」との関連から考察してみる。当然のことながら，TYPE 2 の日常的な文がそのまま辞書項目として内部辞書に登録されていることは考えられない。故に本実験結果を，使用した有意味文音声 (TYPE 2～TYPE 4) に対応する文サイズ辞書項目の存在を仮定し，それら辞書項目の固有な性質に起因するものとして考察することは困難である。また，第 4.4 節における考察と同様に，本実験の結果を「cache 的 STM」に直接結び付けることも適当ではない。

第 4.4 節においては，先行コンテキストによる，意味的関連項目に対するより早い知覚を，辞書検索過程の存在とその動的特性の変化と考察した。本節の結果も同様に，辞書検索過程に起因させることができる。しかしその後の実験より，"早い・容易な知覚" と処理単位/辞書項目単位との関係がより明確になってきており，また第 4.7 節においては，文中における単語間の意味的・統語的な連結性と，辞書項目として確立された文項目内の単語間の連結性の比較が考察されている。そこで本節では，上記の結果に対して更に深く考察することにする。第 4.7 節により，諺や俳句など (TYPE 1) は，"1 辞書項目として存在している" との知見を得てはいるものの，我々の内部辞書に "先天的" に項目として登録されているとは考えられない。これは単語サイズの辞書項目においても言えることであるが，長い年月をかけて使用されることで "後天的に"，単なる音の系列であったものが単語の辞書項目として，単なる単語系列であったものが，句/文の辞書項目として内部辞書に登録されていったものと考えられる。この様に考えると，単語系列として，或は，意味レベルでの系列[19]としての出現頻度が高い項目群は，句・文サイズの一辞書項目として正規に登録されてはいないものの，「単なる単語系列」と「句/文サイズの辞書項目」と

[19] 同義語を一項目として扱った場合の項目系列。

の中間的な性質を有していると考えられる。更にその項目群における各項目間の連結 (項目間の重み付きポインタとして存在？) の強度も，長期の学習/リハーサルにより，連続量として辞書内に記述されていると考えられる。以上の考察 (仮定) を下に本実験結果を眺めた場合，文音声は，辞書項目間で緩やかに結合された単語網を利用して知覚されると推測される。そして，入力文音声が非常に熟知した場面の記述であった場合は，該当する場面の記述に対応する，より結合性の強い単語網が存在する。この単語網を用いることにより，文音声中の各単語は単語以上の大きさで緩やかにまとめられて処理され，その結果，低情報量の提示で辞書内の該当項目への確実な検索/正しい照合が可能となっていると言える。第 3.6 節などで考察した "(間接的) 活性化" も，この単語網の存在によると考えると理解しやすい。また，ここでも第 4.4 節における，"cache 的 STM か否か" と同じ議論を行なうことができる。即ち，この単語網は LTM 内に記述されるものの他に，コンテキストによって動的に形成される単語網が存在すると考えるべきである[20]。本実験のような一文提示の実験パラダイムでは，前者の単語網が優先的に作用すると考えられるが，ターゲット文に対する先行コンテキストを十分に設定し，かつ種々の操作を加えることで，後者の単語網の特性を観察することは可能であろう。今後の課題の一つである。単語間の連結性の強度と言う観点から考察した場合，図 4.18 が，文辞書項目として確立した単語 (文節) 系列と，上記の単語網によりまとめられて処理される単語 (文節) 系列の差を示している。前者は "4 文節全て不正解" 或は "4 文節全て正解" のどちらかのみであり，如何に強い単語間結合を有しているかが分かる。

なお，本実験では，統語的構造の有無による差が殆ど見られなかった。しかし文献 [26] などでは，統語構造の有無による差が認められている。これは英語等の言語と比較して日本語は，語順の制約が小さいためであると考えられ，第 3.6.2 節で述べた仮説を実証することが出来た。

最後に，本実験結果をアクセント型と言う観点から再分析した結果を図 4.19 に示す。文節提示実験から 3 モーラ名詞に対応するものを抽出し,「確実な検索に必要な提示長」をアクセント型別に集計したものである。図 4.19 には，各型における平均及び分散を示してある。分散分析[90]を行なった結果，0 型と 1 型間の分散比は $F^1_{291}=5.64(>F^1_{291}(0.05))$ となり，危険率 5%で有意な差が得られた。一方，0 型と 2 型間には有意な差は見られなかったが，これは 2 型に対応するデータが極端に少なかったからであると考えられる。

[20] 逆に言えば，コンテキストによって生成されるような，単語間の弱い連結性しか要求しないため，文サイズの辞書項目になることが出来ない。

4.8.5 まとめ

直接的結果

- 内容・場面をより熟知している文脈中の文音声では，短時間の提示で該当項目が想起される。これは，低情報量の音響的特徴で正しい辞書検索が可能な状態になっていることを意味する (その結果，知覚・同定が "早く" なる)。
- 諺中の文節はある時間長を閾値として，"全て非正解"→"全て正解" と変化する。その他の刺激文においては，ランダムに正解文節が現れる。
- 統語的整合性と意味的整合性では後者の方がより "知覚の早さ" に影響を与えている。

考察及び知見 (予想)

- 辞書項目間には，その連結の強さを示すポインタが存在し，文音声知覚の際にはそのポインタによって緩やかに結ばれた単語網を利用した知覚が行なわれている。
- 提示文音声がより熟知している事象の場合は，提示文中の各単語は，ある一つの単語網に含まれる単語群であるため，文中の各単語をより大きな範囲で捉えることとなり，早期に抽出される低情報量の音響的特徴で正しい検索・照合が可能となる。
- その結果，語頭音の重要性の低下が予測されるが，追加実験の結果，より熟知している内容・場面の入力音声に対する語頭音の重要性の低下が観測された[86]。

表 4.8. 実験で使用した 4 文節文リスト (TYPE 1~TYPE 5)

TYPE 1	あわてる	こじきは	もらいが	すくない
	へたなか	んがえや	すむに	にたり
	さんしょうは	からいが	びりり	とからい
	はらが	へっては	いくさは	できぬ
	おぼれる	ものは	わらをも	つかむ
	のうある	たかは	つめを	かくす
	いぬむ	あるけば	ぼうに	あたる
	かわいい	こには	たびを	させよ
	しんねん	あけまして	おめでとう	ございます
	ただいまから	しちじを	おしらせ	いたします
TYPE 2	あかちゃんが	ベッドで	ミルクを	のんだ
	きょうしつで	せんせいが	つくえを	たたいた
	こどもがは	なぜを	ひろばで	うちあげた
	おかあさんは	スーパーへ	かいもの	にいきます
	びょういんで	ともだちが	しんさつを	うけた
	ライオンは	えものを	そうげんで	みつける
	おばさんが	おとといは	おこずかいを	くれました
	こうえんで	わたしは	ハンカチを	ひろった
	おとうとが	とんぼを	うらにわで	つかまえた
	にんげんは	いつでも	ことばを	つかう
TYPE 3	がくせいが	さばくで	だいこんを	なげた
	だいがくで	こどもが	おぼんを	ころがした
	ぼくたちは	ぞうきんを	おくじょうで	きざんだ
	たぬきが	なかにわで	みかんを	けった
	ジャングルで	かのじょは	かるたを	やぶいた
	せんばいが	こくばんを	かいがんで	つくります
	いもうとが	そうげんで	ブローブを	もやした
	みずうみで	そんちょうは	ざぶとんを	まわした
	せいとは	せんめんきを	じんじゃで	とかした
	きみたちは	しんりんで	パソコンを	こわせ
TYPE 4	すみれが	かいていで	くるまを	のみこんだ
	くうちゅうで	ずぼんが	はたけを	おぼえる
	ねくたいが	たいように	たんぼで	ぶつかった
	むらさきが	げんかんで	ストーブを	よみます
	デパートで	すいえいが	ステーキを	かんがえる
	ようふくは	めがねを	ヨットで	たべた
	じてんは	がいこくで	ほうそうを	みつけた
	ひきだしで	テレビが	せいかつに	ほほえむ
	しんばいは	れんらくを	みやこで	つかまえた
	しんせつが	かいがんで	じてんしゃを	なめた
TYPE 5	はっきり	えんぴつと	まずい	はしった
	うごくな	かたかなへ	ちいさい	たとえば
	ふかい	さっぱり	チョークも	ようやく
	かいだんで	とても	こたえた	うれしい
	とけいと	あさって	どっさり	くるしい
	どんな	けれども	デパートへ	あたたかい
	こんどは	ぼうしも	はじめた	おとうさんは
	せんぷうきと	たくさん	からだは	こっそり
	どうぶつへ	ねむる	つめたい	おいしい
	あるいた	じっけんは	いっぱい	じてんしゃ

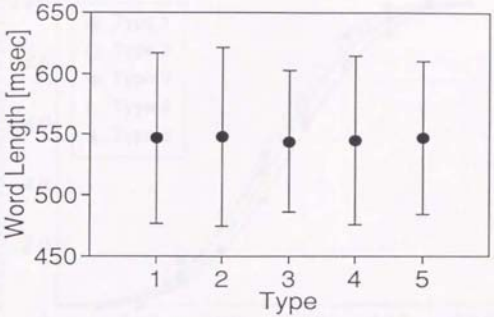


図 4.14. 各 TYPE 別文節長の平均 ([msec]) 及び標準偏差

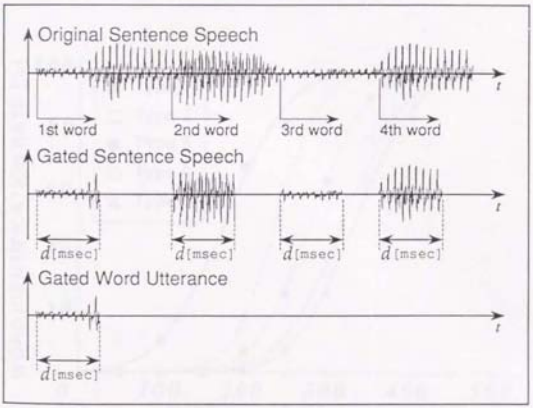


図 4.15. Gating Paradigm

最上段より，原音声波形，ゲートを掛けた文音声，ゲートを掛けた文節音声である。なお，文節当たりの提示長 d は，以下の通りである。

$$d = 50n \quad (n = 1, \ldots, 9) \quad [\mathrm{msec}]$$

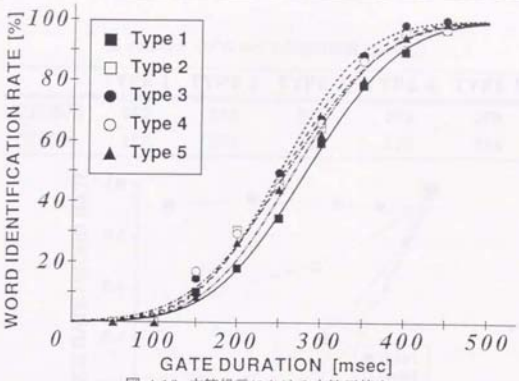


図 4.16. 文節提示における文節正答率

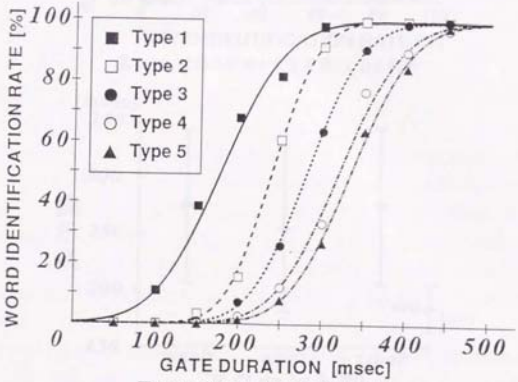


図 4.17. 文提示における文節正答率

表 4.9. 各タイプにおける同定の閾値 ([msec])

	TYPE 1	TYPE 2	TYPE 3	TYPE 4	TYPE 5
文節提示	282	256	251	259	268
文提示	177	239	281	320	334

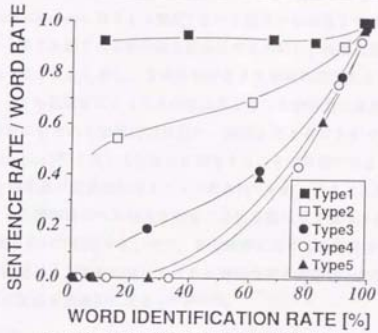


図 4.18. 文節正答率と文正答率/文節正答率

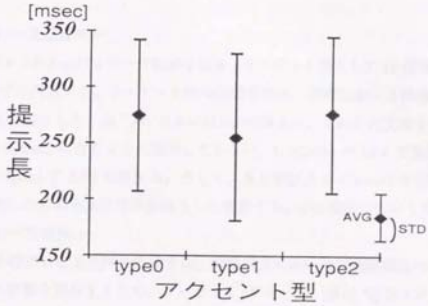


図 4.19. 3 モーラ名詞に対するアクセント型別の集計結果

4.9 談話的情報が文音声知覚過程に及ぼす影響に関する実験

4.9.1 背景と目的

第 4.8 節において使用された音声試料は，提示文音声に含まれる言語的情報量によって 5 つの TYPE に区分されていたが，それはあくまで有/無意味，統語構造の有/無と言った定性的な区分であった。そして，諺以外の有意味文として使用した音声試料は，記述する事象の通常性 (commonness, 以下 c と略記) という観点から高低 2 つのカテゴリに分類されていた。しかし第 4.8 節ではこの分類を容易にするために，極度に日常的な文と非日常的な文のみを扱っていた。しかし，文音声が記述する事象の通常性は本来連続量であり，同一の事象であっても被験者によってその値は異なることが容易に推測される。また，通常性はその文音声に含まれる談話的情報量の一指標と考えることができ，大きさ推定法 (magnitude estimation)[89] を用いて定量的に測定することが可能である。そこで本節では，言語的 (談話的) 情報量の定量的尺度として "通常性" を導入することを考える。そして，文音声を構成する各辞書項目への確実な検索に必要な提示長を第 4.8 節と同様に Gating Paradigm の技法を用いて測定する。次に，各文音声に対する通常性を大きさ推定法を用いて被験者毎に測定し，文音声知覚における音響的情報と言語的 (談話的) 情報の及ぼす影響の定量的相互関係を実験的に求める[87][88][91]。

4.9.2 実験方法

音声試料

• 音声試料 1 —文節音声—

4 モーラ名詞+「が」及び 4 モーラ動詞を夫々，ターゲット語として 18 種類，ダミー語として 12 種類ずつ用意する。ターゲット語の先頭音素は，視察における語頭の検出を容易にすることを目的として，全て/t/あるいは/k/に揃えた。これらの文節を成人男性 1 名に，約 7[mora/sec] になるように発声してもらい，4.5[kHz] の LPF で高域除去した後，16[bit]・10[kHz] で A/D 変換する。そして，各文節頭から d[msec] を保存し，それ以降を無音置換したものを文節音声試料として使用する。d の設定については後述する。

• 音声試料 2 —文音声—

音声試料 1 を接続して文音声を作成する。統語構造の違い及び統語構造の複雑さが音声知覚に及ぼす影響を排除するため，ターゲット/ダミー文，共に "名詞+が+ 動詞" の有意味文に限定して作成した。但し，ターゲット文は名詞/動詞共に音声試料 1 のターゲット音声を用いて作成され，同一単語が異なるターゲット文音声の作成に使用されている。

なお，ターゲット文作成に複数回使用される文節音声は，一度 A/D 変換したものを使用するため，全く同一波形の文節音声が使われることになる。作成された文音声に対して，その一部を無音置換したものを文音声試料として使用する。無音置換の具体的方法については後述する。表 4.10 に全ターゲット文のリストを示す。このようにターゲット文は，記述する事象の通常性が幅広く変化する様に作成されている。ターゲット/ダミー文数は各々39, 36 種類 (合計 75 種類) である。

被験者

成人男性 11 人及び成人女性 2 人，計 13 人。

実験手順

被験者を 7 人と 6 人の 2 グループに分け，前者に対して以下に示す STEP I の実験を，後者に対しては STEP II, III の実験を課す。

1. STEP I —単語 (文節) 提示—

   無音置換を施した文節音声30 個 (名･動詞別) を提示間隔6[sec] で，ランダムにヘッドフォンを通して両耳より提示し，これを 1 セッションとする。但し，名詞と動詞の提示は別個のセッションとして行ない，被験者には前もって提示する単語の品詞を伝えておく。文節頭からの提示長 d は 125[msec] からセッション毎に 25[msec] ずつ増やし，450[msec] まで行なった。被験者には以下に示すインストラクションの下，聴取後即座に無音部を推定し，単語全体を所定の用紙に書き取るよう予め指示しておく。全セッション終了後，各ターゲット語毎に提示長と正答率との関係を求める。但し，動詞の活用の違いは無視して正否を決定した。これを最尤推定法で累積正規分布近似し，同定の閾値 (以下単にθと略す) を各ターゲット語毎に求める。ここで，θ は文脈には依存しない各辞書項目固有の特徴である長期的頻度，親密度などの影響を直接反映していると考えられる。

2. STEP II —文提示—

   ターゲット文 "名詞 n_i+が+動詞 v_j" に対して，夫々の提示長 $d(n_i), d(v_j)$ を以下の 2 つの手法で設定する。どちらの手法でも，各単語の提示長は，その単語のθを基準に決定され，両者の違いは，名詞に対する提示長の設定方法のみである。

   • 手法 1

$$
\begin{aligned}
d(n_i) &= \theta(n_i) + k\Delta d \quad k = -3, -2, ..., 2 \\
d(v_j) &= \theta(v_j) + k\Delta d \quad k = -3, -2, ..., 2
\end{aligned}
$$

• 手法 2

$$\begin{aligned} d(n_i) &= \theta(n_i) + 3\Delta d \\ d(v_j) &= \theta(v_j) + k\Delta d \quad k = -3, -2, ..., 2 \end{aligned}$$

但し，Δd は提示長の増加幅(25[msec])である。ダミー文については$\theta(n_i) = \theta(v_j) = 300$[msec] という仮定の下で提示長を決定した。提示長がθを基準にして設定されている場合，提示音声中に含まれる各項目固有の特徴に関する情報量は異なる項目間でほぼ等しいと考えられる。これを，異なる項目に対して同一の提示長を与えた第 4.8 節と比べると，提示音声に含ませる情報の与え方が項目間でより均一化されていると言える。以上の考察から，手法 1 では名詞及び動詞音声の提示部に含まれる(辞書検索に利用される)情報量はほぼ等しいと仮定される。しかし，手法 2 では名詞の一部は無音置換されてはいるものの，その提示長は辞書アクセスが正しく行なわれるために十分であるように設定されているため，被験者は入力音声を left-to-right に処理することになる。即ち手法 1 では，名詞+が+動詞の同時知覚，手法 2 は動詞の条件付き知覚，と言うことができる。

6 人の被験者を 3 人ずつの 2 グループに分け，一方に手法 1，他方に手法 2 で作成された文音声を，75 文を 1 セッションとして，提示間隔 10[sec] でランダムに，ヘッドフォンを通して両耳より提示する。但し統語解析の負担を軽減することを目的として，提示文の統語的構造は予め被験者に伝えておく。被験者には以下のインストラクションの下，STEP I と同様に無音部を推定し，文全体を所定の用紙に書き取るよう予め指示しておく。

3. STEP III —通常性の測定—

文提示実験で用いた被験者 6 人に対して，各ターゲット文が記述する事象の通常性を大きさ推定法で測定する。刺激文及び中央と両端にのみ目盛りのあるスケールと，カーソルキー操作によってスケール上を左右に動かすことのできるマークをコンソール上に提示する (図 4.20 参照)。被験者にはこのスケールを通常性の程度を表わす尺度であると仮定させ，提示文が記述する事象に対して，主観的に感じる通常性に相当する位置にマークを移動するよう指示した。集計の際には，最左端を 0.0，最右端を 1.0 のリニアスケールとして集計した。なお刺激文中には，標準刺激として極めて日常的な文，及び明らかに実現不可能な事象を記述した文を夫々5 文ずつ含ませた。ターゲット文 39，ダミー文 10，合計 49 個の文提示を 1 セッションとし，これを各被験者に 12 回ずつ繰り返させ，後半の 10 回の平均値をとることで，各ターゲット文の通常性を各被験者毎に求めた。

インストラクション (文節音声)

文節音声がヘッドフォンを通して両耳より提示されます。但し，文節の一部は無音置換されています。聴取後，無音部を埋める形で，聴取した単語全部を所定の用紙に記入して下さい。助詞の部分は省いて結構です。文節音声の提示間隔は 6 秒となっておりますので，聴取後即座に記入を開始し，次提示音声の聴取の妨げにならないよう，注意して下さい。記入時間が限られていますので，無理に考え込む必要はありません。何かしら想起される言葉があれば，それで埋めて頂ければ結構です。分からない場合は「×」を記入して下さい。それでは，宜しくお願い致します。

なお，今回聞いて頂く音声は全て「名詞+が」[21]です。

インストラクション (文音声)

文音声がヘッドフォンを通して両耳より提示されます。但し，文の一部は無音置換されています。聴取後，無音部を埋める形で，聴取した文全体を所定の用紙に記入して下さい。文音声の提示間隔は 10 秒となっておりますので，聴取後即座に記入を開始し，次提示音声の聴取の妨げにならないよう，注意して下さい。記入時間が限られていますので，無理に考え込む必要はありません。何かしら想起される言葉があれば，それで埋めて頂ければ結構です。分からない場合は単語毎に「×」を記入して下さい。それでは，宜しくお願い致します。

なお，今回聞いて頂く音声は全て「名詞+が+動詞」となっております。

インストラクション (通常性)

リターンキーを押すと，2 文節の文，及びスケールとスケール中央部に点が表示されます。表示された文が記述する事象が，どの位日常的であるかを主観的に判断して下さい。そして，表示してあるスケールが，日常性を示す尺度であると仮定し (左/右が日常性がの低/ 高に対応する。初期値 (中央) は中性を意味する。)，提示文の日常性を表す位置まで，カーソルキーで点を移動して下さい。配置が完了したら，リターンキーを押して確定して下さい。その後約 1 秒して次の文が提示されます。以上の事を，文が表示されなくなるまで繰り返して下さい。なお，意味の "有る/無し" で 0/1 に判断することのないよう，気をつけて下さい。それでは，宜しくお願いします。

[21] 動詞音声の場合は「動詞」になる。

本実験の簡単なフローチャートを図 4.21 に示す。ここで，文提示実験で使用する文音声試料における Gating Period は，単語提示実験の結果を参考に作成され，両者の実験には各々異なる被験者が参加している。

4.9.3 実験結果

得られた実験結果に対して次のような後処理を施した。手法 1 による実験結果から，$|D(n_i) - D(v_j)| \leq \Delta d$[msec] を満たす結果を抽出し，その後の分析対象とした。但し，$D(n_i)$ は名詞，$D(v_j)$ は動詞における，該当項目への確実な検索に必要な最短提示長から θを差し引いた値である。この操作は，手法 1 による文音声を提示した際に，名詞及び動詞の提示部に含まれる，辞書検索に使用される情報量が等しいと考えられる実験結果のみに着目するためである。一方手法 2 による実験結果に対しては以上のような後処理が必要無いのは明らかである。次に，通常性 c を切り捨てによって 10 段階に区分し，得られた結果を分類する。そして c の各レベル毎に，"提示長からθを差し引いた値"(オフセット提示長) と文正答率との関係を求める。これを最尤推定法で累積正規分布近似し，文同定におけるθを算出した結果を手法別に図 4.22 に示す。但し横軸は通常性，縦軸は文節提示におけるθからのずれ，θのシフトである。

4.9.4 考察と検討

図 4.22 から，手法 1 においては c の増減に伴って，知覚を助ける方向 (以下，＋方向と記す)，及び知覚を妨げる方向 (以下－方向と記す) への影響は増大していることが分かる。しかし手法 2 では，c の増加に伴う＋方向への影響は増大しているものの，減少に伴う影響は飽和し，－方向へは働いていない。そこで，飽和前 (c=0.2, 0.3)，後 (c=0.6, 0.7)，及びその中間 (c=0.4, 0.5) における，提示長 (孤立単語提示におけるθからの距離) と文正答率との関係を最尤推定法で累積正規分布近似した結果を図 4.23, 4.24 に示す。但し，図 4.23 が手法 1 に，図 4.24 が手法 2 に対応している。ここで，近似曲線の重なりが図 4.22 における飽和に対応している。

実験手順の欄で述べたように手法 1 では，名詞・動詞各々に関する情報がほぼ等しくなるような Gating を施している。また，実験結果の欄で示した後処理により，文節提示で用いた被験者と，文提示で用いた被験者間の個人差をも取り除いている。故に最終的に着目する，手法 1 に対する結果は，名詞・動詞の同時知覚における結果であると言える。一方，手法 2 は先行する名詞が確実に知覚されるように Gating を施しており，動詞の条件付き知覚と言うことができる。即ち本実験結果は，名詞・動詞を同時に知覚する場合，

文の持つ談話的情報は，＋方向にも，－方向にも大きく影響を与えていると言うことができる。これは第 4.8 節の結果と一致するものである。一方，条件付き知覚に対しては，＋方向への影響は談話的情報量の上昇と共に大きくなるが，情報量の下降に伴う影響は飽和し，－方向へは影響しないと言うことができる。なお手法 2 において，低通常性の文音声に対しても僅かながら，文節音声提示時より早い段階で確実な検索が可能となるとの結果が得られているが，これは "名詞＋が＋動詞" 文の有意味性に依るものであると考察できる。

この両手法に対する結果の違いをもたらす原因について，第 4.8 節までの知覚実験結果を基に考察する。本実験で使用した文には諺や熟語の類は無く，句・文単位辞書項目を用いた処理が行なわれる可能性は無い。故に項目固有の性質としての議論は無用である。また，cache 的 STM を導入する必要性や妥当性も見当たらない。さて，手法 1 においても音声を left-to-right に知覚することは物理的には不可能では無い。そして仮に，left-to-right に知覚していたならば，手法 2 と同様の結果が得られるはずである。しかし実際には手法 2 とは異なる処理が働いているとの結果が得られた。この違いは，第 4.8 節で考察した緩やかな辞書項目間結合を考えることで説明できる。即ち人間は，連続的に入力される音声を知覚する場合，緩やかな辞書項目間結合を利用し，より大きなまとまり[22]で音声を捉えようとする。この場合，より大きなまとまりで捉えようとする限り，文音声の記述する場面の通常性は＋方向にも，－方向にも作用するのは至極当然のことである。そして，手法 1 の音声試料に対しては常に，この単語間結合を利用した処理が優先的に働いていたと考察できる。一方，left-to-right と非 left-to-right の処理が可能な手法 1 の場合は，基本的には項目間結合 (単語網) を用いた大きなまとまりで音声を捉えて処理を試みる。その結果，通常性が高い場合は＋方向の作用が生まれ，短い提示長で正しい辞書検索が可能となる。しかし，十分満足する結果を出力出来ない場合，－方向の作用が生じる以前に，即座に left-to-right の処理に切替えられ，各提示単語 (項目) は，前後の語と独立して知覚されるようになると説明できる。

しかし本実験では，"名詞＋が＋動詞" と言う語数及び統語構造が限られた文のみを対象としており，動詞を条件付きで知覚する場合でも，コンテキストとなっているのは先行する名詞ただ一つである。条件付きで知覚する場合，left-to-right 処理への適切な移行により，低通常性は－方向へ作用しないという結果が得られたが，通常性の及ぼす影響はコンテキストの大きさが一つの要因となっていることが十分予想される。つまり，コンテキス

[22] 処理単位と言う言葉は，この場合不適切であると筆者は考える。

トが大きくなると共に通常性の及ぼす影響も大きくなり，条件付きで知覚する場合でも left-to-right 処理への移行が難しくなるなど，低通常性が一方向に働き始めることが予想される。このように，コンテキストサイズを更に大きくした n 語文 (n>2) や，先行文に引き続いて発声された文など，より大きなコンテキストによる通常性の影響に関する実験が今後の課題であろう。

4.9.5 まとめ

直接的結果

- 入力音声を left-to-right に知覚した場合と，そうでない場合ではその文が記述する場面の通常性が及ぼす影響は異なる。
- left-to-right の処理が可能な場合，通常性は知覚を助ける方向には働くが妨げる方向には働かない。
- left-to-right の処理が困難な場合，通常性は知覚を助ける方向にも，妨げる方向にも作用する。

考察及び知見

- 文音声を知覚する場合，辞書項目間の連結の強度を利用し，より大きなまとまりで音声を捉えようとする傾向がある。
- より大きなまとまりで音声を捉える処理で解決できない場合，left-to-right の処理が可能ならば，left-to-right の処理へ速やかに移行し，かつ，以前に知覚された語とは独立して知覚が行なわれる。

表 4.10. 本実験で使用した 2 文節ターゲット文

各文が記述する場面の通常性 (commonness) が幅広く変動するように作成した。なお，波形視察による切り出しを容易にするため，語頭音韻は無声破裂音 (/p/, /t/, /k/) に揃えた。

かんごふが	たすける	としよりが	となえる
かんごふが	こたえる	としよりが	たたえる
かんごふが	きたえる	こうちょうが	つかれる
たれんとが	たすける	こうちょうが	ことわる
たれんとが	こたえる	たんていが	たのしむ
たれんとが	きたえる	たんていが	かぞえる
こうはいが	たすける	てんのうが	とぼける
こうはいが	こたえる	てんのうが	とまどう
こうはいが	きたえる	かんとくが	くるしむ
かがくしゃが	てこずる	かんとくが	こわがる
かがくしゃが	くらべる	たんにんが	たのしむ
かあさんが	かぞえる	たんにんが	くらべる
かあさんが	てこずる	キャプテンが	ためらう
とうさんが	ことわる	キャプテンが	たたえる
とうさんが	となえる	てんさいが	かなしむ
ともだちが	かたらう	てんさいが	とぼける
ともだちが	ためらう	きょうだいが	かたらう
こくじんが	くるしむ	きょうだいが	とまどう
こくじんが	つかれる	こくみんが	かなしむ
		こくみんが	こわがる

図 4.20. 通常性の測定

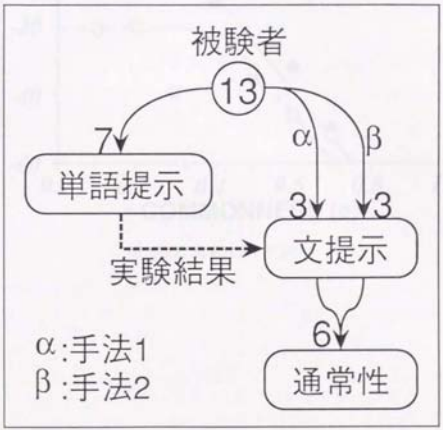


図 4.21. 本実験のフローチャート

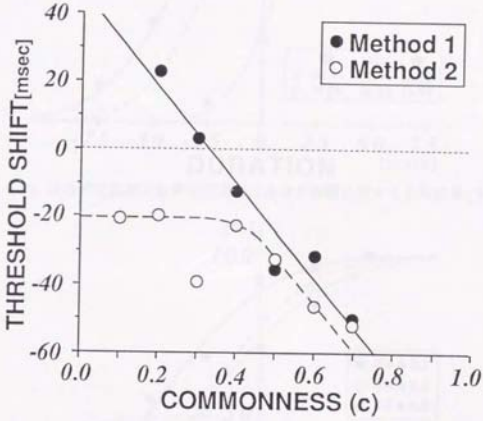


図 4.22. 通常性とθのシフト

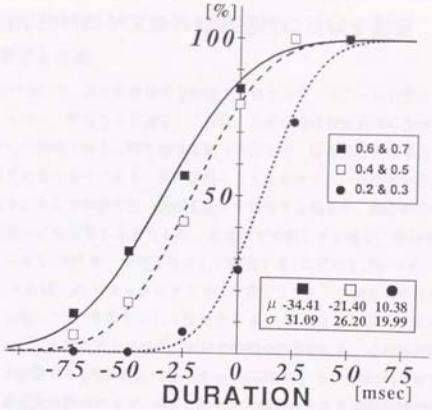


図 4.23. 談話的情報が文音声知覚過程に及ぼす影響に関する実験結果 (手法 1)

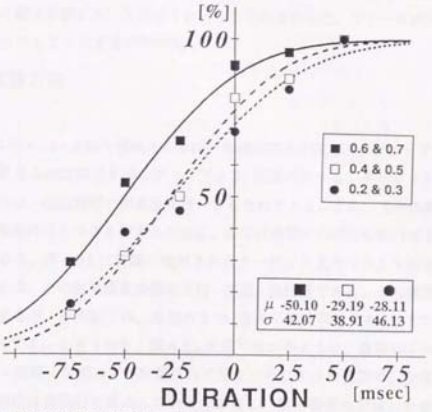


図 4.24. 談話的情報が文音声知覚過程に及ぼす影響に関する実験結果 (手法 2)

4.10 韻律的特徴が文音声知覚過程に及ぼす影響

4.10.1 背景と目的

第 4.5 節において，孤立単語音声を知覚する場合のアクセントの役割について実験的に検討した。しかし，筆者らの目指すところは，人間の音声知覚過程の全体を見渡すことのできるモデルの構築である。即ち第 3.2 節での議論は，韻律的特徴の果たす役割を分析する際にも当てはまるものであり，音声試料として文音声を用いた分析が必要であることは明らかである。そこで本節では，朗読文音声を知覚する場合の，韻律的特徴の果たす役割について実験的に検討することとした。文音声を対象とする場合，単語音声とは異なり，単語よりも大きな "句" を一つの単位として変動する (大局的な)F_0 パターンが存在する。F_0 モデルによれば，F_0 パターンはアクセント成分とフレーズ成分との和の形でモデル化されるが，上記の "句" を単位として変動する F_0 パターンはフレーズ成分に相当するものである。また，当然のことながら文音声中の個々の単語には，その単語固有のアクセントに依存して変動する (局所的な) F_0 パターン (即ちアクセント成分) が存在する。そこで本節では，朗読文音声中のフレーズ/アクセント成分の大きさを数段階で変化させ，各々の大きさにおける文音声知覚の様子を観測することで，文音声知覚時の韻律的特徴の果たす役割，特に第 4.5 節においては全く触れることのなかった，フレーズ成分の果たす役割について分析することにする[80][81][92]。

4.10.2 実験方法

音声試料

図 4.25 に示す，11 文節で構成される同一の統語構造を持つターゲット文を 16 種類，及びこれとは異なる統語構造を持つダミー文を 27 種類用意する。ターゲット文を同一構造に限定するのは，統語解析の難易度を統一するためである。また，文中の単語は，単語間の意味的難易度のばらつきを抑えるために，小学校中学年の国語の教科書を参考に選出されたものである。表 4.11 に実際に使用されたターゲット文のリストを示す。図 4.25 中，太文字の要素は，この複文構造全体の主語・述語・目的語である。この統語構造には 4 つの句が存在するが，本実験では，最初の 3 つ (各々下線で示してある) までを以降，"句" と呼ぶことにする。と言うのも，表 4.11 を見て分かるように，最初の 3 つの句には内部的に,「主語・述語・目的語」が配置されており，その 3 つの文節で，全体の統語構造における，主語或は目的語を構成しているのに対し，第 4 番目の句はいわゆる述部と言われるもので，その統語的性質が他の 3 つとは大きく異なるからである。また，考察と検討

の箇所で文節単位での正答率を算出しているが，これらも全て，最初の 3 つの句に着目して計算したものである。また，ターゲット文中の O_1, O_2の位置には，0~3 型アクセントを持つ 4~6 モーラ名詞単語を均等に配置する。これは，文中に配置された単語知覚におけるアクセントの及ぼす影響を，第 4.5 節の結果と比較するためである。このように作成されたターゲット/ダミー文を東京方言の成人男性話者 1 名に,「一文中に息継ぎが入らないよう」発声してもらい，16[bit]・10[kHz] で A/D 変換する。"息継ぎ" に関する配慮は，F_0 パターン以外の韻律的特徴 (特に休止長) の影響を除去するためである。このため，発話速度は通常の録音時よりもやや早くなり，約 9[mora/sec] となった。得られた音声データに対して種々の F_0 パターンの変形を施して作成される分析合成音を，音声試料として使用する。

LMA フィルタを用いた分析合成

第 4.5 節の実験では音声試料の作成に PARCOR 分析合成方式を用いたが，本節では合成音声の品質向上を狙い，LMA(Log Magnitude Approximation) フィルタを用いた分析合成を行なった。LMA フィルタ[93]及びそれを用いた分析合成系[94]は既に報告されているが，本実験では再合成時に使用する音源波形を，任意の対数振幅特性が近似可能な LMA フィルタの特徴を生かして，図 4.26 の様にして求めた。スペクトル包絡としては，改良ケプストラム係数[96]より求まる包絡を用い，その包絡を近似すべく LMA フィルタ係数を求める。更に，そのスペクトル包絡の逆特性に対する LMA フィルタ係数をも求めておく (逆 LMA フィルタと呼ぶことにする)。再合成の際には，LMA フィルタが合成フィルタとなることを考慮すると，自然音声を逆フィルタに通して得られる信号が誤差最小の意味で最適な音源波形と言うことになる (以下，最適音源波形と呼ぶことにする)。実際，自然音声を逆 LMA フィルタ→LMA フィルタに通して作成される合成音声は，自然音声と殆ど聞き分けが付かない。しかし，実際には音源波形に対して F_0 の制御等が入ってくるため，最適音源波形をそのまま使うことは出来ない。しかし，少くとも無声部分は F_0 の制御とは無関係であり，かつ，破裂音などの無声音は (広く行なわれている様に) 白色雑音を音源としてしまうと，合成音声の劣化を導き易い。そこで本実験ではまず，F_0 自動抽出の際に算出される有声度に基付き，自然音声を有声部と無声部に分割する。そして，無声部分の音源波形として，最適音源波形の該当部分を切り出した波形を用い，F_0 制御の対象となる有声部分に対してのみ，パワーを考慮して人工的に作成される音源波形を用いた。なお，構築した分析合成システムについては，第 6 章についてその詳細が述べられている。

F_0 パターンの操作

F_0 パターン操作の基本的な考えは第 4.5 節と同じである。自然音声から自動抽出された F_0 に対して F_0 モデルを適用し，F_0 パターンをフレーズ/アクセント成分の和として記述する。モデルにおいてこれら 2 つの成分は各々フレーズ/アクセント指令から生成される訳だが，パラメータ自動推定により，両指令の大きさ，時間軸上の位置等が推定される。これらのパラメータを制御・変更することで，種々の F_0 パターンを合成し，分析再合成の際に利用する。

フレーズ成分を取り除いた合成音声は，推定された F_0 モデルパラメータ中，フレーズ指令の振幅を全て 0.0 として再合成すれば容易に得られる。しかし，フレーズ指令のみを単純に 0.0 に置換して (アクセント指令はそのまま) 得られる F_0 パターンを用いた合成音声を聞くと，各々の句に非零のフレーズ成分が残っているように聞こえる。これは，アクセント指令の振幅値が句頭の語では大きく，句尾の語では小さく推定される傾向があることに起因する。つまり，両パラメータは互いに相関を持った形で推定されるため，自動推定によるパラメータに直接基付く制御では，F_0 パターンによる統語構造の表記を，2 つの成分に完全に分離することは困難となる。そこで，両指令のパラメータを，テキスト情報・統語構造情報・有声/無声情報を基にして以下の方法で制御することとした[97]。

(1) アクセント指令は図 4.27 に示すように，各単語のアクセントの "Hight/Low" の二値表記に基付いて付与する。アクセントの上がり/下がりに対応する有声音の開始時点よりも 40[msec] 早い時点に，対応するアクセント指令の上がり/下がりを設ける。

(2) フレーズ指令は図 4.28 に示すように，各句及び残された "副詞+動詞" の計 4 つに対して一つずつ付与する。各句の先頭有声音の開始時刻より 50[msec] 早い時点に，対応するフレーズ指令を設ける。

このように制御した場合，フレーズ成分≡0.0 の合成音声は，単語間を越えて変動する成分が全く無くなるので，孤立発声単語を連結したような音声となる。さて，両成分 (コマンド) の振幅値であるが，本実験ではフレーズ/アクセント各々，振幅値は 0.0〜0.3 まで 0.1 間隔で変化させる。また，図 4.29 に示す様に，一文中のフレーズ指令の振幅値は同一値を付与する。アクセント指令も同様である。なお，合成された F_0 パターンの対数軸上での平均値は常に 110[Hz] となるよう合成している。図 4.29 には自動抽出された F_0(* で表記) と，(2) の方法で生成されるフレーズ成分 (破線表示)，及び (1), (2) の方法で合成された F_0 パターン (両成分の和，実線表示) も同時に示している。

被験者

東京大学工学部学生 8 名。今回の実験は F_0 を操作するため，被験者の出身地が実験結果に影響することが考えられる。特に東京方言に比較して，平坦な F_0 パターンが多く使われる東北以北の出身者は被験者として望ましくない。そこで本実験では，東京近辺に在住する，関東以西の出身者のみ (上記 8 名) を被験者として採用した。

実験手順

アクセント/フレーズ指令の振幅値は各々4 段階で変化するため，韻律的特徴の形態 (phrase，accent) は 4×4=16 種類存在する。16 個のターゲット文を，この 16 種類の形態に各々対応させて F_0 パターンを作成し，分析合成する。即ち，異なるターゲット文は異なる (phrase，accent) を持つことになる。一方，ダミー文中の 7 文節文 (10 個) に対しては F_0=110[Hz] 一定で分析合成する。その他のダミー文の F_0 パターンはランダムに (phrase，accent) を適用して作成する。以上の様に作成された合成文音声 (16+27 種類) を，8 人の成人男性に，直前に提示された文音声と同一長の提示間隔を置いて，ランダムな順序で，スピーカーを通して提示し，これを 1 セッションとする。なお，各音声提示の 1 秒前には，ビープ音を提示し，被験者に対しては予め以下に示すタスクを行なうよう指示しておく。上述したように異なるターゲット文は異なる (phrase，accent) を持つが，(phrase，accent) とターゲット文との対応は，被験者間でも異なる様に計画する。第一セッションの 2 日後，同一被験者にもう 1 セッションを，異なる韻律的特徴の付与 (対応) 方法を用いて行なう。

ターゲット文数 16 に対して，韻律的特徴形態が 16 種類あるため，全体で 16×16=176 個の音声試料がある訳だが，上述したように，これを 8 人の被験者に対して各々，1 セッションで異なる 16 個を提示し，各々2 セッションずつ行なわせる。即ち，176=(8×16)×2 であり，被験者間の個人差を誤差として扱うならば，以上の実験手順で 176 種類全ての文音声が被験者に提示されることになる。

タスクと指標

インストラクションに示すように，被験者には，文音声聴取後即座に，可能な限り多くの単語を口頭再生するよう指示しておく。また，実験中被験者には，“PC の CRT 上を動き回るマークをマウスを用いて追いかける” と言う付加的なタスクを行なわせた。指標としては，各々の (phrase，accent) に対する文節 (単語) 単位での正答率や，O_1, O_2 の位置に限定した正答率などを見る。

インストラクション

これから，2 つの作業を同時に行なって頂きます。1 つ目の作業は，

- 画面上を動き回るマークをマウスで追いかける

作業です。2 つ目の作業は，

- 提示される文音声を聴取後直ちに，出来るだけ正確に口頭再生

することです。以下のことに注意して下さい。

1. 文音声は「ピーッ」と言う発信音の後，1 秒して提示されます。
2. 文中，聞きとれない箇所があった場合は，その箇所を無理に埋めずに，一呼吸置いて，聞きとれた部分の再生を開始するようにして下さい。
3. 聞きとれた部分の再生をし忘れることのないよう，注意して下さい。
4. 文音声は直前に提示された文の時間長と同じ間隔を置いて次々と提示されます (次提示を知らせる発信音が提示されます)。口頭再生音声と次に提示される文音声とが重ならないよう注意して下さい。
5. 文音声の中には，イントネーションのおかしい文音声も含まれています。口頭再生の際には，イントネーションを真似る必要はありません。聴取内容を通常のアクセントで再生して頂ければ結構です。
6. パソコン画面上のマークは，発信音の提示終了時刻から文音声の提示終了時刻の間のみ動きますので，口頭再生中はマウスで追いかける必要はありません。

それでは，宜しくお願い致します。

4.10.3 実験結果

図 4.30 に 16 種類の韻律的特徴形態に対する文節単位での正答率を示す。なお，文節末尾の助詞の違い，用言の活用の違い，及び時制の違いは無視し，明らかに同義語と分かる場合も正解として扱った。この図には 7 文節ダミー文 (F_0=110[Hz] 一定) における単語正答率も示している。図 4.31 は，句としての正答率と単語正答率との比を各韻律的特徴形態に対してプロットしたものである。この [句正答率/単語正答率] の意味付けについては後述する。更に，図 4.33 は O_1, O_2 における単語正答率を各韻律的特徴形態に対して示したものである。

4.10.4 考察と検討

まず，同一音声試料に対して異なる韻律的特徴を付与して分析合成した場合，その分節的特徴 (音韻的情報) の伝搬に影響が無いことを確かめるため，全ての実験の終了後 1 ヵ

月して，実験に参加した被験者の内，4 人に対して，ダミー音声試料からの切り出し文節音声提示実験を行なった。タスクは聴取後の口頭再生である。結果を表 4.12 に示す。条件 **A**, **B** は韻律的特徴形態の違いを指し，**C** は自然音声を意味する。結果より各条件間での差は無い。故に，本実験で用いた分析合成音声試料は，韻律的特徴を操作した場合でも，音韻的情報の伝達には支障を来たしていないと言える。また，この文節提示実験は第 4.5 節で行なった帯域制限などの後処理を行なっておらず，音節単位の小さな処理単位での知覚 (分析的態度) をも許す形となっている。その結果，第 4.5 節の結果とは異なり，F_0=110[Hz] 一定の場合でも自然音声と変わらぬ結果を示していると考えられる。

さて，図 4.30 を見ると，7 文節文の場合非常に高い正答率を示している。この場合，提示される文節項目数は 7chunk と呼ばれる STM の容量と同一数である。よって一旦正しく知覚された場合，その項目は STM 内に安定して保持され，正しい口頭再生も行なわれる，と言える。一方 11 文節の場合，常に STM の容量を越えた (overflow) 情報量が入力され，かつ，必要とされる統語解析の処理量の増加により，口頭再生時には約 50[%] 程の項目しか正しく再生できていないことが分かる。この場合，正しく口頭再生できなかった単語は，1) 該当する辞書項目への検索そのものが失敗したのか，2) 正しく検索され，知覚することは出来たものの，その後に入力された情報によって STM 内で "上書き" される形になったのかは，本実験だけで結論を出すことは出来ない。今後の実験課題の 1 つである。本実験結果から直接的に言えるのは次の通りである。11 文節文音声聴取後 STM 内に保持されている情報の内，正しく辞書検索/照合されているものは全文節数の約 50[%] 程のみであり，かつ，この数の韻律的特徴への依存度は，本実験では観測されなかった。この結果は，孤立単語提示のパラダイムで行なった第 4.5 節とは異なる傾向を示している。これらは，表 4.12 の考察でも述べたような 1) 分析的態度による音声聴取による効果，そして上記した 2) 統語処理量の増加[23]に起因するものと考えている。

単語単位での正答率では韻律的特徴との相関が観測されなかったが，本実験結果を異なる観点から眺めることにより，韻律的特徴の果たす役割の，異なる一面が見えてくる。図 4.31 は，[句正答率/単語正答率] を各韻律的特徴形態毎にプロットしたものである。この比は図 4.32 に示す様に，"正解単語が正解句に含まれる割合 (正解単語の分布の様子)" を示す。図 4.32 では，フレーズ指令の振幅値の増加を "phrase += 0.1" の様に示し，[句正答率/単語正答率] を phrase/word と示している。図 4.30, 4.31 より，アクセント指令

[23] 本実験のターゲット文は，どの音声試料も等しい難度の統語解析が要求される。単語正答率が，音響的提示条件よりも，統語処理の難度により大きく依存するならば，本実験の結果もうなずける。

の値によらずフレーズ指令の増加に伴って，全体の正解単語数には大きな変化は無いものの，その分布の様子が変化し，正解単語は正解句中により多く含まれる傾向にあることが分かる。更に，本実験では振幅値が 0.0 から 0.1 に上がる時に大きな変化が認められている。これは，フレーズ成分が，同一句内に存在する単語をグルーピングする効果を持ち，グルーピングされた単語がSTM 内に安定して保持されることを示唆する。これは人間における音声処理過程において，フレーズ成分が統語構造の解析，特に句境界の検出に有効に作用していることを示唆するものである。これは「文音声音声知覚に対してはフレーズ成分の役割は観測されない」とする参考文献 [98] とは異なる結果である。また，第 4.9 節の考察で述べたような，音声をより大きな塊として捉えるプロセスと，本実験で明らかとなったフレーズ成分による単語間のグルーピング効果との関係が予測されるところであり，その実証は今後の課題の一つである。

図 4.33 は図 4.25 の O_1, O_2に配置された 0~3 型アクセントの 4~6 モーラ単語に対する正答率を示したものである。アクセント指令の振幅値の増加に伴って，正答率が上昇しているものは 1 型アクセントのみである。この結果より，文中においても，1 型単語知覚のアクセント依存の特異性が認められた。

4.10.5 まとめ

直接的結果

- 単語単位での正答率はフレーズ/アクセント指令の振幅値の変化に依らず，ほぼ一定の値を示した。
- しかし，各単語をアクセント型別に集計すると，1 型の単語のみ，アクセント指令の振幅値の増加と共に正答率が上昇している (他は顕著な傾向は無し)。
- また，フレーズ指令の振幅値の増加と共に，正解される単語が，句を単位として集中してくる (フレーズ指令≡0.0→正解単語はランダムに配置) 傾向にある。
- この傾向は本実験で設定した最小振幅値 (0.1) においても観測された。

考察及び知見

- 文音声を知覚する場合，フレーズ成分を一つのキーとして，文を構成する単語系列を幾つかの単語群に分割して処理を行なっている。
- 人間は，フレーズ成分を一つの手がかりとして，統語解析を行なっている。特に句境界検出に対する役割が大きいと考えられる。
- 辞書項目間の緩やかな結合による大きな処理単位は，フレーズ成分によるグルーピング効果により，その効果 (知覚の早さ・容易さ) を増すと予想される。

$$\underline{\mathrm{O+V+}\mathbf{S}} + \underline{\mathrm{S+V+}\mathbf{O}_1} + \underline{\mathrm{O+V+}\mathbf{O}_2} + \mathrm{Adv+}\mathbf{V}$$

図 4.25. ターゲット文の統語的構造

表 4.11. 本実験で使用された 11 文節ターゲット文

しゅくだいを おなかを	やりのこした すかせている	ゆうじんが のらねこに	おじさんが うらにわで	こしらえた たべさせた	おむすびを
がっこうを かなしみを	そつぎょうする かくせない	いもうとが ともだちに	ははおやが きょうしつで	かしてくれた てわたした	はんかちを
きょうかしょを ひるごはんを	うけとった たべおえた	しょうがくせいが せんせいに	にいさんが かいだんで	わすれた あずけた	うでどけいを
すうがくを しけんを	せんこうする のりこえた	がくせいが せんぱいに	ちちおやが きゃくしつで	かってきた さしあげた	ウィスキーを
プレゼントを おもちゃを	さがしていた ほしがる	おばさんが おとこのこに	しりあいが きゅうけいじょで	うっていた あたえた	キャンディーを
にほんごを おべんとうを	はなせない たべている	がいじんが かんごふに	おまわりさんが びょういんで	おしえた たずねた	としょかんを
ぱそこんを キーボードを	どうにゅうした つかえない	ぶちょうが やくいんに	てんいんが かいぎしつで	しめした ひろうした	そうさほうを
ファミコンを おみやげを	たのしんで もってきた	いたきょうだいが りょうしんに	たんにんが しょくたくで	ことずけた わたした	つうしんぼを
ゆうしょうに れんしゅうを	ちかづいた かかさない	かんとくが せんしゅたちと	あばあさんが たいくかんで	つくった ひろげた	たれまくを
けんしゅうに こどもたちを	さんかした しどうする	こうちょうが わかものに	おいしゃさんが こうえんかいで	すすめた くばった	はぶらしを
おさらを レコードを	みがいていた ふいている	いとこが おとうとに	レポーターが さんがいで	ぜっさんした つたえた	コンサートを
しょさいを しんぶんを	そうじした よんでいる	きょうじゅが おじいさんに	おくさんが えんがわで	さくせいした みせている	おりがみを
カレンダーを ぎゅうにゅうを	ながめていた のみたがる	むすめが あかちゃんに	ラジオきょくが ベランダで	ながしている きかせた	おんがくを
かもしかを けがわを	しとめた ほしがっている	りょうしが むらびとに	むすこたちが やすねで	つかまえた うりとばした	きたきつねを
かいものを やきゅうを	すませた かんせんしている	おかあさんが おにいさんに	みんなが だいどころ	だいすきな でつくった	ぜんざいを
みつばちを キャラメルを	こわがる ほおばっている	おいっこが おんなのこに	むらさきが すべりだいで	きれいな さしだした	コスモスを

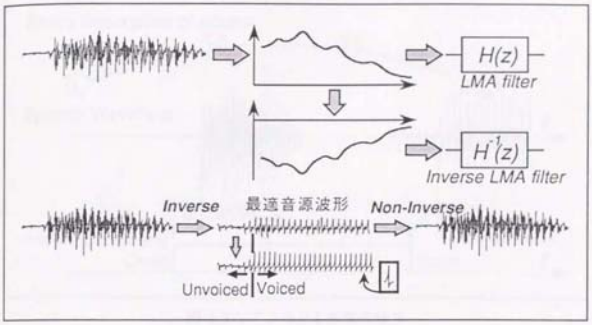


図 4.26. LMA フィルタを用いた音源波形の生成

表 4.12. 切り出し文節音声提示実験結果 ([%])

被験者 ID	条件 A	条件 B	条件 C
4	96.6	97.7	97.7
5	97.7	98.9	98.9
6	98.3	99.4	98.3
8	98.3	98.9	98.3
平均	97.7	98.7	98.3

条件 A: アクセント指令=0.0, フレーズ指令=0.0
条件 B: アクセント指令=0.3, フレーズ指令=0.3
条件 C: 自然音声

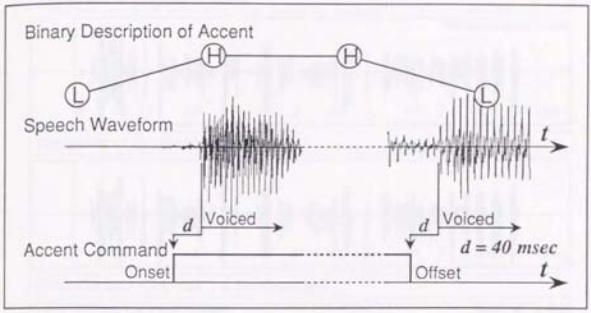


図 4.27. アクセント指令の付与

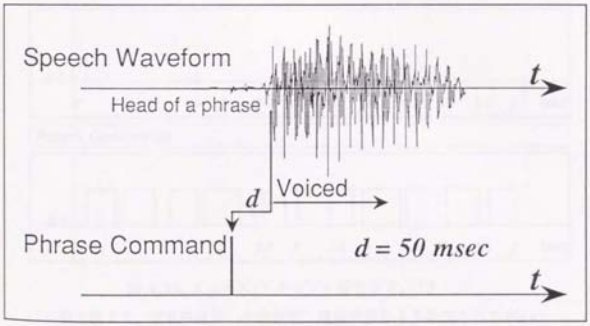


図 4.28. フレーズ指令の付与

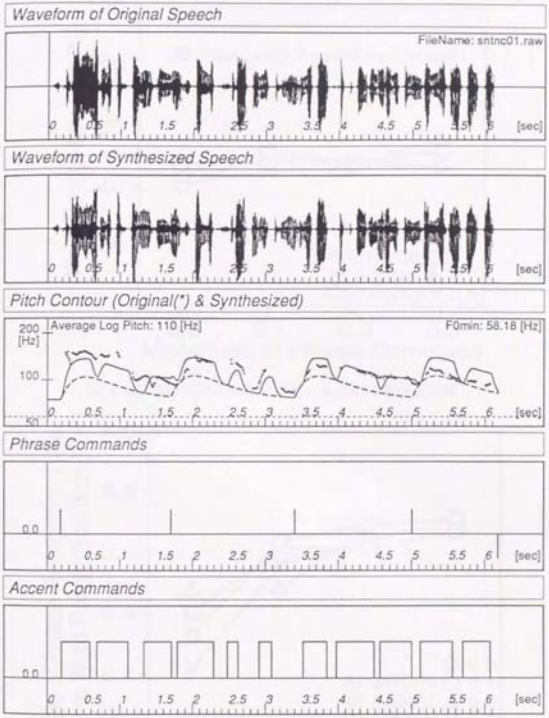


図 4.29. フレーズ/アクセント指令と F_0 パターン

最上段より，原音声波形，合成波形，原音声波形より抽出された $F_0(*)$ と合成された F_0 パターン，アクセント成分，フレーズ成分である。アクセント指令，フレーズ指令の大きさは，文を通して各々一定とした。

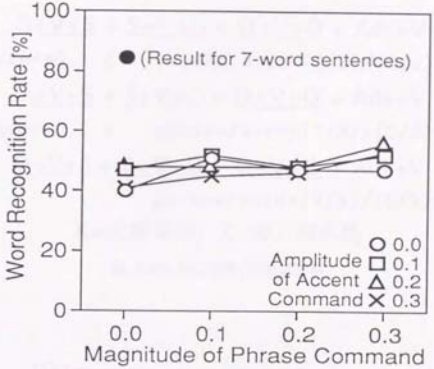


図 4.30. 各韻律的特徴形態に対する文中の単語正答率

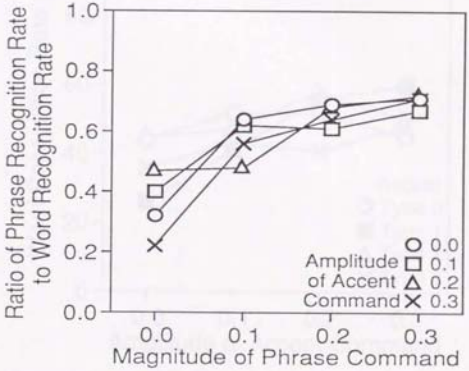


図 4.31. 各韻律的特徴形態に対する句/単語正答率

O+**V**+**S** + **S**+V+O + **O**+V+**O** + Adv+**V**

phrase+=0.1 ⇓ phrase/word=(0/3)/(5/9)=0.0

O+V+**S** + S+**V**+O + **O**+**V**+**O** + **Adv**+V

phrase+=0.1 ⇓ phrase/word=(1/3)/(5/9)=0.6

O+V+S + **S**+**V**+**O** + **O**+**V**+**O** + Adv+V

phrase/word=(2/3)/(5/9)=1.0

X=正解単語，X=非正解単語

図 4.32. 句正答率/単語正答率

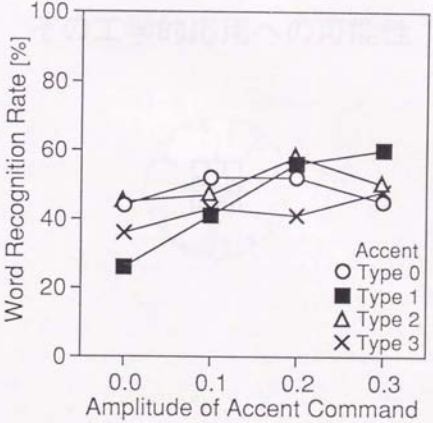


図 4.33. 各アクセント型/指令に対する O_1, O_2における単語正答率

第 5 章

人間の音声知覚過程のモデル化とその工学的応用への可能性

本章ではまず，第 3 章，第 4 章で述べた知覚実験から得られた結果・知見をまとめる。次にこれらの結果・知見を基に，人間における音声言語処理全体像を見渡すことのできる知覚モデルを構築する。但し上記の結果・知見のみでは，下に示すような理由により，定量的かつ精密な工学モデルを構築することは困難である。しかし，定性的な記述が許される認知科学的モデルの構築には，十分な知見が得られたと筆者は考えており，ここでは認知科学的な立場からのモデル化を行なう。

次に，構築された音声知覚モデルの工学的応用を考える。当然のことながら，

- 完全な工学的実現を図るには，十分な (詳細かつ定量的な) 分析結果が揃っていない部分が多い。
- 特に，モデルに記述される，個々の処理部間の相互依存関係については，更なる定量的実験を要する。

などの，認知科学的モデルに起因する (工学モデルとしての) 不完全性[1]と同時に，

- 現在の計算機の処理速度を用いてもまだ，十分なパフォーマンスは得られない。

などの理由により，現在の段階では構築したモデルを直接的に，詳細に，完全にインプリメントすることは非常に困難である。しかし，知覚モデル内の個々の処理部に目を向けた場合，計算機上での処理に応用できる知見も少なくない。そこで，構築された音声知覚モデルの各処理部に対して個別に焦点を当て，そこから得られる知見を生かした認識手法を考案することにする。

5.1 知覚実験より得られた種々の結果・知見

本節では第 3 章，第 4 章で述べた知覚実験から得られた結果・知見をテーマ毎にまとめる。しかし，複数のテーマにまたがった結果・知見もあり，適切な分類を一元的に行なうのは困難であるが，同一結果/知見が複数回記述されないよう，各々を適宜，いずれか一つのテーマに属させて示すことにする。なお，以下では知覚実験より直接的に得られる現象としての結果を "結果" とし，結果に対して種々の考察を行なうことで得られる間接的な結果を "知見" として示すことにする。また，各々の末尾には該当する実験/考察が述べられている節番号を併記している。また，これらの結果/知見には，当研究室で行なわれた先行研究 (第 3.4 節) によって導かれたものも含まれている。

[1] このような部分に，"仮定" を設けて実現すると言う考えも当然ある。

5.1.1 大前提として

知見 1　音声は音響的処理と言語的処理の相互作用によって処理され，不完全にしか受信できなかった情報を互いに補間し合う形で処理される。 (3.6)

5.1.2 言語音としての知覚

知見 2　入力音が言語音として知覚されると，必然的に，情報の形態は連続量 (物理量) から離散量 (言語ラベル) へと変換される。 (3.4.1)

知見 3　離散化されることで，記憶の中に，より安定して保持されることになり，その結果，次入力音との同一性判断への影響がより大きくなる (→範疇化効果)。(3.4.1)

5.1.3 音声処理単位と処理単位長の違いにより生じる処理特性の差異

知見 4　音声処理単位長は複数存在する。 (3.4.2)

知見 5　音声処理単位長は，入力音声の言語的属性 (コンテキスト長を含む) を一つのパラメータとして動的に変化する。 (3.4.2, 4.7)

知見 6　文章音声の場合，単語・文節ほどの大きさの (少なくとも音節よりも長い) 音声長を単位とした処理が主に行なわれる。 (3.4.2)

知見 7　音声処理単位の大きさは句·文にまで及ぶ (但し諺·短歌など，語系列として LTM に登録されていると考えられる項目に限る)。 (4.7)

知見 8　基本的にはまず，より大きな単位での処理を試み，十分に整合性のある候補を限定出来なかった場合に，より小さな単位での処理へと移行する。 (3.4.2, 3.4.3)

結果 9　同一音声を異なるサイズの単位を用いて処理した場合，より大きな単位における処理ほど，“音響的特徴抽出処理”+“音響的照合処理” に必要とされる処理時間は短縮される (早い知覚)。 (3.4.3)

知見 10　より大きな単位を用いた処理ほど，低精度の音響的特徴量による正しい音響的照合が可能となる (容易な知覚)。この低精度の音響的特徴は，より早期の段階で抽出が完了する特徴量と考察される (→早い知覚)。 (4.7)

知見 11　但し，同一単位長における処理が行なわれている場合でも，照合対象となる項目の言語的属性により，正しい知覚に必要な特徴量の精度は異なる。即ち，同一単位長における処理でも，複数の精度の音響的特徴を扱う機構が存在する。(4.2)

知見 12　文音声知覚時には，単なる単語列と句・文サイズの処理単位の中間的性質を持った単語網がコンテキストに依存しながら構築され，その単語網を用いて，音声をより大きなまとまりとして捉えようとする。 (4.8, 4.9)

5.1.4 内部辞書の構成 (と辞書検索処理・音響的照合処理過程)

知見 13 より大きなサイズで内部辞書に登録されている項目は，特徴抽出過程において早期に抽出が完了する低精度の音響的特徴を用いた照合処理においても，辞書検索処理部による検索の対象となり，かつ，入力音声との正しい照合が行なわれる特性を持つ (知見 10)。 (3.4.3, 4.7)

結果 14 ある辞書項目に対して正しく照合処理を行なうために必要な音響的情報量は，項目固有の特徴である，長期的頻度を 1 つのパラメータとして変化し，長期的頻度の高い項目ほど，低情報量で正しい照合が可能となる。 (4.2)

結果 15 長期的頻度の高い項目ほど，(順序的に) 優先されて検索される。 (4.2)

知見 16 「優先的な検索」，及び「低精度の特徴 (早期に抽出が完了する) による正しい照合」と言う特性を持つ長期的頻度の高い辞書項目は，最終的に早く知覚されることとなる。 (4.2)

結果 17 短期的な出現頻度が高くなった項目も同様，辞書検索において優先的に検索される。 (4.3)

知見 18 短期的な出現頻度が高くなった項目も，早期の段階で抽出される低情報量の音響的情報で正しい照合が可能な状態となっており，その結果，いち早く知覚が早く完了する。 (4.3)

知見 19 但し，長期的頻度のもたらす効果と短期的頻度のもたらす効果[2]を実現する機構は異なり，前者は LTM である内部辞書の構造的な要因によるもの，後者は cache 的 STM の存在によるものと考察される。 (4.2, 4.3)

知見 20 1 型アクセント語の場合，その型は語頭のみの入力音声で識別できるため (分節的特徴による照合処理が行なわれる以前の) 検索処理において，検索すべき辞書空間を大幅に限定することができる。その結果，知覚がより容易に/早くなる。 (4.5)

知見 21 1 型アクセントの辞書検索範囲の絞り込み効果は，文中の単語においても適用される。 (4.5)

知見 22 アクセントパターンを記載した内部辞書は，意味・統語的属性などを記載する内部辞書とは独立して存在する。 (4.6)

[2] 上記した様に両者は類似している。

5.1.5 辞書検索処理過程 (特に言語的情報の果たす役割)

結果 23 先行コンテキストに意味的関連性のある項目が存在した場合，優先的に辞書検索が行なわれるようになる。 (4.4)

結果 24 文節間の意味的・統語的整合性の存在は，より低情報量の音響的特徴における該当項目への検索を確実なものとする (容易な知覚)。 (4.7)

結果 25 文節間の意味的・統語的・談話的整合性の存在は，より短時間の入力 (低情報量の音響的情報) における該当項目への検索を確実なものとする (早い知覚)。(4.8)

結果 26 統語的整合性と意味的整合性では，後者の方が知覚の早さに，より大きく影響を与えている。 (4.8)

知見 27 より親密な事象を記述した文に対しては，より大きな範囲で音声を捉えようとする機構が存在する (知見 12)。但しここで言う範囲とは，単位と呼ぶほど確立されたものではなく，単語 (文節) 間の音響的・言語的な緩やかな結合で音声を捉えることを指す[3]。そのため，知見 10 にあるように，早期に抽出される低情報量の音響的情報で正しい照合が可能となる。 (4.8, 4.9)

結果 28 より親密な事象を記述した音声中の単語ほど (大きな範囲で音声を捉えた処理が行なわれるため)，語頭音の重要性が低下する。 (4.8)

結果 29 入力音声を left-to-right に知覚した場合と，そうでない場合では，文が記述する通常性 (談話的情報量) が文知覚に及ぼす影響は異なる。 (4.9)

結果 30 2 文節文において，left-to-right に処理が行なわれる場合，通常性 (談話的情報量) は知覚を助ける方向には働くが，妨げる方向には働かない。 (4.9)

結果 31 left-to-right の処理が困難な (あるいは，行なわれない) 場合，通常性 (談話的情報量) は知覚を助ける方向にも妨げる方向にも作用する。 (4.9)

知見 32 より大きな範囲で音声をとらえた処理の結果 (知見 12, 27)，知覚結果を 1 つに特定出来ない場合，left-to-right の処理が可能であれば，left-to-right に個々の文節を独立に扱った処理へと速やかに移行する (知見 8)。そして，以前に知覚された辞書項目の影響も小さくなる。 (4.9)

5.1.6 韻律的特徴の果たす役割

結果 33 辞書項目本来のアクセント型と異るアクセント型で提示されると，その正答率は減少する。特に 1 型アクセントの場合，処理過程のアクセント依存性が非常に高

[3] 単なる音節列 (単語) が単語 (文) 項目として確立される過渡期における処理，と考えると分かりやすい。

い (知見 20)。 (4.5)

結果 34 F_0 パターンの "上がり/下がり" を単独提示 (2 音節) すると，"上がり" がより早く知覚されるが，既知&未知アクセント語尾に含ませて提示 (4 音節) すると，"下がり"(既知アクセントに対応) への反応の方が早くなる。 (4.6)

結果 35 "上がり/下がり" どちらかに限定して，単独提示と単語 (既知 F_0 パターン) 内提示を比較した場合，単語の有意味性によらず，単語内提示の反応時間が短い。 (4.6)

知見 36 入力音声の有意味・無意味に関わらず，既知アクセント型で捉えようとする機構が存在する。 (4.6)

結果 37 F_0 パターンを構成するアクセント成分とフレーズ成分のうち，後者は文中の単語をフレーズ毎にグルーピングする効果を持ち，しかもその効果はフレーズ成分の値が非常に小さい時に (0.1)，既に現れ始める。 (4.10)

知見 38 フレーズ成分によるグルーピング効果が一句中の単語間を緩やかに結び付け，知見 12, 27 で言う，より大きな範囲による音声処理を引き起こす一要因である考察される。 (4.10)

5.2 人間の音声知覚過程の全体像のモデル化

本章の冒頭で述べたように，人間の音声知覚過程の解明，特に定量的かつ精密な工学的モデルの実現に対しては，まだ数多くの実験が必要である。しかし，人間の全体像を見渡すことのできる，認知科学的モデル[4]の構築を行なうには，十分なだけの知見が得られたと筆者は考えている。そこで本節では，以上得られた知見を基に，まず人間の音声知覚過程の各処理部に対するモデル化を行ない，次にそれらを用いて試験的にではあるが，人間の全体像に相当する知覚モデルを構築することにする[74]。但し第 3.2 節で Logogen, Cohort と言った心理学的/認知科学的モデルに対して,「音響的/言語的処理両方に渡って極度の抽象化が行なわれている。」と書いた。そこで本モデルの構築に当たっては，両処理をより具体的に表記することを心掛ける。

5.2.1 大まかな全体像

各処理部のモデル化を試みる前に，各々をブラックボックスと考え，音声知覚処理の全体を大まかに捉えてみる[5]。音声の持つ意味的内容 (メッセージ) を正確かつ迅速に認識・

[4] 定性的記述が許されるモデル。
[5] Cohort や Logogen モデルのような，非常に抽象化された知覚モデル。

理解する為には，非常に変動の大きい音響的特徴を扱う音響的処理よりも，一旦シンボル化(離散化) された情報を扱う言語 (知識) 的処理が主導権を握り，系全体を制御しながら処理が行なわれていると考えるべきであろう。即ち，前後コンテキストに対する言語的処理結果に基付いて，音響的処理形態 (の一部) が決定され，実行されていると推測される。音響的処理は一般に,「音響的特徴の抽出」及び「抽出された音響的特徴に基付く辞書項目との照合処理 (音響的整合性の算出)」に大別される。末梢神経レベルの非常に低次の処理である音響的特徴の抽出に対しても，抽出すべき音響的特徴の種類など，言語的処理結果が影響を及ぼしていることも考えられない訳ではない。しかし，照合処理の対象となる辞書項目数の莫大なサイズを考慮すると，言語的処理からの制御とは,「膨大な数の辞書項目からどの順序で照合処理を行なわせるか，即ちどの順序で検索を行なうか？」がその多くを占めるであろう。以上の考察の下，人間の音声知覚モデルの概念図を示すと図 5.1 のようになる。この図で，内部辞書が 2 つの箇所に示されている (言語的処理部と内部辞書検索処理部) が，これは，2 次元の図面上でのモデル化を行なうために採った手段であり，以後のモデル化においても本来一つにまとめるべきモジュールを便宜上，複数の箇所に配置しているところがあることを断っておく。

5.2.2 音響的特徴抽出処理部

入力音声から音響的特徴を抽出する。生理学的には抹消神経系における処理である。抽出する音響的特徴は,「音韻情報を伝達する分節的特徴 (segmental feature)」と「韻律的情報を伝達する韻律的特徴 (prosodic feature, 或は超分節的特徴，supra segmental feature)」の 2 つに大別される。辞書項目への検索/項目との照合と言う観点から考えた場合，前者は，入力音声をある特定の辞書項目に対応させるために必要不可欠な情報であり，後者はその対応付けを助けるべく情報であると言える (知見 20, 21)。モデル化の際に，まず問題となるのが，この 2 種類の音響的特徴は各々異なる処理部で抽出されるようモデル化すべきかどうか，と言うことである。この問いに答える手がかりとしては，生理学的に，1) 信号のスペクトルパターンの変化[6]に反応する聴神経と，2) 信号の時間パターンの変化[7]に反応する聴神経が異なる部位に存在するのか否かを観測することなどが考えられる。当然のことながら，上記のような実験は第 2.1 節で述べた，筆者の実験方針とは異なるものであり，第 3 章，第 4 章では行なわれていない。ここで，一般的に知られている以下の生理学的な実験事実を考える。

[6] 具体的にはフォルマントの位置の変化などに相当する。
[7] 具体的には F_0 の変化などに相当する。

1. 聴神経は周波数選択性を示し，特定の周波数 (特徴周波数，Characteristic Frequency，以下 CF と記す。) に対して強く反応する[99]。
2. 聴神経の発火は，音刺激の特定位相 (即ち時間パターン) に同期する傾向があり，phase-lock と呼ばれる[100]。
3. phase-lock の CF 依存性であるが，入力音のフォルマント周波数と離れた CF を持つ聴神経ほど，入力音の時間パターンにより追従した発火パターンを示す[101]。

これらの実験結果は，入力音声のスペクトルパターンに対応して発火する聴神経 (フォルマント周波数に近い CF を持つ聴神経) と，時間パターンに対応して発火する聴神経は各々相補的に配置されていることを示す。即ち，ある入力音声に対する分節的特徴と韻律的特徴は常に離れた部位で抽出されることを意味する。但し上記したように，分節的特徴/韻律的特徴を抽出する部位が固定的に存在している訳ではなく，聴神経が入力音声に依存しながら (動的に対応しながら)，各々異なる部位で両特徴を抽出している，と言うことである。以上の生理学的実験結果を考慮すると，ある入力音声に対する分節的/韻律的特徴は，異なる処理部によって抽出されるようモデル化されるべきである。

また，音響的照合処理では，処理単位長及び提示音声 (及びコンテキスト) の言語的属性に依存しつつ，複数精度の音響的特徴が利用される (知見 10, 11)。これらの知見を導いた実験では，雑音重畳や帯域制限など，主に分節的特徴の伝搬の様子を操作しているが，この高/低精度の音響的特徴と言う考えは，韻律的特徴においても考察することができる。分節的特徴が音声のスペクトルパターンによって記述されること，韻律的特徴が音声の時間パターンによって記述されることを考えると，前者における精度とは周波数分解能として，後者における精度とは時間分解能として考察できる。当然のことながら分節的特徴においては，周波数パターンの時間変化も考慮されるべきであり，この場合は時間分解能も組入れて考える必要がある[8]。

以上のような分節的/韻律的特徴における複数精度の照合処理を実現するためには，音響的特徴抽出部において複数の精度の特徴を出力する必要がある。当然のことながら，より低精度のものがいち早く抽出・出力され，分析時間の経過と共により詳細な高精度の特徴が抽出されることになる。これら複数の精度の音響的特徴の抽出/出力，それを入力とする複数精度の照合処理は，精度に関して縦続的にモデル化することも可能であるが，その場合，処理の時間的順序が固定され，モデルが記述する処理の柔軟性が失われることになる。また，PDP における研究を考慮すると，モデルとしては並列的にこれらの処理が

[8] 但し，分節的特徴における時間分解能は後述する照合処理単位の大小でも，その一部が表現される。

行なわれるよう配置するのが妥当であろう。更に上記した両特徴量の質的違い (不可欠な特徴か否か)，及び第 5.2.4 節で述べる両特徴量に対する内部辞書の構造的差異を考慮すると，時刻 t の韻律的特徴の抽出が時間的に早く行なわれ，それを追従する形で時刻 t の分節的特徴の抽出が行なわれると考察される[102]。以上の考察の下，音響的特徴抽出部をモデル化したものを図 5.2 に示す。

5.2.3 音響的照合処理部

音響的特徴抽出部の出力と内部辞書検索部からの候補項目を入力とし，音響的照合を行なった後，照合結果 (音響的整合性) を出力する。但し，照合処理の前後には夫々蓄積する情報の質が異なる STM を配置する必要がある (連続的情報と離散的情報，知見 2, 3)。また，分節的特徴と韻律的特徴に基付く照合処理部を，各々別個にモデル化すべきであろう。と言うのも，第 5.2.2 節で見たように，分節的/韻律的特徴抽出が異なる処理部としてモデル化されることに加え，アクセント型の辞書と単語辞書[9]とは各々独立して存在している (知見 22) からである。即ち照合処理部への 2 つの入力である，音響的特徴 (入力パターン) 及び照合項目の音響的特徴 (標準パターン) の両者において，分節的特徴と韻律的特徴が各々，別個の処理部/データベースとしてモデル化されると言う知見は，分節的/韻律的特徴による照合処理も十分に異質なものであることを示唆しており，これを同一処理部としてモデル化することは望ましくない。

また，分節的特徴による照合処理の大きな枠組みとして，入力される特徴の精度の高/低 (周波数分解能) に対して，照合処理単位の小/大 (即ち，分節的特徴における時間分解能) を対応させる必要がある (知見 9, 10, 11)。こうすると，より大きな単位での処理はより早く抽出される低精度の特徴を利用し，かつ照合処理においても単位時間当りに処理する情報量が少ないため，照合そのものも早く終了することになり，処理単位長と処理速度/負荷 (知覚の早さ/容易さ) との関係が自然と記述されるようになる。上記の精度と処理単位との関係は，第 5.2.2 節で述べた韻律的特徴の抽出，及び照合処理にも当てはまる。例えば F_0 に着目した場合，入力音声が既知パターンであれば，そのアクセントパターンは単語全体のパターンとして知覚可能であるが，未知パターンであれば，該当する単語全体のパターンはアクセント辞書に存在せず，個々の (2 モーラ単位の)"上がり/下がり/平坦" の連結として知覚される (第 4.6 節参照)。この "上がり/下がり/平坦" のパターンも，2 モーラ単語アクセントとして辞書に登録されてあることを考えると，入力音声の F_0 パターンは $n(n=1,2,3,\ldots)$ モーラ単語アクセントとして，複数サイズの単位で照合さ

[9] 音響的特徴，統語的属性，意味的属性などが記述されている。

れることになる。そして，この処理単位のサイズと必要となる韻律的特徴の時間分解能に，上記の関係があると考察される。また，文章音声を処理する場合の処理単位のサイズとしては，分節的特徴に対して考察した知見 6 をそのまま適用しても，問題は無いであろう。

更に同一単位における処理においても，複数の精度の分節的/韻律的特徴量が扱えるように記述する必要がある (知見 11)。また，大小の単位での処理間の関係を考慮し，“より大きな単位による処理において，十分な整合性を有する結果が得られない場合に，より小さな単位での処理結果を参照する (知見 8)”，と言う処理の流れも汲み込む必要がある。但し上記の知見は，小さな単位の処理が，大きな単位での処理に後続して開始されることを意味するものではない。小さな単位での処理そのものは並列して行なわれており，大きな単位での処理が時間的により早く，最終的な結果を出力する，と言うことである。内部辞書検索処理部と関係してくるが，語頭音韻など (小さな処理単位ではあるが) 時間的にいち早く出力される照合結果が，より大きな単位での処理に影響を与える (第 3.1.3 節，第 3.1.6 節参照) ことも考慮されてモデル化されるべきである。以上の考察の下，音響的照合処理部をモデル化すると，図 5.3 のようになる。この図で，照合処理部の前後に配置されてある STM の形状の違いが，連続的/離散的と言う保有する情報の質的違いを表現している。また，照合結果を保存するよう配置されてある STM 間の矢印に，“あるサイズの処理単位で十分整合性の有る結果が得られない場合，より小さな処理単位での結果を参照する” を言う意味を持たせている。

5.2.4 内部辞書 (心的辞書，Mental Lexicon)

第 5.2.2 節，第 5.2.3 節で述べたように，照合処理過程において，分節的特徴及び韻律的特徴 (アクセントパターン) による照合は別個に行なわれる。そして各処理に対応する辞書も，各々独立して存在している (知見 22)[10]。これらの辞書は LTM として存在しているため，短期的な外界変動 (コンテキスト) による影響は少なく，その特性は全て静的なものである。そして，この辞書に対するアクセス (辞書検索) 方法が動的に変化し，音響的/言語的コンテキストが及ぼす動的影響を創り出していると言える。この静的な特性に関して考慮すべき知見としては，まず “辞書項目のサイズと正しい照合を可能とする音響的特徴の精度 (知見 12)” が上げられる。しかしこれは，第 5.2.3 節で示した照合処理部の各単位に対応した，音韻辞書，音節辞書，単語辞書，(句辞書)，文辞書を想定すれば容易

[10] 但し，単語辞書の各項目には，対応するアクセント辞書項目の ID(あるいはポインタ) が記載されていると考えている。

にモデル化できる。即ち，辞書項目はサイズ別に分類され，蓄積されていることになる。次に "長期的頻度 (知見 14)" 及び，"同一項目に対する複数精度の対応 (知見 11)" が考えられる。長期的頻度は項目固有の特性であり，その影響は，内部辞書の構造そのもので表現されるべきである。即ち内部辞書内に，その静的特性として，優先して検索が行なわれ，かつ，低情報量の音響的特徴で正しい照合が可能となる性質を持つ項目が，固まって存在していると考察される。即ち，上記の特性をより強く持つ項目から弱く持つ項目へと方向性を持った項目の配列が行なわれていると言える。このように考えると，単語提示のようにコンテキストが無く，静的特性のみによって知覚過程が記述される場合は，照合処理単位 (辞書項目単位) が同一であっても，該当単語の持つ特性に応じた個々の精度で単語が知覚されることになる。即ち後者の知見に対しても自然に対応することができるようになる。一方 "短期的頻度 (結果 17、知見 18)" の影響は，cache 的 STM を想定することで説明可能となる (知見 19)。そして，この cache 的 STM に格納された項目は，あたかも，その項目が長期的頻度が高くなったような特性を持っているように振舞う。即ち辞書検索部によって，LTM の検索に先立ってこの STM が検索され，かつ，正しい照合に必要な情報量も低くなる。但し，長期的頻度の影響と短期的頻度の影響の定量的比較は行なわれておらず，両者における優先度などは今後の課題である。

以上の議論の一部は，アクセント辞書に対しても適用することができる。即ち，"サイズ別の辞書の存在"，"照合単位との対応"，及び "長期的頻度[11]" などの要素はアクセント辞書に対しても同様に組入れることができる。但し，cache 的 STM がアクセントに対しても存在するか否かに対しては，十分な説得性のある議論をすることは現在のところできない。即ち，単語アクセントに関する全情報が非常に少ない容量で格納されてしまうことを考慮すると (LTM として存在してはいるものの)，検索速度は単語辞書と比較して非常に速いと考察され，cache 的 STM としてのモデル化に対する必要性はこの点からも疑問視される訳である。上記の点に関しては，無意味語を用いて，既知アクセントパターンの出現頻度を短期的に操作した実験などを行ない解明する必要がある。

以上の考察を下に，分節的照合処理に対応する内部辞書と韻律的照合処理に対応する内部辞書をモデル化したのが図 5.4 である。図では，音節/単音辞書においては長期的頻度の影響を考慮せず，全ての項目が (優先度の意味で) 平等に検索されるよう，モデル化している。しかし，統計的には出現頻度のより高い/低い音節/単音は存在しており，このレベルの辞書においても長期的頻度を考慮すべきか否かは，今後の研究に依らざるを得な

[11] 実験結果としてある訳ではないが，アクセントの偏りは現実に存在する。

い。また，内部辞書項目の各フィールドの音響的/言語的表現方式など不明な点が多いのも事実である。

5.2.5 内部辞書検索処理部

主に先行コンテキストに対する言語的/音響的処理の結果に基付き，照合対象とすべき辞書項目を決定，検索し，音響的照合処理部へと送る。即ち，照合処理部と内部辞書間のインターフェイスであり，第 5.2.1 節で述べた，言語的処理に基付く音響的処理の制御に相当する。第 5.2.3 節，第 5.2.4 節に述べたように，照合処理及び内部辞書は照合単位のサイズ別にモデル化されている。これは，両者のインターフェイスである辞書検索部にも同様の構造を要求するものである。この検索処理の特性を動的に決定する要因としては第 3.6.2 節で考察した種々のものが考えられる。これらを第 5.1 節でまとめた結果/知見に照らし合わせて考えると以下のようになる。まず音響/音声学的要因に対しては，

1. 異なる処理単位間での両方向性[12]の相互作用 (知見 8，第 3.1.3 節，第 3.1.6 節)
2. 韻律的情報処理結果に基付く検索範囲限定 (知見 20, 21)
3. フレーズ成分によるグルーピング (→単語網の生成)(結果 37，知見 38)

などが記述される必要がある。1. における "大きな単位→小さな単位" への方向に対しては，第 5.2.3 節で述べた照合処理結果を格納する STM のモデル化において，その一部は表現されている。しかしこれだけでは不十分であり，大きな単位による照合処理結果によって，先行音韻への予測が生成されること，及びその予測による，より小さな単位での処理過程の変化 (第 3.1.3 節，第 3.1.6 節，第 3.3.1 節) も記述される必要がある。逆に語頭音など，時間的に先行して出力されるより小さな単位での照合結果が，より大きな単位での辞書検索に影響を与える様子も記述される必要がある。

分節的特徴に基付く辞書検索と照合，韻律的特徴に基付く辞書検索と照合，の両者を比較した場合，その検索範囲は後者の方が遥かに狭く，後者の結果がいち早く得られると考察される[102]。特に，語頭音で識別可能な 1 型アクセントなどは顕著にその様子が観測されている (知見 20, 21)。その結果上記 2. にある，先行するアクセント知覚による (分節的) 辞書検索範囲限定による効果は，分節的辞書検索処理には欠かせない一面である。

3. の単語網の生成であるが，これは長期的な音声言語活動の結果 LTM(内部辞書) に生成されているものとは異なり，コンテキストに依存しながら，動的に形成されるものを指し，フレーズ成分によるグルーピングを形成要因の一つと考えるならば，韻律的特徴抽

[12] 即ち,「音韻レベル」⇔「単語レベル」と言うこと。

出/照合処理部との間に情報の交換が行なわれるべくモデル化が必要である。次に辞書構造的要因に対しては，

4. 長期的頻度 (結果 15)
5. 短期的頻度 (結果 17)

が考察されているが，両者について第 5.2.4 節において既にモデル化されており，ここで再度議論する必要は無い。統語的要因，意味的要因，談話的要因に関しては次に述べる言語処理部との兼ね合いもあるが，以下の結果/知見は十分に考慮される必要がある。

6. 意味的関連性の有る先行コンテキストの作用 (結果 23)
7. 現実世界 (常識) との整合性の高い文脈の及ぼす作用 (結果 25)
8. 統語的整合性の作用<意味的/談話的整合性の作用 (結果 26)
9. 単語網を用いた音声処理と left-to-right 処理との相互作用 (知見 32)

ここで，6., 7. は (ある特定の) 言語的整合性の有無 (高低) に対する議論である。一方 8. は言語的情報を 3 つに分けて考えた場合 (統語・意味・談話) の，各情報が及ぼす作用の優先度 (即ち比較) に対する興味深い知見であり，モデルの中に明確に示されるべきである。更に談話的整合性の高低は 9. で言うように，より大きな範囲で音声を捉える処理と left-to-right 処理に対する "舵取り" 的役割を果たしており，第 5.2.3 節で述べた音響的照合結果を格納する STM のモデル化と関連させつつ，モデルの中に表記されるべきであろう。また，第 5.2.4 節に述べたように，内部辞書は LTM であるため，辞書に対する処理手法の変動は全て，この内部辞書検索部によって記述されなければならず，その結果必然的に，本処理部は非常に動的な特性を持つことになる。以上のことを考慮してモデル化したものを図 5.5 に示す。

5.2.6 言語処理部

言語処理は大きく統語解析・意味解析・談話解析に分れる。辞書検索制御に関して，意味解析結果の統語解析結果に対する優位性が述べられている (知見 26) が，これらは言語処理部そのもので実現されるよりも，上記内部辞書検索処理部あるいは言語処理結果を受け付け，辞書検索制御指令を生成する処理部でモデル化されるべきものである。音声言語における言語処理は，音韻性の曖昧さを除けば文字言語における言語処理に対する研究が参考になる[103]。そこで言語処理部に関しては，文字言語に対する知覚実験結果から得られている言語処理モデルを参考にする。また，一般に統語/意味/談話解析の順でより高次な処理となり，その処理量も増すと言われている。このような特性もモデルに組入れるべ

きである[13]。なお，各解析を行なう場合，言語に依存した統語ルール，注目する人間が置かれている社会通念，常識などのデータベースを参照することが必要となる。即ち言語処理部は，統語的/意味的/談話的整合性を見る照合処理部と考えることもできる。ここで当然のことながら，各解析部は対象とするデータベース，及びデータベース間とのインターフェイスを持つことになる。また，言語処理結果に基付いた辞書検索部の制御を作成する処理部も言語処理部とは別にモデルに含める必要があるであろう。以上の考察より，本処理部をモデル化したものを図 5.6 に示す。図では左から右へと時間の流れに沿って統語/意味/談話解析が行なわれる様子及び，辞書検索制御生成においては，より高次の処理結果に基付く制御が優先的に扱われる様子をモデル化している。

5.2.7 総括

以上，各処理部に対する考察とモデル化を行なった。本節はそれらの処理部を有機的に結合することで，人間の音声知覚の全体像をモデル化する。図 5.7 に構築された音声知覚モデルを示す。並列にかつ時間差を伴う処理を 2 次元的に表現する必要があるため，処理の時間的側面が十分に表現されているところ/されていないところが見受けられる。その結果，統合したことにより，逆に分かりづらいモデルとなった感も否めない。しかし，個々の処理部のモデル化に対する考察 (第 5.2.2 節，第 5.2.3 節，第 5.2.4 節，第 5.2.5 節，第 5.2.6 節) を考慮した上で眺めて頂ければ，筆者の意図するところが汲み取れるものと確信している。但し，各節で述べた事柄が全て組み込まれている訳ではないことを予め断っておく。これは全てを組み込むことが逆に，モデルを理解困難なものへとしてしまうと考えたからである。この意味でも，各処理部の説明文とモデルを同時に眺めて戴きたい。

[13] 特に各処理の時間的側面による比較。

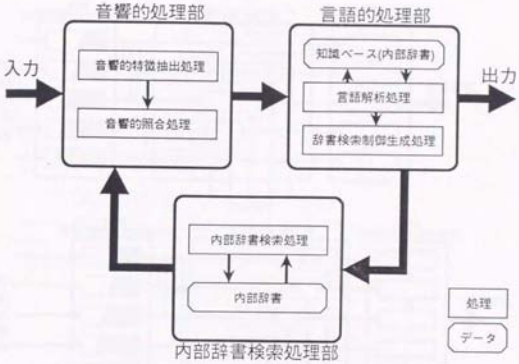


図 5.1. 人間の音声知覚モデルの概念図

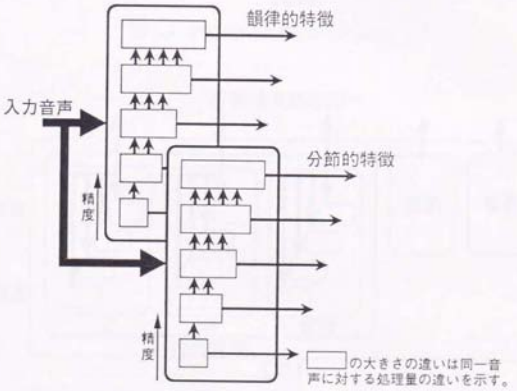


図 5.2. 音響的特徴抽出処理部のモデル化

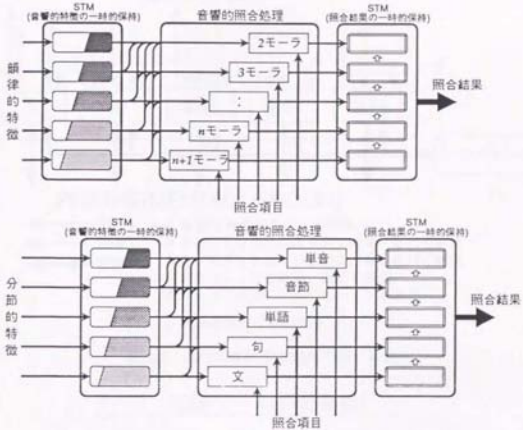


図 5.3. 音響的照合処理部のモデル化

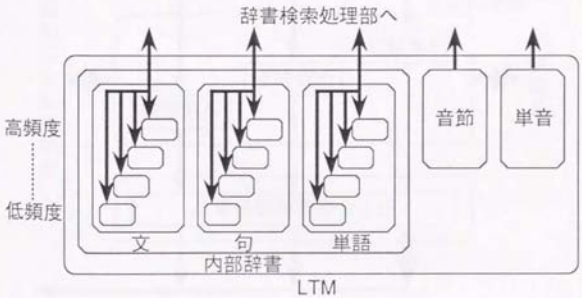


図 5.4. 内部辞書 (心的辞書, Mental Lexicon) のモデル化

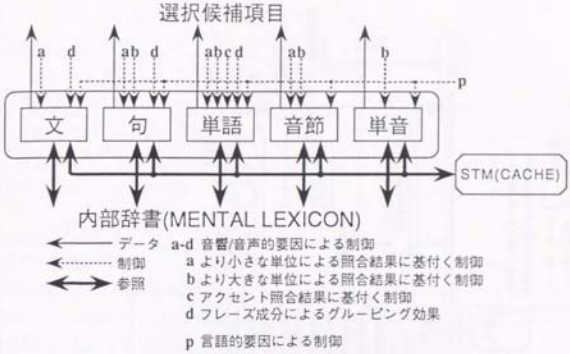


図 5.5. 内部辞書検索処理部のモデル化

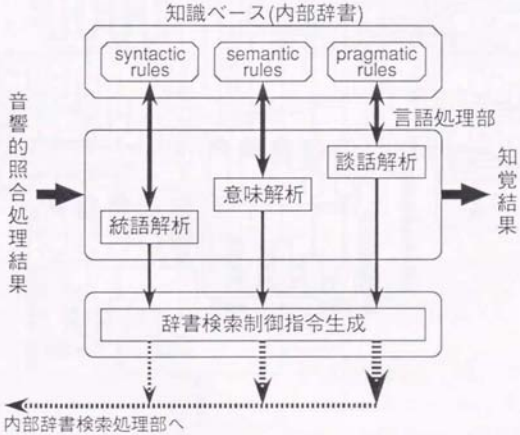


図 5.6. 言語処理部のモデル化

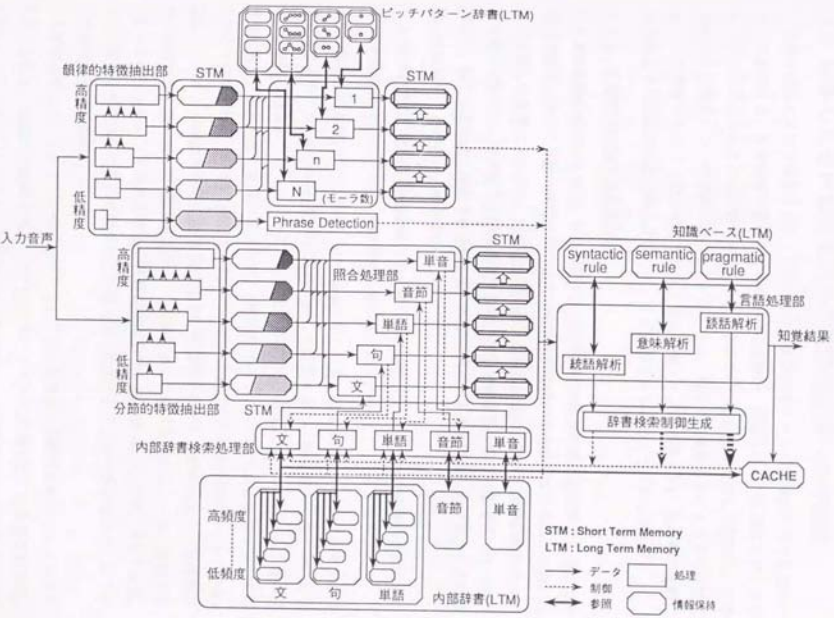


図 5.7. 人間の音声知覚過程のモデル

5.3 構築した音声知覚モデルの工学的応用への可能性

本節では第 5.2 節で構築された，人間における音声知覚モデルの工学的応用の可能性について考察する。本章冒頭で述べたように現在の段階で，直接的に，この知覚モデル全体の工学的実現を図ることは困難である。そこで本節では，各処理部に対して個別に，工学的応用の可能性について考察することにする。なお，本節での議論はあくまでも可能性についての議論であり，『十分に可能である』との考察をした処理部の内，後の章で実際に計算機上での認識手法に反映させるのは，その一部であることを予め断っておく。

5.3.1 音響的特徴抽出処理部

第 5.2.3 節に述べたように，本処理部では入力音声から複数の周波数分解能を持った分節的特徴を，そして複数の時間分解能を持った韻律的特徴の抽出が行なわれる。なお，以下では話しを簡単にするため，分節的特徴に絞った議論を進め，後ほど韻律的特徴についても簡単に触れる。音声を工学的に分析する際，波形を直接扱う代りに短時間パワースペクトル，自己相関関数，線形予測係数など，パワースペクトルに関連したパラメータに変換して取扱うことが多い。その理由として，

- 音声波形は，振幅と位相が時間的に緩やかに変動する正弦波の和で構成される。
- 音韻情報は主にスペクトルの振幅情報に含まれている。

などが挙げられる[1]。この音声を，500[msec] 程の区間で観測すると非定常プロセスの特徴を呈するが，10〜50[msec] 程の区間に区分して観測すると，各区間内では十分に定常信号と仮定し得るようになる。そこで音声信号処理では一般に，音声を 10[msec] 程の区間で区切り，各区間毎の特徴パラメータ (一般的にはパワースペクトルに処理を施して得られる十数次元のベクトル) を抽出し，時系列の形に変換したものを分析対象としている。但し，スペクトルの時間方向の変化 (動的特性) を追跡するため，分析区間を一部重複させたものがよく用いられる (図 5.8 参照)。この切り出した音声区間をフレーム，区間長をフレーム長，フレームを移動させる移動時間長をシフト長と呼ぶ。ここで，フレーム長，シフト長はそれぞれ研究目的に合わせて任意の値に設定することができる訳だが，以下のような問題点がある[105]。

周波数分析によりホルマントやピッチを正しく求めるには，分析のためのフレーム長を大きく設定し，周波数分解能を高くしてパワースペクトルの詳細な形を知る必要がある。しかし，フレームに対して求まる特徴量[14]は，そのフレーム区間内における時間方向の動

[14] 上記したように，"パワースペクトルに関する十数次元のベクトル/フレーム" である。

的変動を平均化したものである。上記したように音声は本来非定常なプロセスであり，注目する音声によっては，スペクトルの時間方向での詳細な変動をも捉える必要が生じる。従ってフレーム長をあまり長くすることは必ずしも適当ではない。しかしその一方で，ある程度の周波数分解能が要求される。

例えば，フレーム長を数 [msec] 程度に小さくした場合，周波数分解能は当然落ちるが，分節的特徴を支配するスペクトル包絡はこの場合でも一応求まり，破裂音など音声のスペクトルの急激な変化を捉えることはできる (時間分解能の上昇)。しかし，得られるスペクトル面は時間方向に対してかなり乱れたものとなってしまう。これはスペクトル分析窓が短時間すぎるために，ピッチに対応した変動や音声の生成過程に含まれるゆらぎの影響を直接受けるためである。一方，フレーム長を数十 [msec] 程度に長くした場合，周波数分解能は上がり，スペクトルの詳細な情報が得られる。しかし，逆にスペクトルの急激な変化には追従できず (時間分解能の低下)，破裂音等は前後の母音の影響をかなり受けることになる。このように，フレーム長をどのように設定しても，時間方向及び周波数方向の両方向に対して，十分に滑らかでかつ十分に分解能があるスペクトル包絡を抽出することは困難になる。

このような状態を打開するためには，複数の分解能の音響的特徴を抽出し，時間と共に変動する音声の特性に動的に，かつ適切に対応しながら，着目すべき特徴を取捨選択していく必要があると考えられる。そしてこれは，本章で構築した音声知覚モデルにおける音響的特徴抽出処理部に良く合致するものである。但し，上記した動的かつ適切な制御は本処理部での範疇ではなく[15]，ここでは，複数精度の音響的特徴を出力することのみが要求される。そして，この要求そのものは，現在の技術でも容易に実現可能である。即ち，

1. 複数のフレーム長，シフト長の音響的特徴を抽出する[104]。
2. 2 次元ケプストラム[105][106]を抽出し，照合処理部において，着目すべき (2 次元ケプストラム内の) 領域を変動させる。
3. LPC 或はケプストラム係数の次数の増減による，(推定) パワースペクトル包絡の詳細化/平滑化を利用する[38]。

など，種々の方法が考えられる。これらは当然のことながら，照合手法に依存して決定されるべきものである。

以上は，分節的特徴を前提としての議論であったが，同様なことが韻律的特徴 (主に F_0) についても言える。即ち，一般的には F_0 の抽出においても同様に窓関数を掛けることで，

[15] 音響的照合処理部において実現されるべき機能である。

音声区間を限定し，その区間内の F_0 を求めることが多い。あるいは，窓長に依存しない抽出を行なった場合でも，F_0 が抽出される周期はある一定の間隔 (上述したシフト長に相当する) となる。この間隔が細かい (当然フレーム長も細かくなる) ほど，時間分解能は高く，局所的な F_0 の動きをも検知することができる。逆に，この時間間隔が大きい場合は，大局的な F_0 の変動のみが観測可能であり，より大きな単位での照合処理においてのみ使用されるべき特徴量となる。そして，これらの精度のコントロールも現段階で可能な処理である。但し，これは分節的/韻律的特徴に限らず言えることであるが，照合処理のどの単位はどの精度の特徴を要求するのかと言った定量的な工学的実現は，図 5.7 を参照するだけでは行なうことができず，本節の議論は，あくまで定性的な工学的実現に留まっていることを記しておく。同様に，他の処理部に対する考察においても，その多くが定性的な工学的実現に関しての議論となっている。

5.3.2 音響的照合処理部

本処理部の目指すところは，図 5.3 にあるように，複数の処理単位 (音韻/単語/句/文) による分節的/韻律的照合処理である。この照合処理単位は，そのまま内部辞書の辞書項目サイズに対応するものであり，内部辞書との考察を含めて議論を進める必要がある。なお，この節でもまず分節的特徴について考察し，韻律的特徴については後述することにする。さて，音声認識の分野における音響的照合単位は，認識対象とする音声長に対応しつつ，以下のような歴史的変遷を経ている。

- 孤立発声単語音声認識が主要な研究対象となっていた時期では，音響的標準パターン (音響モデル) として，“音韻 (節)” 単位，“単語” 単位のものが利用されていた。前者の場合，標準パターンと入力単語音声とを照合する際に，音素 (節) 系列で表記された (テキスト表記の) 単語辞書を参照しながら，音韻 (節) の音響モデルを連結することで，疑似的に単語の音響モデル (標準パターン) を作成して照合を行なう。この方式は音響モデルとして蓄積しておくべきパターン数が認識対象語句数に依らず，音韻 (節) 数に限られると言う利点を持つ一方，音韻 (節) 間の調音結合の影響を十分に記述できないとの欠点を持つ。後者の単語の音響モデルを持つ方式の場合，明らかに調音結合を考慮した (含有した) 音響モデルが作成されるが，認識対象語句の増大に伴い，必然的に蓄積パターンも増え，大語彙の認識システムには不向きであると言う欠点を持つ。
- 大規模単語音声や連続音声を対象とした研究が行なわれるようになると，単語単位の音響モデルを標準パターンとする手法は，蓄積モデル数の爆発的な増大を生み，現在では

殆どが音韻 (節) を単位とした音響モデルが用いられている。この場合，上記の調音結合などが問題となってくるが，第 2.2 節で述べた数理統計的手法によって，調音結合による音響的な "揺れ" もある程度記述できるようになり，また，音素の定常部分で切り出し，CV(Consonant-Vowel) や VC(Vowel-Consonant) を単位とする (半音節と呼ばれる) ことで調音結合の影響を低減させる手法[107][108] なども提案されている。

このように，現在の音声認識においては，使用される音響的標準パターンはその大部分が音素 (あるいはそれに類する単位) であり，テキスト表記の単語辞書を参照し，認識時に音響モデルを連結することで疑似単語モデルを作成している。以上の「単語か音素 (節) か」と言う議論は知覚モデルにおける内部辞書に対しても行なうことができる。即ち，単語辞書の各項目に記述される分節的特徴は，音韻 (節) 辞書の分節的特徴へのポインタとして実現されているのか，あるいは，単語辞書の各項目が，項目固有の特徴としての分節的特徴の記述を持っているのか，と言うことである。しかし，第 5.2.4 節 における内部辞書のモデル化においては上記の議論は行なわれていない。また，第 3 章，第 4 章を見ても，この議論に対して，ある方向性を示すだけの結果を出している実験は無い。即ち，上記 2 通りの音響モデルの作成法に対して，本研究で実施した知覚実験よりその是非を決定することはできない。しかしこの場合においても工学的応用を重視するなら，音韻 (節) 単位による音響的モデル化を採用し，音韻/単語/句/文単位での照合処理は，疑似的に行なわせるべきであろう[16]。そして，この単位の大/小と分節的特徴精度の低/高を対応させれば図 5.3 は工学的に実現される。

一方，韻律的特徴 (F_0) に着目してみる。この場合，分節的特徴と大きく異なり，(アクセント) 辞書サイズは極端に小さくなる。第 4.5 節で示した "High/Low" の 2 値表記をそのまま採用する場合，n モーラ (孤立) 単語辞書は n+1 個のエントリのみを要求する。その結果，10 モーラまでの単語辞書を考えた場合でも 65 エントリ分の容量が必要となるだけである[17]。即ち，F_0 パターンの照合に関する限り，各単語長の辞書項目各々に，項目固有の特徴としての F_0 パターンの記述があると考えてよいであろう。そしてこのようなアクセント辞書の作成は困難ではない。さて，アクセントパターン照合処理部であるが，アクセント辞書の項目サイズ数だけの照合単位を用意し (モーラ数別の処理)，そして，照合単位の大/小と韻律的特徴の精度の低/高を対応させた照合処理を行なわせれば，図 5.3 に示す照合機構の工学的実現は可能である。

[16] 現在の計算機能力に依存する部分である。
[17] 但し，音声生成まで考慮すると，連結規則などのルールベースを蓄積するための容量も必要となる。

5.3.3 内部辞書

第 5.3.2 節で述べたように，分節的特徴 (音韻情報) に対する，音韻/単語/句/文サイズの内部辞書項目に対しては，音韻サイズの辞書のみが分節的特徴そのものの記述を持ち，他サイズの辞書は音韻サイズの辞書へのポインタを持つとの考察をした。一方，韻律的特徴に対するアクセント辞書に関しては，各々のサイズの内部辞書項目が，F_0 パターンの記述そのものを持つとの考察をした。後者の辞書に記述されてある特徴は，F_0 パターンだけであり，その工学的実現は困難ではない。しかし，前者の辞書には分節的特徴の表記の他に，各項目の意味，統語的役割，他項目間の意味的関連性など種々の情報が記載されていると考えられる。第 4.2 節における長期的頻度に関しては辞書項目の配列順序を通して実現可能であるが，上記の言語的属性の記述方式については，何ら直接的な実験結果/知見が無く，現在の段階で工学的実現を図るのは，あまりにも "仮定" に頼る部分が多いように思われる。但し，短期的頻度の作用に関してモデル化した cache 的 STM の工学的実現はそれほど困難ではない。認識結果をある特定のメモリにおいて保存し，次入力の音声に対して優先的に検索されるよう設定すればよい。但しこの場合，STM の時定数的な性質，即ち「どの程度の時間を過ぎると cache 的 STM 内の情報は消滅するのか」更に「時間に依存するのか，後続して認識される語数に依存するのか」などの議論が十分に行なわれておらず，ここでも定性的な工学的実現に留まらざるを得ない。

5.3.4 言語処理部

第 5.2.6 節で述べたように，言語処理を工学的に実現する場合，統語解析・意味解析・談話解析の各解析部を，知覚実験結果/知見から生じる要求を満足させながら，実現する必要がある。各解析結果の及ぼす影響は，その結果を利用する処理部で実現されるべきであり，ここでは純粋に上記 3 種の観点からの解析を行なえばよい。しかし，音韻性の曖昧さが無い文字言語における言語処理においても，現状では,「係り受け」や「指示代名詞の適切な補間」など (即ち統語解析) が主要な研究対象となっており，十分な機能を実現しているとは言えない[18]。意味解析に関しても，有意味/無意味と言った 2 値の世界から抜け出ることが出来ないように思われる。また，電子化辞書作成における，意味属性の設定及び項目間の関係の記述が非常に困難を究めたこと，数回に渡って設定の改訂が行なわれたことなどを考慮すると，単語間の意味的相関を統一的に，適切に記述する方法は，現段階ではまだ編み出されていないようである。人間の行なう連想を模擬した認識シス

[18] 自動翻訳機にかけた英文を見れば一目瞭然である。

テムの考案も実験的に行なわれている[109]が，連想の特性は静的なものと仮定し，後で述べる辞書検索処理部の動的側面は無視した形となっている。談話解析に目を向けた場合，現在試作されている音声応答システムが良い考察対象となる。即ち，現段階での音声応答システムは，「スキー場案内システム[110]」，「ファーストフード店における注文応答[111]」，「大学近辺の案内[112]」と言ったように，対象とする主題が限定されているものが殆どである。これは，十分に人間の自由な会話を談話解析するだけの処理手法が確立されていないことを示唆する。このように現在の技術では図 5.6 に示したモデルを十分に実現することは非常に困難であり，逆にこのため，入力となる音声に統語的/意味的/談話的な制約を課すことが必要となってくる。上記のような現状と，筆者自身言語処理部への知識が不十分であることを考慮し，図 5.7 における言語処理部の工学的実現に関する議論は，本論文ではこれ以上行なわないことにする。

5.3.5 内部辞書検索処理部

内部辞書検索処理は，第 5.2.5 節で述べたように，分節的/韻律的照合結果及び言語処理結果を入力とし，前後コンテキストに動的に適応しつつ，内部辞書を参照/検索して音響的照合処理部へと結果を出力する。第 5.3.3 節及び第 5.3.4 節で考察したように，構築した知覚モデルにおける内部辞書と言語処理部の工学的応用は非常に困難を究める。その結果，本処理部の工学的実現に関してもかなりの制約が伴うことになる。即ち第 3.6.2 節で掲げた要因のうち，統語的要因，意味的要因，談話的要因らについてはその議論を見送らざるを得ない。音響/音声学的要因に関しては，1) 種々の単位長の音響的照合処理の相互作用，2) 先行する F_0 パターン照合結果に基付く，辞書検索範囲の絞り込み[113]，などが実験的考察の可能な項目として挙げられる。辞書構造的要因に対しては以下のように考察できる。まず，長期的頻度が及ぼす影響は内部辞書項目の配列順序として実現できる。即ち，各辞書項目に対して，正しく評価された長期的頻度 (親密度) が定義されれば，容易に実現可能となる。この親密度の定義に対しては，第 3.3.2 節で紹介した研究が今後充実・発展すれば不可能ではない。次に，短期的頻度に関しては第 5.3.3 節に示したように，仕組みそのものは cache 的 STM として実現可能である。しかし，その定量的な設定は最適値を実験的に求める必要がある。

以上，図 5.7 に示した音声知覚モデルの工学的実現可能性について考察してきた。その結果，音響的特徴抽出部・音響的照合処理部・内部辞書に関する知見の一部に限って，計算機上での実現可能性が明らかとなった。第 7 章で，代表的な音声認識手法 (DP& HMM) を考察するが，そこで本章での考察を導入した音声処理手法を再度検討することにする。

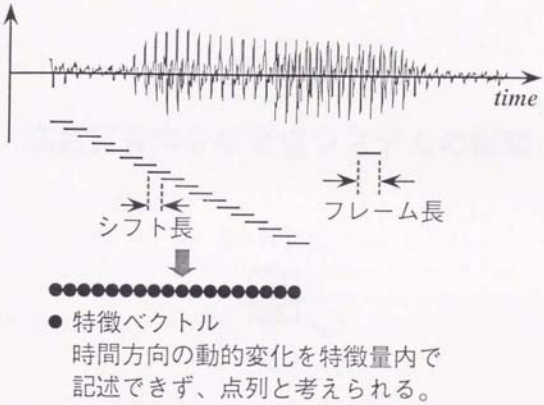


図 5.8. 音声信号の音響的分析方法

第 6 章

高品質音声分析合成システムの構築

本章では，第 4.10 節で述べた知覚実験において，音声試料作成用に構築した音声分析合成システムについて述べる。従来より，分析合成系において用いられる音源波形としては，有声部分に対してはパルス列 (或は，それに類する周期波形，即ち三角波等) を，無声部分に対して白色雑音を利用するものが多い。しかし，破裂音などは 1 パルスを音源として生成される (無声) 子音であり，これらの子音に対しても白色雑音を音源波形として使用している限りは，その品質にも限界があるのは明白である。そこで本システムでは，任意の対数パワースペクトルを，その伝達関数として実現可能な LMA フィルタを用いることにより，自然音声に近い品質が再現されるべく，最適な音源波形の生成を試みる。また，パワースペクトル包絡特性を近似して音声合成フィルタを構成する際にも，この LMA フィルタを使用する。以下本章ではまず，分析合成技術について簡単に紹介し，次に，LMA フィルタについて概説する[93][94]。その後に，提案する音源波形生成法と構築したシステムについて述べる。最後に，簡単にではあるが，本システムによって作成された合成音に対する評価実験も行なったのでその結果についても紹介する。なお，付録 A に本システムのマニュアルを載せておく。随時そちらも参照して頂きたい。

6.1 音声の分析合成

音声の合成・認識・分析の分野に関わらず，音声情報処理の基礎 (大前提) となっているのが，音声生成機構の線形分離等価回路モデルと言われるものである。このモデルは音源 $G(\omega)$ と調音 (共振・反共振特性)$H(\omega)$ を完全に分離し，これらが縦続に接続された形で音声波形 $S(\omega)$ が生成されるとするモデルである (図 6.1 参照)。

$$S(\omega) = G(\omega) \cdot H(\omega)$$

即ち，$G(\omega)$ を周波数特性として持つ音源波形が，$H(\omega)$ を周波数特性として持つ声道フィルタを通して，$S(\omega)$ を周波数特性として持つ音声波形が得られる，と言う訳である。生理学的な分析結果によれば，音源波形のマクロ的な周波数特性 (包絡特性) は一般に-12[dB/oct] であり，唇からの放射伝達特性は+6[dB/oct] であると言われている。上述したように音源波形としては従来より，有声部分に対してはパルス列が，無声部分に対しては白色雑音が使用されているが，当然のことながら，これらの周波数包絡特性は平坦である。即ち，上式の線形分離モデルを使用する場合，声道特性 $H(\omega)$ に音源スペクトルの包絡特性と放射特性を含めた形で[1]モデル化することが多い。

[1] 即ち，その包絡特性が-6[dB/oct] となるように。

この線形分離モデルに基付いて音声を分析し，音源と声道特性 ($G(\omega)$ と $H(\omega)$) に関する特徴パラメータを抽出 (操作 A) し，これらの特徴パラメータを用いて音声を再合成する (操作 B) ことを，音声の分析合成と言う。この場合，概念的には $\mathbf{B}=\mathbf{A}^{-1}$ の関係があると考えてよい。当然のことながら再合成前に，パラメータレベルで特徴量を操作すれば種々の合成音声が得られることになる。さて，ここで抽出されるパラメータであるが，次の 4 種類に集約される。

1. 有声音/無声音の区別 (有声度)
2. 有声音の場合の基本周期 (F_0)
3. 音源の振幅
4. 声道フィルタの共振特性

上記 3 つが音源情報であり，残りの 1 つがスペクトル包絡 (音韻) 情報である。

6.2 LMA(Log Magnitude Approximation) フィルタ

6.2.1 指数関数形の伝達関数を持つフィルタの特性

フィルタを用いて音声や楽器音を合成する場合，合成音の品質はフィルタの極だけでなく，零点にも大きく影響されるため，フィルタの振幅特性は線形目盛や 2 乗目盛より，対数的な目盛で近似するのが望ましい。ある，ディジタルフィルタの伝達特性 $H(z)$ が基礎フィルタと名付けるディジタルフィルタの伝達関数 $F(z)$ によって，

$$H(z) = \exp(F(z)) \tag{6.1}$$

のように表されるものとする。式 (6.1) からこのフィルタの対数周波数特性 $\ln H(e^{j\Omega})$(但しΩは規格化角周波数) は，

$$\ln H(e^{j\Omega}) = F(e^{j\Omega}) \tag{6.2}$$

となる。従って，対数振幅特性 $\ln|H(e^{j\Omega})|$ 及び位相特性 $\arg H(e^{j\Omega})$ は各々，

$$\ln|H(e^{j\Omega})| = \mathrm{Re}\left\{F(e^{j\Omega})\right\} \tag{6.3}$$

$$\arg H(e^{j\Omega}) = \mathrm{Im}\left\{F(e^{j\Omega})\right\} \tag{6.4}$$

で与えられる。ここで，Re, Im は各々実部と虚部をとる演算子である。また，基礎フィルタの伝達関数 $F(z)$ が，

$$F(z) = \sum_{m=0}^{M} F_m(z) \tag{6.5}$$

のように和で表されるとき，指数関数形伝達関数を持つフィルタ $H(z)$ は，

$$H(z) = \exp\left(\sum_{m=0}^{M} F_m(z)\right) = \prod_{m=0}^{M} \exp(F_m(z)) \tag{6.6}$$

となる。即ち，指数関数形伝達関数を持つ幾つかのフィルタの縦続接続は，それらに用いられている基礎フィルタの並列接続の指数関数形に対応する。式 (6.5) のように伝達関数が和の形で表されるものとして，

$$F(z) = \sum_{m=0}^{M} a_m z^{-m} \tag{6.7}$$

のような FIR フィルタを考えた場合，これを基礎フィルタとする指数関数形伝達関数 $H(z)$ の特性は各々，

$$\ln|H(e^{j\Omega})| = \sum_{m=0}^{M} a_m \cos(m\Omega) \tag{6.8}$$

$$\arg H(e^{j\Omega}) = -\sum_{m=0}^{M} a_m \sin(m\Omega) \tag{6.9}$$

で与えられる。式 (6.8) 或は式 (6.9) で，a_m を希望する対数振幅特性のフーリエ余弦係数，或は位相特性のフーリエ正弦係数とすれば，フーリエ係数の性質から，フィルタ $H(z)$ の特性は希望する対数振幅特性あるいは位相特性を 2 乗平均誤差最小の意味で最良に近似するものとなる。即ち，所望のパワースペクトル包絡のフーリエ係数 (ケプストラム) が与えられた時，式 (6.7) を基礎フィルタとする指数関数形伝達関数 (式 (6.1)) がディジタルフィルタとして実現できれば，任意の対数パワースペクトル包絡に対するフィルタが構成されることになる。

6.2.2 指数関数の修正 Padé 近似

本節では，与えられた特性の指数関数形の伝達特性の近似法について述べる。詳しいことは参考文献[93] に譲り，ここでは実際の近似式及び近似条件について記す。複素関数 $\exp(\omega)$ の対数目盛におけるミニマックス近似，即ち修正 Padé 近似式 $\tilde{R}^{(L)}(\omega)(L=1,2,3)$[2] は，

$$\tilde{R}^{(L)}(\omega) = \frac{1 + \sum_{l=1}^{L} \frac{\tilde{A}_l^{(L)}}{l!} w^l \frac{\binom{L}{l}}{\binom{2L}{l}}}{1 + \sum_{l=1}^{L} \frac{\tilde{A}_l^{(L)}}{l!} (-w)^l \frac{\binom{L}{l}}{\binom{2L}{l}}} \tag{6.10}$$

[2] Padé 近似式 $R^{(L)}(\omega)$ は線形目盛による近似。式 (6.10) で ~ を取り除いたもの。

$$\approx \exp(\omega) \qquad (L = 1, 2, 3; |\omega| \leq \bar{r}_L)$$

で与えられる。ここで，係数 $A_i^{(L)}$は対象とする変数ωの絶対値 $|\omega|$ の最大値$\bar{r}_L$によって決まり，通常 1.0 に近い値をとる。参考文献 [93] によれば，$\tilde{A}_i^{(L)}$(L=1,2,3) は，

$$\tilde{A}_1^{(1)} = 1 - 0.013\bar{r}_1{}^4 \tag{6.11}$$

$$\tilde{A}_1^{(2)} = 1 - 0.44 \times 10^{-5}\bar{r}_2{}^8 \tag{6.12}$$

$$\tilde{A}_2^{(2)} = 1 - 0.11 \times 10^{-2}\bar{r}_2{}^4 \tag{6.13}$$

$$\tilde{A}_1^{(3)} = 1 - 0.40 \times 10^{-9}\bar{r}_3{}^{12} \tag{6.14}$$

$$\tilde{A}_2^{(3)} = 1 - 0.50 \times 10^{-4}\bar{r}_3{}^4 - 0.14 \times 10^{-6}\bar{r}_3{}^8 \tag{6.15}$$

$$\tilde{A}_3^{(3)} = 1 - 0.30 \times 10^{-3}\bar{r}_3{}^4 \tag{6.16}$$

$$(\bar{r}_1 < W_1 = 2.000, \quad \bar{r}_2 < W_2 = 3.464, \quad \bar{r}_3 < W_3 = 4.644) \tag{6.17}$$

となる。ただし，上記の W_L(L=1,2,3) は有理関数 $R^{(L)}(\omega)$(Pade 近似式) が零点も極も持たないような変数の範囲の大きさである。これは，次式で求められる対数目盛における誤差 $E^{(L)}(\omega)$ を基準にして算出される。

$$\begin{aligned} E^{(L)}(\omega) &= \ln R^{(L)}(\omega) - \ln(\exp(\omega)) \\ &= \ln R^{(L)}(\omega) - \omega \end{aligned} \tag{6.18}$$

更に $E^{(L)}(\omega)$ に対する近似値も次式で求められる。

$$\begin{aligned} \tilde{E}^{(L)}(\omega) &= E^{(L)}(\omega)|_{R^{(L)}(\omega)=\tilde{R}^{(L)}(\omega)} \\ &\approx (-1)^{(L+1)}\frac{2L!(L+1)!}{(2L)!(2(L+1))!}\omega^{2L+1} \end{aligned} \tag{6.19}$$

6.2.3 LMA フィルタの構成

さて，式 (6.1) を満たす $H(z)$ の近似フィルタ $\tilde{H}(z)$ を実現する場合，第 6.2.2 節より，基礎フィルタ $F(z)$ の修正 Padé 近似式を用いて

$$\tilde{H}(z) = \tilde{R}^{(L)}(F(z)) \tag{6.20}$$

として作成される。さてこの場合式 (6.17) に示した条件が満たされる必要がある。即ち，基礎フィルタの振幅特性は，W_L以下に抑えられなければならない。

$$|F(e^{j\Omega})| < W_L \tag{6.21}$$

そして，基礎フィルタの振幅特性の最大値

$$r_L = \max_{\Omega} |F(e^{j\Omega})| \tag{6.22}$$

が修正 Padé 近似式の対象とする変数ωの範囲の限界$\bar{r}_L$を越えないようにすれば，近似誤差は式 (6.19) より，

$$|\hat{E}^{(L)}(F(e^{j\Omega}))| \leq \frac{2L!(L+1)}{(2L)!(2(L+1))!}\bar{r}_L{}^{2L+1} \tag{6.23}$$

で抑えられる。従って，近似誤差の大きさをε以下にするためには，基礎フィルタの振幅特性の最大値r_Lが，

$$r_L \leq \bar{r}_L \leq \bar{r}_L{}^{\max} \approx \left(\frac{2L!(2(L+1))!\varepsilon}{2L!(L+1)!}\right)^{\frac{1}{2L+1}} \tag{6.24}$$

となるようにする必要がある。具体的には，ε=0.0115 以下にするには，r_Lは，

$$\bar{r}_L{}^{\max} \approx \begin{cases} 0.517 & (L=1) \\ 1.526 & (L=2) \\ 2.740 & (L=3) \end{cases} \tag{6.25}$$

を越えないようにしなければならず，ε=0.0230 の場合は，

$$\bar{r}_L{}^{\max} \approx \begin{cases} 0.651 & (L=1) \\ 1.753 & (L=2) \\ 3.025 & (L=3) \end{cases} \tag{6.26}$$

となる。

さて，基礎フィルタ $F(z)$ の振幅特性の最大値r_Lが明らかに$\bar{r}_L$を越えるような場合はどのようにすればよいのだろうか。ここで，式 (6.6) の特性が役に立つ。LMA フィルタは指数関数形伝達関数のフィルタを近似的に実現したものであり，式 (6.6) の関係も近似的に成立する。即ち，r_Lが$\bar{r}_L$を越えるような場合，$F(z)$ を和の形に分解し，各々の項による幾つかの LMA フィルタの縦続接続を考えることにより，近似誤差を許容範囲内に抑えることができる。この場合，当然のことながら，分解して作成された各項の振幅最大値は$\bar{r}_L$以下になるようにする。

必要な定数を全て導出したので，実際の LMA フィルタ ($\hat{R}^{(L)}(F(z))$) の基本構成を L=1,2,3 の場合について図 6.2 に示す。

6.2.4 LMA フィルタを用いた音声合成フィルタ

第 6.2.1 節，第 6.2.2 節，第 6.2.3 節で LMA フィルタの基本構成を述べた。この節では LMA フィルタを，音声分析合成における合成フィルタとして応用することを考える[94]。音声スペクトルの包絡を求める場合，音声の対数スペクトルの (逆) フーリエ余弦係数で定義されるケプストラム c_m に対して，リフタを用いてその高次係数 ($>M$:ケプストラムの打ち切り次数) を 0 にすることがしばしば行なわれる。この場合，ケプストラムの対称性を保存するために，c_{M+1}〜$c_{N-(M+1)}$ を 0 にすることになる (但し，N はフレーム長)。これは，時間軸を負の方向に拡張し，c_{-M}〜c_M(但し $c_{-m}=c_m$) の $2M+1$ 個以外の係数を 0 にすることと同値である。さて，伝達関数 $F(z)$ が上記 c_m を用いて

$$F(z)=\sum_{m=-M}^{M} c_m z^{-m} \tag{6.27}$$

で与えられる基礎フィルタを考えると，$H(z)(=\exp(F(z)))$ の対数振幅特性は，

$$\begin{aligned}\ln|H(e^{j\Omega})| &= \mathrm{Re}\{F(e^{j\Omega})\}\\ &= \sum_{m=-M}^{M} c_m \cos(m\Omega)\\ &= c_0+\sum_{m=1}^{M}(2c_m)\cos(m\Omega) \qquad (6.28)\\ &= \sum_{m=0}^{M} \tilde{c}_m \cos(m\Omega) \qquad (6.29)\end{aligned}$$

となる。但し，

$$\tilde{c}_m=\begin{cases} c_m & (m=0)\\ 2c_m & (m>0)\end{cases} \tag{6.30}$$

である。ここで式 (6.29) は，

$$F(z)=\sum_{m=0}^{M} \tilde{c}_m z^{-m} \tag{6.31}$$

を基礎フィルタとした場合の $H(z)$ の対数振幅特性でもある。即ち，式 (6.27) を基礎フィルタとする $H(z)$ の対数振幅特性は，式 (6.31) の FIR フィルタを基礎フィルタとすることで実現されることを意味する。以上をまとめると，「着目する音声区間からケプストラムを抽出し，式 (6.30) の変換を施した値を係数とする FIR フィルタ (式 (6.31)) を基礎フィルタとして LMA フィルタを構成すれば，その音声区間のスペクトル包絡を近似したディジタルフィルタが作成される。」ことになる。

実際の LMA フィルタ作成においては，改良ケプストラムを，以下に示すような時間方向での平滑化を行なったもの (平滑化ケプストラム) を上記 c_m として採用し，0 次~32 次までを使用した (M=32)。

$$c_m(t) = \gamma c_m(t-1) + (1-\gamma)C_m(t)$$

但し，$C_m(t)$ は時刻 t における m 次の改良ケプストラム係数，$c_m(t)$ は時刻 t における m 次の平滑化ケプストラムである。なお，γは実験的に 0.5 とした。また，式 (6.31) をそのまま，図 6.2 に適用すると，式 (6.24) で述べた条件を満たすことが困難になる。そこで，実際のコーディングにおいては，$F(z)$ を

$$\begin{aligned} F(z) &= \sum_{m=0}^{32} \tilde{c}_m z^{-m} \\ &= \tilde{c}_0 + \tilde{c}_1 z^{-1} + \tilde{c}_2 z^{-2} + \\ &\quad \sum_{m=3}^{4} \tilde{c}_m z^{-m} + \sum_{m=5}^{7} \tilde{c}_m z^{-m} + \sum_{m=8}^{10} \tilde{c}_m z^{-m} + \\ &\quad \sum_{m=11}^{15} \tilde{c}_m z^{-m} + \sum_{m=16}^{20} \tilde{c}_m z^{-m} + \sum_{m=21}^{28} \tilde{c}_m z^{-m} + \sum_{m=29}^{32} \tilde{c}_m z^{-m} \end{aligned} \tag{6.32}$$

と分割して捉え，

$$\begin{aligned} H(z) = \exp(\tilde{c}_0) &\times \tilde{R}^{(3)}(\tilde{c}_1 z^{-1}) \times \tilde{R}^{(3)}(\tilde{c}_2 z^{-2}) \times \tilde{R}^{(3)}\left(\sum_{m=3}^{4} \tilde{c}_m z^{-m}\right) \times \\ &\tilde{R}^{(2)}\left(\sum_{m=5}^{7} \tilde{c}_m z^{-m}\right) \times \tilde{R}^{(2)}\left(\sum_{m=8}^{10} \tilde{c}_m z^{-m}\right) \times \tilde{R}^{(2)}\left(\sum_{m=11}^{15} \tilde{c}_m z^{-m}\right) \times \\ &\tilde{R}^{(2)}\left(\sum_{m=16}^{20} \tilde{c}_m z^{-m}\right) \times \tilde{R}^{(1)}\left(\sum_{m=21}^{28} \tilde{c}_m z^{-m}\right) \times \tilde{R}^{(1)}\left(\sum_{m=29}^{32} \tilde{c}_m z^{-m}\right) \end{aligned} \tag{6.33}$$

のようにして，個々の指数関数形伝達関数の振幅特性の最大値を抑えた上でフィルタを構成した。本来ならば，上記の分割は音声の特徴の変化に応じて，動的に決定されるべきであるが，本研究ではコーディングを容易にするため，上記の様に固定して分割を行なった。しかしこのために，特に女性音声に対する合成音の品質劣化がしばしば観測された。これは第 6.2.3 節で述べた条件が，上記の静的分割では必ずしも満たされないことに起因すると考えられる。動的な最適分割 (構成) は今後の課題の一つである。

6.2.5 LMA フィルタを用いた最適音源波形の生成

分析合成における再合成のプロセスにおいて，音源波形は，声道特性を近似するフィルタを通ることで合成音声となる。このことを考慮すると，声道特性近似フィルタの逆特性

を持ったフィルタ (以降，逆フィルタと記す) が構成可能であるならば，自然音声を，この逆フィルタに通すことで得られる波形が，声道特性近似フィルタにとって，最適な[3]音源波形と言うことになる (図 6.3 参照)。この，自然音声を逆フィルタに通して得られる波形をここでは，最適音源波形と呼ぶことにする。

第 6.2.4 節において，LMA フィルタを用いた声道特性近似フィルタが式 (6.33) で作成されることが示された。第 6.2.4 節では一貫して，LMA フィルタを自然音声の対数スペクトル包絡 (声道特性) を近似するためにのフィルタと言う観点から説明してきたが，当然のことながら，任意のケプストラムに対応する対数スペクトル包絡を近似することができる。さて，ケプストラムは対数スペクトルの (逆) フーリエ余弦係数として定義されるため，

$$d_m = -c_m$$

で定義される d_m をケプストラムと仮定して求めた対数パワースペクトルは，自然音声のパワースペクトルに対して逆特性を持ったものとなる。即ちこの d_m を LMA フィルタ (式 (6.33)) に適用すれば，逆特性を実現するディジタルフィルタが容易に得られることになる (図 6.3 参照)。

以上の考察の下，成人男性話者 1 名が約 8[mora/sec] で発声した,「来るようになりました」と言う音声に対する最適音源波形を求めたので，図 6.4 に示す。また，最適音源波形を声道特性近似フィルタに通して得られる再合成音波形についても図 6.6 に示す。各々の図の上側に示してあるのは，ソースとなる自然音声である。この文音声には，/k/, /s/, /t/と 3 種類の無声子音が含まれており，それ以外は母音あるいは有声子音と言われているものである。図 6.4 を見て分かるように，有声部分は非常に周期性の高い三角波として推定されているが，各々の無声子音に対応する最適音源波形は，非常に乱れたものとなる。図 6.4 において 0.8 秒付近，即ち/aʃ/付近の波形を拡大したものが，図 6.5 である。有声部から無声部になる際に，波形の周期性が急激に低下していることが分かる。また，3 種類の無声子音に対する最適音源波形は白色雑音とはかなり異なった波形を呈していることも分かる。従来の分析合成においては，このような無声破裂音に対しても白色雑音を利用しており，その場合の合成音の品質劣化がこの図からも伺える。また図 6.6 より，最適音源波形から得られる合成音は原音声である自然音声とほぼ同一波形であることが分かる。図 6.7 は図 6.5 と同じ位置の波形を拡大したものであるが，両波形の一致の度合がよく分かる。これは LMA フィルタにおける近似誤差が無視できる大きさであれば当然

[3] 自然音声と得られた分析合成音との 2 乗誤差最少の意味で。

の結果ではある。しかし，分析合成において使用される合成フィルタ (対数スペクトルを近似するためのフィルタ) に対する最適な音源波形が容易に得られると言うことは，種々のパラメータ操作の後に再合成する場合に，この最適音源波形 (の一部) を利用することで，より高品質の合成音が得られると考察される。

6.3 知覚実験用音声試料作成を目的とした音声分析合成システムの構築

第 6.1 節に述べたように分析合成において，再合成前に音源情報を担う音響的パラメータを操作することにより，同一音韻情報を伝達する種々の音声を合成することが可能となる。ここで，音源情報を操作する場合に第 6.2.5 節の考察にもあるように，最適音源波形の情報を組入れることで，合成音の品質向上が十分に期待できる。当然のことながら，音源情報を操作すると言うことは，最適音源波形をそのまま使うことが出来ないことを意味する。即ち，最適音源波形に関する情報の組入れ方は，音源情報の操作方法に依存することになる。さて,『人間による音声知覚過程の分析とそのモデル化』に関する研究において，韻律的特徴が音声知覚に果たす役割を実験的に検討するため，同一音声から F_0 パターンを操作した種々の合成音声を作成する必要が生じた。そこで，最適音源波形を利用した音声分析合成システムを構築することとした。F_0 の抽出，操作については付録 A に本システムのマニュアルを載せておくので，そちらを参照して頂きたい。本節では，本システムにおいてキーとなる，最適音源波形を利用した音源波形生成方法と，聴取実験による本システムに対する評価について述べる。

6.3.1 分析合成用の音源波形生成

上述したように本システムでは，再合成前に F_0 を操作する，即ち音源情報を操作する必要があるため，最適音源波形をそのまま全て使用することは出来ない。しかし，同一音声から F_0 パターンのみを変形した音声を合成すると言うことは，着目する音声中の有声部のみを操作することを意味する。即ち，本システムにおける音源情報の操作は，有声部分のみがその対象となり，無声音に対しては，そのまま最適音源波形を使用することが可能である。そこで，本システムにおける音源波形作成の基本方針として，

1. 自然音声から F_0 を抽出すると共に，有声度も推定する。
2. 有声部に対する音源波形は指定された F_0 の値に基付き，人工的に作成する。
3. 無声部に対する音源波形は最適音源波形をそのまま利用する。

4. 人工的に作成した有声音源波形と，最適音源波形からの無声音源波形の連結は以下に示すように，連結点を中心に，$2L+1$ サンプルを "重み付け" をした上で足し合わせ，連結する。

と言う方法を採用した。但し，各々以下の点に注意した。まず 1. の有声度であるが，推定された有声度に対し閾値θを設け，有声度がθ以上ならば有声部，θ以下ならば無声部とした。しかし，有声度の推定誤りを回避することを目的として，有声区間及び無声区間の最短時間を設定し，それ以下の時間長で有 (無) 声部が現れた場合は推定誤りとして，無 (有) 声部に修正した。次に 2. の人工的に作成する音源波形 (周期性波形) は，汎用信号処理ソフトウェア ESPS[4]において使用されている音源波形と同一のもの (関数) を用いた。図 6.8 に実際に使用した音源波形の例を示す。更に 4 における，有声部 (人工音源波形) と無声部 (最適音源波形) に対する重み付けであるが，具体的には以下のように行なった。例えば，有声部から無音部に $t=t_1$ で変化する場合，

$$h(t) = \alpha(t)f(t) + (1-\alpha(t))g(t)$$

のようにして両波形を足し合わせる。但し，$f(t)$ が有声部に対する人工的な音源波形 (周期性波形)，$g(t)$ が最適音源波形，$h(t)$ が本システムで採用する音源波形である。重み$\alpha(t)$ は，$\alpha(t \leq t_1 - L)=1.0$，$\alpha(t \geq t_1 + L)=0.0$ であり，$t_1 - L < t < t_1 + L$ に対しては，1.0 から 0.0 へ，時間に対して直線的に変化するよう設定した。無声部から有声部に変化する時も同様である。

6.3.2 聴取実験による評価

本節では，作成した分析合成システムの聴取実験による評価について述べる。上記した，韻律的特徴の果たす役割を検討する知覚実験において，文音声から文節単位で切り出した音声 (自然音声) と，それらに対して 2 通りの F_0 操作を行なって作成した合成音声の聞き取り実験が行なわれているので，その結果を表 6.1 に示す。被験者は成人男性 4 人であり，タスクは聴取後の口頭再生である。条件 A, B は F_0 の形態の違いを指し (各形態の詳細については第 4.10 節を参照)，C は自然音声を意味する。結果より各条件間での差は無い。故に，本実験で用いた分析合成音声試料は，韻律的特徴を操作した場合でも (A のように F_0 の起伏を全く取り除いた場合でも)，音韻的情報の伝達には支障を来たしておらず，F_0 パターンを操作した合成音を必要とする知覚実験の音声試料としては十分に高品質な音声が得られていると言える。

[4] アメリカ Entropic 社製。日本での代理店は (株) アルゴグラフィックス。

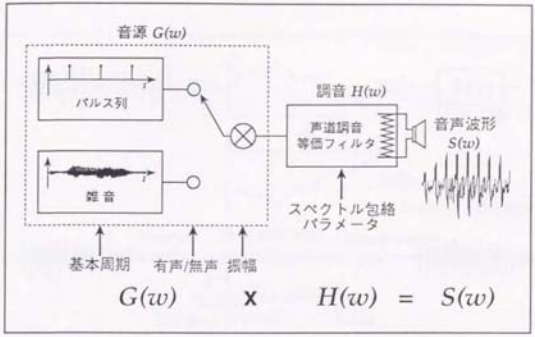


図 6.1. 音声生成機構の線形分離等価回路モデル

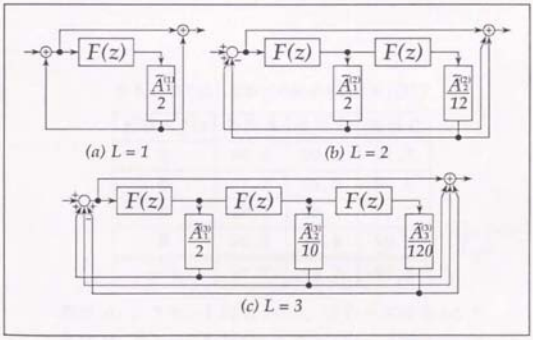


図 6.2. LMA フィルタの基本的な構成

(a), (b), (c) の順に，$\hat{R}^{(1)}(F(z))$, $\hat{R}^{(2)}(F(z))$, $\hat{R}^{(3)}(F(z))$ をディジタルフィルタとして実現したものである。

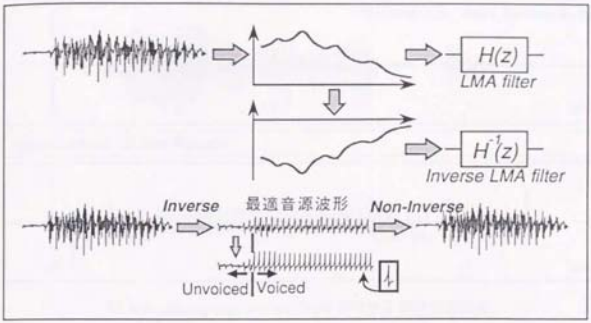


図 6.3. LMA フィルタを用いた音源波形の生成

表 6.1. 切り出し文節音声提示実験結果 (%)

被験者 ID	条件 A	条件 B	条件 C
4	96.6	97.7	97.7
5	97.7	98.9	98.9
6	98.3	99.4	98.3
8	98.3	98.9	98.3
平均	97.7	98.7	98.3

条件 A: アクセント指令=0.0, フレーズ指令=0.0
条件 B: アクセント指令=0.3, フレーズ指令=0.3
条件 C: 自然音声

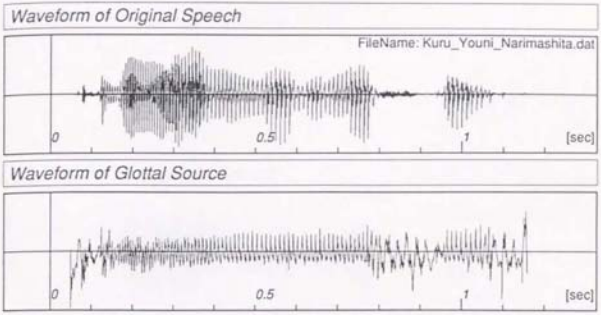


図 6.4. /kuruyooninarimaʃita/に対する最適音源波形

上が，原音声，下が LMA 逆フィルタに通して作成された最適音源波形である。

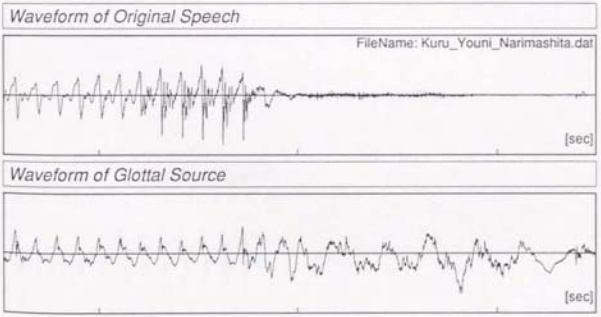


図 6.5. /aʃ/の部分に対する最適音源波形 (拡大)

図 6.4 における 0.8 秒付近の波形，/aʃ/を拡大したものである。

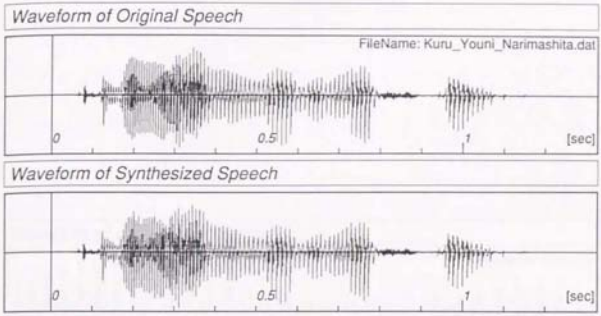


図 6.6. /kuruyooninarimaʃita/に対する最適音源波形を用いた再合成音上が，原音声，下が LMA 逆フィルタ→LMA フィルタに通して作成された再合成音である。

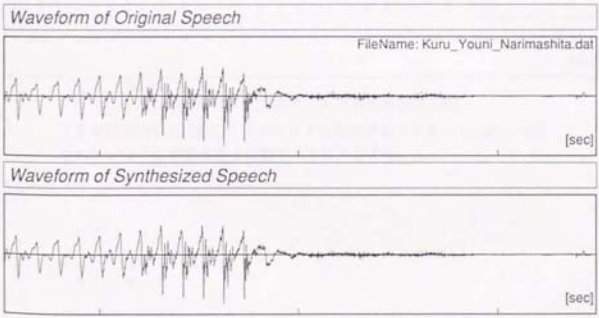


図 6.7. /aʃ/の部分に対する最適音源波形を用いた再合成音 (拡大)

図 6.6 における 0.8 秒付近の波形，/aʃ/を拡大したものである。

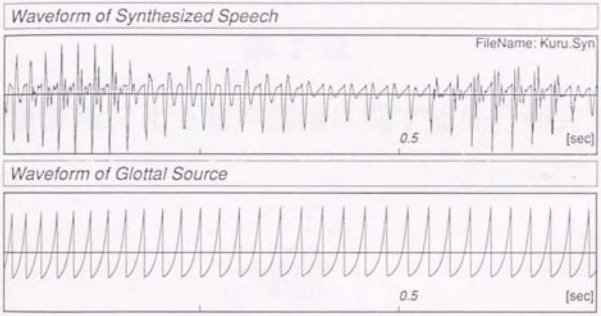


図 6.8. 実際に使用された有声部の音源波形例

上が有声部分の合成波形，下が対応する音源波形である。人工的に作成されていることが図 6.5 と比較するとよく分かる。

第 7 章

計算機による音声認識に関する先行研究

第 1 章で述べたように，音声を媒体とした情報交換は人間にとって非常に "楽な" タスクである。この音声を人間対機械間の情報交換にも，その媒体として利用できないかと，長年に渡り，数多くの研究が行なわれてきた。音声認識に対象を絞って考えてみると，その研究が開始された当初は,「同一シンボル (音素) に対応する音声の音響的特徴には，何らかの不変量があり，それを正確に抽出できれば音声認識は可能となる。」と考えられていた。しかし，研究が進むにつれ，同一シンボルに対応する音声でも，その音響的特徴は時間方向及び周波数方向に広く分布する (揺らぐ) ことが明らかとなり，上記の「不変量の追求」から,「音響的特徴の "揺らぎ" に対する対処方法の追求」へと研究課題が変化して行った[図]。この "時間方向" の揺らぎに対処すべく応用された手法が DP であり，"周波数方向" の揺らぎに対しては HMM が応用されるようになった[1]。

本章ではまず，この代表的な音声認識手法である DP と HMM に関して，第 8 章，第 9 章での議論を理解するために必要になると考えられる側面について説明すると共に，両者の比較をも行なう。そして，従来の研究では十分な議論が行なわれていないと考えられる点，

- 時間軸に沿って動的に変化する音声の音響的特性に対する，特徴表現方式の動的な適応。
- HMM 継続時間長モデルの時間構造記述力の評価。

を取り上げる。そして第 2 章でも述べたように，第 5 章で構築した知覚モデルへの考察を含めて,『計算機における音声の認識手法の高精度化』と言う観点からの本研究の位置付けを行なう。また，本研究に関連する先行研究についても紹介する。

7.1 DP と HMM

本節では，DP[図] 及び HMM[図] を用いた音声認識手法を概説し，両手法の比較を行なう。そして，これらを純粋にパターン認識手法として捉えた場合に，その高精度化に対して議論すべき焦点を明らかにする。

7.1.1 DP を用いた音声認識

同じ人間が，同じ単語を発声しても，その時間長は発声の度に変動し，しかもその伸縮の様子は非線形な特性を持つ。即ち標準パターンが，そのパターン発声時の時間的構造を十分に保存した形で構成されている場合は，標準パターンと入力音声 (以下，入力パターンと呼ぶ) との照合に先立って，両者を非線形に伸縮，対応付ける必要が生じる。この操

[1] 両手法とも音声処理独自の方法ではなく，従来からある基礎理論を音声処理に応用したものである。

作は一般的に DTW(Dynamic Time Warping) と呼ばれている。この DTW を実現する手法として DP(Dynamic Programming) が応用された訳である。以下では話しを簡単にするため，入力パターン中で着目すべき区間が限定された場合 (両端点固定) について DP を説明することにする。

第 5.3.1 節や図 5.8 に示したように，音声認識の分野ではまず，入力パターンから対数スペクトル包絡を記述するために必要な十数次元の特徴パラメータベクトルを各フレーム毎に抽出し，入力パターンをこのベクトル時系列へと変換した後に種々の処理を施す。さて，対応すべき 2 つのベクトル時系列 (入力パターンと標準パターン) を各々以下のように表すことにする。

$$\begin{cases} A = a_1, a_2, \ldots, a_I & (\text{入力パターン}) \\ B = b_1, b_2, \ldots, b_J & (\text{標準パターン}) \end{cases}$$

但し，Iは入力パターン時間長，Jは標準パターン時間長である。ここで図 7.1 のような A, Bからなる平面を考えると，A, B両パターンの時間的対応付けは，その平面上の格子点 $c = (i, j)$ の系列，

$$F = c_1, c_2, \ldots, c_k, \ldots, c_K \qquad (c_k = (i_k, j_k)) \tag{7.1}$$

で表現されることになる。2 つのベクトルa_i, b_i間の (スペクトル) 距離を $d(c) = d(i, j)$ と書くと，経路 Fに沿った距離の総和は，

$$D(F) = \frac{\sum_{k=1}^{K} d(c_k) w_k}{\sum_{k=1}^{K} w_k} \tag{7.2}$$

で表すことができ，この値が小さいほど A と Bとの対応付けがよい (よく似ている) ことになる。なお，w_kは Fに関して定義される正の重み関数である。

ここで，式 (7.2) を次のような制限の下，Fに関して最小化することを試みる。

1. 単調性と連続性の条件

   $0 \le i_k - i_{k-1} \le 1, 0 \le j_k - j_{k-1} \le 1$

2. 境界条件

   $i_1 = j_1 = 1, \quad i_K = I, \quad j_K = J$

3. 整合窓の条件：極端な非線形伸縮を防ぐため，rを正定数として，

   $|i_k - j_k| \le r$

式 (7.2) において，分母$\sum_{k=1}^{K} w_k$が Fに依存しない定数となるよう w_kを定義すると，この式は簡略化される。例えば

$$w_k = (i_k - i_{k-1}) + (j_k - j_{k-1}) \qquad (\text{便宜的に}, i_0 = j_0 = 0) \tag{7.3}$$

とすると，w_kは市街化距離 (city block distance) となり，

$$\sum_{k=1}^{K} w_k = I + J \tag{7.4}$$

となる。そして式 (7.2) は，

$$D(F) = \frac{1}{I+J} \sum_{k=1}^{K} d(c_k) w_k \tag{7.5}$$

となる。この時最小化すべき目的関数が加法的になるので，この最小化は Fの全てのパターンについて総当たり的に調べることなく，効率的に解くことができる (即ち，DP)。部分点列 $c_1, c_2, \ldots, c_k (c_k = (i,j))$ に対する部分和 $g(c_k)$ を考えると，

$$\begin{aligned} g(c_k) = g(i,j) &= \min_{c_1,\ldots,c_{k-1}} \left[\sum_{l=1}^{k} d(c_l) w_l \right] \\ &= \min_{c_1,\ldots,c_{k-1}} \left[\sum_{l=1}^{k-1} d(c_l) w_l + d(c_k) w_k \right] \\ &= \min_{c_{k-1}} \left[\min_{c_1,\ldots,c_{k-2}} \left\{ \sum_{l=1}^{k-1} d(c_l) w_l \right\} + d(c_k) w_k \right] \\ &= \min_{c_{k-1}} [g(c_{k-1}) + d(c_k) w_k] \end{aligned} \tag{7.6}$$

となる。この式を Fに関する条件 1.~3. を考慮して書き直すと，

$$g(i,j) = \min \begin{bmatrix} g(i, j-1) + d(i,j) \\ g(i-1, j-1) + 2d(i,j) \\ g(i-1, j) + d(i,j) \end{bmatrix} \tag{7.7}$$

となる。従って $g(1,1)$=$2d(1,1)$, j=1 として整合窓の範囲内で i を変化させながら上式を計算し，次に jを変化させて，j=Jとなるまで同様の計算を繰り返せば，最後に $g(I,J)/(I+J)$ として A, Bの 2 つのベクトル時系列の非線形時間正規化後の距離が求まることになる。なお式 (7.7) における，局所的な照合パスは図 7.2 のようになるが，Fに関する条件を変更することにより，図 7.3 のような照合パスも提案されている[2]。

[2] 後者の方がより一般的であり，効果があるとされている。

DP における標準パターンは一般的に，標準パターン用に一回発声されたものをベクトル時系列化して用いる場合と，複数回発声されたものを各々ベクトル時系列化して用いる場合とがある。後者の場合は同一音韻列に対して複数の標準パターンを持つこととなり，マルチテンプレート方式と呼ばれる。いずれにしても，多くの発声パターンよりその統計的な広がりを記述した標準パターンと言うのは，計算量の増加を生むため，あまり用いられていない。

このように DP では，照合時に入力パターンと (音声の時間構造が保存された) 標準パターンを，整合窓と言う条件付きで，非線形に時間対応付けすることとなり，時間方向の揺らぎに対しては柔軟に追従することができる。この条件による制約も，整合窓を変えることにより柔軟に対処できる。しかし基本的には，上記したように標準パターンは少数 (あるいは 1 つの) 音声パターンから作成されるため，元来音声が持つ周波数方向の揺らぎに対する記述力は，非常に貧弱なものとなってしまい，不特定話者や連続音声を対象とした場合の認識手法としてはあまり好ましくない特性を持つことになる。

7.1.2 HMM を用いた音声認識

同じ人間が，同じ音韻を発声しても，その時間長は発声の度に変動し，かつ，そのスペクトルパターンも同様に変動する。DP は，このスペクトルパターンの "揺らぎ" に対して柔軟に追従して行くことができないと言う欠点を持っていた。そこで，スペクトル方向での "揺らぎ" を，多くの学習データを用いて数理統計的に記述する HMM(隠れマルコフモデル，Hidden Markov Model) が音声処理に応用された。以下では単一ガウス分布連続 HMM と呼ばれるものを，DP との類似点を踏まえた上で説明することにする。なお，HMM を考える場合，音響モデル (標準パターン) の学習方法と，音響モデルと入力パターンの照合方法の 2 つが必要となるが，ここでは学習方法についてはその概略を述べるにとどめる[3]。また，照合方法についても簡単のため，入力パターン中で照合すべき両単点が固定された場合の処理について述べる。

HMM を用いて音声認識を行なう場合，認識対象のカテゴリ数だけ音響モデル，即ち HMM を作成することになる。音韻認識を行なう場合は，音韻数だけの HMM(標準パターン) を作成する訳である。ここで任意の HMM M は図 7.4 に示すように，有限個の状態とその状態間の遷移 (及び遷移中に出力されるベクトルの分布) によって形成され，次のパラメータによってその特性が記述される[4]($M=(S, A, \mu, \Sigma, \pi, F)$)。

[3] 詳しくは，参考文献 [127] などを参照して頂きたい。
[4] 但し上記したように，ここで説明するのは単一ガウス分布連続 HMM である。離散 HMM の場合や混合分布 (Gaussian Mixture) の場合は若干異なる。

S ：状態の有限集合 ($\{s_i\}$)
A ：状態遷移確率の集合 ($\{a_{ij}\}$)
μ ：平均ベクトルの集合 ($\{\mu_{ij}\}$)
Σ ：分散共分散行列の集合 ($\{\Sigma_{ij}\}$)
π ：初期状態確率の集合 ($\{\pi_i\}$)
F ：最終状態の集合

上記の表現において，s_iは状態 i を表す。a_{ij}は遷移パスが s_iに居る場合に，次に s_jへと遷移する確率であり，$\sum_j a_{ij} = 1$ となる。μ_{ij}は状態遷移 $s_i \rightarrow s_j$の際に出力されるベクトルの平均ベクトル。Σ_{ij}は状態遷移 $s_i \rightarrow s_j$の際に出力されるベクトルの分散共分散行列。π_iは初期状態が s_iである確率であり，即ち$\sum_i \pi_i = 1$ となる。以上のパラメータを定義することで HMM は，初期状態から最終状態へ向かって状態遷移が繰り返され，状態 i から状態 jへの遷移の際には，平均μ_{ij}，分散共分散行列Σ_{ij}で規定される多次元正規分布に基付いてベクトルを一つ出力するモデル[5]と考えることができる。逆にベクトル系列が与えられた時，着目する HMM が，そのベクトル系列をどの程度の確率 (密度) で出力し得るのか，も算出されることになる。

さて，本来ならば上記パラメータで構成される HMM の導出方法，即ち学習方法を示した後に，HMM による認識/照合方法を示すべきところである。しかしここでは，話しを簡単にするため，まず初めに，既に学習によって上記パラメータが算出され，HMM として音響モデルが構築されたものとして，認識/照合手法を説明する。そしてその後に，学習方法についても触れることにする。なお，以下では HMM の構造は図 7.4 であるものとし，初期状態は s_1，最終状態は s_4固定として話しを進める。

HMM を用いたパターン認識の場合 DP とは異なり，パターン間距離ではなく，対象とする標準パターンとの類似度 (尤度) が算出される。即ち上記したように，入力パターン $v = v_1, v_2, \ldots, v_T$($v_t$は時刻 t の特徴ベクトルであり，Tは入力パターン系列長である) が与えられた時に，vが HMM Mから出力される確率 (密度) $P(v|M)$ を求め，それを尤度とする。この $P(v|M)$ は以下のようにして求めることができる。$s = s_{i_0}, s_{i_1}, \ldots, s_{i_T}$を状態遷移系列 (但し，$s_{i_0} = s_1$(初期状態), $s_{i_T} = s_4$(最終状態)) とし，x_m^nを部分列 $x_m, x_{m+1}, \ldots, x_{n-1}, x_n$ とすると，

$$P(v|M) = \sum_{i_0, i_1, \ldots, i_T} P(v|s, M) \cdot P(s|M) \tag{7.8}$$

[5] 即ち，ベクトル出力を伴う状態遷移モデル。

$$
\begin{aligned}
P(\boldsymbol{s}|M) &= \prod_t P(s_{i_t}|s_{i_0}^{t-1}, M) \\
&= \prod_t P(s_{i_t}|s_{i_{t-1}}, M) \quad (\text{単純マルコフ過程の性質より}) \qquad (7.9) \\
P(\boldsymbol{v}|\boldsymbol{s}, M) &= \prod_t P(v_t|\boldsymbol{s}, v_1^{t-1}, M) \\
&= \prod_t P(v_t|s_{i_{t-1}}, s_{i_t}, M) \quad (\text{出力ベクトルにはマルコフ性無し}) \qquad (7.10) \\
P(\boldsymbol{v}|M) &= \sum_{i_0, i_1, \ldots, i_T} \prod_t P(s_{i_t}|s_{i_{t-1}}, M) \cdot P(v_t|s_{i_{t-1}}, s_{i_t}, M) \\
&= \sum_{\boldsymbol{s}} \left\{ \prod_t P(s_{i_t}|s_{i_{t-1}}, M) \cdot P(v_t|s_{i_{t-1}}, s_{i_t}, M) \right\} \qquad (7.11)
\end{aligned}
$$

即ち $s_{i_0}=s_1$, $s_{i_T}=s_4$ となる全ての $\boldsymbol{s}$ に対して $\boldsymbol{v}$ の出力確率密度を算出し，それらを足し合わせたものが式 (7.11) である。これは着目する HMM に対して，理論的に厳密に算出した $\boldsymbol{v}$ の出力確率密度であり，Baum-Welch スコアと呼ばれる。このスコアは可能な遷移全てを対象とするため，当然のことながら計算コストが高くなる。そこで，最大確率密度を示す状態遷移を求め，その遷移における確率密度値を尤度とする方法が提案された (Viterbi スコアと呼ばれる)。即ち，Baum-Welch スコアの近似である。そして，認識率に対するこれら 2 つの尤度の有意差は無い[95]との報告がされていること，及び，認識時の計算における overflow, underflow についても，Viterbi スコアを用いた場合は，“積→log の和” と変換することで容易に回避することが出来るなどの利点から，認識時の尤度計算では殆どの場合 Viterbi スコアが用いられている。

さて，この Viterbi スコアを算出する様子を DP と比較すると HMM と DP の相違が明確になる。HMM における状態を y 軸 (標準パターン) として図 7.1 と同様な平面を示したのが図 7.5 である。但し，DP では格子点でスコアが計算されたが，HMM では x 軸 (入力パターン) のある点 (フレーム) と y 軸 (標準パターン) のある点から点への “遷移” との間でスコアが計算される。この点を除いて考えると両者は非常に良く似た構造を持っていることが分かる。と同時に両者の相違点も明確になる。即ち，

1. HMM はループ遷移を持つため，“横へ横へ” と延びるパスが常に存在する。但し，DP においても局所的照合パスの設定次第でこれは実現可能である。
2. HMM の状態数は一般に，表現する音響的事象の時間長 (フレーム数) よりも少ない (時間方向の圧縮化が行なわれてモデル化されている)。その結果 HMM の状態 (正確には遷移) は，ある区間長 (フレーム数>2) の音声を，平均ベクトルと分散共分散行列を用いて統計的に記述することとなる。

3. 逆に，標準パターンにおける，時間方向での動的変動の記述力が低下し，入力パターンの時間方向の揺らぎに柔軟に対応することが困難になる。
4. また，DP と違って整合窓のような制限がないため，かなり自由度の高い照合 (状態遷移) を許してしまうことになる。

などが挙げられる。

一方，学習方法であるが次の通りである。まず，HMM のパラメータに適当な初期値を設定し，学習データに対して，全遷移を考慮した Baum-Welch スコアを求める。そして，この Baum-Welch スコア (確率密度値) を利用して，各パラメータの期待値を求め (最尤推定法)，その期待値を新たにパラメータ値として設定・更新する。このようにして再設定したパラメータにより求めた学習データ全体の尤度は，決して減少しないことが証明されており，これを繰り返すことで各パラメータは，学習データに対してより高尤度を示す方向へと変動することになる (準最適化)。そして，学習データ全体の尤度の変動がある閾値より小さくなるまで，この操作を繰り返す訳である。しかし，上記の最尤推定法では完全な最適化を行なっている訳ではなく，Local Maximum に陥る危険性を含んでいる。その意味で，初期値の設定は慎重に行なう必要がある。

7.2 本研究の目指す音声認識手法

本研究では第 2.2 節で述べたように，「計算機による音声認識手法の高精度化」を達成すべく，

- 音声の音響的特徴表現方式を動的に変化させた認識手法。
- 継続時間長モデルの時間構造記述力を向上させることを目的とした，学習データのクラスタリング手法。

と言う，異なる 2 つの観点からのアプローチを行ない，各々において新しい手法を提案する。本節では両者について，その背景となるところを述べる。特に前者は筆者が行なった先行研究[77]において残された問題点に対する一つの回答として位置付けられるものである。そこで第 7.2.1 節，第 7.2.2 節において，先行研究の内容を紹介すると共に，前者のアプローチに対する意義を述べることにする。後者のアプローチに関しては，第 7.2.3 節において，その位置付け・意義を述べる。

7.2.1 音声知覚モデルを反映した音声認識手法

第 7.1.1 節，第 7.1.2 節で，音声認識における主要な処理手法である DP と HMM について，両者の比較を含めて考察した。ここで第 5.3 節で述べた，知覚モデルの工学的応

用について再度考えることにする。第 5.3 節では，音響的特徴抽出処理，音響的照合処理及び内部辞書の構成に対する知見の一部が工学的に実現可能な対象として考察されている。これら 3 つの処理系の中で，内部辞書に関する工学的検討は cache 的 STM に限定せざるを得なかった (第 5.2.4 節) が，音響的特徴抽出/照合処理に関しては，音響的特徴精度の高低と照合処理単位の小大を対応させた手法を考案することにより，一度に，両処理部に対する知覚モデルを反映した処理手法が実現されることになる。

さて，上記の DP，HMM を応用した認識手法において，音声は如何なる方式で音響的に表現されているのだろうか。参考文献 [1], [128]–[131] などを見ると，特徴パラメータとしては FFT パワースペクトルのフィルタバンク出力や (メル) ケプストラム，(メル)LPC ケプストラムなど，対数パワースペクトル包絡に関与する種々の方式が提案されている。しかし，実際の認識手法に応用する際に共通して言えることとして，音声事象全てを，一通りの記述方式を用い，画一的に記述していることが挙げられる。これは，精度/単位の両側面において，知覚モデルとは異なった音響的抽出/照合が行なわれていることを意味する。この精度/単位に関する議論は当然のことながら，DP や HMM における照合手法そのものを問題視するものではなく[6]，音声の音響的特徴表現方式を含め，その応用/適用方法に着目するものである。そして第 5.3.1 節，第 5.3.2 節で示したように，DP や HMM の応用/適用方式及び音声の表現方式を知覚モデルに適合させることは十分工学的に可能な項目である。以上の考察の下，筆者は先行研究において，複数の特徴精度/単位の音声認識手法を提案した[77]。上述したように，本論文で提案する認識手法の一つはこの先行研究の直接の延長線上にあるものである。そこで第 7.2.2 節において，該当する先行研究内容・結果について簡単に述べ，残された問題点に触れると共に，第 8 章で詳細に述べる音声認識手法への導入を述べることにする。

7.2.2 複数精度/単位の音響的特徴量を用いた音声認識

本節では，第 5.2.7 節で提案した知覚モデル内の音響的特徴抽出/照合処理部の特徴を反映した認識処理手法の実現について，先行研究[77]で行なわれた実験的検討・結果について紹介する。

第 5.3 節において，分節的/韻律的特徴に対して，

- 複数の精度による音響的特徴抽出処理。
- 複数のサイズの単位による音響的照合処理。

[6] 即ち，人間の知覚モデルから眺めた場合に，DP と HMM のどちらが良いかと言うことではなく。

- 特徴精度の高低と照合単位サイズの小大の対応。

を実現することにより，知覚モデルが上記両処理部に対して表現するところを十分に網羅できると考察した。更にこの中でも，複数の精度の分節的特徴の実現に関しては，

1. 複数のフレーム長，シフト長の音響的特徴を抽出する。
2. 2 次元ケプストラムを抽出し，照合処理部において，着目すべき (2 次元ケプストラム内の) 領域を変動させる。
3. LPC 或はケプストラム係数の次数の増減による，(推定) パワースペクトル包絡の詳細化/平滑化を利用する。

をその実現方法例として挙げた。このように複数精度の特徴量 (或は複数の観点から音声事象を記述すべく，複数種類の特徴量) の利用は，比較的容易に実現が可能である。問題は "単位" と言うものをどのように実現するか[7]である。第 5.2.3 節におけるモデル化では照合処理単位として,「音韻/音節/単語/(分節)/句/文」と言った，いわゆる言語的単位を考えていた。更に第 5.2.4 節では，単語以上の単位を用いた照合処理における標準パターン (音響モデル) は，計算機処理から来る制約のため，音韻レベルの音響モデルの足し合わせによって作成されるべきであるとの考察をした。当然のことながら精度と単位サイズとの関係より，音韻認識に使用される音響モデルは高精度な特徴で記述される必要がある。その結果，これらを単に連結しただけでは，低精度の特徴で記述されるべく単語音響モデルにはなり得ない。即ち，モデルの連結後に，"精度の操作" と言う処理が介在することになる。モデルの連結と言う操作自体第 5.2.7 節で構築した知覚モデルには記述されていない処理であり，これに加えて更に知覚モデルに無い処理を導入するのは望ましいことではない。

次に，入力音声として孤立単語音声を仮定した場合の音韻レベル/単語レベル処理のタスクの違いについて考えてみる。単語認識の場合，入力音声はパワーによるセグメントにより，開始点，終了点が検出され，結局両端点固定の照合になる。一方音韻認識の場合は，第一音韻の開始点及び最終音韻の終了点は固定となるが (上記単語全体の開始点，終了点に相当)，それ以外の各音韻は前置音韻の終了点を始点とすることになり，基本的には (終) 端点自由の照合処理となる。更に，入力音声中の音韻数についても未知のまま処理が開始される。このように，音韻レベル・単語レベルの両処理を純粋にパターン認識の立場から見た場合，音韻認識の方により高いタスクが要求されることになる。各照合処理

[7] どのように考えるか。

において，使用される特徴量精度の差から来る処理量の量的な差と言うのは第 5 章でも考察されていた。しかし上記したような，端点の扱いの違いや，入力音声に対して仮定する単位数の可変性などは両処理間の質的な違いと言える。この両処理をただ単純に並列に走らせただけでは，当然のことながら認識精度の差が生じ，2 つのプロセスを統合して認識結果を求めた場合でも，上記したタスク間の質的な違いが結果に影響してくることが予想される。このタスクの差は当然のことながら，一方の認識プロセスにおける標準パターンは入力全体と照合されるのに対して，他方のプロセスにおいては，より小さな単位へセグメントを行なうと同時に，照合処理をする必要があることに起因する。これは，複数の言語単位を使用した照合処理を計算機上に実現する上では必ず付随してくる問題である。そこで，先行研究においては以上の問題を回避するため，知覚実験より得られている，次の知見の導入を検討した。

- 同一単位の処理においても複数の精度の音響的特徴を扱う機構が存在する。

この知見に対して，今まで議論していた「音響的特徴精度の高低」と「照合 (言語) 単位の小大」の関係を反映させることを試みた。即ち音声知覚過程のマクロ的性質として得られている知見を，ミクロ的な部分に対しても (仮定を置いて) 応用する訳である。音声を音響的に表現する場合，一般にはフレーム単位に作成される十数次元のベクトル時系列が主に使用されるが，この時系列の時間間隔 (シフト長) を "単位" と考え，可変化するのである。つまり，"照合処理における単位 (照合単位)" ではなく，音声の "音響的特徴表現における単位 (以下，表現単位と呼ぶ)" に着目する訳である。そして表現単位が大きい場合は，音声をより低精度で記述すべく特徴を用い，表現単位が小さい場合は，より高精度で記述すべく特徴を使うのである。但しこの場合，どちらに対しても同じフレームを基本とした特徴量を使用していては，第 5.3 節で述べたような問題が自ずと付随してくる。そこで表現単位を大きく設定した場合は，次に説明する 2 次元ケプストラムの適切な領域を使用することによりこの問題の解決を図る。ここで，2 次元ケプストラムから得られる特徴量を，フレームに対してブロックと呼ぶことにすると，

	高精度/小表現単位	低精度/大表現単位
音響的パラメータ	ケプストラム	2 次元ケプストラムの一部
本研究での名称	フレーム	ブロック

となる。そして，この 2 つの特徴表現を用いた音響モデルを各音韻に対して作成する。認識時には両特徴量を随時抽出し，両者を動的に選択しつつ，入力パターンとの照合を行なう訳である。以下，2 次元ケプストラムについて簡単に説明した後に，上記 2 種類の特徴

表現の動的選択に基付く照合方式の有効性について実験的に検討した結果を述べる。なお，ここで述べるのは先行研究結果の要約であり，詳細については参考文献 [77] を参照して頂きたい。その後，残された問題点/改良点についてで触れる。第 7 章冒頭でも触れたように，本論文で提案する認識手法の一つは，この問題点への解決を図るために考案された手法であり，第 8 章にてその詳細が述べられることになる。

2 次元ケプストラム

第 5.3 節でフレーム長はどのように設定しても問題を残すことを述べたが，これは，フレームという特徴量が一般的には対応する音声区間の平均的な特徴のみを表現し，時間方向の変動はフレームの時系列という形で表現させていることに起因している。このため，フレーム長を大きくとると，対応する音声区間の時間方向での変動が平均化され，吸収される結果となる。また，フレーム系列で急激なスペクトル変化に追従するためには，フレーム長を短くする (その結果，周波数分解能は低くなる) 必要があった (図 5.8 参照)。しかし，仮に特徴量レベルで (1 フレーム内で) 時間方向の変動をも含んだ表現が可能であれば[8]，その必要性は無くなる。以上のような考察の下，2 次元ケプストラム法が考案された[105][106]。これは，短いフレームの時系列で表現されたパワースペクトル包絡を 2 次元的に (図形，画像として) とらえて，2 次元フーリエ変換を施すものである。パワースペクトルに対してフーリエ変換を施したものはケプストラムと呼ばれるが，このケプストラムの時系列を時間方向に更にフーリエ変換するものである (図 7.6 参照)。その結果得られる特徴パラメータを利用することで，特徴量レベルで時間方向の動的変化が記述でき，周波数及び時間軸両方向に滑らかで，しかもかなり良い分解能を持ち，子音などの早い変化にも追従できるスペクトル包絡面が得られることになる。具体的な算出法を以下に示す。

適当な時間間隔ΔTでサンプリングした有限区間の離散的な音声信号を

$$x_{n,m} = x_{n+mL} \quad (0 < L \leq N;\ n = 0, 1, \ldots, N-1;\ m = 0, 1, \ldots, M-1)$$

とする。但しLはシフト長，Nはフレーム長，Mはフレーム数であり，$x_{n,m}$は第mフレームの第n番目の信号を意味する。第mフレームの信号$x_{n,m}$の離散フーリエ変換$X_{k,m}$は，

$$X_{k,m} = \sum_{n=0}^{N-1} x_{n,m} W_1^{nk} \quad (W_1 = \exp(-j\frac{2\pi}{N});\ k = 0, 1, \ldots, N-1;\ m = 0, 1, \ldots, M-1)$$

で与えられる。kはフレーム長 N，サンプリング周期ΔTにより決まる角周波数分解能

[8] 当然時系列データとしても表現されることは変わり無い。

$\Delta\omega(=2\pi/N\Delta T)$ の倍数に相当する周波数である。$X_{k,m}$ の対数スペクトル $S_{k,m}$は

$$S_{k,m} = \ln(|X_{k,m}|^2) \quad (k = 0, 1, \ldots, N-1;\ m = 0, 1, \ldots, M-1)$$

で与えられる。通常のケプストラム $c_{q,m}$は，$S_{k,m}$のフーリエ変換，

$$c_{q,m} = \sum_{k=0}^{N-1} S_{k,m} W_1^{kq} \quad (W_1 = \exp(-j\frac{2\pi}{N});\ q = 0, 1, \ldots, N-1;\ m = 0, 1, \ldots, M-1)$$

で与えられる。qはサンプリング周期ΔTの倍数で時間 (quefrency) である。$c_{q,m}$の時間 m に対するフーリエ変換 $C_{q,p}$，即ち対数スペクトル $S_{k,m}$の周波数 kと時間 m に対する 2 次元フーリエ変換，

$$\begin{aligned} C_{q,p} &= \sum_{m=0}^{M-1} c_{q,m} W_2^{mp} \\ &= \sum_{m=0}^{M-1} \sum_{k=0}^{N-1} S_{k,m} W_1^{kq} W_2^{mp} \end{aligned}$$

$$(W_1 = \exp(-j\frac{2\pi}{N});\ W_2 = \exp(-j\frac{2\pi}{M});\ q = 0, 1, \ldots, N-1;\ p = 0, 1, \ldots, M-1)$$

を考えて，$C_{q,p}$を 2 次元ケプストラムと呼ぶ。p は分析の全区間 ML により決まる周波数$(1/ML)/\Delta T$の倍数に相当する周波数であり，図 7.6 では，時間変化周波数と呼んでいる。ここで $C_{q,p}$は，qに対する対称性と p に対するエルミート対称性により，

$$C_{q,p} = C_{N-q,p} = C^*_{q,M-p} = C^*_{N-q,M-p}$$

が成立し，2 次元ケプストラム $C_{q,p}$は $N \times M$の約 1/4 の領域 (0~$M/2-1$,0~$N/2-1$ 次元) を考えれば良いことになる (C^*は Cの共役複素数である)。この領域は，およそ図 7.6 に示すような音響的意味付けが可能となり，どの領域を分析に用いるかで，時間軸・周波数軸夫々独立に平滑化が可能となる。

音声の音響的特徴表現単位/精度を動的に変動させた音声認識

本節で提案した手法を実験的に検証するため，孤立発声の単語認識実験を行なった。使用した音声試料及び実験条件を表 7.1 に示す。また，照合方式としては Stochastic DP 法[35]を用いた。即ち，Stochastic DP 法で要求される標準パターンを，特徴量としてフレームを用いて作成したもの，ブロックを用いて作成したものを用意する。更に，図 7.7 に示すように，3 フレーム毎にその時点に対応するブロックモデルへの遷移を定義し，モデル間遷移確率を付与して両モデルを統合する。実際の照合パスは DP(あるいは Viterbi

Path) と同じように，最大確率 (尤度) を示すパス (遷移) を拾っていく形となる (その結果，モデル間遷移も動的に行なわれることになる)。なお図 7.8 に，実験で用いた局所パスをフレーム・ブロック両モデルについて示す。図 7.9 に認識実験結果を示す。図には，1) フレームのみによる認識結果，2) フレーム/ブロック混合モデルによる認識結果を $n(=1,2,\ldots,5)$ 位までの累積認識率と言う形で示している。本手法の有効性が十分に伺える。

残された問題点

先行実験によって，複数精度の音響的特徴を複数の音響的特徴表現単位に対応させた手法の有効性が示された。しかしこの手法をケフレンシー[9]領域から眺めた場合，次に示すような問題点/改良点が明らかとなる。本手法は基本的には，フレーム (高精度特徴) 及びブロック (低精度特徴) を用いた照合処理間を，入力パターンと標準パターンに依存しながら交互に移行しつつ，最終的な類似度 (累積スコア) を求めるものである。この場合，ある時刻における照合処理に使われる音響的特徴をケフレンシー領域から眺めると，

フレーム 1〜$N(=16)$ ケフレンシーまでのケプストラム。

ブロック 着目するフレーム群から得られる 2 次元ケプストラムの低ケフレンシー部。但し，利用する低ケフレンシー部は時間的にも周波数的にも固定。

となる。即ち，フレームとブロックは，ケプストラム・2 次元ケプストラムと言う特徴量の違いこそあれ，基本的には 1 次から何次までの係数に着目するかの違いであると言える。また，低ケフレンシー部でもってブロックを定義した理由としては，

1. 比較的値の大きなケプストラムは低ケフレンシー部に集中する[10]。

と言う物理的事実の他に，

2. 低ケフレンシー部のケプストラム=スペクトル包絡を一波形と見なした場合の低周波成分，であり，低精度と言う概念に非常に良く適合する。

と言ったことが挙げられる。特徴精度の変化と言う観点からこの手法を考えた場合，上記 2. の仮定が正当なものであるなら，本手法は評価されるべきであろう。しかし，フレーム/ブロックの選択は，“入力パターンに対して着目すべき特徴量の制御” と考えることもできる。このように考えた場合，上記のような静的に定義された 2 種類の領域の移行のみでは，余りにも単純過ぎるように思われる。図 7.10 は，図 7.9 とは異なる実験条件の下，

[9] 理論的には “時間” と等しくなるが，frequency 領域の信号を逆 FFT した結果 (ケプストラム) の単位としては，quefrency(ケフレンシー) と呼ばれる。

[10] 低ケフレンシー部以外で値の大きなケプストラムが観測されるのは，F_0 に相当する部分のみである。

1) フレームのみ，2) ブロックのみ，3) フレーム/ブロックの混合モデルを用いた認識結果を示している。図 7.9 と同様にフレーム/ブロック混合モデルにおいて最高値を示しており，特に Rank 数が低い時にその傾向は顕著になる。しかし，ブロックのみの認識率を見た場合，フレームのみの認識率と比較してかなり低くなっている。ブロックは，フレームの動的変化を平滑化したものに該当する (情報量の低下) ため，ブロックのみにおける認識率低下は十分に予想される結果ではある。しかし，個々のデータを眺めると，正解候補の類似度が低く評価されるのみならず，非正解候補の類似度を極端に高く評価している例が見受けられた。これらは上記したように，着目する特徴量の制御が極端に単純化され，かつ入力音声や標準パターンにも依存せず，静的な形で定義されている[11]ことに起因すると考えられる。入力パターン及び標準パターンに動的に対応した，より柔軟な制御法が望まれるところである。

"着目する特徴量を動的に制御する" と言うことは，フレームやブロックの動的選択よりも更に進んで，フレームの定義そのものを入力パターンや標準パターンなどに依存しつつ，動的に決定することに他ならない。パターン認識の分野においては，入力信号を標準パターン **P** 及び **Q** と比較・照合する場合，入力信号の工学的記述方法は，異なる標準パターンに対して同一の方法を用いることが大前提となっている[12]。また，標準パターン **P** との照合に着目した場合，その記述方法は時間に対して不変であることも前提となっている。しかし，HMM のような数理統計的な手法においては，照合スコア (尤度) は全て確率或は確率密度の形で算出される。すなわち，入力信号の記述方法を時間軸上で変動させた場合，或は，照合する標準パターンによって記述方法を変化させた場合でも，標準パターン **P** に対するスコア **p** と標準パターン **Q** に対するスコア **q** は直接比較可能な物理量となる。これはパターン認識を数理統計的な手法で実現する場合は，上記した前提を必ずしも必要としていないことを意味し，入力信号の工学的記述方法は時間軸上で動的に，或は，照合対象に依存して変化させることが可能であることを意味する。

以上の考察の下，第 8 章では，『音声の音響的特徴表現を動的に制御した認識手法』と言うテーマに基づいて議論を進める。

7.2.3 HMM における継続時間長モデルの高精度化

第 7.2.1 節，第 7.2.2 節では人間の音声知覚モデルとの関連から音声認識手法の高精度化について，その可能性を議論した。しかし，第 7.1.1 節，第 7.1.2 節における DP

[11] フレーム↔ブロックの移行は，入力・標準パターンに依存している。
[12] 即ち,「同じ土俵の上で比較しよう」と言うことである。

とHMMとの比較を通して，両者を純粋に時系列データ間の比較/照合方法，或は時系列データのモデリング手法として捉えた場合は，高精度化の可能性が以下のように示されている。即ち，

- DP に周波数方向の揺らぎの記述を組入れる。
- HMM に時間方向の揺らぎの記述を組入れる。

ことで両手法は，より忠実に時系列データを記述することが可能となる。そして，前者に対して Stochastic DP 法と呼ばれる手法が提案されている[35]。この手法は HMM において状態数を増やし，基本的には "HMM の 1 状態=DP における標準パターンの 1 フレーム" とした手法である。DP の側から考えれば，"標準パターンの各フレームの記述を統計的な分布の広がりを加味して行なったもの" と考えることができる。一方後者に対しては，照合の際に行なわれる状態遷移において，各状態に停留する時間長を学習データから学習・推定し，パラメータとして保持させる継続時間長制御の技術が提案され[36]，広く使われるようになった。本研究では第 2.2 節でも述べたように，認識率の面での HMM の優位性と共に，言語処理を含めた上での数理統計的手法の応用範囲の広さから，後者に着目することにする。

継続時間長制御モデルの作成方法 (学習方法) 及びそれを用いた照合方法は種々のものが提案されている。しかし，作成されたモデルがどの程度，音声の時間構造を正しく記述しているのか，それには音素依存性・話者依存性が存在するのか，そして，継続長時間制御モデルの導入に際して，特有の (前) 処理は必要でないのか，などについての議論が十分に行なわれていないように思う。そこで本研究では，ある特定の継続時間長モデル作成法[37]に着目し，その手法において作られるモデルと，実際の学習データの時間的構造との整合性を音素別に詳細に分析する。そしてその結果を基に，継続時間長モデルの精度を向上させるべく，学習データの新しいクラスタリング手法を考案する[39]。音声認識の従来の研究においても，学習データのクラスタリング手法は数多く提案されている。しかし，それらの多くは音声の周波数軸上での特徴 (スペクトル) に基付くクラスタリングであり，本研究で言う，音声の時間構造及びその標準パターン (継続時間長モデル) に基付くクラスタリングとは質的に異なるものである。

以下第 9 章において，採用した継続時間長モデル作成方法の説明を含め，『クラスタリングによる HMM 継続時間長制御の高精度化』を実験的に検討することにする。

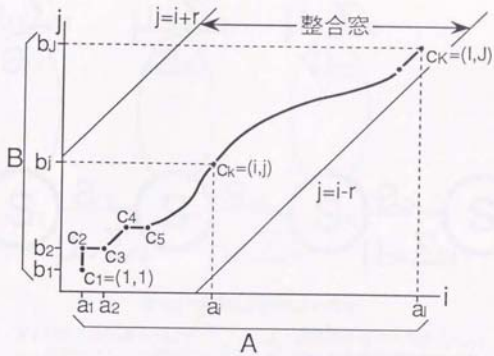


図 7.1. A, B 2 つのベクトル時系列の非線形な時間対応付け

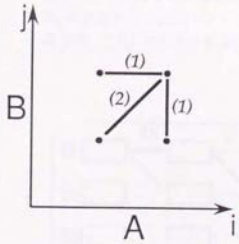


図 7.2. 式 (7.7) における局所的照合パス

括弧内の数字は，各々の局所パスに対して定義される重み付けである。

$$g(i,j) = \min \begin{bmatrix} g(i,j-1) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j) + d(i,j) \end{bmatrix}$$

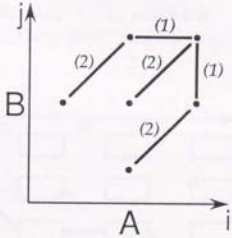


図 7.3. より一般的な局所的照合パス

括弧内の数字は，各々の局所パスに対して定義される重み付けである。

$$g(i,j) = \min \begin{bmatrix} g(i-2,j-1) + 2d(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i-1,j-2) + 2d(i,j-1) + d(i,j) \end{bmatrix}$$

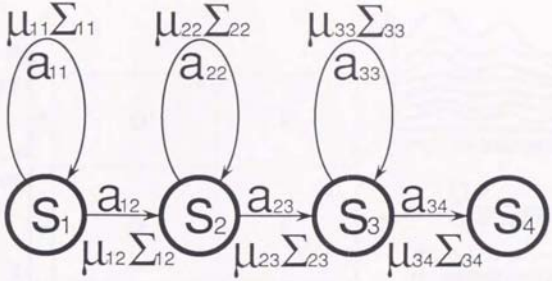


図 7.4. 典型的な HMM の構造

S : 状態の有限集合 ($\{s_i\}$)　　A : 状態遷移確率の集合 ($\{a_{ij}\}$)
μ : 平均ベクトルの集合 ($\{\mu_{ij}\}$)　　Σ : 分散共分散行列の集合 ($\{\Sigma_{ij}\}$)
π : 初期状態確率の集合 ($\{\pi_i\}$)　　F : 最終状態の集合

ベクトルの時系列と本モデルを照合する場合，各ベクトルが状態遷移の際に出力されるものと考える。即ち，ベクトル$\boldsymbol{v}_t$を平均μ_{ij}，分散共分散Σ_{ij}の分布からの出力ベクトルと考える訳である。そして，$\boldsymbol{v}_t$全体を最高確率 (密度) で出力する遷移が最適遷移 (Viterbi Path) となる。

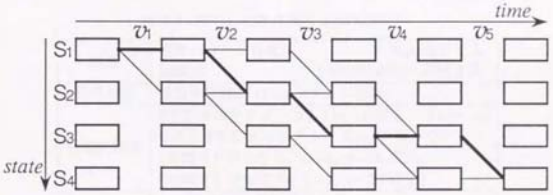


図 7.5. HMM と入力パターン間の Viterbi Path(時間的対応付け)

可能な状態遷移全体を実線で示してある。その中で太線で示してあるのが最適経路 (Viterbi Path) である。なお，上図は概念を示すだけであり，実際の確率値などは記していない。

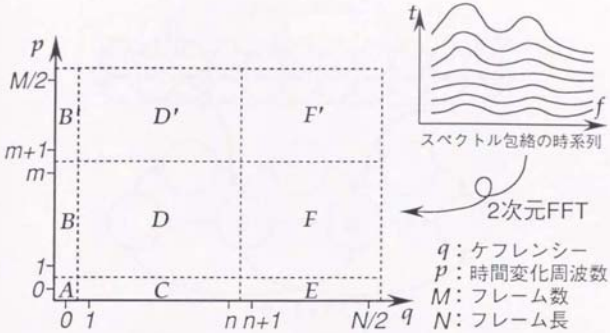


図 7.6. 2 次元ケプストラム

A：対数スペクトルの平均値
B：対数スペクトルの平均の概略的な変化　　B'：対数スペクトルの平均の微細な変化
C：平均的なスペクトル包絡
D：スペクトル包絡の概略的な変化　　D'：対数スペクトル包絡の微細な変化
E：スペクトルの微細構造
F：スペクトル微細構造の概略的な変化　　F'：スペクトル微細構造の微細な変化

表 7.1. 使用した音声試料及び実験条件

話者	標準パターン作成用　成人男性 4 人 認識用　上記以外の成人男性 2 人
発声形態	孤立単語発声 (212 バランス単語)
音響的特徴	FFT メルケプストラム (1~16 次元，フレーム) 2 次元 FFT メルケプストラム (ブロック) (実数マトリックス，0~1, 1~16 次元) (虚数マトリックス，1~1, 1~16 次元)
分析条件	12bit, 12kHz サンプリング フレーム長　256 サンプル フレーム周期　10[msec] ブロック長　4 フレーム ブロック周期　30[msec] (3 フレーム周期)

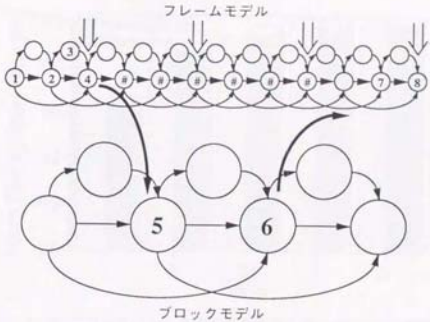


図 7.7. フレーム/ブロックモデルとその間の遷移

フレーム表記における#がブロック表記における状態 5, 6 に相当する。⇓のある状態 (フレーム) には，モデル間遷移確率が算出されており，その状態への遷移は，ブロック及びフレームによる遷移のうち，より高いスコアを示すものが動的に選択される。

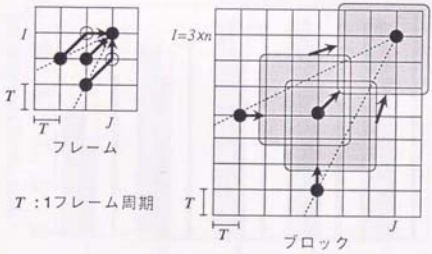


図 7.8. フレーム/ブロックモデルによる照合で使用された局所パス

●は遷移の始点/終点を示し，○及び網掛け部分はフレーム及びブロック照合局所パスの途中で加算される局所スコアの算出に用いられる部分を示す。ブロック照合が行なわれるのは，積分軸 (この場合は標準パターン側の軸) で，$I = 3 \times n$ の時点である。

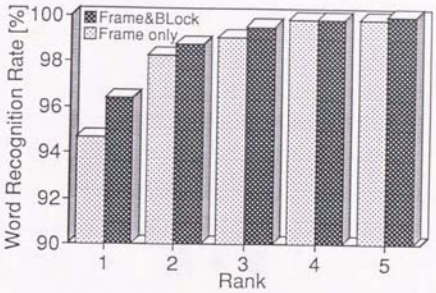


図 7.9. フレーム/ブロックによる音響的特徴表現を用いた単語音声認識結果 1
縦軸が認識率，横軸は第何位まで着目するかを表している。フレームのみによる認識結果と，フレーム/ブロックの混合モデルによる認識結果の両方を示している。

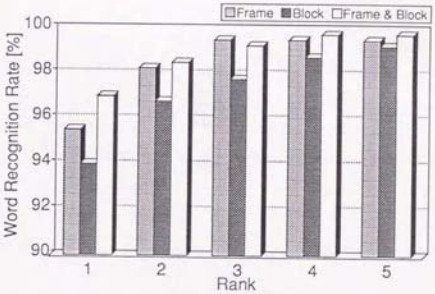


図 7.10. フレーム/ブロックによる音響的特徴表現を用いた単語音声認識結果 2
フレームのみによる認識結果，ブロックのみによる認識結果，及びフレーム/ブロックの混合モデルによる認識結果の 3 種類を各々示している。

第 8 章

音声の音響的特徴表現を動的に制御した認識手法

第 4 章において『人間における音声知覚過程の分析とモデル化』と言うテーマの下，筆者が行なった知覚実験について詳細に報告した。本章より本研究のもう一つの柱である，『計算機における音声の認識手法の高精度化』と言うテーマに基付いて行なった研究に関して，提案する手法，及びその評価のために行なわれた認識実験 (結果) について述べる。なお，計算機上での音声認識に関する研究は第 7 章でも述べたように，大きく 2 つのテーマに対して行なわれている。ここではその一つに触れ，他方のテーマ (クラスタリングによる HMM 継続時間長制御の高精度化) に対する研究は第 9 章で述べることにする。

8.1 本研究の背景と目的

第 5 章において知覚モデルを構築する際に議論したように，人間は音声を認識する際，一般的にはまず低精度の大局的な特徴を用いた照合が行なわれ，次第に高精度の詳細な特徴で照合が行なわれていると考えられる。先行研究において筆者は，複数の精度・時間単位の音響的特徴を用いて音声を表現し，実際の照合に使用する特徴パラメータの表現方式を動的に変化させる手法について検討し，その有効性を示すことが出来た。先行研究では，“大局的特徴≡(2 次元) ケプストラム[105][106]における低ケフレンシー部” と言った単純な仮定の下に，時間的に固定された特徴部分空間[1](低ケフレンシー部) を低精度特徴量として定め，照合時に使用する表現形式を動的に選択していた[114][115]。しかし第 7.2.2 節に述べたように，高/低精度の特徴表現の定義は音響的特徴の状態・様子に依存し，動的に変化すると考える方が妥当であり，入力音声の特性が異なれば，各々の適切な特徴表現方式 (着目すべき特徴部分空間) も自ずと動的に異なってくると考えられる。つまり全音素において，かつ時間軸に沿って静的に，常に “大局的”≡低分解能≡低次係数，と一意に定義するのは危険が伴う可能性がある。実際に先行研究で行なわれた，時間的に固定された定義に基付く大局的 (低精度) 特徴のみによる認識では，非正解項目の照合スコアを予想以上に高める働きもあった (図 7.10 参照)。

そこで本章では，入力パターンから抽出された音響的特徴量に対して，着目する特徴部分空間を動的に変動 (制限) しつつ，照合を行なう手法について検討する。まず初めに動的変動の制御を，1) 入力音声のみに依存，2) 標準パターンのみに依存，3) 両者に依存させながら行なう，と言う 3 種類の異なった制御方法を考え，切り出し音素認識実験を通して，各々の表現方式の特性，及び音声認識に寄与する部分空間の分布について詳細に観察する。次にその結果を基に，残された特徴量の有効利用法について検討していく。

[1] 本章では，N次元特徴空間から $n(<N)$ 個の成分を抽出して定義される，n 次元特徴空間を部分空間と呼ぶことにする。数学の世界で言う部分空間 (この場合，次元数の低下は無い) とは定義が異なる。

なお，本章で検討する入力パターンに対する工学的 (音響的) 表現方式の動的変動と言う考えは，第 7.2.2 節の「残された問題点」でも考察したように，数理統計的なパラダイムでパターン認識を行なうことにより，実現可能となる処理である。つまり，数理統計的に処理を行なうことによって，如何なる種類 (単位)[2]のデータの分布も確率 (密度) と言う一つの尺度に変換されて使用されるからである。そして，全ての異なる種類のパラメータを並べて一つのベクトルを作ることが可能となり，それらの分布 (分散共分散行列) や分布に対する数学的操作も意味のある操作となる[3]。

8.2 着目する音響的特徴空間を動的に制限した音声認識

本節では，照合処理に利用する音響的特徴空間を動的に変化 (制限) させて認識処理を行なわせることで，音声認識に有効に寄与する特徴部分空間の分布の様子を実験的に観測する。ここで特徴部分空間の (時間的に可変な) 定義が結果に大きく影響してくるものと考えられるが，本実験では以下の 3 種類の観点からその定義方法を検討する。

1. 入力音声のみに依存させた制御 (制限)
2. 標準パターン群のみに依存させた制御 (制限)
3. 両者に依存させた制御 (制限)

各々の具体的な定義・制御については第 8.2.1 節，第 8.2.2 節，第 8.2.3 節で説明する。また，比較実験として，部分空間を静的に定めた場合 (先行研究での大局的特徴に相当) も含めて実験を行なうことにする。なお以降の説明では，音響的特徴量としてはケプストラム (或はそれに類する特徴量) を使用することを前提としており，また，全特徴空間の次元数を N，部分空間の次元数を n とする。

8.2.1 入力音声のみに依存した定義

先行研究において，"低精度≡大局的特徴≡(2 次元) ケプストラムの低次係数" と定義するに至った経緯には，

- 分節的特徴を表す音声スペクトル包絡を一波形として見なした場合，低周波の成分がより大きい (絶対値の大きいケプストラムは低次元に集まる) 傾向がある。

と言う物理的事実がある。これは，(スペクトル包絡を一波形と見なした場合) 低周波の成分ほど，音声スペクトルの構成に果たす役割が大きく，"スペクトルに対する記述力"

[2] 長さ，重さ，早さ，強さ …… 。
[3] しかし，計算量の削減を図って，1 つのベクトルとして考慮はするものの，分散共分散行列は各要素の種類毎に作成する (部分的対角化) ことも行なわれている。

のより大きな成分であることを意味する。逆に言えば，N次ケプストラムから$n(<N)$個の係数のみを使用し，他の係数=0.0 と仮定してスペクトルを復元することを考えた場合，絶対値のより大きな係数から参照したスペクトルほど，原スペクトルをより正確に近似できることを意味する[4]。但し，スペクトルを波形 (図形) として観測した場合の復元力の大きさ (2 乗誤差の程度) が，知覚的な復元力の大きさと合致しているか否かについては不明である。ここで言う復元力/記述力とは，スペクトルを単なる一波形 (図形) とみなした場合の議論しているに過ぎない。図 8.1 は，あるフレームの音声スペクトルに対し，次数のより小さいスペクトラム係数から復元したもの (左図。先行研究の大局的特徴に相当する。) と，絶対値のより大きいケプストラムから順に復元した様子 (右図) を示す。当然のことではあるが，後者の方が次元数の少ないうちから元波形 (スペクトル) をより正確に近似できている。以上の考察より，入力音声に依存した音響的特徴空間の動的制限として，「入力音声に対する記述力のより高い成分，即ち絶対値のより大きいケプストラムから参照する照合方式」を定義する (定義 1)。この場合，N次までケプストラムを抽出した後に，絶対値の大きな$n(<N)$個の係数を利用することになり，どの係数が使用されるかは入力音声のみに依存し，フレーム単位で動的に変化することになる。

8.2.2 標準パターン群のみに依存した定義

標準パターンには対応する各音韻固有の特徴が記述されてあるばかりでなく，全標準パターンを群として見た場合，音響的特徴空間の各次元における特徴量分布の粗密の様子をも記述されている。後者は，ある特定の入力音声の観測だけでは得られぬ情報である。標準パターンに依存させて，照合に利用する音響的特徴空間を動的に制限することを考えた場合，標準パターンが持つ，該当音韻の音響的特徴自身に依存した表現方式を考えるよりも，全標準パターンを群として扱い，異なる標準パターン間の相違がより明確な成分からの参照を考えるべきであろう。照合方式として連続 HMM を考える場合，一般的には全音韻モデルに対して共通の構造を与える。このことを利用し，状態遷移毎に，照合に利用すべき成分 (次元) の順序を以下の様にして決定することを考えた。

モデル$m(1\leq m\leq M)$の遷移$ij(1\leq i,j\leq I)$には，平均ベクトルμ_{ij}^{m}，標準偏差ベクトルσ_{ij}^{m}(対角化分散共分散行列を想定) が存在する。各モデルの構造が同一であることを利用し，遷移別にM個のモデルを見た場合，遷移ijにはM個のμとσが存在する。このμ, σは共にベクトルであり，かつ分散共分散行列は対角化されたものを想定しているので，各次元にM個の正規分布が存在していることになる。ある成分 (次元) が各標準パターン間の相

[4] ケプストラムの定義を考えれば当然のことではある。

違をより明確に表現するためには，M個の正規分布がより離れて存在していることが必要である。即ち，与えられた次元数 (n) での照合方式を考えた場合，M個の正規分布がより離れている次元から照合することで，標準パターン群全体が記述する音響的特徴量の分布を適切に捉えることが可能となる。この場合，遷移 ij, 第 k次元における 2 つの正規分布 p, q間 (即ち，モデル p, q間) の距離 $d_{ij}(k)_{pq}$を定義する必要があるが，これを以下のように考える。

$$
\begin{aligned}
d_{ij}(k)_{pq} &= d_{ij}(k)_{qp} \\
&= \frac{|\mu_{ij}(k)_p - \mu_{ij}(k)_q|}{\sigma_{ij}(k)_p} + \frac{|\mu_{ij}(k)_p - \mu_{ij}(k)_q|}{\sigma_{ij}(k)_q}
\end{aligned}
$$

但し，$\mu_{ij}(k)_p$, $\sigma_{ij}(k)_p$は，モデル (分布)p の遷移 ijにおける平均及び標準偏差ベクトルの第 k次元要素を表わす。定義より，$d_{ij}(k)_{pq}(=d_{ij}(k)_{qp})$ は 2 分布の平均間のユークリッド距離を分布 p から見たマハラノビス距離及び分布 qから見たマハラノビス距離に換算して，両者を足したものに相当する (図 8.3 参照)。

次に，分布 p に最も近く存在する p 以外の分布との距離を上位 L 位まで求め，その距離和を $D_p(k, L)$ とする。そして，k次元における分布のばらつきの度合い $V(k, L)$ は，M(モデルの総数) 個の $D_p(k, L)$ の和として

$$
V(k, L) = \sum_{p=1}^{M} D_p(k, L)
$$

の様に定義される。以上の考察より，標準パターン群に依存した音響的特徴空間の動的制限方法として，「$V(k, L)$(第 k次元の分布のばらつき度) のより大きい成分から参照する照合方式」を定義する (定義 2)。以上の説明を図示したものを図 8.4 に示す。

8.2.3 両者に依存した定義

照合方式として対角化分散行列を用いた連続 HMM を考えた場合，出力確率密度の算出途中で，標準パターンに記されている平均ベクトルと入力音声から抽出された特徴ベクトル間のマハラノビス距離が各次元毎に求まる (md_i^2, i は次元)。対角化分散行列を用いているので全次元でのマハラノビス距離 MD^2は単純に

$$
MD^2 = \sum_{i=1}^{N} md_i^2 \quad (i \text{ は次元})
$$

となる。これは標準パターンと入力音声との距離を示す一指標である。当然のことながら，MD^2に対する構成比のより大きな次元におけるマハラノビス距離 md_i^2は，MD^2 のより

良い近似となる。そこで，入力及び標準パターンの両者に依存した次元数 n の音響的特徴空間の動的制限として，「md_i^2のより大きい成分から参照する照合方式」を定義する (定義 3)。更に最終的に算出される各次元毎の出力確率密度に対しても同様な定義を行なうことができ，「出力確率密度のより小さい n 成分から参照する照合方式」を定義する (定義 4)。上述の 2 つの定義は，N次元で構成される音響的特徴の内，照合する標準パターンとの距離がより大きい (より似ていない)n 成分を動的に追跡していくことを意味する。当然のことながら上記の定義とは逆に，「入力パターンと標準パターン間との距離 (類似度) がより小さい (大きい) 成分から参照する照合方式」と言う定義も可能であるが，この定義が実質上無意味であることは第 8.2.4 節で述べる認識実験結果より，明らかとなる。

以上の定義 1~4 に加え，入力にも標準パターンにも依存しない次元数 n の静的な特徴部分空間として，「次元 1~n の n 次元空間による照合方式」を定義 0 として定める。図 8.2 に，定義 0~定義 4 の 5 定義において，照合に使用される n 個の成分が決定されるタイミングを音響的特徴抽出及び照合スコアの算出過程と共に併記する。定義から明らかではあるが，n 成分の決定時刻は定義 0 から 2,1,3,4 の順に遅くなる。特に定義 4 では各次元の照合スコアそのものを用いて選択成分を決定しており，この場合は，従来の音声認識手法を，"全次数での認識能力が動的に変化する n 次部分空間でどの程度記述できるか" と言う観点から見直しているとも言える。他の定義については，最終的なスコア算出の以前で決定される $n(<N)$ 次元空間による認識でも，定義 4 による n 次元空間による認識能力にどの程度迫れるのか，等が着眼すべきところとなる。

8.2.4 音響的特徴部分空間による音素認識実験

定義 0~定義 4 によって構成される n 次元特徴部分空間の音声認識に対する有効性・貢献度を検討するため，特定話者の切り出し音素認識実験を行なった[38][116]-[118]。$n(<N)$ 次元で構成される特徴空間はあくまでも，ある規則に基付いて制限された成分による空間 ($\in N$ 次元特徴空間) であり，全特徴空間を用いた認識結果を上回る結果を期待することは困難である。しかし，定義 0~定義 4 によって構成される部分空間を用いた認識結果を比較検討することで，ケプストラムに基付く特徴量を利用した HMM による[5]音声認識において，より有効的に作用する特徴空間の分布の状況を把握することができる。そしてその分析結果に基付いて，照合処理に対して新しい観点からの特徴操作を導入することができると考えられる。

[5] 本章での分析は特徴量，認識方式にある程度依存するところがあることは否めない。

実験条件と音声試料

表 8.1 に実験条件を，表 8.2 に音声試料について示す。音声試料は ATR 音声データベースの一部であり，音素の切り出しは，対数スペクトル包絡・パワー等の音響的特徴を基に，十分に訓練された labeler が視察により行なったものである。なお，認識用データの 3 種類の発声形態のうち，SA と SB の違いは，例えば「京都国際会議に」を一文節として定義する (SA) か,「京都/国際/会議に」と区分したものを文節として定義するか (SB) の違いである。両者とも，各々の定義に基付いて文節単位に発声した音声である。なお，SC は自然に発声したものである。認識の形態としては，話者 closed/テキスト open の，両端点固定の切り出し音素認識に分類されるものである。

実験結果

音声試料別 (SA, SB, SC) に各々の定義における認識結果を図 8.5 から図 8.7 にかけて示す。横軸が特徴部分空間の次元数 n であり，縦軸は，正解候補が一位として認識された率である。

考察と検討

図 8.5～図 8.7 より定義 1, 2 は，先行研究において定義した低精度 (大局的) 特徴 (定義 0) と比較すると，制限された n 次元で入力音声をより的確に表現していることが分かる。しかし，定義 3, 4 と比較すると n が小さい場合 (n=1～3)，大きな差がある。この定義 3, 4 で構成される特徴部分空間による認識は，n が小さい場合においても著しく認識率が高い。これは非常に効率良く，有効な成分を動的に捉えていることを意味している。しかし，図 8.2 にあるように，これらの定義では，n 次成分の決定タイミングが遅く，特に定義 4 では，最終的に算出される照合スコアそのものを用いて決定していることを考慮すると，当然の結果ではある。とは言うものの，着目する特徴空間を動的にかつ適切に変化させれば，1 次元で図に示すほどの認識能力があることは非常に興味深い。定義 3, 4 は標準パターンとの距離が最も大きい要素から参照している訳だが，これとは逆に，標準パターン間距離が最も小さい (つまり，出力確率密度の最も高い) 要素から参照した認識結果を表 8.3 に示す。表より，標準パターンとの距離の小さい成分の持つ識別能力は著しく低い (殆んど無い) ことが分かる。本実験で示された，音声認識に有効に作用する，標準パターンとの距離がより大きい (出力確率がより小さい) 要素は，情報論的に考えれば，“情報量のより多く含まれている要素[6]” と表現することもできる。更に，図 8.5～図 8.7 のいずれ

[6] いわゆる「情報量=$-p\log(p)$」が p に対して単調減少することを考えれば,「出力確率 (密度) のより低い成分」=「情報量をより多く担った成分」と言うことになる。

においても，次元数 6 辺りで**定義 1** が**定義 3, 4** を上回っている。これは，学習データと入力データとの分布のずれに起因するものと考えられるが，同一話者においてこのような現象が起きていることを考えると，多数話者における音声認識では，**定義 1** のような入力に依存した形で特徴部分空間を動的に制御する方法の有効性が増すことが予想される。

さて，以上の実験結果より，全次元で構成される特徴空間を用いた際の認識能力は，動的に変化する適切な少数の次元から構成される部分空間で，その殆どが記述されることが示された。この結果より以下のことが次なる研究の方向性として考えられる。

- 上述の "適切な" 次元を照合処理以前の段階で動的にかつ的確に予測することは可能か。もし可能であるならば，照合処理時間の大幅な短縮が期待できる。
- "適切な" 次元以外の非選択要素は本当に認識に貢献することは出来ないのか。それとも，処理手法あるいは非選択成分に対して用いている音響的表現手法の問題なのか。

前者の可能性を探る 1 つの手段として，**定義 1, 4** において，n 次元特徴空間を構成する成分の例を図示したものを図 **8.8** に示す。但し，左図が**定義 1**，右図が**定義 4** に相当する。横軸は時間，縦軸はケプストラムの次元を示し，濃淡によって参照の優先度を示している (濃=優先度高)。この例を見て分かるように，**定義 1** における部分空間の構成は大きく変化していないが，**定義 4** において部分空間を構成する成分は，非常に広くかつランダムに分布していることが分かる。また，スペクトル包絡の最も微細な変動を表す $N(=16)$ 次のケプストラムが最優先に選択されている例も見受けられる。これらは動的な予測が非常に困難であることを意味している。そこで本研究では前者については言及を控え，後者のテーマについて以後考察する。

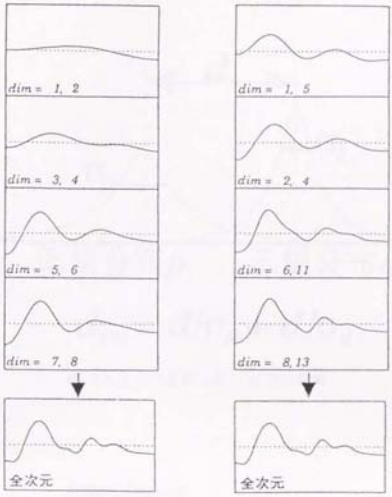


図 8.1. スペクトル包絡の復元

各包絡の左下に示してあるのが新たに加わった次元である。

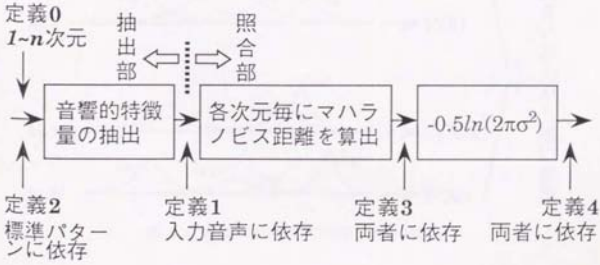
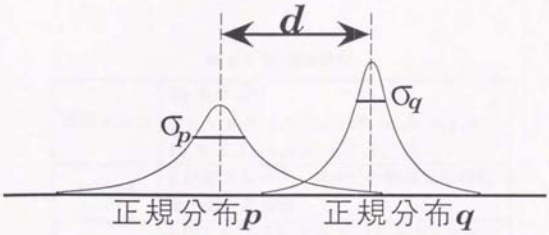


図 8.2. 各定義における n 次成分の決定タイミング



$$d_{pq} = d/\sigma_p + d/\sigma_q$$

図 8.3. 2 つの正規分布 (p,q) 間の距離

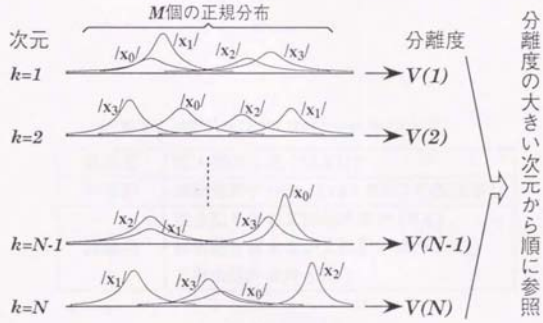


図 8.4. 標準パターン群に依存した動的制御

表 8.1. 認識実験条件

認識タスク	26 カテゴリ (/a,i,u,e,o,p,t,k,ch,ts,b,d,g,s, sh,h,z,dj,m,n,r,w,N,Q,j/)
HMM	4 状態 3 ループ、対角化分散共分散行列、 単一ガウス分布
音響的特徴	LPC ケプストラム 1~16(=N) 次元中 の $n(\leq N)$ 個要素
分析条件	16bit, 10kHz サンプリング、256 点ハミ ング窓、フレーム周期 5[msec]

表 8.2. 音声試料 (ATR 音声データベースより)

発話者	成人男性 1 名 (MAU)
学習用	単語発声データ (5240 単語) の偶数番目
認識用	複合語を含む文節発声音声 (SA) 複合語を含まない文節発声音声 (SB) 文自由発声音声 (SC)

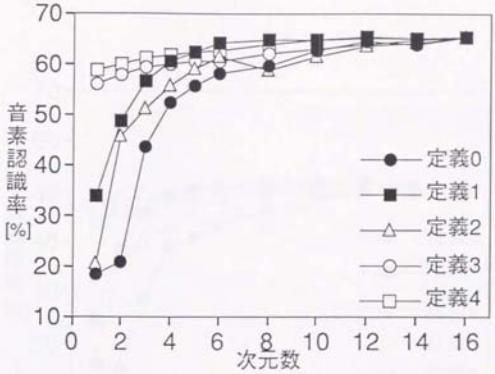


図 8.5. n 次元成分による認識 (SA)

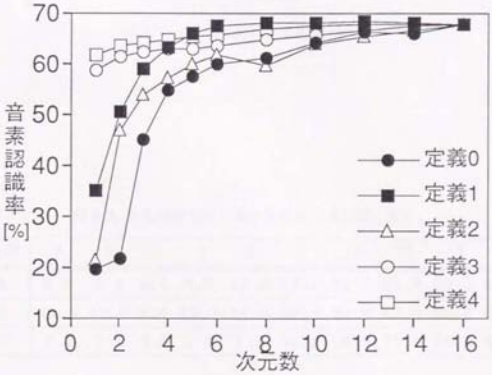


図 8.6. n 次元成分による認識 (SB)

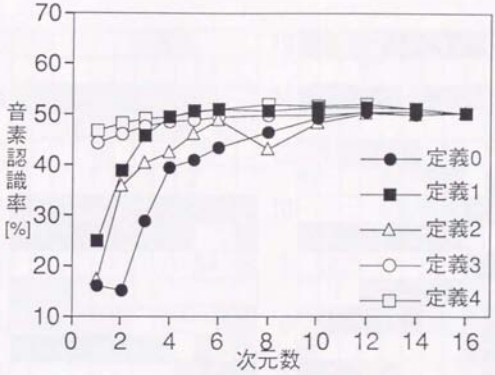


図 8.7. n 次元成分による認識 (SC)

表 8.3. 出力確率密度の高い次元による認識 ([%])

次元数	*1*	*2*	*3*	*4*	*6*	*8*	*10*	*12*	*14*	*16*
SA	9.9	9.4	8.4	8.3	12.8	21.7	31.7	42.8	55.2	65.4
SB	11.6	10.7	9.6	10.0	14.9	23.3	32.6	43.6	55.6	67.8
SC	7.2	7.1	6.0	5.9	7.9	11.1	16.3	24.1	34.8	49.9

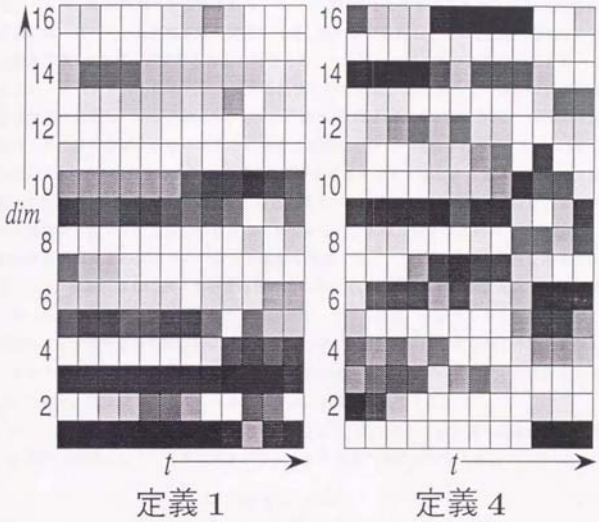


図 8.8. n 次元特徴空間を構成する成分

左側が定義 1，右側が定義 4 における n 次元特徴部分空間の構成の様子である。各時刻において，色の濃淡で選択される次元の優先度を表している (濃=優先度高)。定義 1 においては，低次の成分が選択される傾向にあることが分かるが，定義 4 においては，優先度の分布は非常に大きくばらつき，第 $N(=16)$ 次元の成分が最優先に選択される場合もある。

8.3 非選択成分の有効利用に関する実験的検討

8.3.1 非選択要素の表現方法

第 8.2.4 節で考察したように従来の認識手法では，抽出された音響的特徴はどの成分に対しても同一の表現形式が使用される。この表現形式としてケプストラムを用いた場合，各成分は，スペクトル包絡を直交関数系の足し合わせで表現した際の各重み (即ちフーリエ余弦係数) であり，逆変換も容易に行なえる。つまり，ケプストラムをそのまま用いることは，スペクトル包絡の形状に関する情報を何ら削減することなく，音声の音韻情報を直接的に表現していると言える。さて，**定義 3, 定義 4** によって選択された "適切な" 成分は，その数が非常に少ない場合でも非常に効率良く正しい認識を行なっていた。即ち，上記の "適切な" 成分とはあるカテゴリの音声を他カテゴリと識別する意味において，直接的に記述される必要がある成分であると考察される。逆に言えば，直接的に記述されるべき情報量を担っている成分と言うことができる。これは，第 8.2.4 節の「考察と検討」における情報量との関連から行なった考察を考慮すると，情報論的にも妥当な考えである。一方，非選択成分の担う情報量は相対的に少なく，その結果直接的に記述される必要の無い成分と言うことになる。ここで問題となってくるのが，上記のような性質を持つと考えられる非選択成分を直接的に記述することが，音声認識を妨げる方向に作用しているか否かと言うことである。あるいは，非選択成分に対して適切な (間接的) 表現方式を与えることで，認識能力を上げられるかどうか，と言うことである。

以上の考察の下，本研究では非選択成分に対して，ケプストラムと言う直接的表現に替わる間接的表現として，以下に示すパラメータを導入することを考える。

$$\alpha_i(t) = \frac{c_i(t)}{\sum_{i=1}^{N} |c_i(t)|} \tag{8.1}$$

ここで，$c_i(t)$ は時刻 t のケプストラムの第 i 次要素である。音声からパワーの情報を除去するために 0 次のケプストラム (Log スペクトルの平均値) がしばしば特徴量から除かれるが，上記のパラメータはケプストラム (1〜N次) に対してその大きさの平均値 (あるいは和) の情報を取り除いたものである。つまり，$\alpha_i(t)$ は正負の情報を残した $c_i(t)$ の正規化であり ($\sum_{i=1}^{N} |\alpha_i(t)| \equiv 1.0$)，絶対的な大きさの情報が欠落しているため，スペクトル包絡を間接的にしか表現できない形になっている。また，この$\alpha_i(t)$ は第 8.2 節で議論した "第 i 次成分のスペクトル包絡に対する記述力" と捉えることもできる。なお次節では，ケプストラムをα化することの物理的意味について考える。

8.3.2 $\alpha_i(t)$ 化の物理的意味

$c_i(t) \rightarrow \alpha_i(t)$ への変換は，数学的には「正負の情報を残した正規化」と言えるが，物理的にはどの様な現象として捉えることができるのだろうか。ここでは，$\alpha_i(t)$ そのものを扱う代わりに，$\alpha_i(t)$ の近似として以下で定義される$\alpha_i'(t)$ を考える。

$$\alpha_i'(t) = \frac{c_i(t)}{\sqrt{\sum_{i=1}^{N} c_i(t)^2}} \tag{8.2}$$

つまり，分母として絶対値の和をとる代わりに 2 乗和の平方根を用いるのである。この様に$\alpha_i'(t)$ を定義すると，

$$\sum_{i=1}^{N} \alpha_i'(t)^2 \equiv 1.0 \tag{8.3}$$

が成立する。ここで左辺は Parseval の定理より，$\alpha_i'(t)$ をケプストラムと仮定して復元したスペクトル包絡を一波形として見なした場合の，その波形のパワー密度である。つまり $c_i(t) \rightarrow \alpha_i'(t)$ への変換を施すことで，パワー密度が常に一定となるようにスペクトル包絡波形を上下に伸縮していることになる。以上のことを概念的に図示すると，図 8.9 のようになる。左側の図はある同一のスペクトル包絡を上下に伸縮しながら連ねたものであるが，この一連のスペクトル包絡を$\alpha_i'(t)$ 化すると，右図のように全く同一の波形の連続となり，スペクト包絡の "うねり" の度合を一定にする効果があることが分かる。そして，$\alpha_i(t)$ の場合も，この$\alpha_i'(t)$ と同様な効果があると考えられる。この図より明らかなように，この操作はスペクトル包絡時系列に対して情報量削減の働きを持ち，ケプストラムとして抽出された特徴量を低情報量化するものと捉えることができる。即ち，第 5.3.1 節で言うところの "低情報量" に対応する音響的表現方法の一つであると言える。但し第 5.3.1 節で例示した低情報量化とは次の点で異なることに注意すべきである。即ち，本節で述べた低情報化 (間接表現) は，ある時刻の音声信号から得られる特徴量の一部 (非選択成分) に対して行なう操作であり，特徴量全体に対しての操作として定義される第 5.3.1 節の低情報量化とは質的に異なるものである。以下第 8.3.3 節，第 8.4 節において，本節で提案した，非選択成分をα化することによる効果を実験的に検討することにする。

8.3.3 非選択成分に対する間接的表現の有効性

スペクトル包絡を間接的に表現する特徴パラメータの一つである$\alpha_i(t)$ を，非選択要素に応用することの有効性を検証するために，8.2.4 節と同様，特定話者 (MAU) に対する切り出し音素認識実験を行なった。

実験条件と音声試料

実験条件及び音声試料は第 8.2.4 節で述べたものと同一であるが，本実験を実施するに当たって以下の点に注意した。$\alpha_i(t)$ は第 i 次成分の全成分に対する相対的な大きさであり，全成分との関係を示す一指標であると言える。故に$\alpha_i(t)$ の使用によって何らかの効果が得られた場合，それは "直接的/間接的" と言った音響的表現方式の問題ではなく，分散共分散行列の対角化[7]に対する補間として作用しているとの議論も可能である。そこで，分散行列の非対角成分 (各成分間の相関) が 0.0 になるように音韻毎に KL 展開し (軸を回転させ)，軸を再設定する[119]ことによって得られる $c_i(t)$(F，但しこの場合の $c_i(t)$ は正確には主成分値である)，及び F に対して$\alpha_i(t)$ を求めて特徴量とする方法 (G, H, I) をも含めて実験を行なった。更に比較実験として，混合数 8 の連続 HMM[8](J) についても実験を行なった (J 以外の実験は全て単一ガウス分布による HMM である)。結局，音響的特徴空間の動的制御及びその表現方式として表 8.4 の 10 通りの方法を用いて比較した。ここで C~E 及び G~I において，選択要素に $c_i(t)$ を非選択要素に$\alpha_i(t)$ を適用している。なお C~E，G~I では，HMM の学習は特徴パラメータを $2N$次元ベクトルとして行ない，認識時に $2N$次中の N次 ($c_i(t)$ $N/2$ 次，$\alpha_i(t)$ $N/2$ 次) を使用して照合を行なう。

実験結果

SA~SC に対する A~J の制御方法・表現方式による認識実験結果を累積認識率 (1 位・3 位) の形で図 8.11~図 8.13 にかけて示す。なお，各々の図で縦軸 (音素認識率) のレンジが異なっていることを断っておく。

考察と検討

A, B より，$c_i(t)$ 全次元に対して単純に$\alpha_i(t)$ を考慮すると，認識を妨げる方向に働いてしまうことが分かる。一方第 8.3.1 節で考察した様に，非選択要素のみに$\alpha_i(t)$ を適用した場合 (C, D, E) はいずれも，全次元を $c_i(t)$ で記述した A より認識率の向上が見られる。特に，学習データと発声形態が最も異なる SC での向上が大きい。

また，音韻毎に無相関化させる様に軸を設定し，照合の度に軸変換して認識させた場合 (F) の効果が参考文献 [119] と同様，ここでも観測されている。しかしこの場合でも間接的表現形式 ($\alpha_i(t)$) の導入で，さらに認識率が向上している (G)。

以上の結果は，$\alpha_i(t)$ が，分散共分散行列の対角化を補正する形で働いているのではな

[7] HMM の計算に対角化分散共分散行列を使用しているため。
[8] 一状態における学習データの分布の様子を単一多次元正規分布で近似するのではなく，複数個の正規分布の線形結合で表現するものである。

く，抽出された音響的特徴の各成分の特徴を適切に反映した表現方式であることを意味している。その結果，スペクトルを直接的に記述すべき成分に対しても$\alpha_i(t)$のような間接的表現を導入すると認識率は低下する (B)。

また，状態数を 8 に増やすことにより，認識率は A から J へと上昇しており，これは学習データの特徴分布が単一の正規分布では十分に表現できていないことを示している。しかしこの結果と，KL 展開+$\alpha_i(t)$ 化の結果である G~I と比較すると後者の方が高い値を示している。J と同様に F も図 8.10 に示すように，学習データの特徴分布に対して単一正規分布を直接当てはめただけでは，十分に近似出来ないとの考察の下，提案された手法である。そして各々,「軸の方向に自由度を与えた (即ち軸の回転を許した) 単一正規分布で近似する (F)」,「複数の正規分布の和として確率密度関数を定義して近似する (J)」ことを実現した手法である。そして，これらが十分に学習データの特徴分布を近似できていると仮定するならば，G~I での認識率向上をもたらした非選択成分に対する間接的表現方式は,「学習データの特徴分布を如何に効率良く近似するか」をテーマとして行なわれている従来の研究とは質的に異なる,「音声を音響的に表現する際の特徴パラメータの各成分の担う情報量」に着目した，新たな観点からの特徴操作方法であると言える。

更に非選択成分のみに着目した場合は，どのような差が生じるのだろうか。各時刻において$\alpha_i(t)$ の小さい ($|c_i(t)|$ の小さい) $N/2$ 個の成分のみを動的に求め，照合処理においてはその $N/2$ 個のみを使用して行なった認識実験結果を図 8.14 に示す。但し，$c_i(t)$ をそのまま用いた場合と$\alpha_i(t)$ に変換して用いた場合の 2 通りの結果が示されている。図より，$\alpha_i(t)$ 化することにより認識力が大幅に向上していることが分かる。これは，非選択成分に対して間接的表現を用いる本手法の有効性を十分に示しているものである。

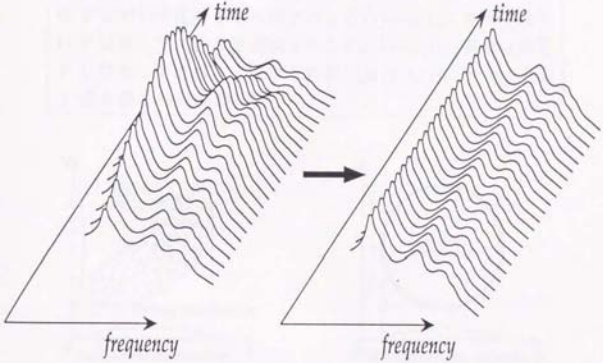


図 8.9. スペクトル包絡波形の正規化 (概念図)

左図に示されるスペクトル包絡の時系列に対して，$\alpha'_i(t)$ 化することにより，各スペクトル包絡のパワー密度は一定となる (右図)。この図ではその様子がよく分かるように，正規化後のスペクトル時系列が，同一包絡波形の連続となるよう，原スペクトルを設定している。

表 8.4. 音響的特徴の表現方法

A $C_i(t)$, N次元 (従来法)
B $C_i(t)+\alpha_i(t)$, $2N$次元
C 定義 1 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
D 定義 3 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
E 定義 4 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
F 音素毎に KL 展開した $C_i(t)$, N次元
G F に対して定義 1 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
H F に対して定義 3 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
I F に対して定義 4 で選択される $C_i(t)+\alpha_i(t)$, 各々$N/2$次元
J 混合数 8 の連続 HMM

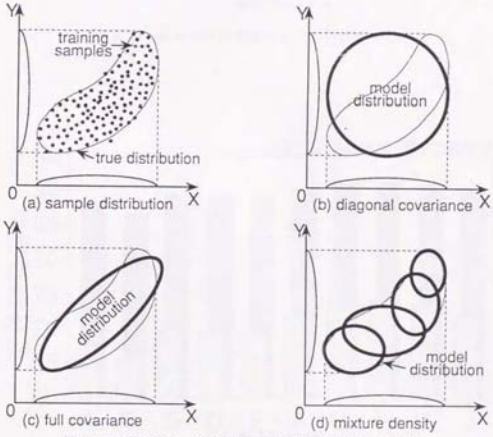


図 8.10. 学習データ分布と種々の分布関数との関係

対角化分散共分散行列を使用した単一正規分布 ((b)) では，軸の方向が x, y方向に固定されてしまう。そこで軸を回転させる (KL 展開) ことで柔軟に学習データの分布を近似する ((c) と同意) 方法や，複数の正規分布の足し合わせで学習データの分布を近似する方法 ((d)) が提案された。

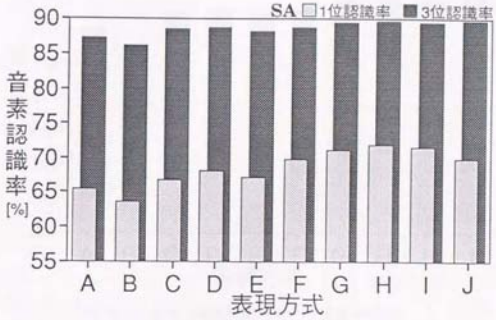


図 8.11. 種々の特徴表現の下での認識率 (SA)

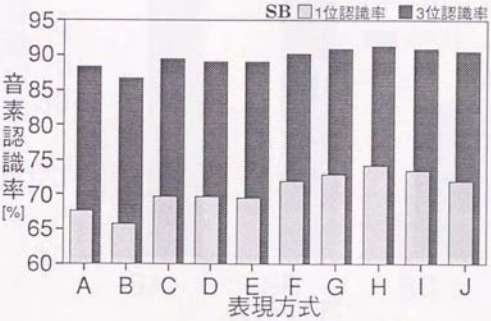


図 8.12. 種々の特徴表現の下での認識率 (SB)

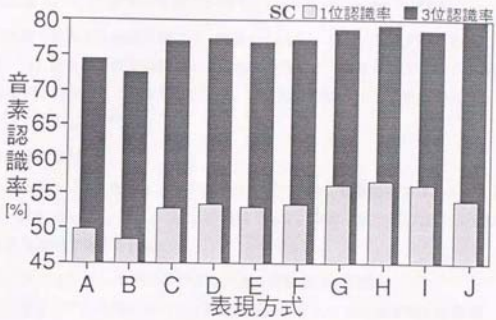


図 8.13. 種々の特徴表現の下での認識率 (SC)

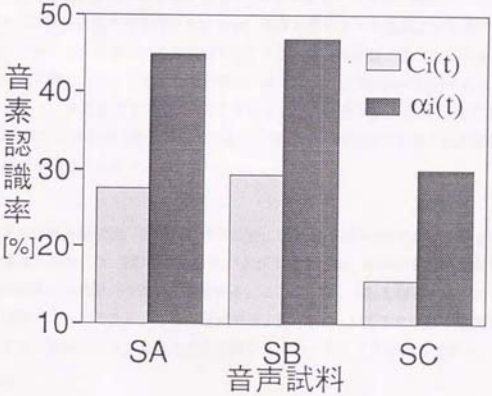


図 8.14. 記述力の小さい成分のみによる認識結果 (SA~SC)

8.4 提案した手法の話者依存性に関する実験的検討

第 8.3.3 節の切り出し認識実験では，話者を MAU に限って実験が行なわれていた。そこで本節では，第 8.3.1 節で定義された正規化特徴量の，複数の話者に対する有効性，或は話者依存性を観測することを目的として，特定話者の切り出し音素認識実験を，複数の話者に対して行なった[120]。

実験条件と音声試料

表 8.5 に実験条件を，表 8.6 に音声試料について示す。複数話者ではあるが，各々話者に対して，話者 closed，テキスト open の実験である。なお，選択/非選択の決定は Viterbi スコア計算の途中で算出される，

1. LPC ケプストラムの絶対値のより大きな n 個要素 (定義 1)
2. 入力/標準パターン間のマハラノビス距離のより大きな n 個要素 (定義 4)
3. 出力確率密度のより小さな n 個要素 (定義 0)

の 3 通りについて実験的に検討する。また，選択要素と非選択要素を組合わせて照合する場合，"1~3 で動的に決定される選択要素 $N/2$ 次元+正規化特徴量へ変換される非選択要素 $N/2$ 次元" の計 N 次元の形で用いられるが，学習の際は 2 つの表現形式を合わせた $2N$ 次元のベクトルとして音響モデルを作成する。また，第 8.3.1 節で述べたような，非選択成分を直接的に扱うことにより生じる認識力の低下について実験的に観察できるよう，N 次元特徴パラメータに全て LPC ケプストラムを用いた認識実験，及び 1~3 で選択される $N/2$ 次元のみによる認識実験も併せて行なう。最終的に本実験で使用された音響的特徴量の表現形式を表 8.7 に示す。

実験結果

一連の実験結果を話者別，音声試料の形態別に図 8.15~図 8.29 にかけて示す。なおスペースの関係上図中には，縦/横軸の説明がされていないが，縦軸は音素認識率を，横軸は音響的特徴量の表現形式の違いを意味する。ここで，M_1, M_2 とは各々B~D, E~G の最大値を抽出したものである。また，認識率は 1 位・3 位・5 位までの累積認識率の形で示されており，縦軸のレンジも各々の図で異なっていることを予め断っておく。

考察と検討

まず，A と B~D(+M_1) を比較すると，MMS の SA, SB を除いたケースにおいて非選択成分を間接的に表記することの効果が見られる。特に学習/認識データ間の発話形態

が最も異る SC における効果が大きい。また，B~D 間の差であるが，最大値を示すものが話者によって異るなどの現象はあるが，E~G に見られるような大きなばらつきはない。

次に，A と E~G(+M_2) を比較すると，N次元全体を使用した認識が必ずしも，その部分空間のみを使用した認識を上回る訳ではなく，MHT, MTK, MMY において，適切な$N/2$ 次元を用いた認識が N次元の認識を上回る結果を出している。しかも，中には B~M_1の認識率さえも上回る結果を示しているものもある。これは第 8.3.1 節で考察したように，全ての次元の成分が，直接的に表現されるべく情報量を担っている訳ではなく，かつ，一部の要素を照合処理に直接反映させると，音響モデルの識別能力を低下させる方向に働くことを意味する。しかし，どの成分に限定すべきか (どの成分を除くべきか) に関しては MTK と MMY を比較すれば分かるように，話者依存性が観測され，一意には決定できない。しかも E~G 間の差が非常に大きいため，安易に特徴量の制限を行うのは非常に危険である。これらに比べて，選択要素を直接的に，非選択要素を間接的に表記した B~D での結果はどれも安定した値を示している。

第 8.3.3 節で述べたように本章で提案した手法は，「如何にすれば，学習データの特徴分布を効率的に，精密に近似できるのか」と言う，従来行なわれてきた研究とは，その性質を異にしており，特徴パラメータの各成分が「ある話者，あるカテゴリを識別するに当たり，どの程度の情報量を担っているのか，どの程度着目されるべきなのか」と言う観点から考案・提案した手法である。しかし，本節での実験結果を見る限りでは，より多くの情報量を担っている成分の分布に対する話者依存性は否定できず，話者性を含めた動的適応の実現が今後の課題の一つである。

表 8.5. 認識実験条件

認識タスク	26 カテゴリ (/a,i,u,e,o,p,t,k,ch,ts,b,d,g,s, sh,h,z,dj,m,n,r,w,N,Q,j/)
HMM	4 状態 3 ループ、対角化分散共分散行列、単一ガウス分布
音響的特徴	16 次 LPC ケプストラム及びその正規化特徴量
分析条件	16bit, 10kHz サンプリング、256 点ハミング窓、フレーム周期 5[msec]

表 8.6. 音声試料 (ATR 音声データベースより)

発話者	成人男性 5 名 (MAU,MHT,MTK,MMY,MMS)
学習用	単語発声データ (5240 単語) の偶数番目
認識用	複合語を含む文節発声音声 (SA) 複合語を含まない文節発声音声 (SB) 文自由発声音声 (SC)

表 8.7. 音響的特徴量の表現形式

A	$C_i(t)$、N次元 (従来法)
B	1. で選択される $C_i(t)+\alpha_i(t)$、各々$N/2$次元
C	2. で選択される $C_i(t)+\alpha_i(t)$、各々$N/2$次元
D	3. で選択される $C_i(t)+\alpha_i(t)$、各々$N/2$次元
E	1. で選択される $C_i(t)$、$N/2$次元
F	2. で選択される $C_i(t)$、$N/2$次元
G	3. で選択される $C_i(t)$、$N/2$次元

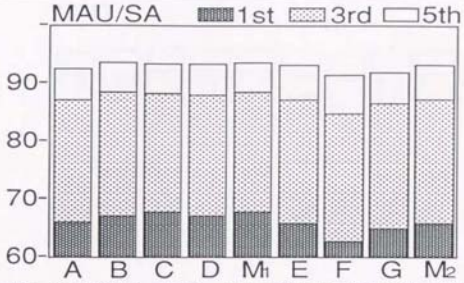


図 8.15. 発話者 MAU・音声データ SA に対する切り出し音素認識結果

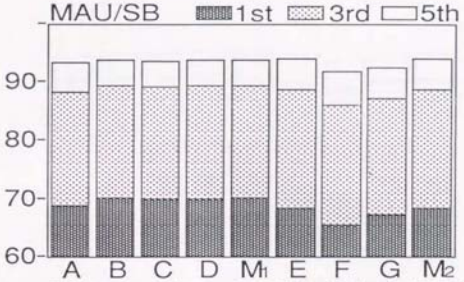


図 8.16. 発話者 MAU・音声データ SB に対する切り出し音素認識結果

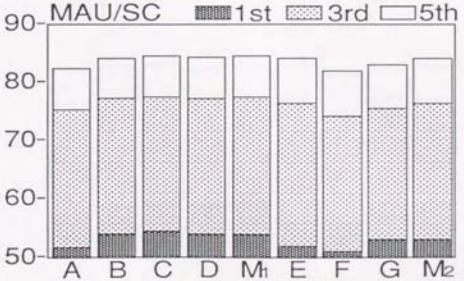


図 8.17. 発話者 MAU・音声データ SC に対する切り出し音素認識結果

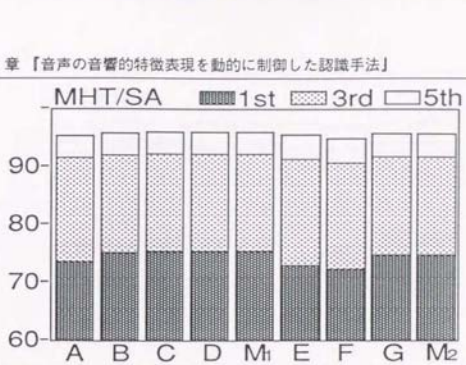


図 8.18. 発話者 MHT・音声データ SA に対する切り出し音素認識結果

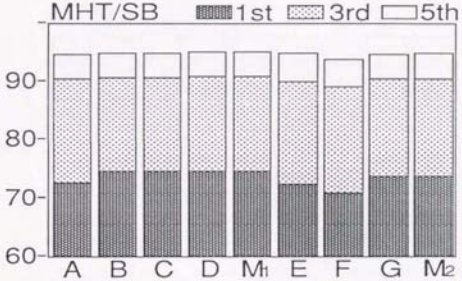


図 8.19. 発話者 MHT・音声データ SB に対する切り出し音素認識結果

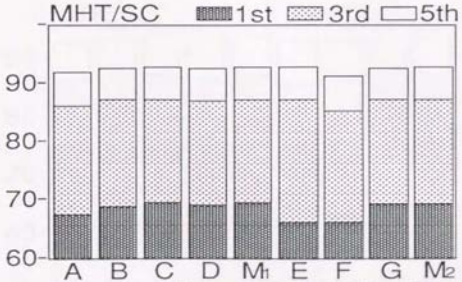


図 8.20. 発話者 MHT・音声データ SC に対する切り出し音素認識結果

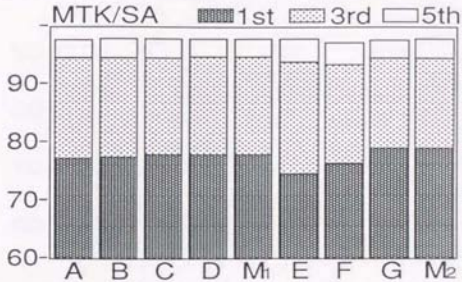


図 8.21. 発話者 MTK・音声データ SA に対する切り出し音素認識結果

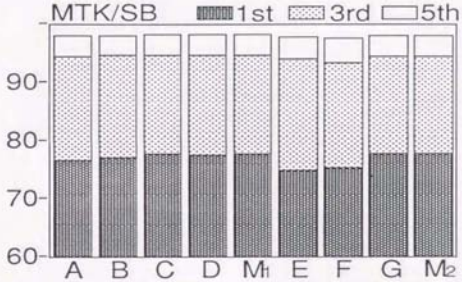


図 8.22. 発話者 MTK・音声データ SB に対する切り出し音素認識結果

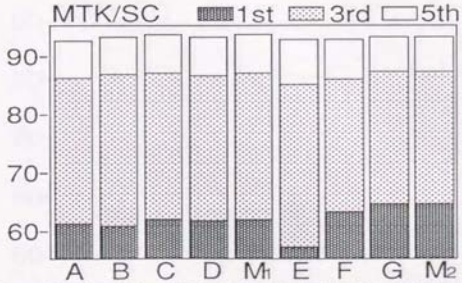


図 8.23. 発話者 MTK・音声データ SC に対する切り出し音素認識結果

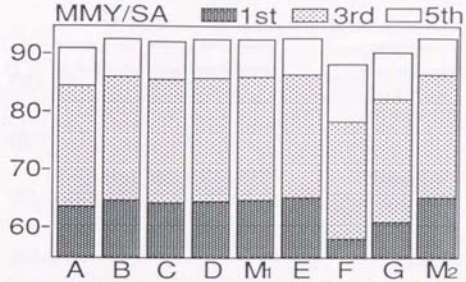


図 8.24. 発話者 MMY・音声データ SA に対する切り出し音素認識結果

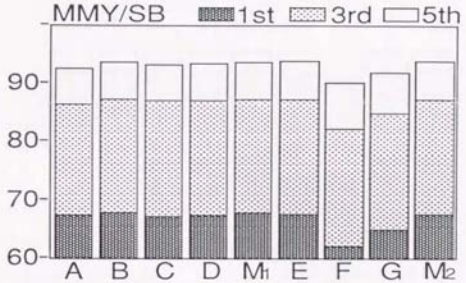


図 8.25. 発話者 MMY・音声データ SB に対する切り出し音素認識結果

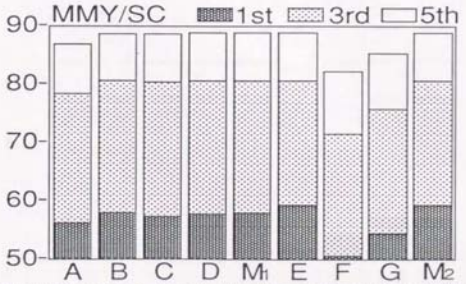


図 8.26. 発話者 MMY・音声データ SC に対する切り出し音素認識結果

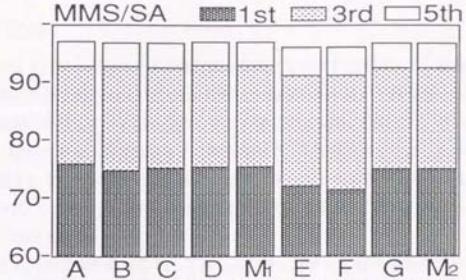


図 8.27. 発話者 MMS・音声データ SA に対する切り出し音素認識結果

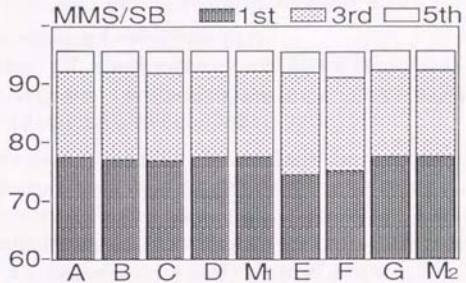


図 8.28. 発話者 MMS・音声データ SB に対する切り出し音素認識結果

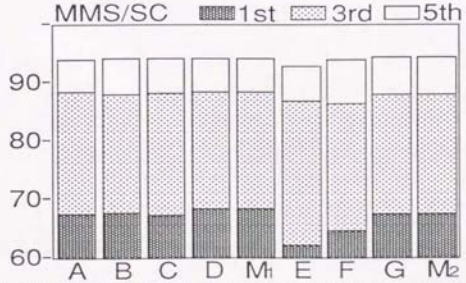


図 8.29. 発話者 MMS・音声データ SC に対する切り出し音素認識結果

8.5 まとめ

本章ではまず，先行研究に対する問題点を解決することを目的として，音声認識に使用される音響的特徴空間を $n(<N)$ 次元に制限した部分空間を用いて照合させることで，音声認識に有効に寄与している部分空間の分布の様子を実験的に検討した。その結果，動的には大きく変化するが，非常に限られた部分空間で，全特徴空間を使用した際の認識力の殆どが記述されることが示された。特に，標準パターンとの距離がより大きな成分 (出力確率のより低い，即ち，より多くの情報量を担っている成分) に認識力が偏って分布していることが観測された。

次に残された特徴成分に着目し，これらの成分に対する音響的特徴表現形式を，音声を間接的にしか表現できない形式 (正規化特徴量) へ変換することで認識率の向上を図ることに成功した。更に，複数話者に対して話者依存性を検討したところ，

1. 正規化特徴量の効果はおおよそ認められた。
2. 特徴量の一部は，直接的に記述されることで，認識を妨げる方向に作用している。特徴量の適切な制限 (正規化特徴量の導入は無し) により認識率の向上が観測された。
3. 上記の適切な制限方法は話者によって異る。

などの知見が得られた。話者 MAU に対しては HMM における出力確率分布の混合化以上の効果が見られており，異なる話者に対しても同様の結果が期待できる。しかし，**2.** **3.** に関しては HMM における出力確率の分布数との関連は否めず[9]，動的な話者適応も含め，今後の研究課題である。また本章では扱わなかったが，認識能力の高い少数成分 ("適切な" 成分) の動的予測の可能性も非常に興味深い課題である。

[9] 単一ガウス分布と言う実験条件がこのような結果を導いた可能性は否めない，と言うこと。

第 9 章

クラスタリングによる HMM
継続時間長制御の高精度化

本章では第 7 章で述べたように，HMM を純粋にパターン認識の立場から捉え，その認識能力を向上させるための一手法である，継続時間長モデルの導入に対して新しい観点からの分析を行なう。そしてその結果に基付き，学習データのクラスタリングによる継続時間長モデルの高精度化手法を提案する[30][121]。なお以下では継続時間長モデルのことを簡単のため，継続長モデルと呼ぶことにする。

9.1 本研究の背景と目的

第 7 章に述べたように音声認識の分野では，音声の時系列性に基付いて，left-to-right の形態を持つ HMM を用いた音響的モデリングが広く行なわれてきた[122](図 7.4 参照)。しかし第 7.1.2 節で概説した，いわゆる基本 HMM[1] と呼ばれる HMM を標準パターンとした場合，入力パターンに対して (原理的には) 無制限の非線形マッチングを許してしまうことになり，時間方向の揺らぎを極端に吸収してしまう恐れがある。その一方で基本 HMM は，以下に示すような歪んだ時間構造を持っている[123]。即ち，状態遷移パスが状態 i に時間 $t_i(=0,1,2,\ldots)$ だけ留まる確率 $P_i(t_i)$ は，

$$P_i(t_i) = {a_{ii}}^{t_i} \times (1 - a_{ii})$$

となり，指数関数的に減少することとなる[2]。但し上式で，a_{ij} は状態遷移パスが状態 i に到達している場合に，次に状態 j に遷移する確率である。基本 HMM における後者の特性は，母音など，その中心的特徴に定常部 (即ち一状態で表現されるべき特徴) が存在する音韻に対しては不適切なモデリングを行なうことになる。つまり，基本 HMM は時系列データが持つ時間構造を反映することが困難なモデリング手法と言える。

このような状況を打破すべく，音声の時間構造を基本 HMM に導入する技術が考案され，広く使用されるようになった。即ち，標準パターンと入力パターン間の照合パスに相当する状態遷移が，HMM 内の各状態においてどの程度停留するのかを学習データを用いて推定し，継続長モデルとして基本 HMM に組み込むのである。

しかし従来の研究においては，最終的な認識率の向上[3]ばかりが着目され，評価の対象となっていた感がある。即ち，作成された継続長モデルと学習データの持つ時間構造がどの程度整合しているのか，即ち継続長モデルの正当性に対して詳細に，定量的に分析した例が少なく，そのため，継続長モデルを導入したことで認識率が低下した場合，その理

[1] ここでは，継続時間長制御が行なわれていない HMM のことを指す。
[2] 即ち，状態 i に時間 t_i だけ停留する確率として，二項分布を仮定していることと等しい。
[3] 時には低下。

由付けが曖昧なままとなっている。そこで本研究ではまず，比較的簡単に算出することができる継続長モデル作成方法を取り上げ，その継続長モデルが表現する時間構造と実際の(学習データ用) 音声が持つ時間構造との整合性を幾つかの指標を定義した上で定量的に分析する。この場合，整合性の評価が継続長モデル作成法に依存することは厳密には否めない。しかし，その作成法に依らず，継続長モデルは状態 i に停留する時間長を (ある仮定の下に) 推定した分布関数であり，ある特定の作成法によるモデルが示した傾向は，継続長モデルの作成法に依らずおおよそ当てはまるものと考えられる。以上の考察の下，本論文では継続長モデルの作成法は第 9.2 節で述べる方法に限定し，仮定する確率分布関数の形態の変化，或は学習データのクラスタ数の変化による継続長モデルの効果の増減を比較することをその主旨とする。即ち，従来の継続長モデル作成法に対して不十分であったと思われる側面を指摘し，新しい観点からの継続長モデルの高精度化について論じることを目的とする。

作成された継続長モデルと学習データとの時間的整合性の分析結果より，幾つかの音素モデルでは，その時間的対応におけるばらつき (以下，時間構造変動と示す) が非常に大きくなっていることが示される。当然のことながら，多くの音韻モデルでは時間構造変動は許容範囲内であり，この場合は従来の継続長モデル作成アルゴリズムを適用することで，十分に学習データの時間構造を反映したモデルが作成される。しかし，継続長モデルは基本的には使用する学習データが持つ時間構造の平均 (及びその分散) を用いて作成されるため，時間構造変動が極端に大きな音韻モデルに対しては，平均化の結果，不適切な継続長モデルが作成されてしまう可能性が十分にある。この場合，継続長モデルの有効性が下がるばかりか，継続長モデルの導入による認識率の低下も予測される。そこで本研究では，このような時間構造変動の大きな音韻に対しても，十分に継続長モデルが有効に作用するよう，2 つの方法を考案し，各々の有効性について実験的に検討することにする。

9.2 本研究で使用する継続時間長モデル

継続長モデルと学習データとの時間的整合性に対する分析を行なう前に，本研究で使用する継続長モデルの構造・作成法・利用法などについて説明する。一般に，基本 HMM に対して音声 (学習データ) の時間構造を反映させる方法として以下の 3 つが考えられる。

1. 状態数を多くする方法
2. 後処理法
3. HMM 内に継続時間長モデルを組み込む方法

元来 HMM は十数~数十フレームの音響的特徴を，数個の状態及びそれらの間の遷移で表現すると言う，時間方向での情報圧縮に基付くモデリングであるため，手法 1 のようにその状態数を上げれば自ずと学習データの時間構造をより直接的に反映するようになる。しかしこの場合，単に状態数を上げるだけでは，パラメータ数の増加を招き，その結果，推定精度を落すだけになる。そこで，ある条件の下，異なる遷移間で遷移確率を等しくしたり，或は，異なる状態間で出力確率を等しくするなど，いわゆる「結び」の技法を導入することが不可欠となる。図 9.1 に，状態 i と状態 j の遷移の間に更に状態を付け加えたモデルを示す。この場合状態 i，j 以外で，かつ隣り合う状態間を「結び」の関係にすることが多い。なお，この「状態数の増加」と言う考えの延長線上にあるのが，1 状態 =1 フレームとする Stochastic DP 法である。

手法 2 の後処理法とは，間接的に継続長を制御する方法である。本手法では，学習データと作成した基本 HMM を再度照合させ，その結果得られる最適パス (Viterbi パス) に基付いて，各状態での停留時間の分布を推定する。この場合，分布をヒストグラムを用いて表現したり[37]，ガンマ分布やガウス分布を用いて近似する方法が行なわれている[39]。但し，前者の場合はスムージング等の処理を必要とする場合がある。作成された継続長モデルの認識時における適用法は，以下の通りに行なわれる。まず基本 HMM と入力パターンとをマッチングし (尤度計算)，その後，Viterbi パスから得られる各状態における停留時間と継続長モデルを考慮して，尤度を後処理的に修正する (図 9.2 参照)。しかし，認識に Viterbi アルゴリズムを利用する場合は，継続長モデルによる尤度修正を認識アルゴリズム中に直接組み込むことが出来，この場合手法 3 との違いは，継続長モデル作成方法のみとなる。但し，参考文献 [37] によれば，Viterbi アルゴリズムに直接組み込んだ場合，計算量が 15~20 倍に増大したものの，効果は単純な後処理法と全く同じであったと報告されている。

手法 3 においては，Baum-Welch アルゴリズムを用いて学習データから HMM を作成する際に，継続長モデルも全く同一の方法で学習・推定してしまおうと言うものである。その結果，HMM の他のパラメータは継続長の分布の影響を受けて学習・推定されることになる。但しこの場合，ガンマ分布，Gaussin 分布などを仮定せずに直接分布形を推定してしまうと，凸凹の激しい分布が得られることもあり，スムージングなどの後処理も必要となる。音響モデルの各パラメータが同一の方法で，かつ，その相互作用を考慮して計算されることは望ましいことである。しかし，この手法ではその計算量はかなり増える。また本手法の場合，認識における Viterbi アルゴリズム中に直接継続長モデルを組み込ん

で尤度計算を行なうことが多いが，この場合も，基本 HMM による認識と比較すると，その計算量はかなり増えることになる (図 9.3 参照)。

さて，従来の研究において用いられた継続長モデルでは，ある状態 i に時間 t_i だけ停留する確率を $P_i(t_i)$ のように，状態独立に各々一変数の確率 (密度) 関数として推定することが多い。しかし，あるカテゴリに属する音声の長さがある一定範囲内に収まっているとするならば，状態 i での停留時間が長ければ，状態 $j(\neq i)$ での停留時間が短くなるなどの (負の) 相関関係が生まれてくるのは必須のことであろう。即ち，各状態独立に継続長モデルを推定するのではなく，各状態における停留時間長の相関を考慮し，同時確率 $P(t_1, t_2, ..., t_n)$ として継続長モデルを構築することにより，より一層の効果が期待できる。また，ある分布形を仮定して推定する場合でも，HMM におけるスペクトル領域の特徴分布近似では常套手段となっている混合分布の考えを導入した例も見ない。継続長の分布が分布関数一つで十分に近似可能であれば，あえて混合形にする必要もないであろうが，一分布関数による近似可能性 (適合性) についての十分な議論もあまりされていないように思う。

本研究における継続長モデルは上記の考察の下に，その基本構造が決定された。まず，同時確率 $P(t_1, t_2, ..., t_n)$ とすることで，各状態間における相関を考慮したモデル化を試みる。この場合上記の手法 1 を用いると，推定すべきパラメータ数は更に増えることになり，あまり望ましい選択とは言えない。一方手法 3 による方法を用いた場合を考えると，学習及び認識処理における計算時間の一層の増加を招くこととなり，これも積極的に支持できる方法とは言えない。そこで，モデルの推定及び適用方法としては，上記の手法 2 の後処理法を採用することにする。この場合まずカテゴリ **C** の基本 HMM を求め，使用した **C** の学習データと **C** の HMM をマッチングし，Viterbi パスを求める (以降，この操作を "rematching" と呼ぶ)。参考文献 [37] では，各状態に n 回留まった Viterbi パスの数を数え，継続長モデルをヒストグラムの形で求めていた。しかし，この場合分布における凸凹の発生を抑えることができない。そこで本研究では，得られた Viterbi パスより，状態 i に停留する平均時間及び分散を求め，ある分布関数を用いて，状態 i に停留する時間長の分布を推定することにした。従来継続長モデルにおいて利用される分布関数としてはガンマ分布が代表的であり，その効果もガウス分布より高いことが報告されている[124]。しかし，多次元ガンマ分布として定式化されているウィシャート (Wishart) 分布は，

$$f(W) = \frac{c|W|^{(n-p-1)/2}}{|\Sigma|^{n/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\Sigma^{-1}W\right)$$

$$\text{但し}, \boldsymbol{W} > 0, \quad \boldsymbol{\Sigma} > 0, \quad p \leq n$$

のような分布形をとり，確率変数として行列を要求する[125]。ここで，$\boldsymbol{W} > 0, \boldsymbol{\Sigma} > 0$ とは各々の行列が正値であることを意味し，そうでない場合は，$f(\boldsymbol{W}) = 0.0$ である。また，c は定数であり，

$$c = \left[2^{np/2} \pi^{p(p-1)/4} \prod_{j=1}^{p} \Gamma\left(\frac{n+1-j}{2}\right) \right]^{-1}$$

となる。しかし，入力音声と基本 HMM との照合結果 (Viterbi パス) から得られる確率変数の実現値は，各状態での停留時間を一要素としたベクトルであり，多次元ガンマ分布を継続長モデルに直接導入するのは困難であることが分かる[4]。そこで，本研究では，同時確率化することによる効果を観察することを第一義とし，ガウス分布を仮定した推定を行なうこととした。また，従来の研究では考慮されていない，混合分布形による継続長モデルについても実験的に検討することにする。

[4] このために，従来継続長モデルに対して多次元の同時確率が用いられなかったとも考察される。

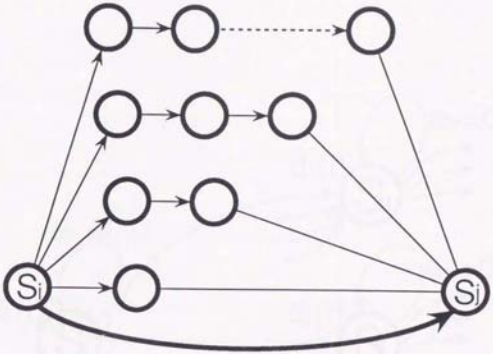


図 9.1. 状態数を上げることによる時間構造の加味

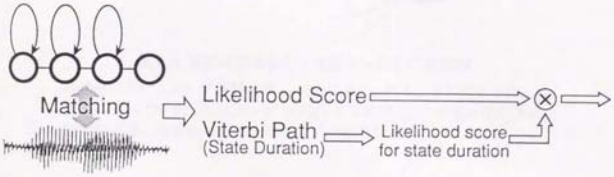


図 9.2. 尤度計算の後処理として継続時間長を考慮する方法

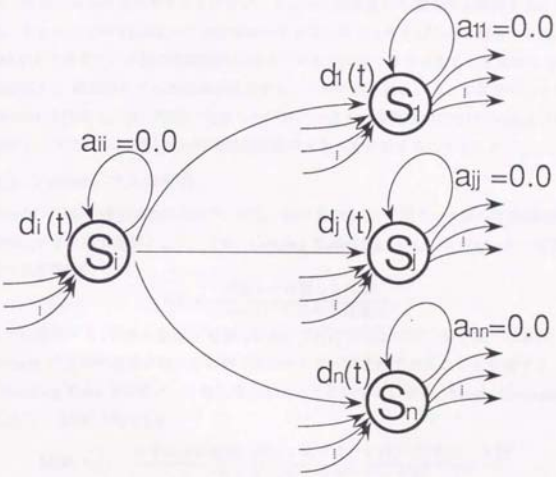


図 9.3. 継続時間長制御を直接組み込んだ HMM

各状態におけるループ遷移確率 a_{ii} は全て 0.0 となる。その代わりに，状態 i において時間 t だけループ (停留) する確率 $d_i(t)$ が各々推定され，認識時にはこれを加味した上でスコアが算出される。

9.3 基本 HMM と学習データ間の時間的対応付け

第 9.1 節でも述べたように，実際に作成された継続長モデルと学習データの時間的整合性を詳細に分析した例は少ない。本研究で採用した後処理法では，学習データと基本 HMM との照合結果より得られる Viterbi パスの分布，即ち，状態 i での停留時間 t_i の平均・分散が直接継続長モデルに反映される。その結果，Viterbi パスの分布を観測することで，作成される継続長モデルと学習データ間の時間構造的な整合性を検証することになる。そして，この Viterbi パスの分布が一カテゴリ内で大きくばらつく場合，それは基本 HMM と学習データ間の時間的対応付けに大きなばらつきがあることを意味し (時間構造変動大)，継続長モデルの効果が減少することが予想される。そこで本節ではまず基本 HMM を作成し，次に作成に使用した学習データとの時間的対応付け (Viterbi パス) を観測し，カテゴリ (音素) 別に時間構造変動の大きさを分析することとした。

9.3.1 Viterbi パスの分布

Viterbi パスの分布の変動の大きさ，即ち，基本 HMM と学習データ間の時間的対応の分布の広がりを示す指標として，まず，**Looping Rate**:$\theta(i), i = 1, 2, \cdots, n$($n$ = ループ遷移を持つ状態数) を

$$\theta(i) = \frac{\text{状態 } i \text{ に停留した回数}}{\text{Viterbi パス中の全遷移数}}$$

のように定義する (図 9.4 参照)。状態 j において $\theta(j)$ が 1.0 に近い値を取った場合，その Viterbi パス中の遷移は殆どが状態 j におけるループ遷移であることを意味する。なお，**Looping Rate** は学習データ毎に算出されることになる。更に，**Single Occupancy Rate**(以下，**SOR** と略す) を

$$\mathbf{SOR}(/\mathrm{x}/) = \frac{\text{いずれかの状態で } \theta(i) > \theta_0 \text{ を満たす } /\mathrm{x}/ \text{ の学習データ数}}{\text{カテゴリ } /\mathrm{x}/ \text{ の全学習データ数}}$$

のように定義する (図 9.5 参照)。即ち **SOR** は，あるカテゴリの学習データに対して，いずれかの状態で，条件 $\theta(i) > \theta_0$ を満たすデータを，全ての学習データ数で割ったものである。なお本実験では，θ_0=0.7 としている。図では条件 $\theta(i) > \theta_0$ を満たすループ遷移を太線で示している。なお，以下の説明ではこのような状態を "支配的状態 (dominant state)" と呼ぶことにする。

9.3.2 分析条件と音声試料

表 9.1 に本分析で使用した音声試料について示す。表にあるように，5 人の成人男性によって発声された 5240 単語音声を 2 つに区分し，片方を用いて HMM の学習を行な

う。基本 HMM の学習に使用された学習データ数は，各音韻最大 300 個である。その後，Viterbi パスを求め，HMM と学習データ間の時間的対応付けを求める。また，その際の分析条件については表 9.2 に示す。なお，表 9.1 にある認識用とは，第 9.6 節における音韻認識実験における入力音声のことである。また，表 9.2 にある 4 状態 3 ループの HMM を，図 9.6 左上に模式的に表してある。

9.3.3 分析結果

図 9.6 から図 9.10 にかけて，各話者における照合結果より算出される SOR を幾つかの音韻に対して示す。これらの音韻は，最も高い SOR を示した 8 個の音韻及び最も低い SOR を示した 8 個の計 16 個の音韻を，各話者別に選んだものである。各棒グラフは最大 3 つのブロックにより構成されているが，これらは各々状態 1, 2, 3(ループ遷移を保有する状態) が支配的状態となる割合である。そして，SOR はこれらの和として算出される。

9.3.4 考察と検討

図より，音韻によって SOR の値はかなりばらつくことが分かる。また，高い SOR を示す音韻の中には，MAU の/m/, /n/，或は，MAU の/g/, /h/のように複数のブロックによって比較的等分割されている音韻もあれば，MAU の/s/や MTK の/q/のように殆ど一つの支配的状態によって，非常に高い SOR を示している音韻もある。

低 SOR の場合，或は，高 SOR の場合でも一つの支配的状態によってその大部分が占められている場合は，Viterbi パスの分布，即ち時間構造変動は小さいと言える。このような音韻に対しては，学習データの時間構造のモデリングは従来のように状態独立に行なっても十分な近似が可能であろう。しかし高 SOR で，しかも支配的状態が複数になって分布している音韻は，図 9.11 に示すように，本来異なる特徴を持った複数のグループを一カテゴリとして (強引に) まとめて，HMM が作成されていると解釈できる。その結果，極端な平均化が行なわれる従来の状態独立の継続長モデルでは，図 9.12 に示すような問題を引き起こす。図では各学習データを支配的状態によって振り分け，各グループ毎に基本 HMM に対する時間的対応付けを模式的に示している。各々のグループでは，特定の状態 (支配的状態) でループ遷移の殆どが発生するため，継続長の分布はその状態以外では，ほぼ $t_i = 0.0$ 付近に集中する。しかし，これらグループ全体の時間構造を表す継続長モデルを作成すると，当然のことながら平均化の作用を受け，図 9.12 に示すように，

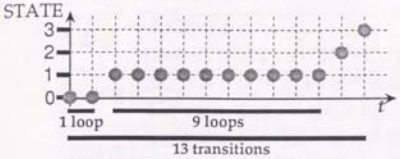
- 各状態でほぼ同じような分布を持つようになる。
- $t_i = 0.0$(i=0,1,2) の部分に山を持った分布を持つようになる。

と言う，本来の時間構造とはかけ離れた特性を持った継続長モデルが作成される。

これら分析結果より従来行なわれてきた，状態独立，しかも単一分布による継続長モデルでは十分に近似出来ない音韻があることが示された。継続長モデルの導入により，認識率が減少した例がいくつか報告されているが，上記の分析結果はその原因の一つとして十分に考えられるものである。さて，本研究では上記の問題を解決する手段として 2 つの観点からの異なる方法を考えた。一つは，時間構造変動の大きな音韻に対しても十分な近似が出来るよう，継続長分布に対して，より柔軟性のある関数系を導入するものである。即ち，多次元確率密度 (同時確率密度) 関数の混合型による近似であり，基本 HMM と学習データ間の時間的対応付けをより正確に表現しようとする試みである。もう一つは時間構造変動を低減するような学習データのクラスタリングである。即ちクラスタリングを行なうことで，1 カテゴリ内の学習データの時間構造 (変動) の統一を図ろうと言うものである。そしてその結果，従来の継続長モデル作成法で十分近似可能となれば，継続長モデル導入による効果も向上すると予測される。

❑ Tracing Viterbi Paths

Temporal correspondences between a phoneme HMM and its training data can be obtained by tracing the Viterbi paths.



ex.)
$\theta(0) = 1/13 \fallingdotseq 0.077$
$\theta(1) = 9/13 \fallingdotseq 0.69$
$\theta(2) = 0/13 = 0.0$
$\theta(3) = 0/13 = 0.0$

Looping Rate at state i : $\theta(i)$

$$= \frac{\text{[Number of loop transitions at state } i\text{]}}{\text{[Total number of transitions]}}$$

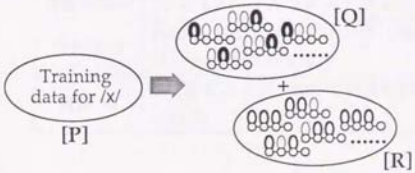
図 9.4. Viterbi パスの分布と **Looping Rate**

❑ Single Occupancy Rate (SOR)

SOR(/x/) =

$$\frac{\text{[Number of training data satisfying condition (A)]}}{\text{[Total number of training data of /x/]}}$$

Condition (A) : $\theta(i) > 0.7$ for a certain i (i=0,1,2,3).



$$\text{SOR(/x/)} = \frac{[Q]}{[P]}$$

図 9.5. **Single Occupancy Rate**

表 9.1. 音声試料 (ATR 音声データベースより)

話者	成人男性 5 人 (MAU, MHT, MTK, MMY, MMS)
学習用	孤立発声された 5240 語の偶数番目から得られる切り出し音韻。1 モデル当たり最大 300 個。
認識用	孤立発声された 5240 語の奇数番目から得られる切り出し音韻。1 モデル当たり最大 300 個。

表 9.2. 実験条件

認識タスク	26 カテゴリ (/a,i,u,e,o,p,t,k,ch,ts,b,d,g,s,sh,h,z,dj,m,n,r,w,y,N,Q,j/)
音響的特徴	16 次 LPC メルケプストラム
分析条件	16bit, 10kHz サンプリング、256 点ハミング窓、フレーム周期 5[msec]
HMM	4 状態 3 ループ、full 分散共分散行列、単一ガウス分布

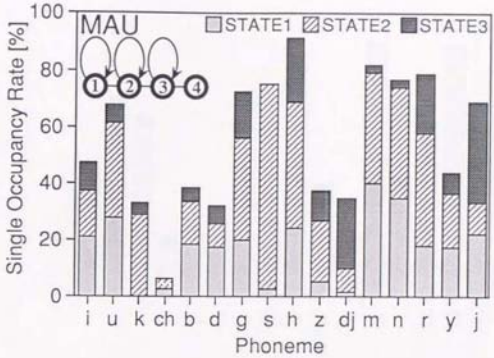


図 9.6. 話者 MAU における Single Occupancy Rate

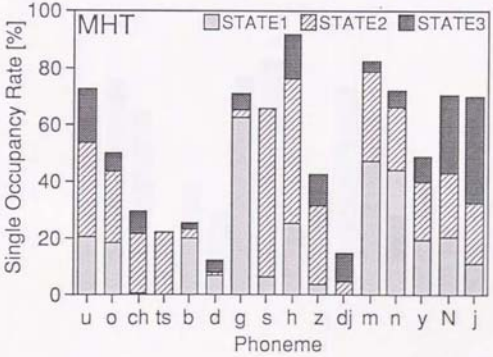


図 9.7. 話者 MHT における Single Occupancy Rate

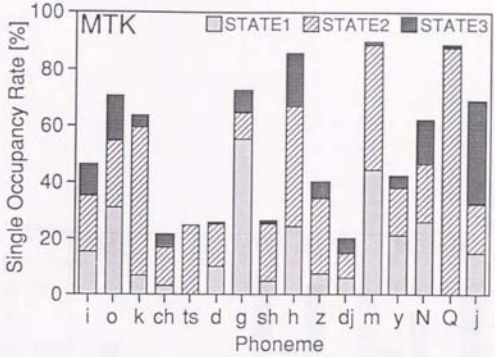


図 9.8. 話者 MTK における Single Occupancy Rate

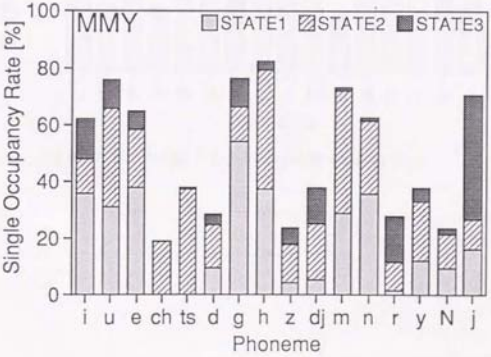


図 9.9. 話者 MMY における Single Occupancy Rate

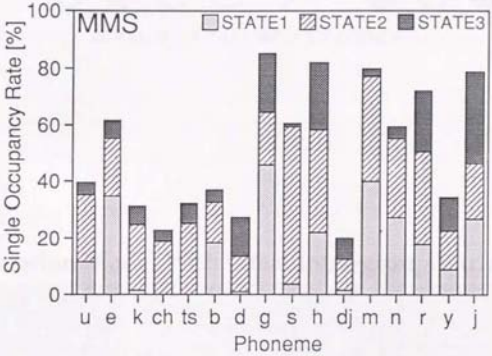


図 9.10. 話者 MMS における Single Occupancy Rate

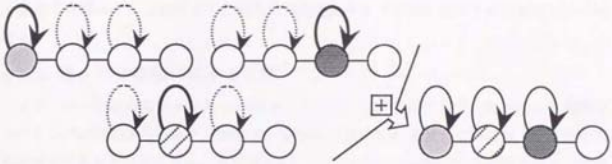


図 9.11. サブモデルの "和" としての HMM

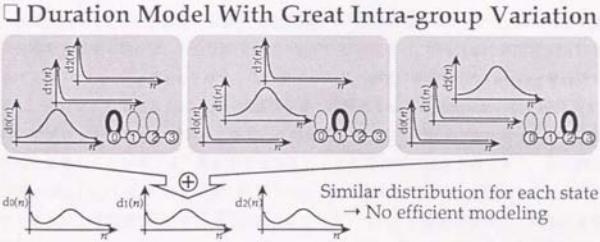


図 9.12. 高 Single Occupancy Rate かつ複数の支配的状態を持つ音韻における継続時間長モデル

9.4 多次元分布関数の混合型を分布関数とした継続時間長制御

第 9.3 節で示したように，従来行なわれてきた「状態独立かつ単一分布」に基付く継続長モデルでは，十分にその時間構造を近似できない音韻が存在する。本節では，より柔軟な関数系を導入し，上記のような音韻の時間構造をより正確に近似する方法について考える。

9.4.1 種々の分布型に関する考察

まず，単一分布から混合分布への移行を考える。こうすることで，確かにより精密なモデリングが可能となるが，状態独立に継続時間長分布を近似している限り，十分な近似が行なわれるとは言えない。再度図 9.12 を見てみる。この図では平均化の結果どの状態においても，類似した分布，即ち t=0.0, $T(\neq 0.0)$ の 2 箇所に山を持つ分布，として近似されることを示している。混合型による状態独立なモデリングとは，この 2 箇所の山を正確に近似することを意味する。しかし，明らかなようにこれは適切なモデリング手法ではない。$P(t_1=0.0,t_2=0.0,t_3=0.0)$ は本来非常に低い値を示すはずであるが，この方法では，$P(t_1=0.0,t_2=0.0,t_3=0.0)$ に Local Maximum が存在することになってしまう。同じようなことが，各グループのもう一つの山の中心 (上記の T) についても言える。本来 $d(t_1=T,t_2=T,t_3=T)$ は低く抑えられるべきところが，実際にはある程度の値を持つようになる。

次に，状態独立な分布関数から多次元分布関数への移行を考える。しかし，この場合も単一分布で近似する限りは，十分な精度を見い出すには至らない。図 9.13 に発話者 MAU における/h/の **Looping Rate** を示す。この図において，各軸は各状態 (の **Looping Rate**) を示し，プロットしてある点は，ループ遷移を持つ状態の **Looping Rate** を $(\theta(1),\theta(2),\theta(3))$ の形でプロットしたものである。さて，この分布から $(\theta(1),\theta(2),\theta(3))$ に対する確率密度関数を求めることを考える。最も簡単な例として 3 次元ガウス分布を考えると，最大確率密度は $(\theta(1),\theta(2),\theta(3))$ の平均値で観測されることになる[5]。しかし，図 9.13 を見て明らかなように，平均値 $(\overline{\theta(1)},\overline{\theta(2)},\overline{\theta(3)})$ 付近にプロットされている点は少なく，対応する確率密度値も低く抑えられなければならない。このように，図 9.13 で示される分布に対して，強引に単一の多次元分布関数で，その確率密度分布を近似することは非常に危険であることが分かる。以上の考察の結果，基本 HMM との時間的対応付けを正確に近似するためには，

[5] ガウス分布以外の分布関数においても，最大確率密度が平均値で観測されると言う性質は広く適用される。

- 各軸 (状態) 毎の相関を考慮した多次元分布を導入する。
- かつ，混合分布による近似を導入する。

の 2 つが必要であることが分かる。

9.4.2 混合分布多次元確率密度関数によるモデリング

HMM におけるスペクトル領域のパラメータの分布を混合分布で表現する場合，その推定方法には，Baum-Welch のアルゴリズムが適用されることが多いが，本研究では第 9.2 節で述べたように，継続時間長分布の推定に Baum-Welch のアルゴリズムを使用しないため，継続長分布を混合分布で表現する場合も自ずと Baum-Welch の方法を用いることは出来ない。そこで，以下に示す簡便法を採った。まず，基本 HMM と学習データとを rematching し，(t_1, t_2, t_3) を各学習データ毎に算出する。そして (t_1, t_2, t_3) 空間を K-means 法を用いて M 分割する (本実験では M=3，即ち混合数 3 とした)。次に，分割された各グループ毎に (t_1, t_2, t_3) の平均値及び $N \times N$ の分散共分散行列を求める。また，第 9.2 節で述べたように分布形としてはガウス分布を仮定する。最終的に点 $d(t_1, t_2, t_3)$ に対する確率密度 $G(d)$ は，各グループに対して得られた平均値及び分散行列を用いて，

$$G(d) = \sum_{m=1}^{M} \gamma_m G_m(d)$$

と表される。但し，γ_m は重み係数であり，K-means 法による分割の結果，グループ $m(<M)$ に属する学習データ数の総データ数に対する割合である。また，G_m はグループ m に対して算出された多次元ガウス分布である。

9.4.3 実験条件

混合分布多次元確率密度関数を用いた継続長モデルの有効性を検討するため，以下に示す 4 種類の HMM による，話者 closed，テキスト open の切り出し音韻認識実験を行なった。継続長モデルの作成は第 9.2 節に述べた通り，各学習データに対する Viterbi パスから，各状態に停留する時間長 t_i を，ベクトル (t_1, t_2, t_3) の形で求め，その平均及び分散共分散行列を音素毎に求める。そして，この 2 つのパラメータを用いて継続長分布を推定する。なお，用いた音響的パラメータ，HMM の構造，音韻の種類などは第 9.3 節における分析条件と同じである。

CASE 1 基本 HMM

CASE 2 CASE 1+対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 3 **CASE 1**+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 4 **CASE 1**+非対角化分散共分散行列による，混合ガウス分布を用いた継続時間長モデル

CASE 2 は状態独立に (相関=0.0 とすること，即ち対角化と同値)，しかも単一の分布関数で各状態の継続長分布を近似している従来法である。これに対し，**CASE 3, 4** では，相関を考慮して継続長のモデルを行ない (**CASE 3**) かつ，混合分布化することで，より精度良く近似している (**CASE 4**) ものである。なお，状態独立の分布関数を混合化することは本実験では行なわなかった。

なお，照合スコア算出に当って，継続長モデルは具体的には以下のように考慮された。まず，基本 HMM による尤度 Q を Viterbi アルゴリズムを用いて計算し，と同時に Viterbi パスをも記録する。そして，得られた Viterbi パスと継続長モデルとを照合し，継続長分布における尤度 Dを計算する。そして最終的な尤度$\hat{Q}$は以下の式で求める。

$$\log(\hat{Q}) = \log(Q) + \kappa \log(D)$$

ここでκは，重み係数であり，本研究では 4 に設定した。なおこの値は，**CASE 2** に対して最高認識率を示す整数を 0 から 10 の範囲で予備実験において求めたものである。

9.4.4 実験結果

各 **CASE**・話者毎の結果を図 9.14 に示す。

9.4.5 考察と検討

CASE 1 と **CASE 2** を比較すると，全ての話者において後者が上回っている。これは，状態独立な単一分布関数による継続時間長分布の近似でも，認識率の向上に有効に作用していることを意味する。更に **CASE 2** と **CASE 3** との比較により，各状態 (軸) 間の相関を考慮した多次元分布の導入による効果が全ての話者において現れていることが分かる。しかし，混合分布化した **CASE 4** を見ると，必ずしも **CASE 3** を上回る結果を出している訳ではない。混合分布化することで，図 9.13 のような分布に対しても，より正確な近似が可能となっているのは明らかである。しかし，認識率と言う点から見た場合，必ずしも望ましい結果を出している訳ではない。即ち，継続時間長を正確に近似したことが逆に，異なる音韻間に共通して観測される特徴をより強調してしまったものと考えられる。特に図 9.13 のように各軸毎に分布を持つような音韻が複数存在した場合，3 次元と言う次元の低さでは，それらを十分に分離するだけの識別能力が無かったと考察される。

以上の結果及び考察より，もう一つのアプローチ，即ち時間構造変動を低減し，単一分布でも十分近似できるよう，学習データをクラスタリングする方法を，以降考えることにする。

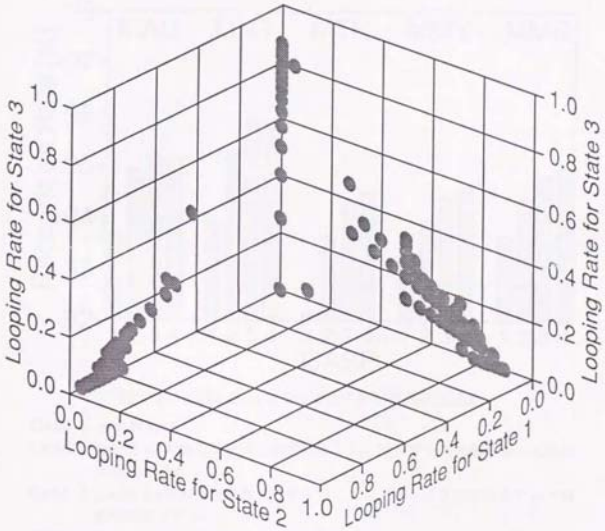


図 9.13. 発話者 MAU の音韻/h/に対する Looping Rate の分布

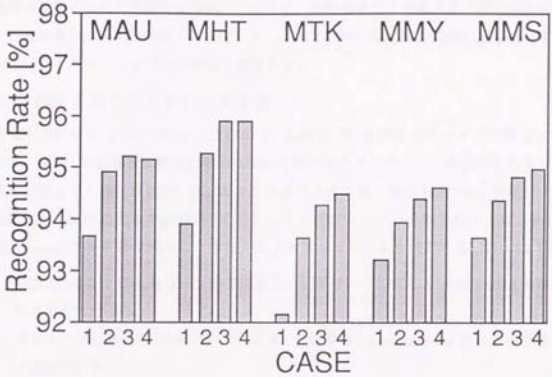


図 9.14. CASE 1~4 における切り出し音韻認識結果

CASE 1 基本 HMM

CASE 2 CASE 1+対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 3 CASE 1+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 4 CASE 1+非対角化分散共分散行列による，混合ガウス分布を用いた継続時間長モデル

9.5 時間構造変動の抑制を目的としたクラスタリング手法

同一カテゴリ内の変動を抑えることを目的としたクラスタリング手法は，幾つかの研究において提案され，非常に広く使用されている方法もある[40]。しかし，これらの多くは音声のスペクトル領域での特徴分布変動をカテゴリ内で抑えることを目的としており，必ずしも音声の時間構造変動を十分に抑えているとは言えない。この時間構造変動の大きさは第 9.3 節において実験的に検討したように，継続長モデルを導入する際には特に注意する必要がある。以上の考察の下本節では，カテゴリ内の時間構造変動を直接抑制することを目的としたクラスタリング手法を提案する。

9.5.1 提案するクラスタリング手法

適切なクラスタリングを行なうためには，閾値などの適切なパラメータの設定が必要となる。特に，時間構造変動の定量的な表記方法は本クラスタリングを実現する上で非常に重要になってくると考えられる。本クラスタリングでは，基本 HMM と学習データ間の時間的対応付けに大きな変動が生じないような学習データの分割を行なえばよいので，rematching 結果を直接反映するような，以下のアルゴリズムを提案する (図 9.15 参照)。

1. 基本 HMM を Baum-Welch 法で作成し，学習データに対して rematching 処理を行なう (図 9.15 左)。
2. 得られた Viterbi パスに基付いて，状態 i に対する **Looping Rate**:$\theta(i)$ を各学習データ毎に求める。
3. 求まった$\theta(i)$ が状態 i に対して，$\theta(i) > \theta_0$の条件を満たしている場合，その学習データを "状態 i グループ" として分類する (図 9.15 中央)。ここで，同一学習データが異なる複数の状態グループに属さないよう，θ_0は 0.5 以上とする。以上のプロセスの結果，ある音素の学習データは $N+1$ 個 (N=ループ遷移を持つ状態数) のグループに分割される。即ち N個の状態グループと，いずれの状態においても$\theta(i) > \theta_0$を満たさなかったデータ群のグループ (仮に，非状態グループ (non-state group) と呼ぶ) である。
4. 上記で求まった最大 N個の状態グループの，学習データ全体に対する割合を求め，これを$\lambda(i)$ とする。即ち，状態 i グループに属する学習データ数を $\mathsf{L_i}$ (i=1,2,3)，非状態グループのデータ数を $\mathsf{L_0}$，全学習データ数を $\mathsf{L}(=\sum_{i=0}^{3}\mathsf{L_i})$ とした場合，$\lambda(i)=\mathsf{L_i}/\mathsf{L}$ (i=1,2,3) である。また，第 9.3.1 節で定義した **SOR** を$\lambda(i)$ で定義すると，**SOR**=$\sum_{i=1}^{3}\lambda(i)$ となる。

5. 求まった$\lambda(i)$に対して，$\alpha < \lambda(i) < \beta$を満たす状態 i グループを "サブグループ" として独立させる。またこれとは別に，非状態グループと，条件$\alpha < \lambda(i) < \beta$を満たさない状態グループの和集合も "サブグループ" として独立させる。図 9.15 では，3 つの状態グループ及び (1 つの) 非状態グループから 3 つのサブグループが独立する様子を表している。即ち，状態 1 グループ，状態 2 グループ，状態 3 グループと非状態グループの和集合である。さて，上記の条件における閾値αは，少数の学習データ数のみ属する状態グループがサブグループとして選択されるのを防ぐ役割を持ち，閾値βは，大多数の学習データが属する状態グループがサブグループとして選択されることで，残された "非状態グループ+条件を満たさない状態グループの和集合" が，学習データが少ないにも拘らず，サブグループとして独立することを抑える役割を持つ。

なお，上記のアルゴリズムは繰り返し行なうことが理論的には可能であるが，本研究では学習データの量を考慮して 1 回のみ行なうこととした。クラスタリングの後，新たに独立して定義されたサブグループ毎に基本 HMM，及び継続長モデルを作成する。

9.5.2 提案した手法の時間構造変動低減に基付く評価

提案したクラスタリング手法を施して作成された HMM による認識実験を行なう前に，同一カテゴリ内の時間構造変動低減効果による評価を行なう。ここで時間構造変動の大きさを示す定量的指標が必要となってくるが，ここでは以下のように定め，これを **Temporal Variation Rate**(以下 TVR と示す) と呼ぶことにする。

$$\mathrm{TVR} = \frac{\text{サブグループとして独立した学習データ数}}{\text{全学習データ数}}$$

即ち TVR が大きい音韻ほど時間構造変動の大きな音韻と言うことになる。なお，図 9.15 の場合，TVR= $(L_1+L_2)/L$ となる。

図 9.16 に本手法の適用前及び適用後の各話者毎の TVR の変化を示す。なお，TVR そのものは音韻別に求めることが可能であるが，ここで示す値は全音韻での平均値である。また，適用後の TVR とは第 9.5.1 節で述べたアルゴリズムを再度繰り返し (但し分割はしない)，仮想的に TVR を求めたものである。また図 9.16 中，左側の図は$(\varepsilon,\alpha,\beta)$=(0.7,0.3,0.7) に設定して行なった結果であり，右側の図は，$(\varepsilon,\alpha,\beta)$=(0.7,0.25,0.75) に設定した行なった結果である。図より明らかなように，本クラスタリング手法はカテゴリ内の時間構造変動の大きさを効果的に抑制していることが分かる。なお，ここで示した 2 つのパラメータ

組は何らかの最適化の結果算出された値ではない。時間構造変動をより効率的に抑制するパラメータ値の算出方法は今後の課題の一つである。

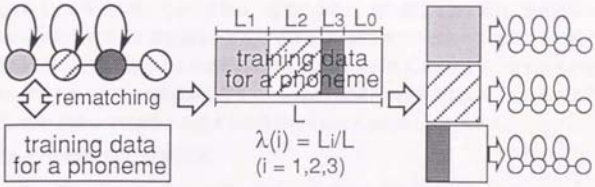


図 9.15. 本研究で提案するクラスタリング手法

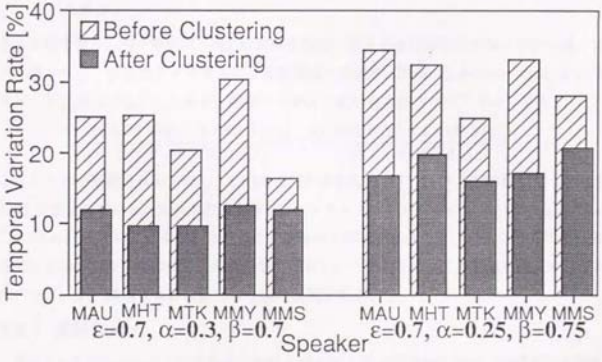


図 9.16. 本研究で提案したクラスタリング手法による時間構造低減効果

9.6 提案した手法の音韻認識実験による評価

第 9.5 節において，提案するクラスタリング手法が時間構造変動の低減に有効的に作用していることを示した。しかしこれは，認識率の向上，即ち第 9.3 節で言う，継続長モデルによる効果の向上を直接意味するものではない。継続時間長制御の効果向上を検証するためには，認識実験による評価が必要である。そこで，第 9.4.3 節と同一の分析条件を用いて切り出し音韻の認識実験を行なった。なお，認識用の音声資料は表 9.1 にあるように，5240 単語中の奇数番に位置する単語から得られる切り出し音素である。

9.6.1 実験条件と音声試料

本節の認識実験においては，以下に示す 4 つの種類の HMM が使用された。

CASE 1 基本 HMM

CASE 3 **CASE 1**+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 5 提案するクラスタリング手法を施して作成された基本 HMM

CASE 6 **CASE 5**+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

第 9.3 節で言う，同一カテゴリ内の時間構造変動の低減が継続時間長制御に有効に働くのであるならば，継続長モデル導入による認識誤り低減率 (Error Reduction Rate) をクラスタリング前後で比較した場合，後者の方がより高い値を示すはずである。即ち，

$$\frac{(\textbf{CASE 3}[\%] - \textbf{CASE 1}[\%])}{(100.0 - \textbf{CASE 1}[\%])} < \frac{(\textbf{CASE 6}[\%] - \textbf{CASE 5}[\%])}{(100.0 - \textbf{CASE 5}[\%])}$$

となることが期待される。但し，**CASE 3**[%]>**CASE 1**[%]，及び，**CASE 6**[%]>**CASE 5**[%] を仮定している。なお本認識実験では，クラスタリングパラメータとして $(\varepsilon,\alpha,\beta) = (0.7,0.3,0.7)$ を用いて行なわれた。但し，第 9.5.2 節でも述べたように，これらの値は最適化などの特別な処理の結果得られた値ではない。クラスタリングの動作を決定する種々のパラメータ・閾値などの最適化は今後の課題である。

9.6.2 実験結果

図 9.17 に上記 4 条件下の各話者における切り出し音素認識率を示す。各話者においてクラスタリングを施すことにより，カテゴリ数は 26 から約 40 へと増えた。これらは，クラスタリングで細分化されなかった音素クラス及び細分化されたサブグループの和である。そして約 40 各々のカテゴリに対して HMM 及びその継続長モデルを構築し直した。

図中，話者名の下に括弧付きで表示してある数字がクラスタリング後のカテゴリ (HMM) 数である。複数の HMM を持つ音韻に対する認識処理は，マルチテンプレート方式と同様の処理を行なった。

9.6.3 考察と検討

図 9.17 において，**CASE 1** から **CASE 5** への認識率の向上は，本クラスタリング手法が継続長モデル導入の有無に拘らず有効であることを示している。そして **CASE 6** に示されるように，継続長モデルを導入することで認識率は更に向上している。**CASE 3** 及び **CASE 6** の棒グラフ上に示してある数字は，各々継続長モデル導入による認識率の向上 (絶対値) である。即ち，(**CASE 3**[%]−**CASE 1**[%]) と (**CASE 6**[%]−**CASE 5**[%]) である。絶対値による比較でも，後者の方により大きな認識率向上が観測されていることが分かる。更に，この結果を継続長モデル導入による認識誤り低減率と言う観点からプロットし直したものが，図 9.18 である。この結果は，第 9.6.1 節での予測と合致するものであり，時間構造変動の低減に基付く本クラスタリング手法の継続時間長モデルに対する有効性を裏付けるものである。

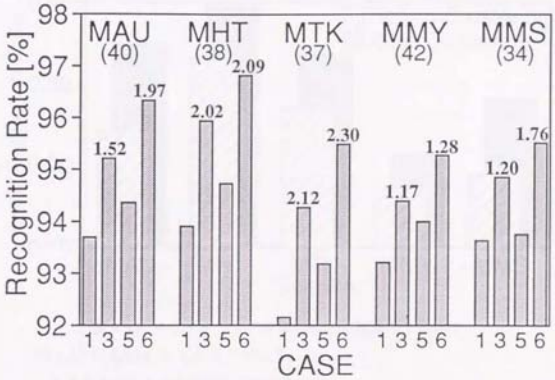


図 9.17. CASE 1, 3, 5, 6 における切り出し音韻認識結果

CASE 1 基本 HMM

CASE 3 CASE 1+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

CASE 5 提案するクラスタリング手法を施して作成された基本 HMM

CASE 6 CASE 5+非対角化分散共分散行列による，単一ガウス分布を用いた継続時間長モデル

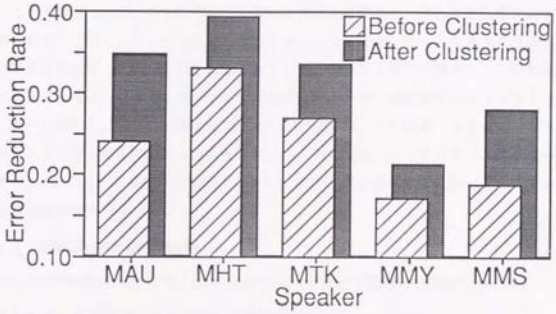


図 9.18. クラスタリング前後における認識誤り低減率

図 9.17 の CASE 1, 3, 5, 6 に対して,

- クラスタリング前の認識誤り低減率

$$\frac{(\text{CASE 3}[\%] - \text{CASE 1}[\%])}{(100.0 - \text{CASE 1}[\%])}$$

- クラスタリング後の認識誤り低減率

$$\frac{(\text{CASE 6}[\%] - \text{CASE 5}[\%])}{(100.0 - \text{CASE 5}[\%])}$$

として算出される。

9.7 混合分布 HMM に対する有効性の検討

さて前節までの議論において，基本 HMM の構造は 4 状態 3 ループ固定，出力確率密度分布の形態は，非対角化分散共分散を用いた多次元正規分布ではあるが，単一の分布を用いて推定を行なっていた。第 9.3 節による Viterbi パスに対する分析で，同一カテゴリであっても学習データ間に大きな時間構造変動が存在する音韻があることを示した。しかし，この現象が HMM の (スペルトル領域における特徴量の) 出力確率分布に対する不適切な推定・近似によるものであると議論することもできる。即ち，多次元正規分布の混合分布を出力確率とした HMM を用いることで，時間構造変動は減少し，上記の問題は解決するとの考え方である。更にこの考えを押し進めれば，混合分布における各々の分布が，単一分布化した場合には異なる状態として実現され，その結果，大きな時間構造変動を示すようになったとの議論も可能である。そこで，提案したクラスタリング手法の混合分布 HMM に対する有効性を検証するため，第 9.6 節と同一分析条件の下，再度切り出し音韻認識実験を行なった。

9.7.1 実験条件と音声試料

本節の認識実験においては，以下に示す 4 つの種類の HMM が使用された。

CASE 1 単一多次元ガウス分布による基本 HMM

CASE 7 混合多次元ガウス分布による基本 HMM(混合数=3)

CASE 8 **CASE 7**+単一多次元ガウス分布による継続時間長モデル

CASE 9 **CASE 7**+クラスタリング+単一多次元ガウス分布による継続時間長モデル

なお上記の 4 条件において，スペクトル領域の特徴量に対する出力確率分布で使用される分散共分散行列，及び，継続長モデルで使用される分散共分散行列は共に，非対角化行列である。

9.7.2 実験結果

図 9.19 に **CASE 1, 7, 8, 9** の結果を各話者毎に示す。また各話者名の下の数字は，**CASE 9** におけるクラスタリング後の HMM の数を示す (クラスタリング前は 26 種類)。

9.7.3 考察と検討

CASE 1 と **CASE 7** を比較すると，全ての話者に対して，分布数の増加 (1→3) が認識率の向上を引き起こしていることが分かる。そして，**CASE 7** と **CASE 8** の比較より，継続長モデルの導入が更に認識率を向上させていることも分かる。**CASE 7** の混合多次元

ガウス分布 HMM に対して，提案したクラスタリング手法を適用することで作成された HMM の数を図 9.17 に示してある数字と比較すると，全ての話者において下回っていることが分かる。図 9.17 の結果は単一多次元ガウス分布 HMM によるものであることを考慮すると，本節の冒頭に述べたように，混合数の増加によって時間構造変動はある程度抑えられていると考察できる。しかしクラスタリング処理を施すことで，認識率は更に上昇している (**CASE 9**)。これは混合数 3 の多次元ガウス分布による HMM においても，時間構造変動はまだ無視できない状況にあることを意味している。当然のことながら更に混合数を上げることで，この変動の大きさは更に抑制できると考えられるが，混合数の増加は要求する学習データの大幅な増加を招いてしまい，望ましい方向性であるとは必ずしも言えない。

以上の結果・考察より，本研究で提案したクラスタリング手法は，継続時間長制御を基本 HMM に導入する場合に，

『基本 HMM 及び対応する学習データが継続長制御に適した時間構造を保有しているかを検証する。』

そして，保有していない場合に，

『クラスタリングにより，1 カテゴリに対応する学習データが継続長制御に適した時間構造を持つよう，カテゴリ群を再構築する。』

手法であると結論付けることができる。

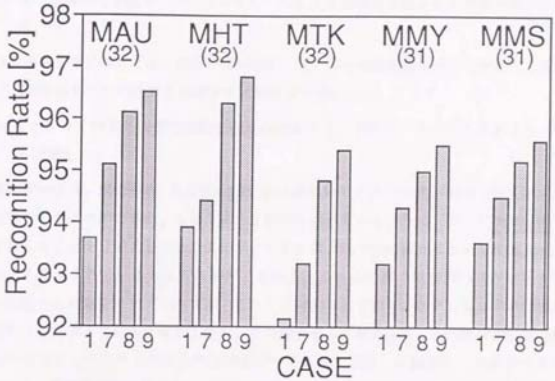


図 9.19. CASE 1, 7, 8, 9 における切り出し音韻認識結果

CASE 1 基本 HMM
CASE 7 混合多次元ガウス分布による基本 HMM(混合数=3)
CASE 8 CASE 7+単一多次元ガウス分布による継続時間長モデル
CASE 9 CASE 7+クラスタリング+単一多次元ガウス分布による継続

なお，使用する分散共分散行列は全て非対角化行列である。

9.8 まとめ

本章では HMM を純粋にパターン認識の立場から捉え，その認識能力を向上させるための一手法である，継続時間長モデルの導入に対して新しい観点からの分析を行なった。即ち，Viterbi パスの分布を調べることにより，継続時間長モデルと学習データの時間構造的な整合性を定量的に分析した。その結果，

- 同一音韻内の学習データにおいても，大きな時間構造変動を示す音韻が幾つか観測された。

このような音韻に対しては，従来の状態独立，単一分布の継続長モデルでは正しく近似できず，継続長モデルの効果を低減させる恐れが十分にある。そこで，

- 1 カテゴリ内の時間構造変動を低減させるべく，学習データのクラスタリング手法を提案した。

認識実験の結果，混合分布 (混合数=3)HMM においてもその効果が観測され，本手法の有効性を示すことができた。なお，第 9.5 節でも述べているように，同一カテゴリ内の変動を抑えることを目的としたクラスタリング手法は，幾つかの研究において提案され，非常に広く使用されている方法もある[40]。しかし，これらの多くは音声のスペクトル領域での特徴分布変動をカテゴリ内で抑えることを目的としており，必ずしも音声の時間構造変動を十分に抑えているとは言えない。その意味で，本章で提案した時間構造変動低減を直接の目的としたクラスタリングは学習データに対する新しい観点からのクラスタリングであり，継続時間長制御に対して従来とは異なる観点から行なったアプローチであると言える。しかし，本研究では終始，時間構造変動のみに着目しており，スペクトル領域での分布の広がり/変動に対しては何も考察を行なっていない。従来行なわれてきたスペクトル領域での検討と今回行なった時間領域での検討の両方を考慮に入れ，周波数・時間両方向において十分に変動を低減できるクラスタリング手法が今後望まれるところである。

第 10 章

結 論

本論文で行なってきた種々の実験を再度概観し，得られた結果/知見をまとめるとともに，その将来的展望を述べることで，本論文の結論とする。本論文では『音声を媒体とした情報の受容』と言う事象に関して，人間及び機械の両者に焦点を当て，

1. 人間による音声の知覚過程の分析とそのモデル化
2. 計算機による音声の認識手法の高精度化

と言う 2 大テーマの下，一連の基礎研究を行なった。前者における研究では，認知科学的なパラダイムに基付き，以下の項目を対象とした音声知覚実験を行ない，人間の音声言語処理過程を観測・分析した。

- 単語以上の処理単位の存在の是非。
- 存在する場合，その処理単位を用いた処理の特性。
- 複数精度による (照合) 処理と複数単位による (照合) 処理との関係。
- 内部辞書の構成・構造。
- 辞書検索過程へ影響を及ぼす諸要因。項目固有の特性がもたらす作用 (単語知覚) と複数の単語が一文中に存在することにより生じる単語間の作用 (文知覚)。
- 韻律的情報処理と音韻情報処理の単語内/文内における相互作用。
- 言語的情報処理が音声知覚過程に及ぼす影響。特に統語的・意味的・談話的情報処理による影響の差異。

特に，内部辞書検索過程に影響を与える要因に関しては深く考察し，音響レベル，辞書 (意味) レベル，談話レベルにおける要因を考え，種々の実験を行なった。また，現在の音声認識においては十分に利用されていない韻律的特徴が，音声知覚過程において果たす役割についての実験も行なった。その結果，数多くの結果/知見が得られた。次に，これらの知見を基に，人間による音声言語処理全体を見渡すことの出来る，(認知科学的な) 知覚モデルの構築を行なった。しかし，個々の実験における『検討と考察』の項を見て分かるように，各々が課題として残している問題がある。そして，この一つ一つを解決していくことで，知覚モデルもより洗練されたものとなる訳だが，特に筆者は以下の観点からの観測・分析が今後重要な課題となると考えている。即ち，コンテキストが無く，項目固有の特徴に支配された (静的な) 知覚過程と，コンテキストの存在により，辞書検索過程の特性が動的に変化しながら行なわれる (動的な) 知覚過程との，定量的な比較である。と言うのも，筆者の見る限りにおいて，知覚過程の特性の動的変化を定量的に分析した実験がまだまだ少ないように思うからである。それ故，“連想” などを工学的に模擬する場合

でも，静的な連想のみを対象としている研究例を多く見かける。本論文で紹介した大規模データベースを使った親密度の測定や，筆者が行なった談話的情報量 (通常性) の定量的測定及びそれに基付く実験結果の評価などは，静的特性と動的特性，及び両者の相互関係を分析する第一歩として位置付けることができる。そして，このような定量的な実験/考察を積み重ねることで，工学的モデルとしても十分応用できるものが構築されると筆者は考えている。

『計算機による音声の認識手法の高精度化』と言うテーマの下に行なった研究では，

- 音声の音響的特徴表現方式を動的に変化させた認識手法
- 継続時間長モデルの時間構造記述力を向上させることを目的とした，学習データのクラスタリング手法

と言う 2 つのテーマの下，実験的検討を行ない，各々において新しい手法を提案した。前者は，知覚モデル内の一処理部を考慮して行なった先行研究の延長線上に位置するものである。即ち先行研究において残された問題に対する一アプローチであり，『抽出された音響的特徴量の各成分が，入力音声に関する情報をどの程度担っているか』と言う観点に基付いて各成分の表現方式—直接的表現/間接的表現—を動的に変化させようと言うものである。従来行なわれてきた音声認識手法の研究の多くは，『如何にすれば，学習データの特徴分布を効率的に，精密に近似できるのか』と言う観点の下に行なわれてきたことを考慮すると，本研究で提案した手法は，新しい観点から音声認識手法を見直したものと考えることができる。即ち，特徴パラメータの各成分が『ある話者，あるカテゴリを識別するに当たり，どの程度の情報量を担っているのか，どの程度着目されるべきなのか』と言う観点から考案されているものである。特定話者の切り出し音素認識実験結果より，その有効性が確認されたが，話者依存性が観測されるなど，話者性を含めた更なる動的適応の実現が今後の課題の一つである。

後者では HMM を純粋にパターン認識の立場から捉え，その認識能力を向上させるための一手法である，継続時間長モデルの導入に対して新しい観点からの分析を行なった。即ち，Viterbi パスの分布を調べることにより，継続時間長モデルと学習データの時間構造的な整合性を定量的に分析した。その結果，『同一音韻内の学習データにおいても，大きな時間構造変動を示す音韻』が幾つか観測された。このような音韻に対しては，従来の状態独立，単一分布の継続長モデルでは音声の持つ時間的構造を正しく近似できず，継続長モデル導入の効果を低減させる恐れが十分にある。そこで，『1 カテゴリ内の時間構造変

動を低減させるべく，学習データのクラスタリング手法」を提案した。認識実験の結果，混合分布 HMM においてもその効果が観測され，本手法の有効性を示すことができた。なお，同一カテゴリ内の変動を抑えることを目的としたクラスタリング手法は，幾つかの研究において提案され，非常に広く使用されている方法もある。しかし，これらの多くは音声のスペクトル領域での特徴分布変動をカテゴリ内で抑えることを目的としており，必ずしも音声の時間構造変動を十分に抑えているとは言えない。その意味で，本章で提案した時間構造変動低減を直接の目的としたクラスタリングは学習データに対する新しい観点からのクラスタリングであり，継続時間長制御に対して従来とは異なる観点から行なったアプローチであると言える。しかし，本研究では終始，時間構造変動のみに着目しており，スペクトル領域での分布の広がり/変動に対しては何も考察を行なっていない。従来行なわれてきたスペクトル領域での検討と今回行なった時間領域での検討の両方を考慮に入れ，周波数・時間両方向において十分に変動を低減できるクラスタリング手法が今後望まれるところである。

付録 A

高品質音声分析合成システム *PROSODY*

A.1 高品質音声分析合成システム *PROSODY*

知覚実験で使用する (合成) 音声試料の作成を目的として，(自然) 音声に対して，

- 基本周波数パターンの抽出，及び F_0 モデルに基付く F_0 パターンの編集。
- 分析結果及び編集結果に基付く，分析合成法による音声 (再) 合成。

を行なうためのシステム ***PROSODY*** を試作した。本ドキュメントは，***PROSODY*** を使用するために必要となる知識を提供するものである。なお，分析合成法による音声合成，F_0 モデル及び f0calc〜pacedit の一連のツール[1][126]に関して既知であることを前提に話しを進める。また，f0model が出力する，拡張子が.PAC のファイルを **PAC** ファイル，その中に記述されている F_0 モデルパラメータを初めとする種々の情報を **PAC** 情報と呼ぶことにする[2]。

A.1.1 起動する前に ……

1995 年 3 月 13 日現在，広瀬研究室には，2 セットの AD/DA コンバータが各々，diana, apollo に接続されている[3]。***PROSODY*** を起動する前に，どの AD/DA コンバータ (即ち，どちらのマシンを通して) 音を出力するかを以下の方法で設定する必要がある。

```
% setenv PRO_HOST hostname ↵
```

当然のことながら，*hostname* としては diana 或は apollo のみが有効である。それ以外のマシン名を設定した場合，当然音は出力されない。また，以上の設定をせずに ***PROSODY*** を起動した場合は，デフォルトの DA 用マシンを使用することになる。デフォルトの DA 用マシンは，***PROSODY*** が起動されたマシンに依存する。また，***PROSODY*** を通して DA した場合，DA 中に，

```
--- DA on hostname (== PRO_HOST) ---
```

の表示がされ，どのマシンを使用しているかが分かるようになっている。なお，1995 年 3 月 13 日現在の設定では，以下の場合を除いて全マシンより (PRO_HOST を設定することで)DA 可能である。即ち，DA 用に指定したマシン以外で，***PROSODY*** を起動した場合，そのマシンローカルのファイルを DA することはできない。つまり artemis 上で，diana を PRO_HOST として指定した場合，artemis の/tmp/speech.raw は DA できないような設定になっている。DA できるのは，diana からも artemis からも同じファイル名でアクセスできるファイルに限られている。よく分からない人は管理者に聞くか，/home あるいは/share?以下のファイルのみを使用するように。

[1] 瀬戸氏 (現東芝) によるマニュアルが研究室にあるはずである。技官の高橋さんに聞くと良い。
[2] **PAC** 情報として何が記述されてあるのか，についても高橋さんに聞くと良い。
[3] 今後 AD/DA コンバータが接続されるマシンは "増える/変更される" 可能性大である。注意するように。

A.1.2 起動方法

至って簡単である。

```
% prosody ↵
```

とすれば，ファイル選択用のウィンドウが開き，そのウィンドウで，オリジナルとなる音声試料を選択すればよい。このファイル選択用のウィンドウは ***PROSODY*** で作成した合成音声をファイルに出力する時などにも使用される。また，コマンドライン上で，

```
% prosody RAWfile ↵
```

として，音声データファイルを指定してもよい。なお 1995 年 3 月 13 日現在，入力ファイルとしてサポートしている音声データファイルのフォーマットは，

- サンプリング周波数 10 [kHz]
- 16 [bit]
- ヘッダ無し

のファイルのみである。広瀬研では 3 通りの音声データフォーマットを使用しているが，このうちの 2 つ，ESPS や VAX フォーマットのファイルは使用できないと言う訳である。予め，bhd(ESPS→RAW) や pcm2raw(VAX→RAW) と言ったコマンドを使用して，RAW file に落してから使用するように。ファイルを指定，或は選択して起動した直後の様子を図 A.1 に示す。図 A.1 を見て分かる様に，***PROSODY*** は，

- 5 つのメニュー
- 5 つのウィンドウ

によって構成されている。メニューについては第 A.1.3 節以降で説明することとして，ここでは各ウィンドウについて簡単に説明しておく。

1. *Waveform of Original Speech*
   このウィンドウには，起動時に選択，或は指定した音声データが波形表示される。なお，ウィンドウの物理的な大きさはマウスで変更するまで変化しない (当然) が，論理的な大きさ (何 [sec] から何 [sec] までが表示されるか) は音声データを基に自動的に設定される[4]。また，現在の音声データファイル名がウィンドウ右上に表示される。以降，このウィンドウをオリジナルウィンドウと呼ぶ。
2. *Waveform of Synthesized Speech*
   このウィンドウには合成結果が波形表示されるが，合成結果には主に 2 種類ある。一つは，オリジナル音声を "LMA 逆フィルタ→LMA フィルタ" の順に通した結

[4] 当然この値は変更することもできる。

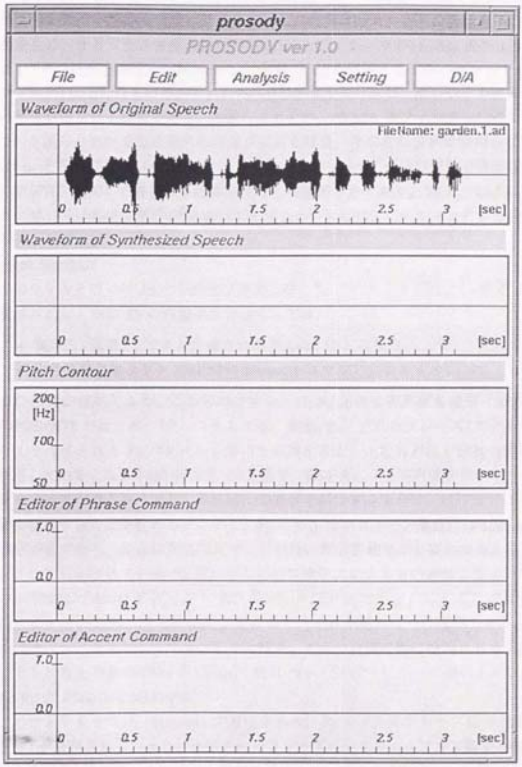


図 A.1. 起動直後の PROSODY

果(分析→音響的パラメータの操作何も無し→合成)であり，他方は，フレーズ/アクセント指令(F_0 モデルパラメータ)を推定し，それに変更を加えて作成される F_0 パターンを用いて分析合成した結果である。前者は，スペクトル特性を $H(z)$ とした場合に，オリジナル音声を逆特性 $H(z)^{-1}$ を持つフィルタ(LMA 逆フィルタ)に通し(その結果音源波形が得られる)，再度 $H(z)$ を特性として持つフィルタ(LMA フィルタ)に通したもので，フィルタの近似誤差が無ければ，オリジナル音声と同一の信号が得られるはずである。注意しておくが，"LMA 逆フィルタ→LMA フィルタ" を通しただけの合成音声の品質が劣悪な場合，その自然音声は使用しない方がよい。と言うのも "LMA 逆フィルタ→LMA フィルタ" のみの操作で得られる合成音が ***PROSODY*** で得られる最高品質の合成音だからである。なお，LMA フィルタに関しては参考文献[93][94]を参照して頂きたい。以降，このウィンドウを合成ウィンドウと呼ぶ。

3. *Pitch Contour*

   このウィンドウには F_0 の自動抽出結果及び，F_0 モデルによる推定(+変更)結果が表示される。なお F_0 の自動抽出方法としては，

   - 瀬戸氏(現東芝)により作成された `f0calc` によるもの。
   - 信号処理汎用ソフト ESPS[5]の `formant` コマンドによるもの。

   の 2 種類が使用できる。これらのコマンドは F_0 以外に有声度も推定・出力する。***PROSODY*** では，F_0 パターンとしては，変形/修正されたフレーズ/アクセント指令より作成される F_0 パターンを用いて合成するが，上記有声度も分析合成の際に利用している。なお 1995 年 3 月 13 日現在，ESPS による有声度を用いた方が高品質の合成音が得られており，最終的に合成音を録音する場合は，ESPS による有声度を使用することを勧める。しかし，F_0 モデルのパラメータ推定には `f0calc` の結果が必要であり，結局は両方のコマンドを用いた分析結果が必要となる。このウィンドウに表示されている F_0 が，どちらのコマンドによるものかは，ウィンドウ上の "*Pitch Contour*" の表示が "*Pitch Contour (Extracted by F0CALC)*" 或は "*Pitch Contour (Extracted by ESPS)*" となるので，それを見れば一目瞭然である。以降，このウィンドウをピッチウィンドウと呼ぶ。また，合成の際に使用される有声度もウィンドウ上の表示 ("*by F0CALC*" 或は "*by ESPS*") に従って選択される。

4. *Editor of Phrase Command*

   このウィンドウには，`f0model` で推定される，F_0 モデルのフレーズ成分(位置&大きさ)が表示される。また，表示されたフレーズ成分をマニュアルで編集(追加・削除も含めて)することもできる。編集の方法については後に触れる。以降，このウィンドウをフレーズウィンドウと呼ぶ。

[5] アメリカ Entropic 社製。日本での代理店は(株)アルゴグラフィックス。

5. *Editor of Accent Command*

このウィンドウには，f0model で推定される，F_0 モデルのアクセント成分 (位置&大きさ) が表示される。また，表示されたアクセント成分をマニュアルで編集 (追加・削除も含めて) することもできる。編集の方法については後に触れる。以降，このウィンドウをアクセントウィンドウと呼ぶ。

A.1.3 *File* メニュー

このメニューでは基本的にはファイルの入出力を行なう。以下順に説明する。

- *Open RAW File*

新たに音声データを取り込む。サポートしているフォーマットは第 A.1.2 節 で述べた通りである。***PROSODY*** の内部で使用されている種々のパラメータは，その殆どが初期化される。

- *Open PAC File*

PAC ファイルを読み込む。このメニューを選択することで，任意の **PAC** ファイルを任意の音声データに適用することができる。

- *Save as RAW File*

合成結果を RAW file フォーマットで出力する。ファイルに出力する時も，ファイルの入力と同様の選択メニューが使用される。

- *Save as PCM File*

合成結果を PCM(VAX) file フォーマットで出力する。入力ファイルのフォーマットとしては PCM(VAX) フォーマットはサポートしていないが，出力は何故か可能である。なお，ファイル番号は 100 番固定である。

- *Save as PAC File*

"推定+マニュアル修正+種々の変更" を加えた **PAC** ファイルを出力する。

- *Save as EPS File*

現在の ***PROSODY*** の状況を EPS ファイルとして出力する。X-Window 上で動作する ***PROSODY*** とほぼ同一のイメージが出力されるはずである。もちろん，LaTeX の中にも取り込むことができる。一例として，図 A.2 に示す。但し epsf.sty で取り込み，かつ，EPS ファイルのヘッダー部において，

```
%%BoundingBox: 28 28 567 814
```

を

```
%%BoundingBox: 28 105 567 814
```

に変更している。

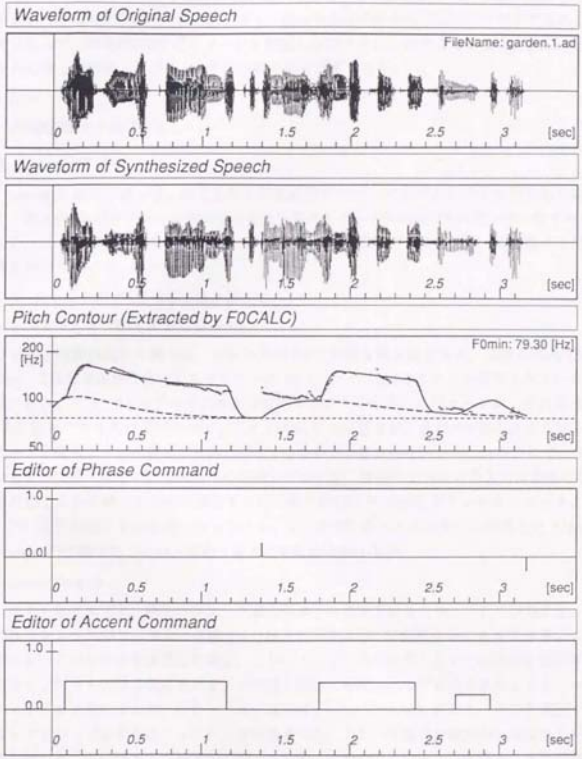


図 A.2. EPS ファイルによる出力と LaTeX への読み込み

- *Preview EPS File*

出力した EPS ファイルを preview する。*PROSODY* の中から ghostview を適当なオプションで呼んでいるだけである。しかし，ghostview の中からプリントアウトすることができるので，*PROSODY* のイメージを奇麗に印刷することができる。なお，ghostview から印刷した場合，ほぼ A4 一杯に印刷されるはずである。

- *Quit*

PROSODY を終了する。

A.1.4 *Edit* メニュー

pacedit などと違って，推定結果を測定結果にマニュアルで合わせるのが目的ではなく，種々の F_0 パターンへと変形することを目的とする *PROSODY* の売りの一つである。このメニューを使用することで F_0 パターンを種々に変形できる訳だが，2 種類の変形方法を利用している。各々，

1. カレントの PAC 情報を書き換える。
2. F_0 モデルの式を多少いじくる。

である。両者の大きな違いは，カレントの PAC 情報を書き換えると，現在の仕様では，Undo に相当する操作としてオリジナルの F_0 パターンに戻ることしか用意されていない点である。一方，F_0 モデルの式を多少変形することで作成したパターンは，その前後で PAC 情報そのものは固定なので，モデルの式を元に戻せば，直前の (修正された)F_0 パターンに戻ることが出来る。言葉で言ってもなかなか分からないかもしれないので，色々いじくって見て欲しい。なお，一つ注意しておくが，数値やテキストの入力は最後に ↵ を押して入力を終了するのではなく，その殆どが *OK* や *Apply* ボタンをクリックすることで，入力が完了する仕様となっている。この仕様は本ツールの作成に使用した Athena Widget の仕様であるため，変更する方法を筆者は知らない。

- *Show Phrase*

ここを選択すると，現在のフレーズ成分の様子を数値で示したウィンドウが現れる。そのウィンドウの中で，フレーズ成分を "値を代入する形" で変更することができる。ここでフレーズ成分の値を変更した場合，フレーズウィンドウなど，フレーズ成分を表示する他ウィンドウも同時に変更される。その逆も同様に同期をとって変更される。また，一度ウィンドウを表示すると，このメニュー項目は，*Hide Phrase* に変わり，そこを選択することでフレーズ成分表示ウィンドウは消去される。フレーズ成分の数が少い時はフレーズを数値でモニターできるため便利だが，フレーズの数が多くなると，画面内に入り切れなくなり，ちと困ることになってしまう(^_^;)。

- *Show Accent*

上記と同じように，アクセント成分表示ウィンドウが現れる。数値によるアクセントの操作，同期のとれた表示変更，ウィンドウの消去方法など，フレーズ成分表示ウィンドウと全く同じである。

- *Add Phrase*

フレーズを新たに一つ追加する。どの時点に追加されるかであるが，現在フレーズウィンドウが表示している音声区間の左端に，立ち上がりのフレーズ指令と，立ち下がりのフレーズ指令が各々一つずつ加えられるはずである。なお，立ち下がりのフレーズは多くの場合，入力音声末尾に固まって推定されるはずである。そこで *Add Phrase* を行なった後に，何も考えずに，立ち下がりのフレーズはすぐ入力末尾へと移動することを勧める (^_^;)。フレーズを 1 つ消す場合の方法は後述。

- *Add Accent*

アクセントを新たに一つ追加する。どの時点に追加されるかであるが，現在アクセントウィンドウが表示している音声区間の左端に，新たに 1 つアクセントが立ち上がるはずである。アクセントを 1 つ消す場合の方法は後述。

- *F0min*

F_0 モデルパラメータの一つである，F_0min を変更する。F_0min の値はピッチウィンドウの右上に表示される。

- *All Phrase -> 0*

フレーズ成分の大きさ (通称 ap) を全て 0.0 に変更する (カレントの PAC 情報を変更する)。フレーズウィンドウ中の各コマンドの大きさも全て 0.0 になる。

- *All Accent -> 0*

アクセント成分の大きさ (通称 aa) を全て 0.0 に変更する (カレントの PAC 情報を変更する)。アクセントウィンドウ中の各コマンドの大きさも全て 0.0 になる。

- *Back to Original*

以上の変更は，全てカレントの PAC 情報を変更するものであり，その結果，フレーズウィンドウ或はアクセントウィンドウに表示される各コマンドの様子も変更されることになる。が，*Back to Original* を選択することで，カレントの PAC 情報としてオリジナルの PAC 情報がセットされる。ここで，オリジナルの PAC 情報とは，

- `f0calc`, `f0model` 等で抽出された PAC 情報。
- *File* メニューの *Open PAC File* で選択された PAC ファイルに記述されている PAC 情報。

のいずれかを指す。いずれを指すかは，上記2つのオペレーションのうち，最も最近に行なわれたオペレーションに依存する。

なお，以下に説明する項目は，カレントの PAC 情報を書き換えるものではなく，PAC 情報と F_0 モデルから F_0 を推定する場合の式を一時的に変更するものである。そのため，元に戻す (Undo) 場合，上記の *Back to Original* を利用するのではなく，各々の項目に応じた Undo が存在する。

- *Constant Pitch*

F_0 を一定にする。この項目を選択すると，F_0 を指定するためのウィンドウが開く。数値を入力して *OK* をクリックすると，入力操作は完了する。F_0 を一定にすると，この項目は，*Constant Pitch* から *Undo Constant* に変更されるので，そこを選択すると Undo できる。また，指定した値は，ピッチウィンドウ中に表示される。

- *+Constant*

現在の F_0 推定値に，指定された周波数を加算する。この項目を選択すると，周波数指定用のウィンドウが開くので，数値を入力し，*OK* をクリックすると，入力操作は完了する。Undo は，再度ウィンドウを開き，*0[Hz]* をクリックすると，加算されるべき周波数が 0[Hz] にリセットされる。加算されるべき周波数が 0 以外の値をとると，ピッチウィンドウの右上の F_0min の下に，現在加算されている値が表示される。

- *+log(Constant)*

現在の F_0 推定値に，指定された周波数を log で加算する。つまり，指定された周波数倍する。この項目を選択すると，周波数指定用のウィンドウが開くので，数値を入力し，*OK* をクリックすると，入力操作は完了する。Undo は，再度ウィンドウを開き，*1[Hz]* をクリックすると，log で加算されるべき周波数が 1[Hz] にリセットされる。log で加算されるべき周波数が 1 以外の値をとると，ウィンドウの右下 (F_0min のライン上) に現在加算されている値が表示される。

- *Average Pitch*

Edit メニューで種々の編集が可能であるが，この項目を利用することで，種々の編集の結果得られる推定パターンの F_0 平均値を常に一定にすることができる。但し，ここでの平均とは，絶対的な値の平均値であり，また有声度が 0.7 以上のフレームを参照して得られる平均である。Undo は，再度ウィンドウを表示し，*Undo* をクリックすればよい。指定した平均値は，ウィンドウ左上付近に表示される。

- *Average Log Pitch*

上記の項目とほぼ同じ機能を持つが，平均値として，$\log(F_0)$ の平均値が一定となるように操作するものである。平均の求め方は log をとる以外は上記と全く同じである (有声

度 0.7 以上のフレームのみ考慮)。Undo の方法も同じであり，指定した平均値もウィンドウ左上付近に表示される。

なお，*Average Pitch* と *Average Log Pitch* は当然のことながら，一度に同時には指定できない。一方を設定すると，他方の設定は無効となる。

- *0.0 x Phrases*

  F_0 モデルは，

$$\log(F_0) = \log(F_0 min) + [\text{Phrase Component}] + [\text{Accent Component}]$$

のように定式化されているが，この項目を選択することで，フレーズ成分を

$$[\text{Phrase Component}] \rightarrow \alpha \times [\text{Phrase Component}]$$

と考え，αの値を $0.0 \rightarrow -1.0 \rightarrow 1.0$ へと変更することができる。なお，ここで言うαを 0.0 にする操作は，*All Phrase -> 0* でも可能である。両者の相違は *All Phrase -> 0* がカレントの **PAC** 情報を書き換えてしまう (Undo は Original へしか戻れない) のに対し，この項目でαを 0.0 にした場合は，直前の (種々の編集を施した)**PAC** 情報へ戻ることができる (つまりお手軽)。

- *0.0 x Accents*

  上記の Accent 成分版である。説明は必要無いだろう。

- フレーズウィンドウでの編集作業

ここで，フレーズウィンドウでの編集作業について説明する。フレーズ成分は，*PROSODY* の *Analysis* メニュー内の項目を用いて分析・抽出するか，*File* メニュー内の項目を用いて **PAC** ファイルを取り込むことで表示される。任意のフレーズ成分の所にマウスポインタを持っていき，マウスの左・中・右をクリック&ドラッグして戴きたい。それぞれ以下の編集が可能である。

- 左クリック&ドラッグ
  任意のフレーズ成分の頭をクリック&ドラッグすると，フレーズ成分の大きさ&位置を任意に変更することができる。頭以外をクリック&ドラッグすると，フレーズの位置のみ変更できる。
- 中クリック&ドラッグ
  任意のフレーズ成分の頭をクリック&ドラッグすると，フレーズ成分の大きさのみを任意に変更することができる。頭以外をクリック&ドラッグすると，フレーズの位置のみ変更できる。
- 右クリック
  任意のフレーズ成分上でクリック (頭でもどこでもよい) すると，指定したフレーズ

に関する情報が表示されたウィンドウが現れる。このウィンドウのスライダーを用いることで指定したフレーズの各パラメータを動かすことができる。この時ピッチウィンドウの F_0 パターンも動的に変動する (結構面白い)。また，このウィンドウの右下に Delete と言うボタンがあるが，ここをクリックすることで指定したフレーズを消去することができる。フレーズの追加・消去の方法に統一がとれていないが，勘弁してもらいたい(^_^;)。

- アクセントウィンドウでの編集作業

アクセントウィンドウでの編集作業もフレーズウィンドウでの作業とほぼ同じである。アクセント成分も，*PROSODY* の *Analysis* メニュー内の項目を用いて分析・抽出するか，*File* メニュー内の項目を用いて PAC ファイルを取り込むことで表示される。任意のアクセント成分の所にマウスポインタを持っていき，マウスの左・中・右をクリック&ドラッグして戴きたい。それぞれ以下の編集が可能である。

- 左クリック&ドラッグ
  任意のアクセント成分の辺をクリック&ドラッグすると，上辺の場合は大きさを，それ以外の時は位置を変更することができる。角をクリック&ドラッグすると，アクセントの大きさ&位置を変更できる。なお，アクセントの上がり時間と下がり時間が交差した場合，マウスボタンを離すまでは $t_1 > t_2$ となるが，マウスボタンを離すと同時に，各々の値が入れ替わり，$t_1 < t_2$ となる。
- 中クリック&ドラッグ
  アクセント成分上でクリック&ドラッグすると，アクセント成分が水平移動する。中クリック&ドラッグでは大きさを変更することはできない。
- 右クリック
  任意のアクセント成分上でクリックすると，指定したアクセントに関する情報が表示されたウィンドウが現れる。このウィンドウのスライダーを用いることで指定したアクセントの各パラメータを動かすことができる。この時 F_0 パターンも動的に変動する。また，このウィンドウの右下に Delete ボタンで指定したアクセントを消去することができる (図 A.3 参照)。

A.1.5 *Analysis* メニュー

本メニューでは，F_0 の抽出，F_0 モデルパラメータの推定，及び合成を行なう。

- *Extract F0 & Esitimate Fujisaki Model*

瀬戸氏 (現東芝) 作の `f0calc`，`f0model` を使用して，F_0 の自動抽出及び F_0 モデルパラメータの自動推定を行なう。`f0calc`，`f0model` の出力するファイル名のフォーマットは固定されているが，*PROSODY* ではその中から必要な情報のみを集めて，新たな別のファイル名で出力する仕様となっている。その結果，

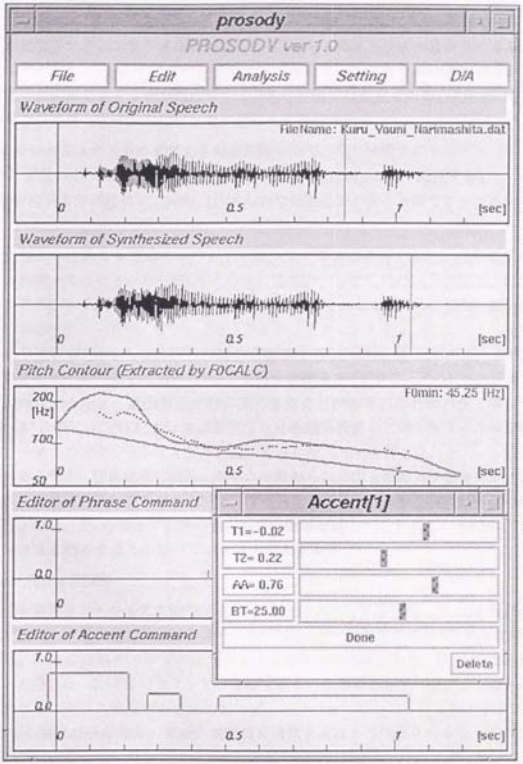


図 A.3. アクセント指令上の右クリックにより現れるウィンドウ

オリジナル音声ファイル名	speech.10kHz.ad
PAC ファイル名	speech.10kHz.PAC
F_0 & 有声度ファイル名	speech.10kHz.fpro

のような拡張子で各々のファイルが作成される (もちろんオリジナル音声ファイルは入力ファイルである)。なお，オリジナル音声ファイルの ".ad" と言うのは固定ではなく，任意の拡張子が使用できる。その他のファイル名の拡張子は固定である。*PROSODY* では分析結果の出力用ファイル名は以下のようにして決定される。

1. オリジナルファイルの末尾より先頭に向かって，"." を探す。
2. "." が見つかった場合，先頭からその場所までを *body* として記憶する。
3. 分析結果出力の際には，*body*.[決められた拡張子] と言う名前でファイルをオープンする。

即ち，拡張子のみ異なる複数のオリジナル音声ファイルが存在すると，分析結果の出力ファイルは同一名となり，*PROSODY* のその後の動作は予想外の結果に陥る。結局，オリジナル音声ファイルの拡張子も各自適当に決めておくことを勧める。分析・抽出が終了すると，ピッチウィンドウ，フレーズウィンドウ，アクセントウィンドウにその結果が表示され，"*Pitch Contour*" の表示が "*Pitch Contour (Extracted by F0CALC)*" に変わる。ここで以下のことに注意して頂きたい。"分析して結果を (グラフィカルに) 表示する" と言うプロセスにおいて，画面表示が妙に遅くなることがある。これはバグではなく "仕様" である[6]。*PROSODY* における波形表示や分析結果表示などのグラフィカルな表示は全て，ユーザーのアクション (マウスを動かすだとか，キーを押すだとか …) によって行なわれる。即ち，分析結果の表示の際もユーザからの，何らかのアクションが必要となる。恐らく，画面表示が遅くなった時も，マウスをちょっと動かすと表示が行なわれることと思う。結局，F_0 の抽出や音声の合成など，計算処理が行なわれている時も，ちょこちょこマウスを動かすよう心掛けていれば大丈夫である。

- *Extract F0 by ESPS*

汎用信号処理ソフトウェア ESPS の中の `formant` コマンドを使用して，F_0 及び有声度その他を抽出し，その出力ファイルより，*PROSODY* で必要となる情報 (F_0&有声度) を抽出する。本来 ESPS のコマンドはライセンス管理がなされており，誰かが ESPS を使用している場合は，ESPS のコマンドは使用できない仕様になっているが，この `formant` コマンドはライセンス管理がなされておらず，いつでもこのメニューを選択することで，F_0 その他の抽出が行なえる。なお，本項目を選択することで作成されるファイルは，

F_0 & 有声度ファイル名	speech.10kHz.epro

[6] 作者の "プログラミングテクニックの乏しさ" とも言われている。

だけである。分析・抽出が終了すると，ピッチウィンドウに抽出された F_0 が表示され，"*Pitch Contour*" の表示が "*Pitch Contour (Extracted by ESPS)*" に変わる。

- *Synthesis with Inverse Filter*

本項目を選択することにより，

オリジナル音声→LMA 逆フィルタ→LMA フィルタ→再合成音

と言うプロセスを通って作成される合成音が作成される。単に特性 $H(z)^{-1}$ と $H(z)$ のフィルタを通すだけなので，近似誤差がなければオリジナルと全く同一の音声が合成音として得られるはずである。なお，本項目を選択することで作成されるファイルは，

Cepstrum ファイル	speech.10kHz.cep
理想音源ファイル	speech.10kHz.glt
合成音声ファイル	speech.10kHz.syn

である。理想音源ファイルとは，"オリジナル音声→LMA 逆フィルタ" の段階で得られるデータを指す。合成が終了すると，合成ウィンドウにその結果が表示される。なお *PROSODY* では，一つのオリジナル音声データに対してデータ固有 (unique) の分析・抽出結果は，基本的に小文字の拡張子 (PAC ファイルは例外) のファイルに保存される。そして，新たに分析・抽出する際には該当する拡張子のファイルの有無を調べ，ファイルが無い場合のみ分析・抽出が行なわれる。

- *Synthesis with Estimated Pitch*

本項目を選択することで，*PROSODY* のメインの機能である分析合成を行なうことができる。推定+マニュアル修正+変更された F_0 パターンを元に音源波形を生成し，その音源を声道特性を近似した LMA フィルタに通す訳である。その結果作成されるファイルは，

Cepstrum ファイル	speech.10kHz.cep
理想音源ファイル	speech.10kHz.glt
有声部ファイル	speech.10kHz.VCD
修正 F_0& 有声度ファイル	speech.10kHz.PRO
修正音源ファイル	speech.10kHz.GLT
合成音声ファイル	speech.10kHz.SYN

である。上の 2 つは *Synthesis with Inverse Filter* などを通して作成済みであれば，新たに作成されることはない。下の 4 つは本項目が選択される度に書き替わるファイルである。また，合成が終了すると当然のことながら，合成ウィンドウにその結果が表示される (勿論，ユーザー側のアクションに反応して)。

- *Marked Synthesis with Estimated*

Synthesis with Estimated Pitch はオリジナル音声全体を通しての分析合成だが，オリジナル音声長が 4[sec] ほどあると (.cep, .glt ファイルが抽出済みの場合でも) 約 1 分ほど時間がかかる。そこで，注目したい部分のみの分析合成を行なうために本項目を付け加えた。オリジナルウィンドウで右クリック・左クリックをすると，クリックした場所に破線が描かれる。本項目を選択すると，オリジナルウィンドウで選択した部分のみに注目して分析合成し，既に合成結果が表示されてある場合は，注目した部分の分析合成結果を，既にある合成結果と置換することになる。当然のことながら，繋ぎの部分は若干不連続になることもあり，最終的には，*Synthesis with Estimated Pitch* で全体の分析合成を行なうことを勧める。

なお，*Analysis* メニューで行なう処理はいずれも多少時間がかかる。項目選択後，すぐに新たな項目が選択可能な状態になるが，処理中はこの *Analysis* メニューの各項目は選択できない仕様となっている。これは中途半端に作成されたファイルを元に次の処理を行なってしまわないようにとの配慮からである。

A.1.6 *Setting* メニュー

本メニューは第 A.1.2 節で述べた 5 つのウィンドウの

- X 軸方向の最小値・最大値。
- Y 軸方向の最小値・最大値。

など，各ウィンドウの設定に関する項目が並んでいる。

- *Max & Min of X-Axis*

本項目を選択することで，5 つのウィンドウに共通した X 軸の最小・最大値を設定することができる。入力ウィンドウが現れるので，最小・最大値を各々入力し，*Apply* ボタンをクリックすれば適用される。また，再度ウィンドウを出し *Default* をクリックすればデフォルトの値に戻る。

- *Between Original Marks*

Max & Min of X-Axis は数値を入力する形での指定であったが，本項目はオリジナルウィンドウにおいて指定したマークを元に，5 つのウィンドウに共通した X 軸の最小・最大値を設定するものである。オリジナルウィンドウにおいて右・左クリックすると破線が描かれる (マーク)。これを元に X 軸の最小・最大値を決定する。デフォルトに戻す場合は，*Max & Min of X-Axis* の *Default* を利用するように。

以上 2 つ項目は 5 つのウィンドウ全ての最小・最大値を変更したが，以下 5 つの項目は各ウィンドウ毎の設定である。

- *Orisinal Speech*

オリジナルウィンドウにおける，X軸の最小・最大値，Y軸の最小・最大値を数値を代入する形で変更する。*Default* ボタンは当然のことながら，オリジナルウィンドウのみをデフォルトに戻すものである。

- *Synthesized Speech*

合成ウィンドウにおける，X軸の最小・最大値，Y軸の最小・最大値を数値を代入する形で変更する。*Default* ボタンは当然のことながら，合成ウィンドウのみをデフォルトに戻すものである。

- *Pitch Contour*

ピッチウィンドウにおける，X軸の最小・最大値，Y軸の最小・最大値を数値を代入する形で変更する。*Default* ボタンは当然のことながら，ピッチウィンドウのみをデフォルトに戻すものである。

- *Phrase Command*

フレーズウィンドウにおける，X軸の最小・最大値，Y軸の最小・最大値を数値を代入する形で変更する。*Default* ボタンは当然のことながら，フレーズウィンドウのみをデフォルトに戻すものである。

- *Accent Command*

アクセントウィンドウにおける，X軸の最小・最大値，Y軸の最小・最大値を数値を代入する形で変更する。*Default* ボタンは当然のことながら，アクセントウィンドウのみをデフォルトに戻すものである。

- *Back to Defaults*

5 つのウィンドウに対して，X軸及び Y 軸の最小・最大値をデフォルトの値に戻す。

- *Monitor Current Marks*

オリジナルウィンドウ及び合成ウィンドウにおいて右・左クリックすることで破線が描かれる (マーク) が，本項目を選択すると，現在のマークの位置をモニターするためのウィンドウが現れる。なお単位は [sample] であり，マークが未設定の場合は空欄となる。

A.1.7 *D/A* メニュー

オリジナル音声及び表示されている合成音声を DA する。「オリジナル音声か合成音声か」,「全体かマーク間だけか」により 4 通りの組合せがある。また，***PROSODY*** から DA する場合第 **A.1.1** 節に述べたように，基本的には全ホストから DA できるように設定している (但し完全ではない)。

- *Whole Original*

オリジナル音声全体を DA する。

- *Marked Original*

  オリジナル音声をマーク間だけ DA する。

- *Whole Synthesized*

  表示されている合成音全体を DA する。

- *Marked Synthesized*

  表示されている合成音をマーク間だけ DA する。

A.1.8 その他

その他，*PROSODY* を使用して気づいた点をいくつか挙げる。

- 女性の声に対してもそれなりの結果は出すが，多少不安定な所があるようにも思う。具体的には合成音声がサチッテしまう場合があった。これは *PROSODY* における LMA フィルタの構成が，入力音声に対して動的に適用するようには作成されていない点に起因すると考えられる[7]。
- 男性の声の場合でもオリジナル音声のレンジがギリギリ一杯のところまでとってあると (~0.8×MAXSHORT)，サチルことがあった。まあ，~20,000 程度で AD して下さい。
- 第 A.1.5 節でも述べたように，種々のパラメータファイルを出力するので，結構ディスクを食う。
- 理想音源から修正音源を作成する場合，有声度の高い部分は人工的に作られた音源波形 (ほぼ三角波) を，有声度の低い部分は理想音源波形をそのまま使用している。しかし，母音部分の有声度が低く推定されることがあり，その場合 F_0 パターンをいじっても LMA フィルタの入力として，理想音源波形がそのまま使用されるため，F_0 が変わらない部分が生じることがある。F_0 一定の合成音などを作る場合，極端に低い/高い F_0 パターンの合成音を作って，有声部が指定した F_0 で合成されているかを確かめる必要がある。
- そのような場合は，可能ならもう一度音声を録り直すのが最も良いように思う。少くとも，ソースファイルととっ組合いをするより，時間的には早く解決するだろう。

いずれにしても完全なツールではなく，やはり "癖" のようなものが存在する。何度か使ってみて，その "癖" と上手に付き合うことも必要だろう。

また，バグの可能性も無いとは言い切れない(^_^;)。妙な動作を発見した場合は，

mine@gavo.t.u-tokyo.ac.jp

まで連絡して頂ければ解決するかもしれない …… 。

[7] 今後時間がある時に... と言った具合である。

謝 辞

卒論研究から博士課程に至る約5年半の長きに渡り，研究においてはもちろん，生活面においても，様々な助言・御指導を賜りました広瀬啓吉教授に深く感謝申し上げます。また，人間の音声知覚過程の研究に関しては，この分野の第一人者でもあられる東京大学名誉教授 (現東京理科大学教授) の藤崎博也教授から昼夜問わず，熱心な御指導を賜り，ここに深謝申し上げます。知覚実験の実施に際しては，広瀬研究室の方々を始め，東京理科大学・藤崎研究室の方々にも参加して戴きました。また，実験計画及び実験後のデータ収集の際には，伊藤みか君 (1991 年度広瀬研究室卒業生，現ソニー)・松下哲君 (1994 年度広瀬研究室卒業生，現ソニー) にも協力して戴きました。計算機上での認識実験では，石金正明君 (1994 年度広瀬研究室卒業生，現東京大学大学院今井研究室) に協力して頂きました。ここに感謝申し上げます。更に，本研究を通して常に良きアドバイザーとして相談に応じて頂いた雷海清氏 (1991 年度広瀬研究室博士課程修了，現東芝)，大野澄雄氏 (1992 年度広瀬研究室博士課程修了，現東京理科大学助手)，浅野康治氏 (1992 年度広瀬研究室博士課程修了，現ソニー)，瀬戸重宣氏 (1990 年度広瀬研究室修士課程修了，現東芝) の諸先輩方には感謝の言葉が見つかりません。その他，研究室における研究環境を常に整備して頂いた高橋登技官，履修登録等の事務処理以外にも，5年半にも及ぶ研究室生活に様々な意味で "色" を添えて頂いた，電気/電子/電子情報学科の事務室の方々にも感謝致します。最後に，28 歳になるまで，息子の我儘を心良く聞き入れてくれた両親に感謝致します。

本当に有難うございました。

発表論文一覧

投稿論文

1. N.Minematsu and K.Hirose, "Duration Modeling with Decreased Intra-group Temporal Variation for HMM-based Phoneme Recognition," *IEICE Trans. Fundamentals*

2. N.Minematsu and K.Hirose, "Role of Prosodic Features in the Human Process of Perceiving Spoken Words and Sentences in Japanese," *J. Acoust. Soc. Jpn. (E)*

国際会議論文

3. H.Fujisaki, K.Hirose, S.Ohno and N.Minematsu, "Influence of Context and Knowledge on the Perception of Continuous Speech," *Proc. ICSLP'90*, **10.9.1**, pp.417-420 (1990).

4. N.Minematsu, S.Ohno, K.Hirose and H.Fujisaki, "The Influence of Semantic and Syntactic Information on Spoken Sentence Recognition," *Proc. ICSLP'92*, **Tu.fPM.4.5**, pp.153-156 (1992).

5. K.Hirose, N.Minematsu and M.Ito, "Experimental Study on the Role of Prosodic Features in the Human Processes of Spoken Word Perception," *Proc. ESCA Workshop*, Working Papers **41**, pp.200-203 (1993).

6. N.Minematsu and K.Hirose, "Speech Recognition Using HMM with Increased Uniformity in the Temporal Structure," *Proc. ICSLP'94*, **7.1**, pp.187-190 (1994).

7. N.Minematsu and K.Hirose, "Role of Prosodic Features in the Human Process of Speech Perception," *Proc. ICSLP'94*, **21.8**, pp.1151-1154 (1994).

8. N.Minematsu and K.Hirose, "Influence of Prosodic Features on the Human Process of Speech Perception," *Proc. Japan-China Symposium on Advanced Information Technology*, pp.69-75 (1994).

国内学会・研究会発表資料

9. 大野澄雄，峯松信明，広瀬啓吉，"連続音声のための人間の音声知覚過程の検討"，連続音声シンポジウム資料，**SPREC-91-2**，pp.21-24 (1992).

10. 峯松信明，大野澄雄，広瀬啓吉，藤崎博也，"連続音声知覚における高次言語情報の及ぼす影響"，日本音響学会聴覚研究会資料，**H-92-56**，pp.1-6 (1992).

11. 峯松信明，広瀬啓吉，"動的に制御された音響的特徴を用いた音声認識"，音声入出力機器の現状とヒューマンインターフェイスの課題シンポジウム資料，**SPREC-93-2-13** (1993).

12. 峯松信明，広瀬啓吉，"音響的特徴量の動的制御に基づく音声認識"，電子情報通信学会信学技報，**SP93-105**，pp.9-16 (1993).

13. 峯松信明，広瀬啓吉，“連続音声知覚における韻律的特徴の果たす役割に関する実験的検討”，電子情報通信学会信学技報，EA94-46，pp.25-32 (1994).

14. 藤崎博也，広瀬啓吉，大野澄雄，峯松信明，“人間の内部辞書の構成と検索方式に関する検討”，日本音響学会春季講演論文集，3-3-17，pp.103-104 (1990).

15. 藤崎博也，広瀬啓吉，大野澄雄，峯松信明，“人間の内部辞書の構成・検索に関する実験的検討”，日本音響学会秋季講演論文集，3-8-2，pp.97-98 (1990).

16. 藤崎博也，広瀬啓吉，大野澄雄，峯松信明，“知識が音声知覚過程に及ぼす影響に関する実験的検討”，日本音響学会春季講演論文集，1-8-14，pp.309-310 (1991).

17. 広瀬啓吉，大野澄雄，峯松信明，藤崎博也，“音声認識における音響的特徴表現の時間単位に関する検討”，日本音響学会秋季講演論文集，2-P-8，pp.153-154 (1991).

18. 大野澄雄，峯松信明，広瀬啓吉，藤崎博也，“音素系列辞書に基づく連続音声中の語句認識における鳥駆動的照合処理の検討”，日本音響学会秋季講演論文集，2-P-19，pp.175-176 (1991).

19. 峯松信明，大野澄雄，広瀬啓吉，藤崎博也，“複数の時間単位・精度の音響的特徴表現を用いた音声認識”，日本音響学会春季講演論文集，1-1-16，pp.31-32 (1992).

20. 大野澄雄，峯松信明，広瀬啓吉，藤崎博也，“連続音声中の語句の照合における種々のレベルの辞書情報の利用”，日本音響学会春季講演論文集，3-1-14，pp.97-98 (1992).

21. 峯松信明，大野澄雄，広瀬啓吉，藤崎博也，“意味的内容が音声知覚過程に及ぼす影響に関する実験的検討”，日本音響学会秋季講演論文集，2-9-16，pp.377-378 (1992).

22. 峯松信明，広瀬啓吉，伊藤みか，“音声知覚過程における韻律的特徴の果たす役割”，日本音響学会秋季講演論文集，2-9-18，pp.381-382 (1992).

23. 峯松信明，広瀬啓吉，“音声認識における大局的特徴の利用に関する一考察”，日本音響学会春季講演論文集，2-Q-4，pp.92-93 (1993).

24. 峯松信明，広瀬啓吉，“音声認識において使用される音響的特徴空間の動的変化”，日本音響学会秋季講演論文集，1-8-9，pp.17-18 (1993).

25. 峯松信明，広瀬啓吉，“音声認識における正規化特徴量の利用に関する実験的検討”，日本音響学会春季講演論文集，1-Q-19，pp.147-148 (1994).

26. 峯松信明，広瀬啓吉，“HMM に対する音声の時間構造の記述に関する検討”，日本音響学会秋季講演論文集，1-R-4，pp.199-200 (1994).

27. 峯松信明，広瀬啓吉，“文音声知覚過程における韻律的特徴の果たす役割”，日本音響学会秋季講演論文集，2-5-18，pp.311-312 (1994).

28. 峯松信明，松下哲，広瀬啓吉，“文音声の多重提示実験における韻律的特徴のもたらす効果”，日本音響学会春季講演論文集，3-9-4 (1995).

参考文献・図書

以下に，本研究において参照した文献・図書について記す。なお，研究或は論文作成を進めるに際に，非常に有益であった図書については，分野を問わず記すことにする[1]。本論文に目を通した方にとって参考になれば幸いである。

[1] 古井貞熙，"デジタル信号処理"，東海大学出版会 (1985).

[2] 嵯峨山茂樹，鷹見淳一，永井明人，H.Singer，竹沢寿幸，谷戸文廣，鈴木雅実，森元逞，樽松明，"自動翻訳電話実験システム ASURA の概要"，日本音響学会春季講演論文集，**3-4-17**，pp.83-84 (1993).

[3] 中津良平，"音声認識・合成技術の市場動向"，日本音響学会誌 **48**，No. 1，特集—音声—，pp.60-65 (1992).

[4] H.Kuwahara and H.Sakai, "Perception of Vowels and CV Syllables Segmented from Connected Speech," *J. Acoust. Soc. Jpn.*, **28**, pp.225 (1972).

[5] 嵯峨山茂樹，"数理統計モデルによる音声認識の現状と将来"，日本音響学会誌 **48**，No. 1，特集—音声—，pp.26-32 (1992).

[6] 小畑秀文，"音声認識のはなし"，日刊工業新聞社 (1983).

[7] S.E.Levinson, L.R.Rabiner and M.M.Sondhi, "An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition," *AT&T. Tech. J.*, **62**, 4 (1983).

[8] 天野成昭，"単語知覚モデルの研究動向"，日本音響学会誌 **48**，No. 1，特集—音声—，pp.20-25 (1992).

[9] 粕谷秀樹，"音声特集号の編集にあたって"，日本音響学会誌 **48**，No. 1，特集—音声—，pp.2 (1992).

[10] 増山英太郎，"感性に関する計量法の研究"，平成 4 年度重点領域研究「感性情報処理の情報学・心理学的研究」，第 2 回全体会議予稿集，pp.25-28 (1993).

[11] 徃住彰文，原田悦子，"高次知識構造が関与する感性的認知への記号計算モデル的研究"，重点領域研究「感性情報処理の情報学・心理学的研究」，平成 5 年度成果報告書，pp.49-52 (1994).

[12] 広瀬啓吉，藤崎博也，柳田益造，"音声コミュニケーションにおける感性情報の表出・受容過程の定量的解析とモデル化"，重点領域研究「感性情報処理の情報学・心理学的研究」，平成 5 年度成果報告書，pp.227-230 (1994).

[13] 中村敏枝，"メディアにおける "間 (ま)" の心理学的研究"，重点領域研究「感性情報処理の情報学・心理学的研究」，平成 5 年度成果報告書，pp.111-114 (1994).

[1] 但し，論文からの引用が必ずしもある訳ではない。

[14] 額賀雅夫，河原達也，堂下修司，"声質の感性的評価の処理モデル"，情報通信学会技術報告，HC93-67 (1994).

[15] H.Rubenstein and I.Pollack, "Word Predictability and Intelligibility," *J. Verbal Learning and Verbal Behavior*, **2**, pp.147-158 (1963).

[16] R.M.Warren, "Perceptual Restoration of Missing Speech Sounds," *Science*, **167**, pp.392-393 (1970).

[17] P.E.Rubin, M.T.Turvey, P.van Gelder, "Initial Phonemes Are Detected Faster in Spoken Words Than Spoken Nonwords," *Perception and Psychophysics*, **19**, pp.394-398 (1976).

[18] R.A.Cole, "Listening for Mispronunciations : A Measure of What We Hear during Speech," *Perception and Psychophysics*, **13**, pp.153-156 (1973).

[19] 今泉敏，森浩一，米田孝一，桐谷滋，宮城島一明，湯本真人，"複合音聴覚誘発脳磁図のディコンボルーションの試み"，日本音響学会秋季講演論文集，2-9-21，pp.423-424 (1993).

[20] 平原達也，"母音知覚と聴覚系内スペクトル表現"，日本音響学会春季講演論文集，2-7-13，pp.341-342 (1993).

[21] 筧一彦，"音声知覚過程の研究"，NTT R&D，41，12 (1992).

[22] 柏野牧夫，"閉鎖区間の前後に分散する手掛かりに基づく日本語閉鎖子音の知覚"，日本音響学会誌 48，No. 2，pp.76 (1992).

[23] 加藤和美，筧一彦，"話者情報の音素知覚における役割"，聴覚研究会資料，H-91-42 (1991).

[24] 天野成昭，"単語内音韻知覚における心的辞書の役割"，聴覚研究会資料，H-87-71 (1987).

[25] M.A.Blank and D.J.Foss, "Semantic Facilitation and Lexical Access during Sentence Processing," *Memory and Cognition*, **6**, pp.644-652 (1978).

[26] A.Salasoo and D.B.Pisoni, "Interaction of Knowledge Sources in Spoken Word Identification," *J. Memory and Language*, **24**, pp.210-231 (1985).

[27] 峯松信明，"人間の音声知覚過程の分析とそのモデル化"，東京大学工学部電子工学科，卒業論文 (1990).

[28] D.B.Pisoni, H.C.Nusbaum, P.A.Luce and L.M.Slowiaczek, "Speech Perception, Word Recognition and the Structure of the Lexicon," *Speech Communication*, **4**, pp.75-95 (1985).

[29] 広瀬啓吉，高橋登，酒井敦正，薩崎博也，村田博士，西山誠一，"音声の韻律的特徴における発話意図の知覚"，日本音響学会春季講演論文集，2-8-15，pp.223-224

(1993).

[30] 藤崎博也，村田博士，西山誠一，広瀬啓吉，高橋登，酒井敦正，"音声の韻律的特徴による発話意図の表現"，日本音響学会春季講演論文集，**2-8-16**，pp.225-226 (1993).

[31] 上床弘幸，小林豊，新美康永，"音声の感情表現の分析とモデル化"，電子情報通信学会信学技報，**SP92-131**，pp.65-72 (1993).

[32] 今泉敏，志村洋子，首藤敏元，"乳児音声における感性情報表出の発達"，電子情報通信学会第二種研究会資料，**LK92-9**, pp.1-8 (1992).

[33] 宇田川博文，藤崎博也，"音声知覚における辞書照合に関する実験的検討"，日本音響学会秋季講演論文集，**3-5-20**，pp.151-152 (1987).

[34] 大河内正明，"Hidden Markov Model に基づいた音声認識"，日本音響学会誌 **42**，**No. 12**，pp.936-941 (1992).

[35] 中川聖一，"ストキャスティック DP 法および統計的手法による不特定話者の英語子音認識"，電子通信学会論文誌，**J70-D**，**1**，pp.155-163 (1987).

[36] 橋本泰秀，平田好充，中川聖一，"連続出力分布型 HMM による日本語音韻認識の検討"，電子通信学会技術研究報告，**SP89-48** (1989).

[37] L.R.Rabiner, B.H.Juang, S.E.Levinson and M.M.Sondhi, "Recognition of Isolated Digits Using Hidden Markov Models with Continuous Mixture Densities," *AT&T. Tech. J.*, **64**, **6**, pp.1211-1231, (1985).

[38] 峯松信明，広瀬啓吉，"音響的特徴量の動的制御に基づく音声認識"，電子情報通信学会信学技報，**SP93-105**，pp.9-16 (1993).

[39] N.Minematsu and K.Hirose, "Speech Recognition Using HMM with Increased Uniformity in the Temporal Structure," *Proc. ICSLP'94*, **7.1**, pp.187-190 (1994).

[40] 嶮峨山茂樹，"音素環境クラスタリングの原理とアルゴリズム"，音声研究会資料，**SP87-86** (1987).

[41] 甘利俊一，中川聖一，鹿野清宏，東倉洋一，"音声・聴覚と神経回路網モデル"，オーム社 (1990).

[42] J.Morton, "Interaction of Information in Word Recognition," *Psychological Review*, **76**, **2**, pp.165-178 (1969).

[43] 御領 謙，"読むということ"，認知科学選書 5，東京大学出版 (1987).

[44] D.E.Meyer, R.W.Schvaneveldt and M.G.Ruddy, "Loci of Context Effects in Visual Word Recognition," in: P.M.A.Rabbit and S.Dornic, eds., *Attention and Performance*, **5**, Academic Press, New York (1975).

[45] M.K.Tanenhaus, H.P.Flanagan and M.S.Seidenberg, "Orthographics and Phonological Activationn in Auditory and Visual Word Recognition," *Memory and Cognition*,

8, pp.513-520 (1980).

[46] L.M.Slowiaczek, H.C.Nusbaum and D.B.Pisoni, "Acoustic-Phonetic Priming in Auditory Word Recognition," *Cognitive Psychology* (1984).

[47] D.E.Broadbant, "Word-Frequency Effect and Reponse Bias," *Psych. Rev.*, **74**, pp.1-15 (1967).

[48] R.A.Cole and J.Jakimik, "A Model of Speech Perception," in: R.Cole, ed., *Perception and Production of Fluent Speech*, Erlbaum, Hillsdale, N.J., pp.133-163 (1980).

[49] W.D.Marslen-Wilson and A.Welsh, "Processing Interactions and Lexical Access during Word Recognition in Continuous Speech," *Cognitive Psychology*, **10**, pp.29-63 (1978).

[50] W.D.Marslen-Wilson and L.K.Tyler, "The Temporal Structure of Spoken Language Understanding," *Cognition*, **8**, pp.1-71 (1980).

[51] S.G.Nooteboom, "Lexical Retrieval from Fragments of Spoken Words: Beginnings vs. Endings," *J. Phon.*, **9**, pp.407-424 (1981).

[52] W.D.Marslen-Wilson, "Functional Parallelism in Spoken Word Recognition," *Cognition*, **25**, pp.71-102 (1987).

[53] H.C.Nusbaum and D.B.Pisoni, "Human Speech Perception: Implications for Computer Speech Recognition," in: W.A.LLea,ed., *Towards Robustness in Speech Recognition* (1984).

[54] D.B.Pisoni and P.A.Luce, "Speech Perception: Research, Theory and the Principal Issues," in *Pattern Recognition by Humans and Machines*, **1**, Speech Perception, E.C.Schwab and H.C.Nusbaum, Eds., pp.1-50 (1986).

[55] D.W.Shipman and V.W.Zue, "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition," *Proc. ICASSP'82* (1982).

[56] D.P.Huttenlocher and V.W.Zue, "A Model of Lexical Access Based on Partial Phonetic Information," *Proc. ICASSP'84*, **2** (1984).

[57] D.H.Klatt, "Speech Perception: A Model of Acoustic-Phonetic Analysis and Lexical Access," in: R.Cole, ed., *Perception and Production of Fluent Speech*, Erlbaum, Hillsdale, N.J., pp.243-288 (1980).

[58] J.L.McClelland, "The TRACE Model of Speech Perception," *Cognitive Psychology*, **18**, pp.1-86 (1986).

[59] 津崎実, "音声知覚モデルへの考察", マルコフモデル・ニューラルネットワークを包含する新しい音声認識手法の総合的研究, 第7回研究討論会資料 (1991).

[60] L.M.Slowiaczek, H.C.Nasbaum and D.B.Pisoni, "Phonological Priming in Auditory Word Recognition," *J. Exp. Psychol. Learn. Mem. Cognit.*, **13**, pp.64-75 (1987).

[61] H.Fujisaki, "On the Modes and Mechanisms of Speech Perception-Analysis and Interpretation of Categorical Effects in Discrimination," *B.Lindblom and S.Ohman, des., Frontiers of Speech Communication Research* (1979).

[62] 藤崎博也，広瀬啓吉，大野澄雄，峯松信明，"人間の内部辞書の構成・検索に関する実験的検討"，日本音響学会秋季講演論文集，**3-8-2**，pp.97-98 (1990).

[63] S.Amano, "Lexical and Coarticulatory Effects on Phoneme Monitoring before and after a Word Identification Point in Spoken Japanese Words," *Proc. ICSLP'90*, **10.4**, pp.397-400 (1990).

[64] 天野成昭，近藤公久，筧一彦，"日本語単語の親密度の大規模評定実験"，日本音響学会春季講演論文集，**3-4-1**，pp.345-346 (1994).

[65] 近藤公久，天野成昭，筧一彦，"視覚呈示と聴覚呈示による日本語単語の親密度評価値の差異"，日本音響学会春季講演論文集，**3-4-2**，pp.347-348 (1994).

[66] H.Fujisaki, K.Hirose, H.Udagawa and N.Kanedera, "A New Approach to Continuous Speech Recognition Based on Considerations on Human Processes of Speech Perception," *Proc. ICASSP'86*, pp.1959-1962 (1986).

[67] 宇田川博文，大野澄雄，藤崎博也，"人間の音声知覚過程の知見に基づく音声認識方式"，電子通信学会技術研究報告，**SP87-90** (1987).

[68] 藤崎博也，広瀬啓吉，大野澄雄，峯松信明，"人間の内部辞書の構成と検索方式に関する検討"，日本音響学会春季講演論文集，**3-3-17**，pp.103-104 (1990).

[69] S.Takeda and A.Ichikawa, "Analysis of Prosodic Features of Prominence in Spoken Japanese Sentences," *Proc. ICSLP'90*, **12.3**, pp.493-496 (1990).

[70] J.Azuma and Y.Tsukuma, "Prosodic Features Marking the Major Syntactic Boundary of Japanese: A Study on Syntactically Ambiguous Sentences of the Kinki Dialect," *Proc. ICSLP'90*, **11.9**, pp.453-455 (1990).

[71] J.J.Venditti and H.Yamashita, "Prosodic Information and Processing of Temporarily Ambiguous Constructions in Japanese," *Proc. ICSLP'94*, **21.6**, pp.1147-1150 (1994).

[72] A.Salasoo and D.B.Pisoni, "Sources of Knowledge in Spoken Word Recognition," *J. Verbal Learn. Verbal Behav.*, **23**, pp.210-231 (1984).

[73] 藤崎博也，広瀬啓吉，宇田川博文，金寺登，平田恭二，"音声知覚における文脈の役割に関する検討"，日本音響学会春季講演論文集，**3-5-7** (1987).

[74] H.Fujisaki, K.Hirose, S.Ohno and N.Minematsu, "Influence of Context and Knowledge on the Perception of Continuous Speech," *Proc. ICSLP'90*, **10.9**, pp.417-420 (1990).

[75] 佐久間英, "日本人の姓", 六藝書房

[76] W.D.Marslen-Wilson, "Linguistic Structure and Speech Shadowing at Very Short Latencies," *Nature*, **244**, pp.522-523 (1973).

[77] 峯松信明, "人間の音声知覚過程の分析・モデル化とその工学的応用" 東京大学工学部電子工学科, 修士論文 (1992).

[78] 峯松信明, 広瀬啓吉, 伊藤みか, "音声知覚過程における韻律的特徴の果たす役割", 日本音響学会秋季講演論文集, **2-9-18**, pp.381-382 (1992).

[79] K.Hirose, N.Minematsu and M.Ito, "Experimental Study on the Role of Prosodic Features in the Human Processes of Spoken Word Perception," *Proc. ESCA Workshop*, Working Papers **41**, pp.200-203 (1993).

[80] N.Minematsu and K.Hirose, "Role of Prosodic Features in the Human Process of Speech Perception," *Proc. ICSLP'94*, **21.8**, pp.1151-1154 (1994).

[81] 峯松信明, 広瀬啓吉, "連続音声知覚における韻律的特徴の果たす役割に関する実験的検討", 電子情報通信学会信学技報, **EA94-46**, pp.25-32 (1994).

[82] 平山輝男, "全国アクセント辞典", 東京堂出版 (1955).

[83] 安居院猛, 中島正之, "コンピュータ音声処理", 産報出版 (1980).

[84] 藤崎博也, 須藤寛, "日本語単語アクセントの基本周波数パターンとその生成機構のモデル", 日本音響学会誌 **27**, pp.445-453 (1971).

[85] F. Grosjean, "Spoken Word Recognition Processes and the Gating Paradigm," *Perception and Psychophysics*, **28**, pp.267-283, (1980).

[86] 藤崎博也, 広瀬啓吉, 大野澄雄, 峯松信明, "知識が音声知覚過程に及ぼす影響に関する実験的検討", 日本音響学会春季講演論文集, **1-8-14**, pp.309-310 (1991).

[87] N.Minematsu, S.Ohno, K.Hirose and H.Fujisaki, "The Influence of Semantic and Syntactic Information on Spoken Sentence Recognition," *Proc. ICSLP'92*, **Tu.fPM.4.5**, pp.153-156 (1992).

[88] 峯松信明, 大野澄雄, 広瀬啓吉, 藤崎博也, "連続音声知覚における高次言語情報の及ぼす影響", 日本音響学会聴覚研究会資料, **H-92-56**, pp.1-6 (1992).

[89] 田中良久, "心理学的測定法", 東京大学出版会

[90] 石村貞夫, "分散分析のはなし", 東京図書 (1992).

[91] 峯松信明，大野澄雄，広瀬啓吉，藤崎博也，"意味的内容が音声知覚過程に及ぼす影響に関する実験的検討"，日本音響学会秋季講演論文集，2-9-16，pp.377-378 (1992).

[92] 峯松信明，広瀬啓吉，"文音声知覚過程における韻律的特徴の果たす役割"，日本音響学会秋季講演論文集，2-5-18，pp.311-312 (1994).

[93] 今井聖，"対数振幅近似 (LMA) フィルター"，電子通信学会論文誌，J63-A，12，pp.886-893 (1980).

[94] 今井聖，北村正，"対数振幅特性近似フィルタを用いた音声の分析合成系"，電子通信学会論文誌，J61-A，6，pp.527-534 (1978).

[95] N.Merhav and Y.Ephraim, "Hidden Markov Modeling Using a Dominant State Sequence with Application to Speech Recognition," *Computer Speech and Language*, 5, 4, pp.327-339 (1991).

[96] 今井聖，阿部芳春，"改良ケプストラム法によるスペクトル包絡の抽出"，電子通信学会論文誌，J62-A，4，pp.217-223 (1979).

[97] 広瀬啓吉，藤崎博也，河井恒，山口幹雄，"基本周波数パターン生成過程モデルに基づく文章音声の合成"，電子情報通信学会論文誌，J72-A，1，pp.32-39 (1989).

[98] 北原義典，武田昌一，市川熹，東倉洋一，"言語音声認知における韻律の役割"，電子情報通信学会論文誌，J70-D，11，pp.2095-2101 (1987).

[99] E.F.Evans, "The Sharpening of Cochlear Frequency Selectivity in the Normal and Abnormal Cochlea", *Audiology*, 14, pp.419-442 (1975).

[100] J.E.Rose and et al., "Patterns of Activity in Single Auditory Nerve Fibers of the Squirrel Monkey," *Hearing Mechanisms in Vertebrates*, eds. A.V.S. de Reuck and J.Knight, Churchill, London (1968).

[101] B.Delgutte, "Representation of Speech-like Sounds in the Discharge Patterns of Auditory-Nerve Fibers," *J. Acoust. Soc. Amer.*, 68, 3, pp.843-857 (1980).

[102] J.W.Mullennix and D.B.Pisoni, "Stimulus Variability and Processing Dependencies in Speech Perception," *Perception and Psychophysics*, 47, pp.379-390 (1990).

[103] 藤崎博也，広瀬啓吉，森川由博，亀田弘之，宇田川博文，森田敏生，"人間の言語処理過程のモデルに基づく自然言語理解システムの構築"，言語情報処理の高度化のための基礎的研究，研究課題番号 68101004 (1989).

[104] 大野澄雄，"人間の音声言語処理過程のモデル化とその音声認識への応用"，東京大学工学部電子工学科，修士論文 (1990).

[105] 今井聖，北村正，"2 次元ケプストラムを利用する音声分析"，電子通信学会論文誌，J59-A，pp.1096-1103 (1976).

[106] 水田忍，有田康雄，坂井利之，"2 次元ケプストラム上でのリニアマッチングによる不特定話者単語認識"，電子情報通信学会信学技報，SP87-24，pp.17-24 (1987).

[107] 古賀真二，吉田和永，渡辺隆夫，"半音節を単位とした HMM による音声認識の評価実験"，日本音響学会秋季講演論文集，2-P-23，pp.247-248 (1988).

[108] 古賀真二，吉田和永，渡辺隆夫，"半音節を単位とした HMM による音声認識"，日本音響学会秋季講演論文集，2-P-24，pp.249-250 (1988).

[109] 松本和教，神部知明，関口芳廣，重永実，"単語予測機能を備えた不特定話者連続音声認識システム"，日本音響学会秋季講演論文集，1-5-18，pp.35-36 (1991).

[110] Y.Asano and K.Hirose, "A Dialogue Processing System for Speech Response with High Adaptability to Dialogue Topics," *IEICE Trans. Information and Systems*, E67-D, 1, pp.95-105 (1993).

[111] 竹林洋一，坪井宏之，金澤博史，貞本洋一，山下泰樹，瀬戸重宣，永田仁史，新居孝章，橋本秀樹，新地秀昭，"不特定話者音声対話システム TOSBURG の開発"，日本音響学会春季講演論文集，1-P-16，pp.135-136 (1992).

[112] V.Zue, J.Glass, D.Goodine, H.Leung, M.Phillips, J.Polifroni and S.Seneff, "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation," *Proc. ICASSP'90*, pp.73-76 (1990).

[113] 石金正明，"HMM を用いた単語アクセント型の識別"，東京大学工学部電子工学科，卒業論文 (1995).

[114] 広瀬啓吉，大野澄雄，峯松信明，藤崎博也，"音声認識における音響的特徴表現の時間単位に関する検討"，日本音響学会秋季講演論文集，2-P-8，pp.153-154 (1991).

[115] 峯松信明，大野澄雄，広瀬啓吉，藤崎博也，"複数の時間単位・精度の音響的特徴表現を用いた音声認識"，日本音響学会春季講演論文集，1-1-16，pp.31-32 (1992).

[116] 峯松信明，広瀬啓吉，"音声認識における大局的特徴の利用に関する一考察"，日本音響学会春季講演論文集，2-Q-4，pp.92-93 (1993).

[117] 峯松信明，広瀬啓吉，"音声認識において使用される音響的特徴空間の動的変化"，日本音響学会秋季講演論文集，1-8-9，pp.17-18 (1993).

[118] 峯松信明，広瀬啓吉，"動的に制御された音響的特徴を用いた音声認識"，音声入出力機器の現状とヒューマンインターフェイスの課題シンポジウム資料，SPREC-93-2-13 (1993).

[119] D.Rainton and S.Sagayama, "A New Tied Continuous Mixture Density HMM Via Orthogonalisation of the Full Covariance Observation Matrix," *Reposts Autumn Meet. Acoust. Soc. Jpn.*, 2-5-15, pp.77-78 (1991).

[120] 峯松信明，広瀬啓吉，"音声認識における正規化特徴量の利用に関する実験的検討"，日本音響学会春季講演論文集，1-Q-19，pp.147-148 (1994).

[121] 峯松信明，広瀬啓吉，"HMM に対する音声の時間構造の記述に関する検討"，日本音響学会秋季講演論文集，1-R-4，pp.199-200 (1994).

[122] F.Jelinek, "Continuous Speech Recognition by Statistical Methods," *Proc. IEEE*, 64, 4, pp.532-556 (1976).

[123] B-H.Juang and L.R.Rabiner, "Mixture Autoregressive Hidden Markov Models for Speech Signal," *IEEE Trans., ASSP*, 33, 6, pp.1404-1413 (1985).

[124] S.E.Levinson, "Continuous Variable Duration Hidden Markov Models for Automatic Speech Recognition," *Computer Speech and Language*, 1, pp.29-45 (1986).

[125] R.A.Johnson and D.W.Wichern, "Applied Multivariate Statistical Analysis," *Prentice-Hall, Inc* (1988).

[126] 瀬戸重宣，"連続音声における基本周波数の高時間分解能抽出とその時間パターンの自動分析"，東京大学工学部電子工学科，修士論文 (1991).

[127] 中川聖一，"確率モデルによる音声認識"，電子情報通信学会 (1988).

[128] 斉藤収三，中田和男，"音声情報処理の基礎"，オーム社 (1981).

[129] 新美康永，"音声認識"，共立出版株式会社 (1979).

[130] 中田和男，"音声"，日本音響学会 (1977).

[131] 上坂吉則，尾関和彦，"パターン認識と学習のアルゴリズム"，文一総合出版 (1990).

[132] 宇田川博文，"パソコンデジタル信号処理"，工学社 (1987).

[133] 前田渡，"ディジタル信号処理の基礎"，オーム社 (1980).

[134] 宮川，辻井，"ディジタル信号処理"，電子情報通信学会 (1975).

[135] 辻井重男，久保田一，"ディジタル信号処理"，オーム社 (1986).

[136] 安居院猛，中島正之，"FFT の使い方"，秋葉出版 (1986).

[137] 小池慎一，"C による科学技術計算"，CQ 出版社 (1987).

[138] 吉川敏則，"C 言語実用数値処理プログラム集"，近代科学社 (1988).

[139] 奥村晴治，"C 言語による最新アルゴリズム事典"，技術評論社 (1991).

[140] Leendert Ammeraal 著，小山裕徳 訳，"C—データ構造とプログラム—"，オーム社 (1990).

[141] 木下凌一，"X-Window Ver.11 プログラミング"，日刊工業新聞社 (1989).

[142] 安居院猛，永江孝規，"X アプリケーション・プログラミング，Xlib 編"，新紀元社 (1992).

[143] 安居院猛，永江孝規，“X アプリケーション・プログラミング，Athena ウィジェット編”，新紀元社 (1992).

[144] Adrian Nye and Tim O'Reilly 著，今泉貴史 監訳，“X ツールキット・イントリンシクス，プログラミング・マニュアル”，ソフトバンク株式会社 (1992).

[145] Adrian Nye and Tim O'Reilly 著，今泉貴史 監訳，“X ツールキット・イントリンシクス，リファレンス・マニュアル”，ソフトバンク株式会社 (1992).

[146] 芝祐順，渡部洋，石塚智一，“統計用語辞典”，新曜社 (1984).

[147] 池田央，“統計ガイドブック”，新曜社 (1989).

[148] 田中豊，脇本和昌，“多変量統計解析法”，現代数学社 (1983).

[149] 有馬哲，石村貞夫，“多変量解析のはなし”，東京図書 (1987).

[150] 蓑谷千凰彦，“統計学のはなし”，東京図書 (1987).

[151] 石村貞夫，“統計解析のはなし”，東京図書 (1989).

[152] 蓑谷千凰彦，“推定と検定のはなし”，東京図書 (1988).

[153] 林周二，“統計および統計学”，東京大学出版会 (1988).

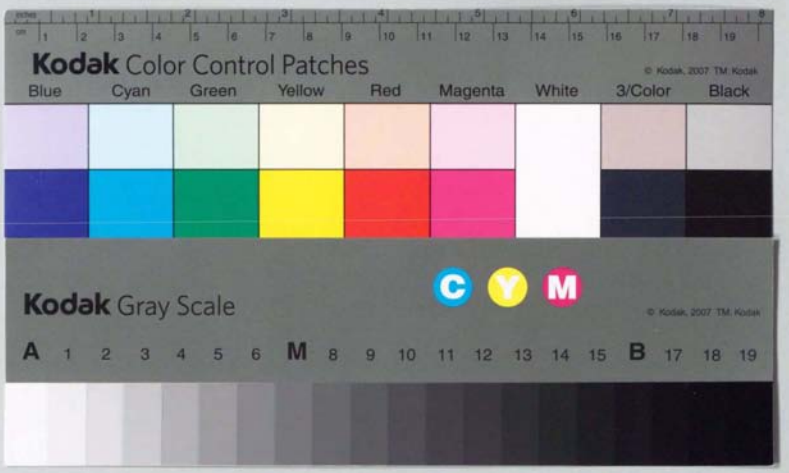
[154] 野寺隆志，“楽々LaTeX”，共立出版株式会社 (1990).

[155] 伊藤和人，“LaTeX トータルガイド”，秀英システムトレーディング株式会社 (1993).

[156] 奥村晴治，“LaTeX 美文書作成入門”，技術評論社 (1991).

[157] 磯村秀樹，“LaTeX 自由自在”，サイエンス社 (1992).

[158] 岩熊哲夫，古川徹生，“LaTeX のマクロやスタイルファイルの利用”，ftp:ftp.tohoku.ac.jp//pub/TeX/latex-styles/bear_collections/styleuse.dvi.gz (1995年 3月 13日現在) (1993).


Kodak Color Control Patches
© Kodak, 2007 TM Kodak
Blue
Cyan
Green
Yellow
Red
Magenta
White
3/Color
Black
Kodak Gray Scale
C
Y
M
© Kodak, 2007 TM Kodak
A 1 2 3 4 5 6 M 8 9 10 11 12 13 14 15 B 17 18 19